THE BELL SYSTEM TECHNICAL JOURNAL

VOLUME XLIII NOVEMBER 1964 NUMBER 6

# Effectiveness of Error Control in Data Communication over the Switched Telephone Network

By R. L. TOWNSEND and R. N. WATTS

(Manuscript received September 24, 1963)

*This article describes the results of a data communication experiment designed to investigate the effectiveness of error detection and retransmission in providing high-accuracy data transmission over the switched telephone network. Data were encoded into a Bose-Chaudhuri (31,21) error-detecting code and transmitted at 2000 bits per second by a DATA-PHONE data set 201A over a variety of dialed long-distance connections. Transmitted and received data were compared to obtain error data which were analyzed to obtain an estimate of the error performance of the data set and the effectiveness of the code. The results of this analysis are presented.*

*During the test approximately $6.36 \times 10^7$ 31-bit code words or $1.97 \times 10^9$ bits were transmitted. Of these, 63,002 bits appearing in 29,731 different code words were received incorrectly. Thus, the over-all bit error rate was $3.19 \times 10^{-5}$ and the word error rate $4.67 \times 10^{-4}$. The decoder was successful in detecting all but two of the erroneous code words, resulting in an average undetected word error rate of $3.14 \times 10^{-8}$ or an average of $9.85 \times 10^8$ bits between undetected word errors. These results demonstrated that very low undetected error rates can be obtained in practice using an error detection and retransmission system of modest complexity.*

## I. INTRODUCTION

Much attention has been focused recently on the problem of transmitting digital data over the switched telephone network with a high degree of accuracy. Selection and evaluation of error control schemes by which the desired high accuracy can be achieved require detailed information about the digital error statistics. Because of the complexity of the switched telephone network, the only feasible way to obtain this information is through the analysis of experimental data.

A method of error control which offers promise for use with telephone facilities is error detection and retransmission. An experiment has been performed to explore the feasibility of this type of error control and to obtain useful statistical information about the switched telephone network. In this experiment, a DATA-PHONE data set 201A,[1] which is a 4-phase unit designed for synchronous operation at 2000 bits per second, was used to transmit data over a variety of connections in the direct distance dialing network. The transmitted data were encoded into the Bose-Chaudhuri[2] (31,21) code described in Appendix A, which had been selected on the basis of a computer study. Transmitted and received data were compared to obtain error data from which digital error statistics were derived.

The over-all results of the test are shown in Table I. This indicates that the decoder was successful in detecting all but two of the 29,731 words received containing transmission errors. These results demonstrate the feasibility of providing high accuracy data transmission over the direct distance dialing network by using an error detection and retransmission system of modest complexity. A description of the error control equipment is given in Appendix A.

A description of the test and an analysis of the numerical error data are presented in the remainder of the article.

## II. DESCRIPTION OF THE TEST

The test was conducted between March 13 and August 31, 1962, during which time approximately $1.97 \times 10^9$ bits were transmitted. A portable transmitter was used to transmit data from various locations throughout the Continental United States to a stationary receiver located at Murray Hill, New Jersey from March 13, 1962 until May 1, 1962, and then at Holmdel, New Jersey for the remainder of the test. All performance measurements were made at the receiving terminal.

At both Murray Hill and Holmdel three foreign exchange lines were installed, one each to a No. 5 crossbar, No. 1 crossbar, and step-by-step central office. The characteristics of these lines are outlined in Table II. Dialed connections were originated from the receiving terminal, which was so arranged that it could be connected to any of the three lines. At both receiving stations calls were distributed equally, as nearly as pos-

Table I — Experimental Results of Data Communication over the Switched Telephone Network

| | |
|---|---|
| Number of transmitter locations | 28 |
| Number of calls | 548 |
| Number of hours of transmission | 273 |
| Total bits transmitted | $1.97 \times 10^9$ |
| Information bits transmitted | $1.33 \times 10^9$ |
| Words transmitted | $6.36 \times 10^7$ |
| Number of bits in error (total) | $6.30 \times 10^4$ |
| Number of words in error | $2.97 \times 10^4$ |
| Number of undetected word errors | 2 |
| Bit error rate | $3.19 \times 10^{-5}$ |
| Word error rate | $4.67 \times 10^{-4}$ |
| Undetected word error rate | $3.14 \times 10^{-8}$ |
| Factor of improvement (word) | $1.49 \times 10^4$ |
| Average bits between undetected word errors | $9.85 \times 10^8$ |

sible, among the three foreign exchange lines. The duration of each call was approximately 30 minutes.

The transmitting terminal was moved to the locations listed in Table III. These were selected on the basis of their distance from the receiving terminal, types of connecting facilities and type of end switching office. Since one objective of the experiment was to collect and to analyze data transmitted over typical connections, the locations selected were in or near large metropolitan areas where data traffic is likely to be heaviest.

A pseudo-random sequence generator was used to produce a repetitive pattern of 511 distinct 31-bit code words. These were transmitted serially at 1000 bauds or 2000 bits per second by a DATA-PHONE data set 201A. Received data were demodulated with another data set 201A and then compared with the output of a synchronized, duplicate sequence generator. The output of the receiver and system performance information were recorded on magnetic tape. Error data also were recorded by means of electronic event counters. A test log was kept which

Table II — Receiving End Test Lines

| Recvr. Location | Type of End Office | Location of CO | Line No. | Line Loss to CO | Type of Line to CO |
|---|---|---|---|---|---|
| MH | #5XB | New Providence, N. J. | 4641116 | 3.4 db | H-88 |
| MH | #1XB | Plainfield, N. J. | PL68684 | 9.8 db | H-88 |
| MH | SXS | Carteret, N. J. | 5414054 | 13 db | H-88 |
| HO | #5XB | Holmdel, N. J. | 9464674 | 5.3 db | H-88 |
| HO | #1XB | Rahway, N. J. | 3814270 | 10.4 db | H-88 & N carrier |
| HO | SXS | Monmouth Junction, N. J. | DA96550 | 11 db | H-88 & N carrier |

TABLE III—LOCATIONS OF TRANSMITTING TERMINALS

| Transmitter Location City | CO Prefix | Date | Office Type | No. of Calls | Transmission Time (Hours) | Total Bits Transmitted |
|---|---|---|---|---|---|---|
| Rahway, N. J. | FU8 | 3/15/62 | #1XBAR | 11 | 5.5 | $3.96 \times 10^7$ |
| Passaic, N. J. | PR8 | 3/13/62 | #1XBAR | 9 | 4.62 | $3.32 \times 10^7$ |
| Paterson, N. J. | MU3 | 3/22/62 | #1XBAR | 7 | 3.22 | $2.32 \times 10^7$ |
| Ridgewood, N. J. | 444 | 3/27/62 | #5XBAR | 10 | 4.88 | $3.52 \times 10^7$ |
| Manhattan (N.Y.C.), N.Y. | 349 | 4/18/62 | #1XBAR | 10 | 4.83 | $3.48 \times 10^7$ |
| Manhattan (N.Y.C.), N. Y. | HA5 | 4/23/62 | #1XBAR | 13 | 6.5 | $4.68 \times 10^7$ |
| Manhattan (N.Y.C.), N. Y. | LT1 | 4/20/62 | #5XBAR | 12 | 6.17 | $4.44 \times 10^7$ |
| Manhattan (N.Y.C.), N. Y. | RI9 | 5/11/62 | #1XBAR | 14 | 7 | $5.04 \times 10^7$ |
| Manhattan (N.Y.C.), N. Y. | UN1 | 5/10/62 | #5XBAR | 9 | 4.45 | $3.20 \times 10^7$ |
| Brooklyn, N. Y. | JA2 | 5/16/62 | #1XBAR | 12 | 6 | $4.32 \times 10^7$ |
| Queens, N. Y. | 445 | 5/18/62 | #5XBAR | 12 | 6 | $4.32 \times 10^7$ |
| Freeport, N. Y. | FR9 | 5/22/62 | #5XBAR | 13 | 6.5 | $4.68 \times 10^7$ |
| Central Islip, N. Y. | CE4 | 5/21/62 | SXS | 11 | 5.67 | $4.08 \times 10^7$ |
| Trenton, N. J. | LY9 | 3/29/62 | SXS | 9 | 4.53 | $3.27 \times 10^7$ |
| Camden, N. J. | WO4 | 4/2/62 | #1XBAR | 9 | 4.63 | $3.33 \times 10^7$ |
| Manahawkin, N. J. | LY7 | 3/20/62 | SXS | 10 | 5.02 | $3.61 \times 10^7$ |
| Atlantic City, N. J. | 823 | 4/6/62 | SXS | 11 | 5.64 | $4.06 \times 10^7$ |
| Bridgeton, N. J. | GL1 | 4/4/62 | #5XBAR | 7 | 3.68 | $2.65 \times 10^7$ |
| Hartford, Conn. | 247 | 6/25/62 | SXS | 10 | 4.97 | $3.58 \times 10^7$ |
| Washington, D. C. | 232 | 7/20/62 | #1XBAR | 14 | 6.75 | $4.86 \times 10^7$ |
| Washington, D. C. | 333 | 7/18/62 | #5XBAR | 14 | 7.0 | $5.04 \times 10^7$ |
| Washington, D. C. | 392 | 7/19/62 | SXS | 12 | 6.08 | $4.38 \times 10^7$ |
| Washington, D. C. | 393 | 7/17/62 | #1XBAR | 12 | 6.0 | $4.32 \times 10^7$ |
| Newton, Mass. | 244 | 6/27/62 | #1XBAR | 11 | 5.65 | $4.07 \times 10^7$ |
| Waltham, Mass. | 899 | 6/26/62 | #5XBAR | 8 | 3.95 | $2.84 \times 10^7$ |
| Quincy, Mass. | 773 | 6/28/62 | #1XBAR | 9 | 4.63 | $3.33 \times 10^7$ |
| South Boston, Mass. | 268 | 6/29/62 | #5XBAR | 14 | 7 | $5.04 \times 10^7$ |
| Atlanta, Ga. | 231 | 8/29/62<br>to 8/30/62 | #5XBAR | 32 | 16 | $11.52 \times 10^7$ |
| Atlanta, Ga. | 237 | 8/28/62 | SXS | 11 | 5.23 | $3.76 \times 10^7$ |
| Atlanta, Ga. | 457 | 8/31/62 | SXS | 13 | 6.42 | $4.62 \times 10^7$ |
| Atlanta, Ga. | 521 | 8/27/62 | #5XBAR | 13 | 6.5 | $4.68 \times 10^7$ |
| Atlanta, Ga. | 525 | 8/28/62 | SXS | 13 | 6.5 | $4.68 \times 10^7$ |
| Hammond, Ind. | 844 | 8/24/62 | #5XBAR | 9 | 4.5 | $3.24 \times 10^7$ |
| Libertyville, Ill. | 686 | 8/21/62 | #5XBAR | 8 | 4 | $2.88 \times 10^7$ |
| Lafayette, Ill. | 247 | 8/23/62 | #1XBAR | 12 | 6 | $4.32 \times 10^7$ |
| Wabash, Ill. | 431 | 8/22/62 | #5XBAR | 12 | 6 | $4.32 \times 10^7$ |
| Superior, Ill. | 467 | 8/20/62 | #1XBAR | 7 | 3.25 | $2.35 \times 10^7$ |
| Los Angeles, Calif. | 234 | 8/7/62<br>to 8/8/62 | SXS | 22 | 10.33 | $7.44 \times 10^7$ |
| Los Angeles, Calif. | 273 | 8/6/62 | #5XBAR | 12 | 6 | $4.32 \times 10^7$ |
| Los Angeles, Calif. | 385 | 8/8/62 | SXS | 10 | 4.91 | $3.54 \times 10^7$ |
| Los Angeles, Calif. | 620 | 8/9/62 | #5XBAR | 22 | 11.57 | $8.33 \times 10^7$ |
| Los Angeles, Calif. | 655 | 8/10/62 | #5XBAR | 6 | 3.06 | $2.21 \times 10^7$ |
| San Francisco, Calif. | 399 | 8/16/62 | SXS | 13 | 6.58 | $4.74 \times 10^7$ |
| San Francisco, Calif. | 981 | 8/17/62 | #5XBAR | 2 | 1 | $0.72 \times 10^7$ |
| San Francisco, Calif. | YU1 | 8/14/62 | #5XBAR | 11 | 5.38 | $3.87 \times 10^7$ |
| San Francisco, Calif. | 982 | 8/16/62 | #1XBAR | 12 | 5.67 | $4.08 \times 10^7$ |
| San Francisco, Calif. | 982 | 8/15/62 | #1XBAR | 4 | 2.06 | $1.48 \times 10^7$ |
| San Francisco, Calif. | 982 | 8/13/62 | #1XBAR | 11 | 5.65 | $4.07 \times 10^7$ |

summarized the results of each call. Included were descriptions of any unusual transmission or operational conditions which caused the test to be interrupted. Appendix A contains a description of the test system. The complete test procedure is given in Appendix B.

During the test 548 calls were completed, each containing approximately $3.6 \times 10^6$ bits. The magnetic tape data were reduced and analyzed for the 412 completed calls which contained errors. The other 136 completed calls were error free. Fifty-nine attempted calls were not completed for reasons which are summarized in Appendix C.

## III. ERROR RATES

During the course of the test approximately $1.97 \times 10^9$ bits were transmitted. Of these, 63,002 bits were received incorrectly, giving an over-all bit error rate of $3.19 \times 10^{-5}$. As has been mentioned earlier, data were transmitted in 31-bit code words. Of the $6.36 \times 10^7$ code words transmitted, 29,731 were found to contain one or more bit errors. This gives an over-all word error rate of $4.67 \times 10^{-4}$. The decoder was successful in detecting all but two of the erroneous code words, thus yielding an average undetected word error rate of $3.14 \times 10^{-8}$. This is equivalent to an average of $9.85 \times 10^8$ bits or 136 hours of transmission between undetected word errors.

The over-all distribution of bit error rates per call is plotted in Fig. 1. Also plotted in this figure are the corresponding distributions observed by Alexander, Gryb, and Nast[3] for transmission rates of 600 bits per

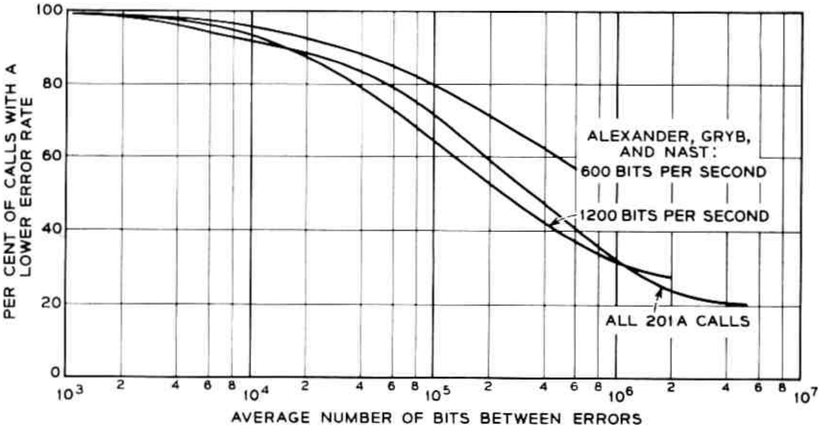


Fig. 1 — Bit error rate distribution for all calls: Alexander, Gryb, and Nast, 600 and 1200 bits/sec.

second and 1200 bits per second in a different test. The distributions for the Alexander, Gryb, and Nast 1200 bits per second data and the 201A data (2000 bits per second) show a remarkable similarity in view of the fact that the two tests employed different types of modulators operating at different speeds.

A question of major interest is "What factors have the greatest effect on error rate?" An attempt to answer this question was made by sorting the call error rates by all of the known parameters of the call, such as types of central offices at the transmitting and receiving ends, time of day, day of week, etc. Since none of the calls was actually traced, factors such as types of carrier systems in the circuit, types of intermediate central offices, etc., for any given call were generally not known. The only call parameter examined which showed a clear relationship with error rate was distance over which the call was made. Although none of the other parameters showed a definitive effect on error rate, this does not necessarily imply that these other parameters do not affect performance. It is likely that the data recorded did not allow adequate separation of the effects of these parameters.

Calls were classified into exchange, short-haul, and long-haul categories. Following the definitions used by Alexander, Gryb, and Nast, exchange calls are those within a single dialing area; short-haul calls are interarea calls between points separated by an airline distance of 400 miles or less; and long-haul calls are those exceeding 400 airline miles. Distributions of bit and word error rates per call for these categories are shown in Figs. 2 and 3. Again the bit error rate distributions of these

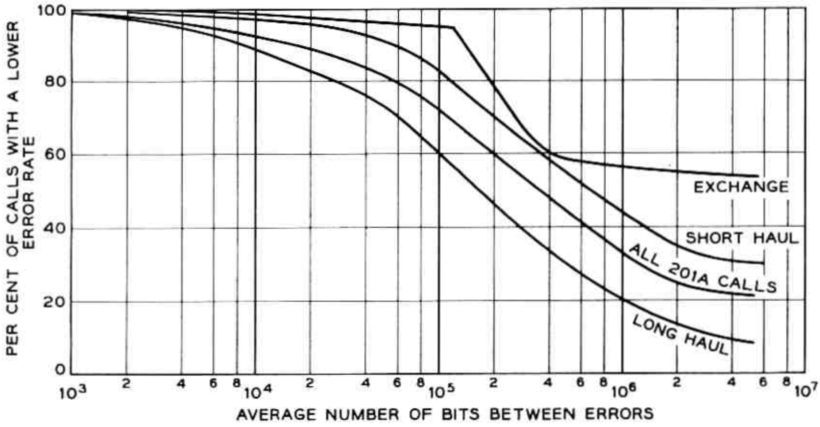


Fig. 2 — Bit error rate distribution for all calls: exchange, long-haul, and short-haul.

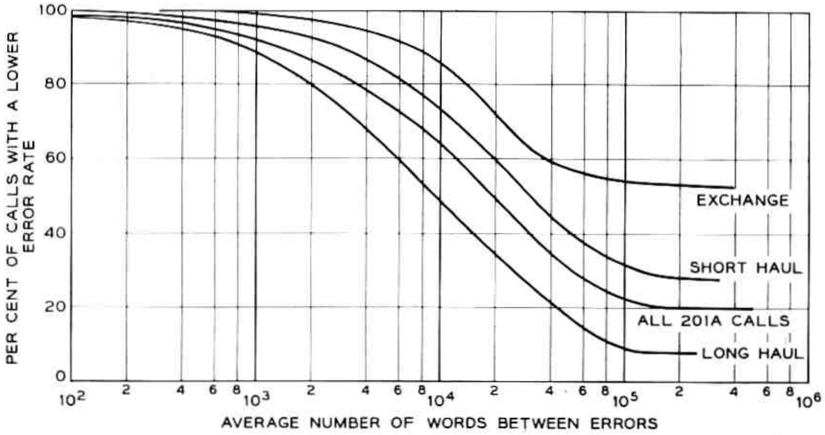


Fig. 3 — Word error rate distribution for all calls: exchange, long-haul, and short-haul.

three categories are very similar to the corresponding distributions of Alexander, Gryb, and Nast. As a rule, word error rates varied quite uniformly with bit error rate, indicating that the parameters studied had little effect on the density of error bits in an error word.

## IV. CORRELATION BETWEEN ERRORS

It is well known that digital data errors in telephone circuits tend to be bunched together,[3] but little is known about the exact nature of their correlation. One measure of the degree of correlation between errors is the autocorrelation function of the bit error sequences of the calls. Here we shall define the sequence $\{X_{ji}\}$, $i = 1,2, \cdots ,N_j$ of call $j$ to be the binary sequence having 1's in positions corresponding to the positions of bits incorrectly received, and 0's in positions corresponding to error-free bits. The number $N_j$ of terms in the sequence is equal to the number of bits transmitted in the call. We shall define the normalized autocorrelation function $\varphi(k)$ of the bit error sequences of any collection $M$ of calls to be:

$$\varphi(k) = \frac{\sum_{j\epsilon M} \sum_{i=1}^{N_j-k} X_{ji}X_{ji+k}}{\sum_{j\epsilon M} \sum_{i=1}^{N_j-k} X_{ji}}.$$

As the number of terms in the above expression becomes large, $\varphi(k)$ will converge to the conditional probability that, given an error bit, the bit $k$ positions later will also be in error.

The normalized bit error autocorrelation function for all calls containing errors is plotted in Fig. 4. This curve shows the presence of a very strong periodic component. The fact that the oscillations occur at a rate of exactly 120 cycles per second strongly suggests 60-cycle power line interference with the circuit. This periodic component was traced to three calls from a single location. The bit error autocorrelation function for all calls except the three containing a 120-cycle component is shown in Fig. 5. The general shape of this curve is similar to that observed in other studies.[4] The three periodic calls are excluded from the remaining distributions.

The autocorrelation function was also tabulated individually for each call. Efforts to find relationships between the autocorrelation and the known parameters of the calls were generally unsuccessful. It was noticed, however, that the initial shapes of the autocorrelation functions were similar for most calls, but the sizes of the tails varied widely. For most individual calls the autocorrelation function decreased considerably more rapidly than did the autocorrelation for all calls, which

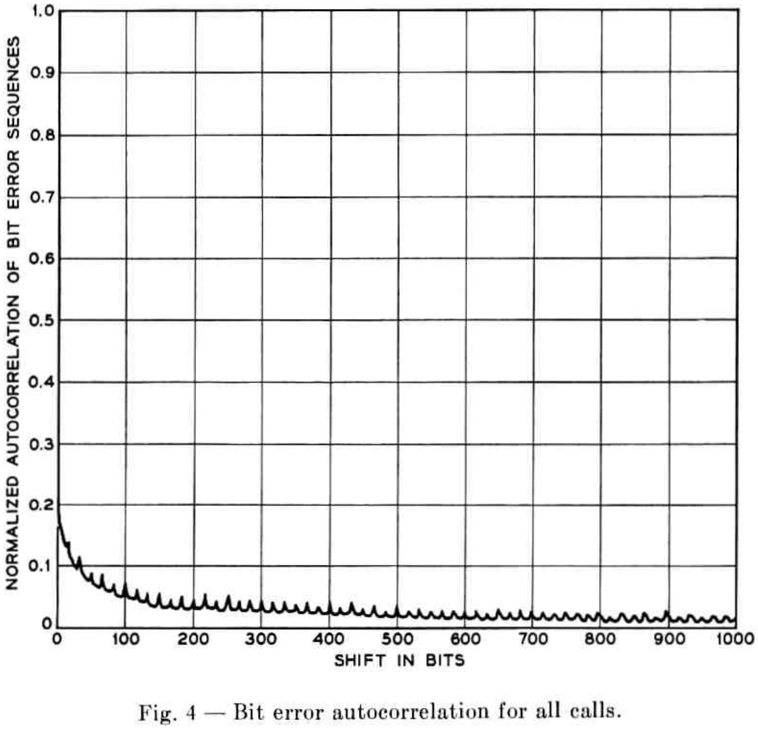


Fig. 4 — Bit error autocorrelation for all calls.

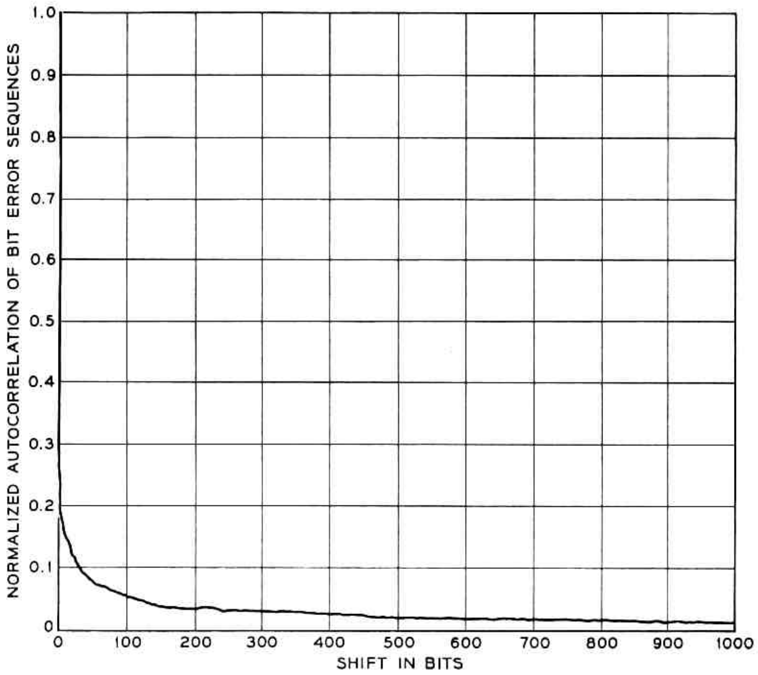


Fig. 5 — Bit error autocorrelation for all calls except those with a 120-cycle component.

is shown in Fig. 5. As one would expect, those calls whose autocorrelations had large tails were found to contain short periods of very high error rates. Surprisingly, there was not a very strong correlation between the call error rate and the size of the tail, but there was an apparent relationship between the variance of the error rate over one-minute intervals within the call and the size of the tail. This suggests that a long tail on the autocorrelation function probably was due to short dropouts or very noisy periods which were more or less independent of the over-all error rate.

The autocorrelation of the word error sequences was similarly computed. Here the autocorrelation is defined analogously, with error bits being replaced by error words. The word autocorrelation for all error calls except the three previously mentioned 120-cycle calls is plotted in Fig. 6. As one would expect, this curve is very much flatter than the corresponding error bit autocorrelation.

Further insight into the nature of the bunching of the errors can be obtained from the distribution of error-free bits between errors. The

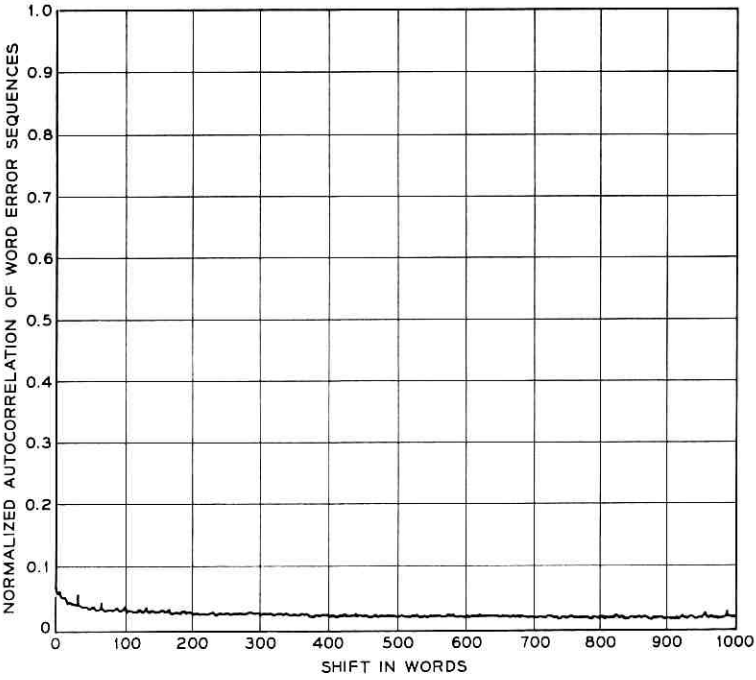


Fig. 6 — Word error autocorrelation for all calls except those with a 120-cycle component.

empirical cumulative probability distribution function of the number of error-free bits between bit errors is shown in Fig. 7. In this curve the ordinate gives the fraction of the total number of bit errors whose proximity to the previous bit error was equal to or less than the value given on the abscissa. As the number of occurrences becomes large, the empirical probability distribution function will converge to the true probability distribution function. It is interesting to note that the curve has a rather sharp knee at about ten bits on the abscissa and levels off to an ordinate value of approximately 0.65. This suggests that roughly one-third of the bit errors are separated by at least 200 bits (100 ms) from the previous error and that the remaining two-thirds of the errors are usually separated by not more than ten good bits. The errors are therefore observed to be bunched together in groups. The distribution of the lengths of these groups will be discussed in the next section.

The corresponding empirical probability distribution function for error-free 31-bit words between word errors appears in Fig. 8. The fact that the derivative of this curve changes comparatively slowly implies that small groups of errors are themselves bunched together, since the separation of errors would be much greater if they were randomly distributed.

## V. ERROR BURSTS AND DROPOUTS

Knowledge of the duration and error density of a burst of errors is important, since it is desirable to avoid combining bits into words in such a way that a substantial fraction of the bits in a single word is likely to be in error. Let us define an error burst of density $1/b$ to be any sequence of bits starting with an error bit and at least $b$ bits long such that every block of $b$ bits within the sequence will contain at least one error bit. In other words an error burst is a sequence which begins with an error bit and does not contain $b$ or more consecutive correct bits. We shall define the length of the burst to be the length of the

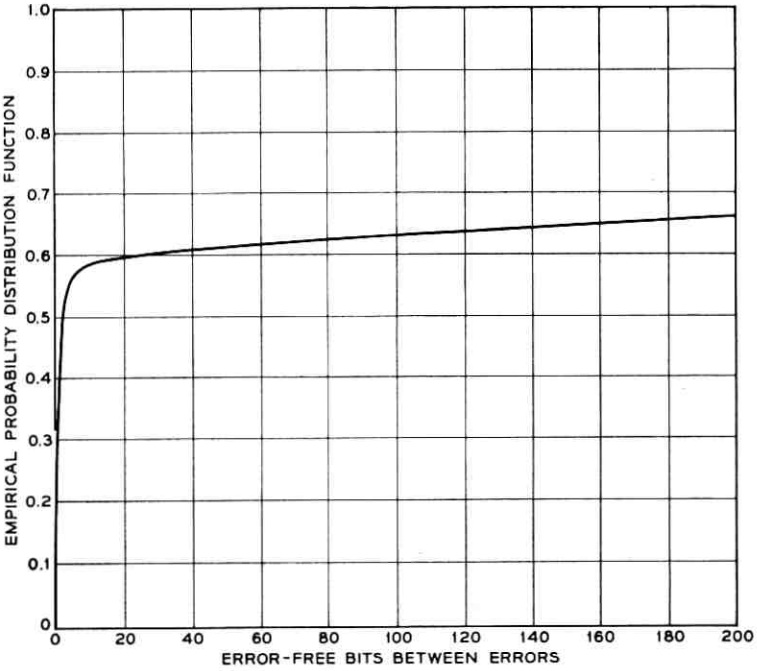


Fig. 7 — Empirical probability distribution function for error-free bits between errors.

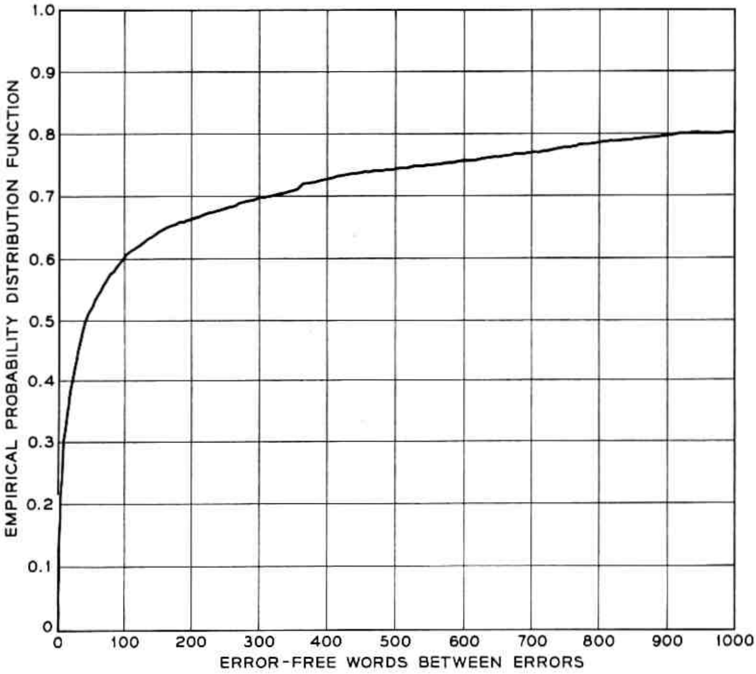


Fig. 8 — Empirical probability distribution function for error-free words between errors.

longest sequence consistent with the above definition. For example, consider the following sequence:

$$00000000001011000001000000000.$$

Let us assume that the 0's represent bits which were correctly received and the 1's represent bits which were incorrectly received. According to the above definition this sequence contains two bursts of density 1/5. The first burst begins with the eleventh digit in the sequence and is eight bits long. The second burst begins with the twentieth digit of the sequence and is five bits long. The sequence could also be thought of as containing a single burst of density 1/10 beginning with the eleventh digit and 19 bits long.

The empirical probability distribution functions of the lengths of bursts of densities 1/5, 1/10, and 1/31 were calculated and are plotted in Figs. 9–11. It can be seen that most high-density bursts are fairly short. We observe that the bursts become considerably longer for error densities of less than 1/10, which is in agreement with Fig. 7.

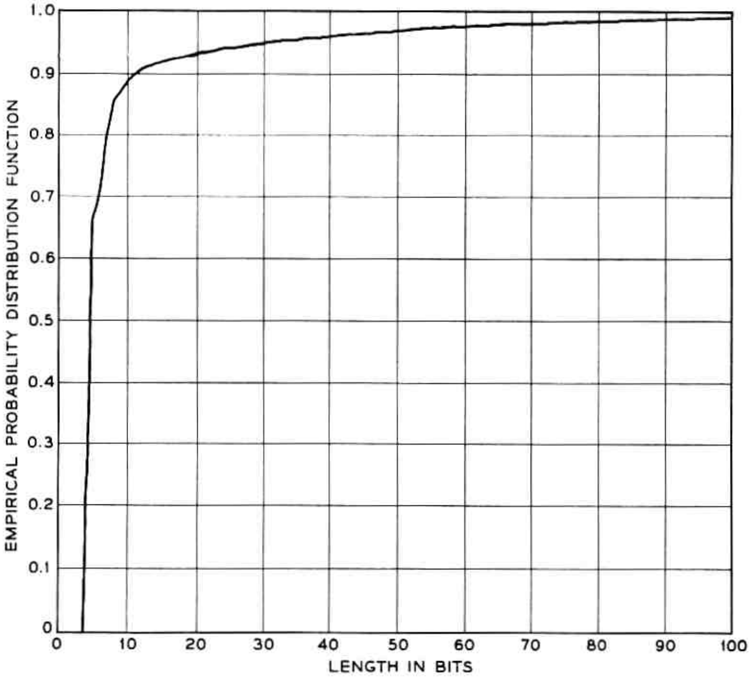


Fig. 9 — Distribution of lengths of bursts of density 1/5.

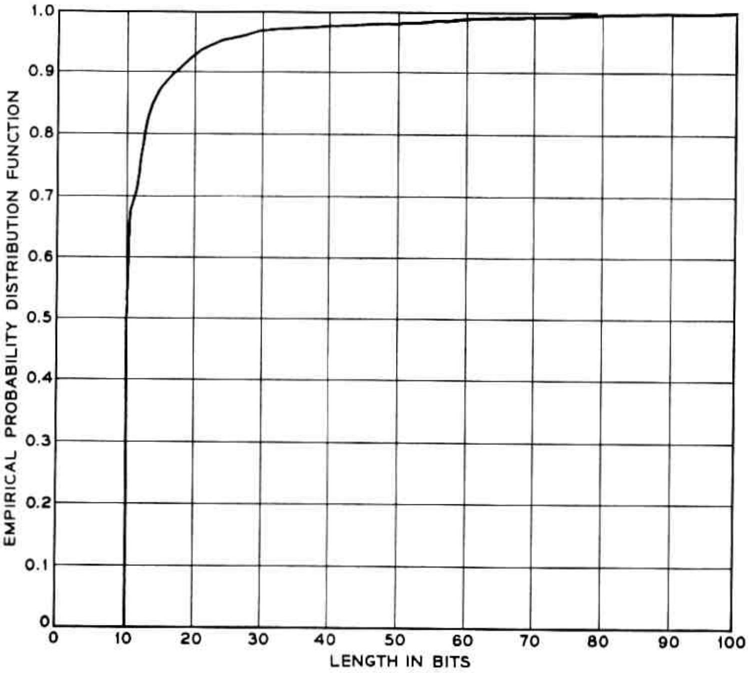


Fig. 10 — Distribution of lengths of bursts of density 1/10.

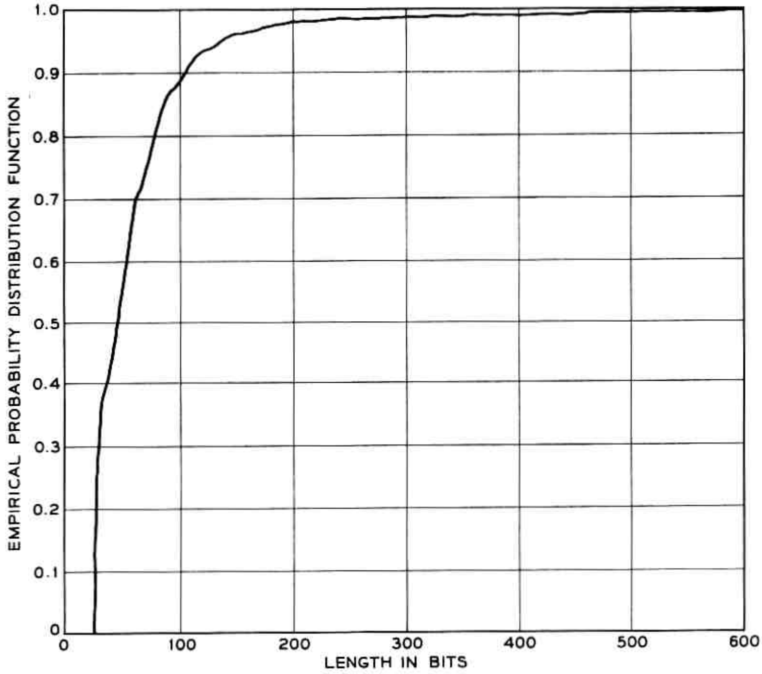


Fig. 11 — Distribution of lengths of bursts of density 1/31.

A dropout is a phenomenon whereby the connection is temporarily interrupted and the line signal drastically attenuated or completely lost for a fraction of a second. In the system tested a dropout caused only 1's to be received regardless of the transmitted message. Any sequence containing at least ten bit errors and in which only 1's were received was deemed to be a dropout. On the basis of this definition, about two per cent of the high-density bursts were found to be dropouts. These were contained in 44 different calls.

The empirical probability distribution function of the lengths of the observed dropouts is shown in Fig. 12. It should be pointed out that this distribution may be biased, since some of the longer dropouts probably caused the system to lose synchronization and were not included in the distribution. The large jump in the curve in the neighborhood of 145 bits was contributed entirely by four transcontinental calls. One plausible explanation for the occurrence of dropouts of this length is that they were caused by echo suppressors. Since the echo suppressors were not disabled during transmission, a high-energy noise impulse in the reverse channel could momentarily activate an echo suppressor. This would cause a dropout of approximately 145 bits duration.

The empirical probability distribution function of the number of bits between dropouts is shown in Fig. 13. It can be seen that the dropouts exhibit some tendency to be bunched together in time. This apparent bunching suggests fading rather than other possible causes.

The error data exhibited some asymmetry. There were about 15 per cent more 0 → 1 errors than 1 → 0 errors, a result consistent with the effect of dropouts. The fact that 1 → 0 errors were slightly more prevalent than 0 → 1 errors in calls not containing dropouts supports the conclusion that dropouts caused the asymmetry.

A more convenient distribution for some purposes is the distribution of the number of error bits appearing within a block of a given length. Following Elliott's[5] notation we shall define the function $P(m,n)$ to be the probability that exactly $m$ bits will be in error in a block of $n$ bits. The functions $P(m,n)$ for $n = 10, 15, 21, 23, 31, 63, 115$, and 230 are plotted in Fig. 14. These curves demonstrate quite vividly the effect of

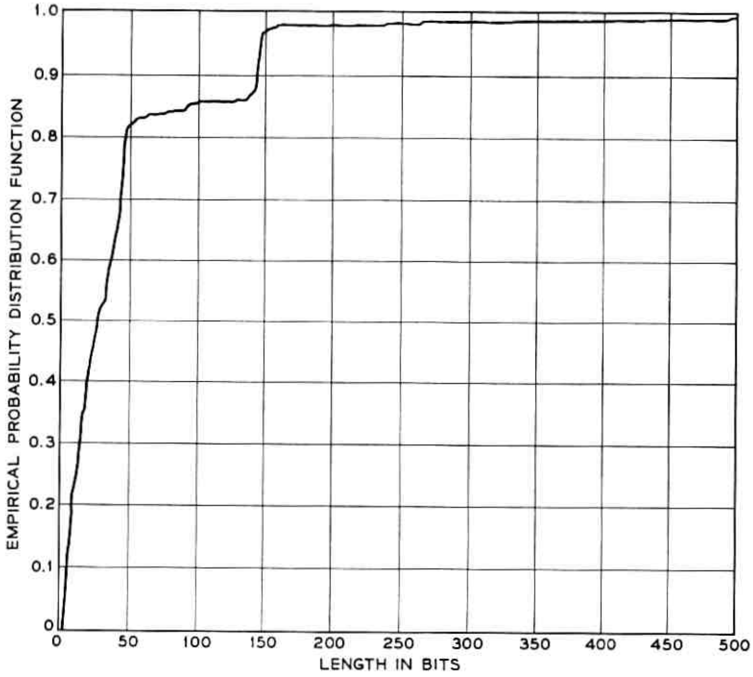


Fig. 12 — Distribution of dropout lengths.

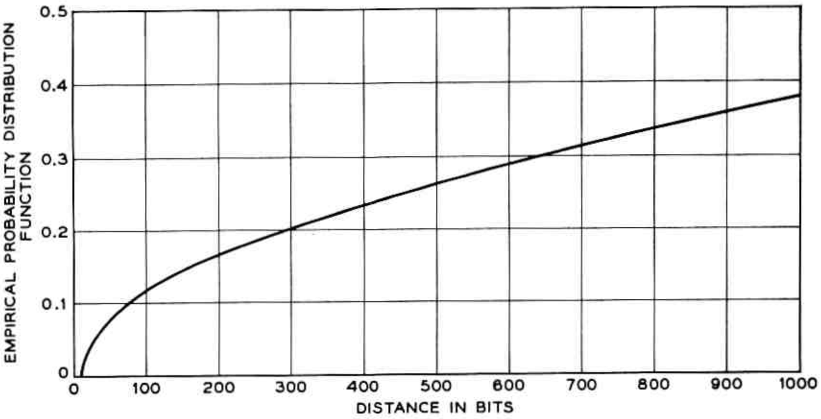


Fig. 13 — Distribution of distances between dropouts.

dropouts. Most of the curves exhibit a local maximum at about $n/2$. This is due mainly to the occurrences of dropouts longer than $n$. On the assumption that 0's and 1's were transmitted with approximately equal probabilities, dropouts of at least $n$ bits in length would contribute a component in the form of a symmetrical binomial distribution to the $P(m,n)$ function. This is illustrated in Fig. 15, in which the function $P(m,31)$ is plotted with and without dropout components.

Elliott[5] has suggested that a good approximate evaluation of the performance of a code can be made by assuming that all permutations of any given number of error bits in a block are equally likely. Using his methods and the function $P(m,31)$ the estimated number of bits between undetected errors was calculated to be $8.55 \times 10^8$. As stated previously, an average of $9.85 \times 10^8$ bits between undetected errors was actually observed. This is excellent agreement, although it should be remembered that the number of observed undetected errors was too small to assure good convergence of the observed average to a true average.

The function $P(m,n)$ changes radically as $n$ becomes very much larger. Calls were divided into 1-minute and 5-minute time intervals. The cumulative empirical probability distribution functions of the numbers of bit errors and word errors occurring within these time intervals were calculated and are plotted in Figs. 16 and 17. It is interesting to note that the distributions for 5-minute intervals are almost identical to the corresponding distributions for 1-minute intervals except for a scale factor. This suggests that the numbers of errors occurring in suc-

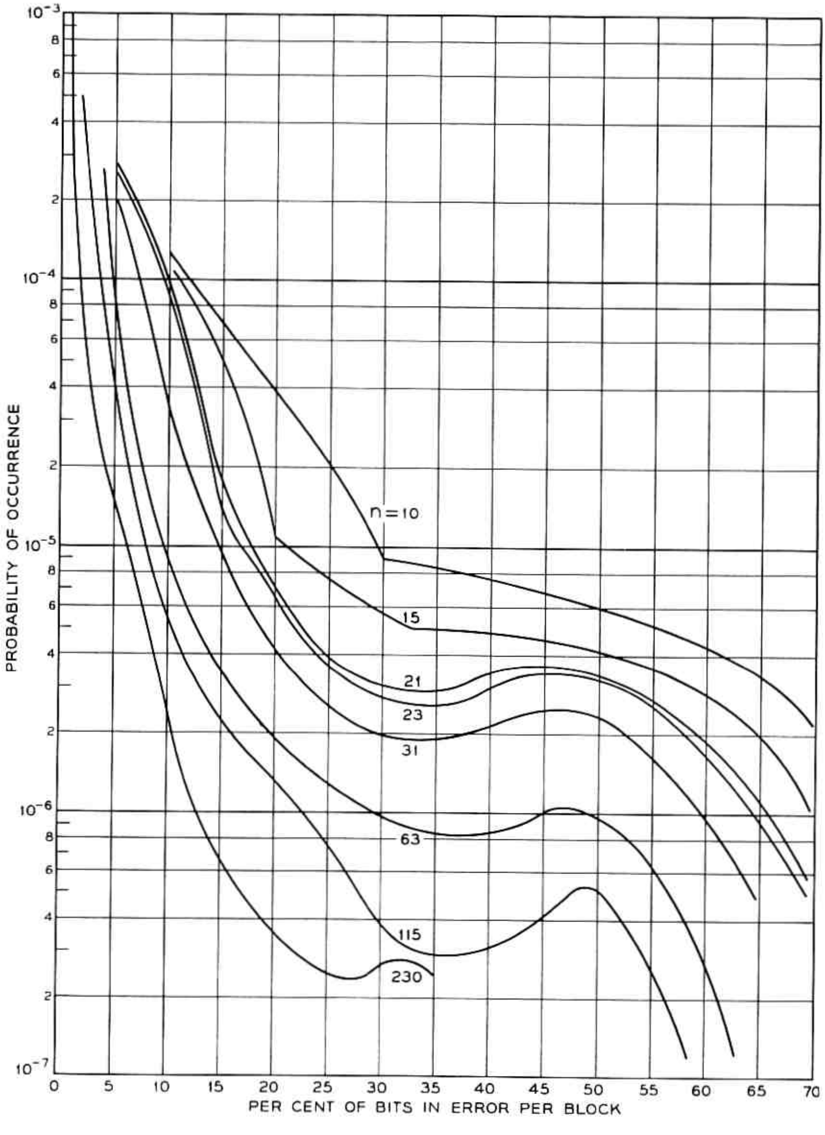


Fig. 14 — $P(m,n)$: the probability that exactly $m$ bits will be in error in a block of $n$ bits.

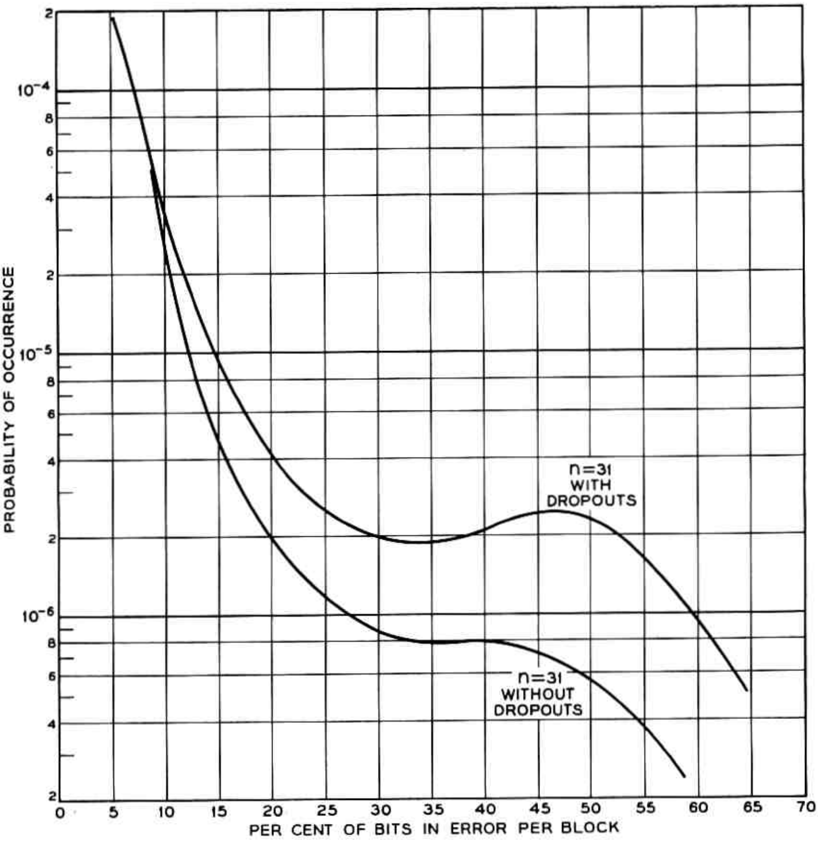


Fig. 15 — $P(m,31)$, with and without dropout components.

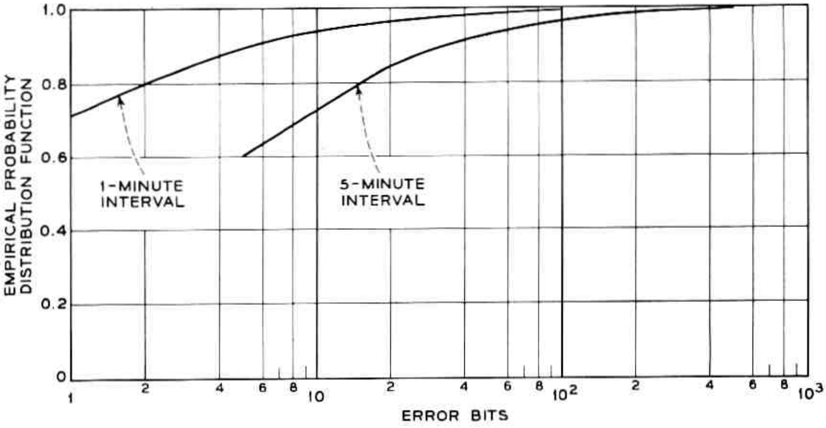


Fig. 16 — Distribution of error bits per one- and five-minute time interval.

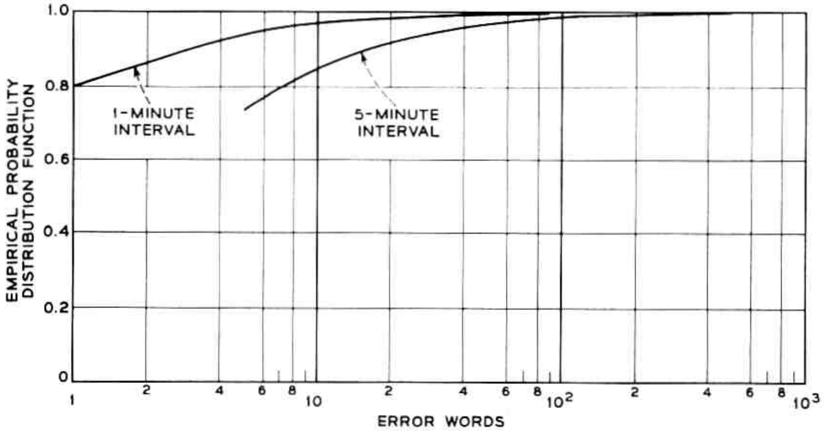


Fig. 17 — Distribution of error words per one- and five-minute time interval.

cessive intervals as long as a minute are essentially independent. Also, there is no noticeable effect of dropouts.

## VI. SUMMARY

The test demonstrated that it is possible to provide data transmission over the switched telephone network with extremely low undetected error rates by means of a coding technique of moderate complexity.

The statistical properties of the error data appear to be similar to those observed in other tests. Distributions of bit error rates without regard to coding showed a strong similarity to the results of Alexander, Gryb, and Nast, despite the fact that different modems operating at different speeds were used in the two tests. The digital errors were strongly correlated with each other, and the error rates were highly nonstationary. Bit errors were observed to occur in groups of two or three and generally had a density of at least one error bit per ten good bits. These small groups of errors were themselves bunched together. Dropouts occurred frequently in certain calls, but it was difficult to determine their cause.

## VII. ACKNOWLEDGMENTS

The authors wish to acknowledge the efforts of a number of people who have contributed to the results given here. A. B. Covey of the American Telephone and Telegraph Company was instrumental in making arrangements for the field tests. L. P. McRae, G. S. Robinson, and W. J. Wolf assisted in the testing program, and V. Koll assisted in the reduc-

tion of the test data. The authors also wish to thank A. B. Brown, W. R. Cowell, E. O. Elliott, M. D. Fagen, F. E. Froehlich, and R. Morris for their many helpful suggestions and technical advice. Personnel of the various operating companies involved were most cooperative and helpful.

## APPENDIX A

### *The Error Detecting System and Performance Measuring Apparatus*

A preliminary requirement on the code used for the error control experiment was that it be capable of detecting approximately 99.9 per cent of all transmission errors occurring in data transmitted by means of a DATA-PHONE data set 201A over switched, long distance telephone connections. Ease and economy of implementation were other factors affecting the selection of a code. Computer studies of a Bose-Chaudhuri[2] (31,21) code indicated that this code, subsequently used in the experimental system, had the desired error detecting ability. The above notation indicates that data were transmitted in blocks 31 bits long consisting of 21 information and 10 check bits.

The code is cyclic with a minimum distance of five* and is therefore capable of detecting any four or fewer bit errors in a 31-bit block. Furthermore, all single-error bursts† of length 10 bits or fewer, 511/512 of all 11-bit error bursts and 1023/1024 of all error bursts 12 to 31 bits in length are detected. The generator polynomial of this code is $h(X) = X^{10} + X^7 + X^6 + X + 1$, which is equivalent to saying that the code is the null space of the matrix:

$$
H = \begin{bmatrix}
1000000000100110101001000011111 \\
0100000000110101111101100010000 \\
0010000000011010111110110001000 \\
0001000000001101011111011000100 \\
0000100000000110101111101100010 \\
0000010000000011010111110110001 \\
0000001000100111000010111000111 \\
0000000100110101001000011111100 \\
0000000010011010100100001111110 \\
0000000001001101010010000111111
\end{bmatrix}.
$$

* I.e., every code word (block) differs in at least five places from every other code word.

† The length of a "burst" in this context is the number of bits between and including the first and last bits in error in a 31-bit block.

Examination of row 10 of the $H$ matrix (in its canonic form) reveals that the first check bit transmitted is the modulo 2 sum of information bits $d_1$, $d_2$, $d_3$, $d_4$, $d_5$, $d_6$, $d_{11}$, $d_{14}$, $d_{16}$, $d_{18}$, and $d_{19}$. Information bits are numbered in the order of their transmission—i.e., $d_1$ is the first information bit in the block, $d_2$ the second, etc.

The encoder was implemented by the feedback shift register shown in Fig. 18,[6] which operates as follows. With the feedback path closed (i.e., S in position 1), 21 information bits were shifted from the data source into the encoder and simultaneously transmitted. After the twenty-first bit had been encoded and transmitted, the feedback path was disabled by setting S to position 2, and the contents of the shift register (i.e., the 10 check bits) were transmitted. Thus, each block consisted of 21 information bits transmitted as a group in their original order followed by the 10 associated check bits.

Decoding was accomplished by the same circuit. The decoder was synchronized with respect to the received data and thus was able to distinguish between information and check bits. To decode a 31-bit

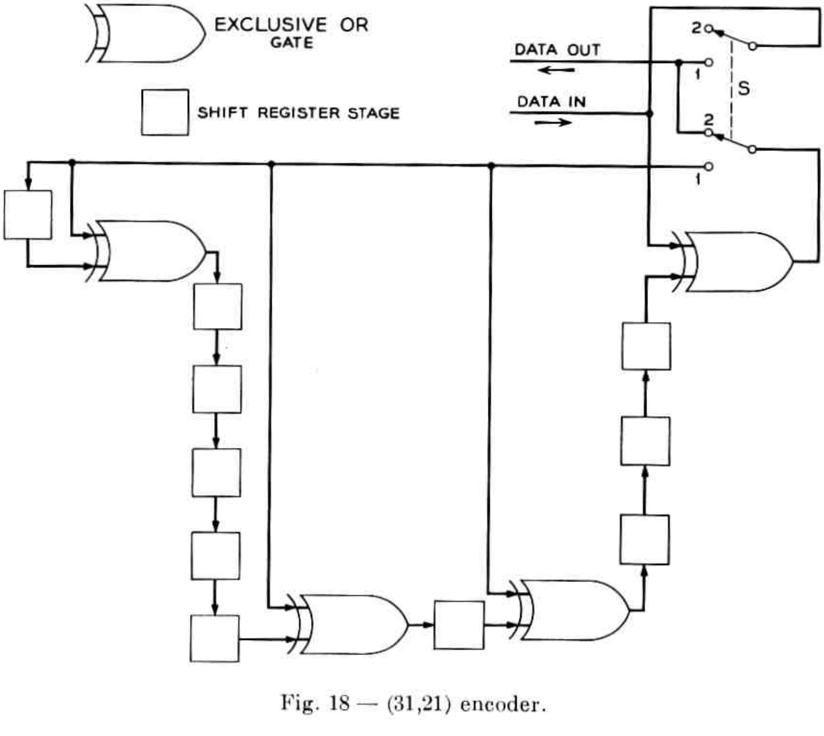


Fig. 18 — (31,21) encoder.

block the information bits were shifted from the demodulator into the encoder, which generated 10 check bits. These were compared to the 10 check bits received following the 21 information bits. Any difference between the two sets of check bits indicated the occurrence of a transmission error.

A block diagram of the error detection system is shown in Fig. 19. Both the source of data and the reference source are provided by a pseudo-random sequence generator. The output of this device is a repetitive 511-bit sequence containing every 9-digit binary sequence except the all-0 sequence. Since these data are meaningless so far as information content is concerned, a continuous chain of timing pulses shifts both the encoder and sequence generator. The output of the source is disregarded while check bits are shifted from the encoder. Since 31 and 511 are relatively prime, all possible 21-bit sections of the 511-bit sequence are transmitted as data. In practice the data source must stop after delivering 21 bits, while the 10 associated check bits are transmitted. For purposes of the test this would be impractical, since 511 and 21 have a common factor of 7, and therefore only 73 of the possible 511 21-bit

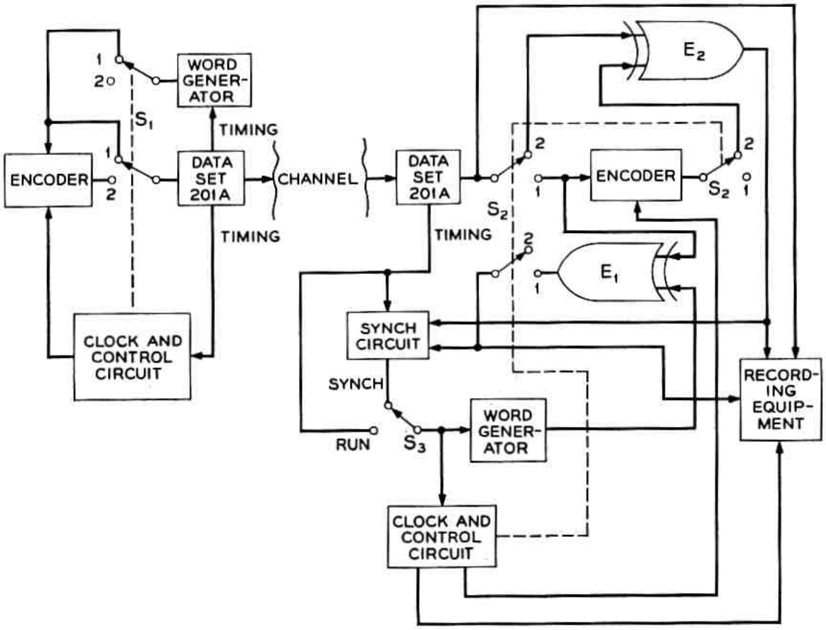


Fig. 19 — The test system.

sequences would be encoded. Timing is provided by the data set 201A, which generates 2000 timing pulses per second. The function of the clock and control circuit is to operate the encoding circuits at the transmitting and receiving terminals in the manner described earlier. The clock is driven by timing pulses from the data set and produces a periodic output signal which is "on" for 21 bits and "off" for 10. This signal is used by the control circuit to operate S (see Fig. 18), S1, and S2.*

At the transmitting terminal S and S1 were set to position 1 while information bits were shifted into the encoder and data set, then switched to position 2 while check bits were shifted from the encoder. Switches S and S1 then were reset to position 1 and the process was repeated for the next block of information.

At the receiver S and S2 were set to position 1 while information bits were received and encoded. During this time the received information was examined for transmission errors by E1, which produced an output whenever a received information bit differed from the output of the synchronized reference sequence generator. (The method of synchronization will be described in the next paragraph.) After the 21 information bits had been received and encoded, S and S2 were switched to position 2 while the check bits were received. Each of the 10 check bits in a block was compared with the output of the local encoder by E2. An output signal from E2 indicating the occurrence of detected errors was produced whenever a received check bit differed from the corresponding locally generated check bit. The outputs of E1 and E2, the received data, and timing information from the clock and control circuit were delivered to the recording equipment.

The clock and sequence generator at the receiving terminal were synchronized with respect to the demodulated data by means of the synch circuit. This circuit was activated manually by switching S3 to "synch." With S3 set to "synch" the phase of the sequence generator and clock was automatically shifted by one bit with respect to the received data in response to each output pulse from E1. When the outputs from E1 and E2 were observed to remain constant for one second or more, S3 was switched manually to "run" and the recording equipment started. Error data could not be recorded with S3 in the "synch" position.

The performance measuring apparatus was located at the receiving terminal as indicated in Fig. 19. Signals from E1, E2 and the control circuit were combined to indicate the occurrence and type of block errors.

* Switches S, S1, and S2 are implemented with solid-state circuits.

If, for a given 31-bit block, an error indication was received from E2 then transmission errors occurred and were detected. Undetected errors occurred in a block when errors were indicated in the information section by E1 but not in the check section by E2.

The received data and the error information derived from the outputs of E1 and E2 were recorded on dual-channel magnetic tape. The output of the demodulator was transformed into a series of positive and negative pulses and recorded on one channel. Following each 31-bit block one, two or three framing pulses were recorded on the second channel to indicate that the preceding block contained no errors, detected errors or undetected errors respectively. The inputs to both channels of the tape for each of the three conditions are shown in Fig. 20. Data were recorded on channel 1 and block framing on channel 2. In each case only one framing pulse follows block $K - 1$, which is assumed to be error free.

Cumulative error data for each call were recorded on five electronic event counters. The two types of bit errors (0 → 1 and 1 → 0) occurring in user information were derived from the output of E1 and recorded separately on two counters. A third counter was incremented whenever a block was received containing any errors in the information section. The fourth and fifth counters recorded detected and undetected block errors respectively. The counters were photographed automatically at 20-second intervals during each call. A clock was included at the camera's

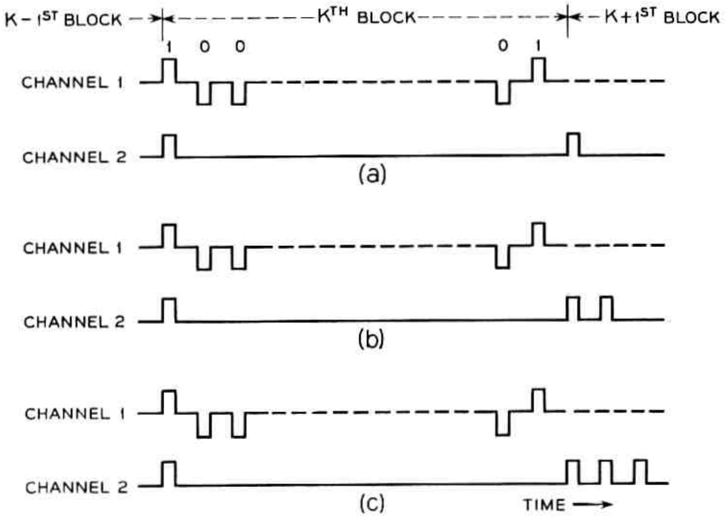


Fig. 20 — Block framing; three cases.

field of view and the film dated so that photographic data could be correlated with log book records.

## APPENDIX B

### *Field Test Procedure*

The purpose of this section is to describe the field test procedure in some detail. To review briefly, the receiving error control terminal remained fixed at Murray Hill or Holmdel and the transmitting terminal was carried to the locations listed in Table III. One day was spent at each location, during which time data were transmitted over a number of dialed connections established from the receiving terminal and used for approximately 30 minutes each.

Before initiating a series of calls between two locations, the line loss from the transmitter to the central office was measured with a 12B transmission measuring set. The data set's transmission level then was adjusted so that the signal strength at the central office was approximately −8 dbm and the data set then placed in the on-hook automatic answer mode. The location, telephone number, local loop loss, transmitting level, and type of central office serving the transmitting terminal were recorded in the test log. The encoding equipment was started and ran continuously throughout the series of calls (i.e., data were generated and encoded continuously and transmitted automatically whenever a connection was established from the receiving terminal).

After the transmitting terminal had been readied for the series of calls as described in the preceding paragraph, the receiving terminal was attached to a foreign exchange line. A connection was dialed to the transmitting terminal, which answered the call automatically and started transmitting encoded data immediately. Then the receiving error detection system was synchronized with respect to the demodulated data as described in Appendix A and the recording equipment started. Thirty minutes later the recording equipment was stopped and the call terminated from the receiving end. The data recorded on event counters, the times (local) at which the call was started and terminated, and a description of any unusual transmission or operating conditions were entered into the test log. The recording equipment was reset and the receiving terminal attached to a different foreign exchange line for the next call.

The error control equipment, data sets, and error recording apparatus were checked periodically throughout the test. This was done to insure that the data collected during the experiment would not be affected by marginally operating test equipment.

## APPENDIX C

### *Summary of Incomplete Calls*

During the test, 59 calls could not be completed for reasons which fall into three general categories. These are:

(1) long dropouts or fades resulting in loss of synchronism between the transmitting and receiving data sets,

(2) the inability to achieve initial synchronization of the terminals within a reasonable length of time, and

(3) lost connections.

These conditions will be more fully described in the following paragraphs. In addition to these 59 calls, 6 calls were interrupted due to human errors and two calls were lost as a result of local power failures.

Loss of synchronism between the transmitting and receiving data sets during otherwise normal communication caused the interruption of 30 calls. This condition usually was caused by long dropouts, particularly on long-haul connections. Dropouts lasting more than approximately 100 milliseconds generally caused the transmitting and receiving data sets to lose synchronism, since timing in the demodulator was derived from the line signal. Within 20 milliseconds of a loss of line signal the timing reverted to the natural resonant frequency of the high-$Q$ circuit in the demodulator's bit synch recovery circuit. The natural frequency of this high-$Q$ circuit was within one cycle of the transmitter frequency. Thus, if the modulator and demodulator remained decoupled long enough, synchronism between the two was lost. This situation was detected easily but resulted in some loss of data, as the terminals had to be resynchronized before data transmission could be resumed. Intense channel noise was observed to have approximately the same effect as dropouts.

Twenty-one dialed connections were sufficiently noisy that the test apparatus could not be synchronized within a reasonable length of time. To recapitulate, initial synchronization was obtained by automatically shifting the phase of the receiving end clock and sequence generator by one bit whenever the synchronization circuit was enabled and a received bit differed from the corresponding locally generated information bit. When the system was synchronized, but the synchronization circuit was not yet disabled, any transmission errors occurring in the information section of a block caused the synchronization procedure to be repeated. Therefore, when the channel was unusually noisy the receiving terminal would not remain synchronized long enough for the synchronization circuit to be disabled manually. If synchronization could not be established within 2 or 3 minutes after the call was placed the connection was dropped and the transmitting terminal called a second time using the same foreign exchange line. The semiautomatic synchronization procedure could have been fully automated. Had this been done, synchronization might have been achieved over a few of the connections for which the semiautomatic method described in Appendix A was unsuccessful. However, since all 21 of these calls were exceptionally noisy it appears doubtful that data set synchronism could be maintained for a full 30-minute period.

Eight connections were lost entirely during data transmission and dial tone was returned to both terminals. This situation was easily detected by the test apparatus. On at least three of these occasions the lost connections appeared to be associated with telephone maintenance operations.

The conditions described above are transmission impairments which cannot be integrated directly into the error rate data. These are, however, situations with which the data communicator must contend and are included here to provide an estimate of their frequency of occurrence. These data are of importance in the design of error control systems which must recognize such transmission impairments and allow for some type of remedial action to be taken, such as manual intervention or automatic resynchronization.

## REFERENCES

1. Baker, P. A., Phase-Modulation Data Sets for Serial Transmission at 2000 and 2400 Bits per Second, AIEE Trans., Part 1, Comm. and Elect., **61,** July, 1962, pp. 166–171.
2. Bose, R. C., and Ray-Chaudhuri, D. K., On a Class of Error-Correcting Binary Group Codes, Information and Control, **3,** 1960, pp. 68–79.
3. 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, pp. 431–476.
4. Townsend, R. L., Error Characteristics of a Data Communication System, IEEE Trans. Comm. and Elect., March, 1963, pp. 69–72.
5. Elliott, E. O., Estimates of Error Rates for Codes on Burst Noise Channels, B.S.T.J., **42,** Sept., 1963, pp. 1977–1997.
6. Peterson, W. W., *Error Correcting Codes*, The MIT Press and John Wiley & Sons, 1961.
7. Cowell, W. R., The Use of Group Codes in Error Detection and Message Retransmission, IRE Transactions on Information Theory, **IT-7,** July, 1961, pp. 168–171.
8. Cowell, W. R., and Burton, H. O., Computer Simulation of Use of Group

Codes with Retransmission on a Gilbert Burst Channel, AIEE Trans., Part 1, Comm. and Elect., **58,** Jan., 1962, pp. 577–585.

9. Fontaine, A. R., and Gallager, R. G., Error Statistics and Coding for Binary Transmission over Telephone Circuits, Lincoln Laboratory Report 25G-0023, Oct. 30, 1960.
10. Gilbert, E. N., Capacity of a Burst-Noise Channel, B.S.T.J., **39,** Sept., 1960, pp. 1253–1266.
11. Mertz, P., Error Burst Chains in Data Transmission, Rand Corporation Memorandum RM-3024-PR, Feb., 1962.
12. Morris, R., Further Analysis of Errors Reported in "Capabilities of the Telephone Network for Data Transmission," B.S.T.J., **41,** July, 1962, pp. 1399–1414.

# A Simulation Study of Routing and Control in Communications Networks*

By J. H. WEBER

(Manuscript received March 23, 1964)

*A set of studies has been undertaken to develop guidelines for the design and operation of communications networks with automatic alternate routing. Comparisons are made of engineered costs and overload capability of networks using several alternate routing configurations, and employing a number of different operating and control procedures. The traffic model selected consists of a 34-node network abstracted from the U.S. telephone toll network, with basic load levels obtained from field data. The overload evaluations were made using a simulation program prepared for the IBM 7090 computer.*

## I. INTRODUCTION

In a recent paper[1] the results of some preliminary comparisons of two alternate routing configurations for communications networks were reported. Those results indicated that for small networks (six or fewer nodes) with low traffic densities a symmetrical or unrestricted routing pattern is superior to a hierarchy similar to that in use in the U.S. toll network, while for higher traffic densities there appeared to be little difference in the network behavior in terms of economy and reaction to overloads.

Subsequently, a new simulation program has been constructed[2] and substantially larger networks have been examined to provide a more meaningful guide to network design under various circumstances of geography and load level. An additional configuration, called the "gateway," as well as several operating and control variations, has been examined. The latter include stage-by-stage operation with and without crankback (return of routing control to a previous node for rerouting when blocking is encountered at an intermediate switching point), limitation of number of links per call in symmetrical networks, and trunk reservation for first-routed traffic only.

* Presented at the Fourth International Teletraffic Conference, London, July, 1964.

The results show that:

(1) There is little difference in network cost or overload capability between hierarchical and symmetrical networks at the load densities considered.

(2) The gateway network (a two-level hierarchy with no interregional high-usage groups) requires substantially more trunking and switching than either the hierarchical or symmetrical networks and shows no significant difference in overload performance.

(3) Restriction of alternate routing in symmetrical networks improves performance at all levels of load.

(4) The use of crankback is a disadvantage for symmetrical networks with a high traffic density, at all levels of overload. It offers a slight advantage for symmetrical networks with lower traffic intensities and does not appear to have any significant effect on the performance of hierarchical networks.

(5) Trunk reservation for first-routed traffic on a dynamic basis improves the performance of almost all networks examined, for all load conditions, and displays no detrimental effects.

## II. THE SIMULATION

The simulation program used in these studies is described in Ref. 2. It has many of the capabilities of the program described in Ref. 1, but has been reprogrammed to accept networks with heavier loads and to operate more efficiently. A number of additional features have also been provided.

The program is basically capable of simulating networks of up to 63 nodes, with arbitrary alternate routing patterns and stage-by-stage call forwarding. There is no congestion or delay allowed at switching points, all congestion being assumed due to trunk shortages. Calls which fail to complete initially may be abandoned with a fixed probability or retried after a constant or exponentially distributed interval. Any prespecified number of trunks can be reserved for first-routed traffic only, and calls may "crank back" or return to a prior node if blocked at some point in the network. The maximum-size network which can be accommodated is largely determined by the number of simultaneous calls in progress, which may have a maximum of about 6000. Traffic loads are specified on a point-to-point basis, with arbitrary proportions in each direction, and may be changed linearly at any time during the run. That is, mean arrival rates can change linearly in time during the run at any rate and between any bounds. (Another modification of the program allows the use of two priorities of traffic and mixtures of direct and store-and-forward traffic, with trunk reservation by traffic type and priority. This version was not used for the studies described herein, however.)

In order to accommodate larger networks more efficiently, the program was written in several sections. The first of these accepts the basic load inputs (mean point-to-point loads and holding times) and generates call arrival times and holding times, which are then stored on magnetic tape. This tape is then used as input to the simulation program, which processes the calls through the simulated system and prints out raw data on trunk utilizations and call histories on two magnetic tapes. These tapes are presented to the output processor programs, which provide the appropriate reduced outputs.

For convenience in preparing the input data, the main section of the program has been arranged to determine its own routing for symmetrical and hierarchical networks, given the numbers of trunks and the distances for symmetrical networks, or the homing arrangements for hierarchical networks.

The output statistics are reported at prespecified time intervals, and these subinterval results may then be used as samples for a final output containing both means and standard deviations of all relevant quantities. The quantities which are printed out are as follows:

(1) point-to-point traffic loads at the end of the run (input data).

(2) routing tables for all point-to-point traffic items.

(3) means and standard deviations of the following measured quantities for each point-to-point traffic item:

(a) blocking probability

(b) average delay and distribution of delays for retried calls

(c) average number and distribution of number of links per call.

(4) means and standard deviations of the following measured quantities for each trunk group (obtained by switch count measurements):

(a) number of trunks present in each group (input data)

(b) number of trunks reserved for first-routed traffic in each group (input data)

(c) total carried load in erlangs on each group (and per cent occupancy)

(d) first-routed carried load on each group (and per cent of total)

(e) alternate routed carried load on each group (and per cent of total).

(5) means and standard deviations of measured over-all network quantities as follows:

(a) over-all average blocking probability, $\bar{B}$, given by

$$\bar{B} = \frac{\sum_{ij} a_{ij} B_{ij}}{\sum_{ij} a_{ij}}$$

where

$a_{ij}$ = offered load between nodes $i$ and $j$, and
$B_{ij}$ = blocking probability of calls offered between nodes $i$ and $j$

(b) average number and distribution of number of links per call
(c) weighted average delay and delay distribution for retried calls
(d) total number of trunks in the network (input data)
(e) total trunks reserved for first-routed traffic (input data)
(f) total carried load and over-all occupancy
(g) total carried first-routed load
(h) total carried alternate routed load.

(6) the "space dispersion," $D$, of the blocking probability, delay distribution and links per call distribution, given by

$$D_B = [(\sum_{ij} a_{ij} B_{ij}^2 / \sum_{ij} a_{ij}) - \bar{B}^2]^{\frac{1}{2}}$$

for the blocking probability, and similar expressions for the other quantities. This serves as a measure of the variation in grade of service provided to various traffic items in the network depending upon their origin and destination.

(7) the following input parameters:

(a) maximum number of links allowed per call

(b) number of stages of crankback allowed (That is, the number of steps a call is allowed to back up before progressing forward after having reached a point of congestion.)

(c) percentage of calls to be retried
(d) retrial time distribution and mean value
(e) holding time distribution and mean value
(f) number of nodes
(g) number of reporting intervals and their lengths
(h) number of reporting intervals to be collected for final processing
(i) routing pattern (hierarchical or symmetrical)

(j) interval between switch counts for determination of trunk information.

The simulation runs quite rapidly, processing about 500,000 calls per hour using the IBM 7090 computer for a 34-node network with a total load of about 5000 erlangs, excluding the traffic generation. If traffic generation time is included, the processing rate drops to about 375,000 calls per hour. If several networks are evaluated using the same traffic input, as was done in these studies, however, the traffic need be generated only once, and the same tape can be used as input to the simulation any number of times. Several dozen simulation experiments were made for the studies described below, but only eight traffic tapes were generated.

## III. GENERAL PROCEDURE

The evaluation procedure encompasses the following steps:

(1) select a geographical area, including switching center locations;

(2) select a basic traffic model on which to base network engineering;

(3) engineer networks to a given grade of service using each of the routing procedures to be considered;

(4) determine the costs of each of the networks so engineered;

(5) change the loads to correspond to reasonable patterns of overload or shifting load;

(6) using simulation, measure the performance of each of the networks under the load changes used in step (5);

(7) repeat steps (5) and (6) for each of the control and operating variants considered.

These steps will be described in detail in the following sections.

### 3.1 *The Geographical Region*

Although it is not possible to select a geographical region (or regions) which will be typical of all situations, it is desirable to find an area which at least has the capability of accommodating a sufficient number of nodes to adequately exercise the various routing patterns to be examined and of reacting to realistic load fluctuations. The region should also contain both densely and sparsely populated sectors, which to some extent must exist in all real networks. (A uniform or arbitrarily variable traffic distribution would probably not be a valid test, since actual telephone traffic varies with the population density, higher-density areas having large amounts of traffic within and among them, and sparsely populated areas being lower in traffic to all destinations.) Since the geographical region will utimately require a traffic model to be superimposed upon it, an area for which actual traffic data is obtainable is again more likely to represent reality than one for which traffic quantities need to be invented.

A single region which appears to meet most of the criteria specified above exists on the West Coast of the U.S.A. The states of California, Washington, Oregon and Nevada are almost entirely administered by two local telephone operating companies for toll purposes and represent a region which ranges from sparsely populated areas such as Nevada and eastern Oregon to sections such as Southern California, which contains the Los Angeles and San Diego metropolitan areas. Fig. 1 is a map of this region, showing the 34 switching centers used. Although there are many more than 34 toll switching offices in this region, only the control switching points (CSP's), which make up the offices of the top three levels of the U.S. toll network hierarchy,[3] are included. (Las Vegas, though actually a toll center, is assumed to be a primary center.) All traffic which both originates and terminates in the region, however, is included, as will be discussed below.

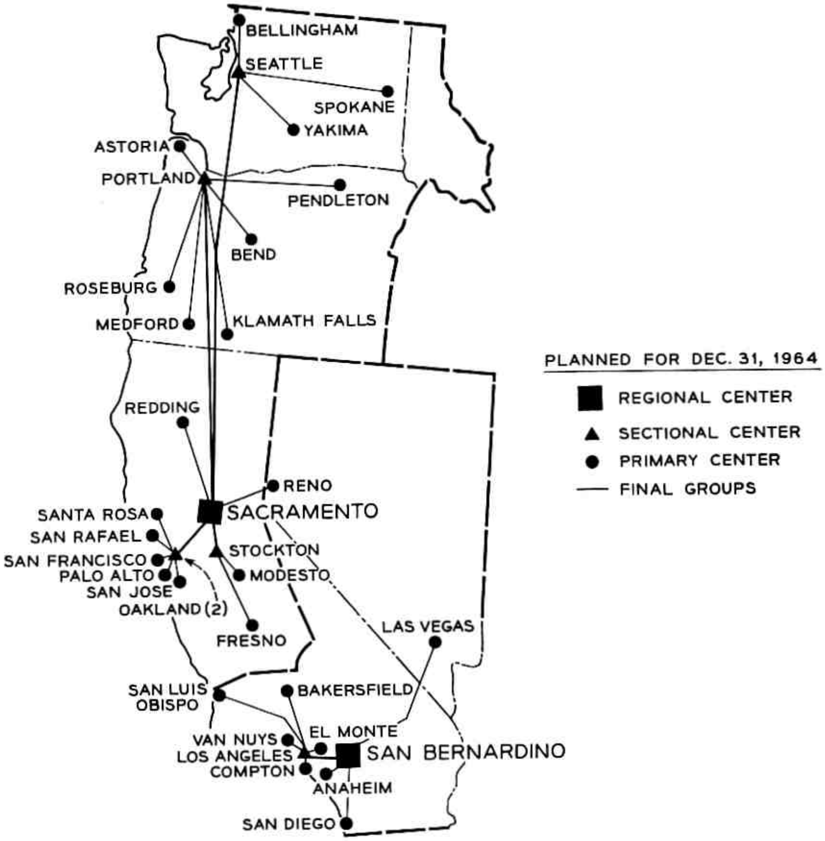


Fig. 1 — U.S. West Coast traffic region.

### 3.2 *The Basic Traffic Model*

The basic traffic model used for engineering the various networks was developed from actual message records of the Pacific Telephone and Telegraph Company and the Pacific Northwest Bell Telephone Company. These records include the total messages for a period of ten consecutive business days during June 1962. They provide total messages and message minutes from every toll switching center in the area to every other. Traffic which originates or terminates in other than the four-state area is not considered, nor is traffic which is not carried on the toll (or long distance) network. Traffic originating and/or terminating at offices of connecting companies, but which is carried on the toll network, is included.

In order to obtain a busy hour traffic base from the ten-day records, it was assumed that ten per cent of the total traffic was offered during each day and ten per cent of the day's traffic was presented during the busy hour. Therefore the busy hour traffic load was assumed to be one per cent of the total ten-day message load.

Traffic between toll centers of the fourth rank, which are not explicitly included in the 34-node model, is handled in two ways, giving rise to two networks with different traffic densities. In the first of these, called the "full-traffic" or "full-load" network, all traffic between toll centers is added to that between the centers on which they home. Traffic between toll centers homing on the same control switching point is eliminated. For example, referring to Fig. 2, traffic between toll centers A and B is added to the traffic between control switching points D and E, as is the traffic between A and E, and between B and D. Traffic between toll centers B and C and between points A and D, and B and E is eliminated.

In the second network, called the "reduced-traffic" or "reduced-load"

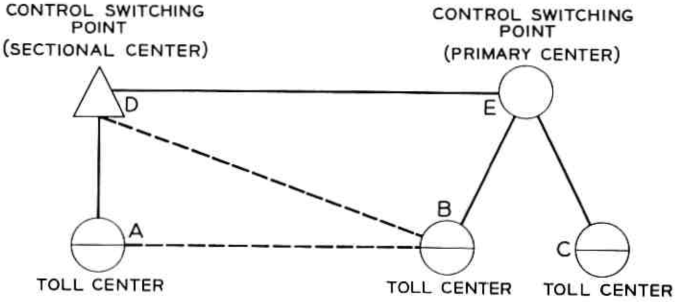

CONTROL SWITCHING POINT (SECTIONAL CENTER)
CONTROL SWITCHING POINT (PRIMARY CENTER)
D
E
B
A
C
TOLL CENTER
TOLL CENTER
TOLL CENTER

Fig. 2 — Disposition of toll center traffic.

network, it is assumed that some of the traffic between toll centers is carried on high-usage groups, and only the overflow is carried on the groups between the CSP's. Therefore traffic between toll centers such as A and B, in which there may be two routes before the CSP-CSP trunk group is reached, is assumed to overflow only 10 per cent to the D-E group. Traffic between a toll center and a CSP, such as that between A and E, may have only one route before the CSP-CSP route is reached, and 20 per cent of this traffic is then assumed to be offered to route D-E. Traffic between toll centers such as B-C, or between toll centers and the CSP on which they home, such as A-D, is again eliminated.

The net effect of these assumptions is to develop a total network load of 4764 erlangs for the full-load case, and 2031 erlangs for the reduced-load network. The maximum point-to-point load for the full traffic network is 158 erlangs, and the maximum load originating and terminating at any node is 848 erlangs. The minimum point-to-point load is 0.01 erlang, and the smallest node has 26 erlangs originating and terminating at it. For the reduced network the maximum point-to-point load is 84 erlangs and the minimum is zero. The total traffic originating and terminating at the largest node is 288 erlangs, and at the smallest is 15 erlangs. A tabulation of the total loads originating and terminating at each point in both networks is given in Table I.

### 3.3 *The Network Configurations*

Five specific networks of three configuration classes for the full-traffic model and two networks in two classes for the reduced-traffic model were examined. The first class of networks is hierarchical in structure, similar to that in use in the Bell System toll network. In these networks, trunk groups are defined as high-usage, which may overflow traffic to alternate routes, or final, which may not. The apportionment of trunks among high-usage and final routes is decided on an economic basis.[4] Both two- and three-level hierarchies were examined in the full traffic model, while only two levels were used for the reduced traffic case. The routing for these networks is shown in Fig. 3.

In Fig. 3(a) the basic elements of a two-level hierarchy are shown. Calls from node 1 to node 2 will, if unable to use the direct route, attempt to reach node 4, from which the only allowable choice is the final route 4-2. If unable to reach node 4, calls will then attempt to reach node 3, from which they will attempt the direct route 3-2, finally overflowing to the final route 3-4. Calls from 2 to 1 will reverse the procedure, attempting to reach node 3 and overflowing to node 4. Calls initially routed

TABLE I — SWITCHING CENTER LOADS

| Switching Center | Total Originating and Terminating Traffic In Erlangs | |
|---|---|---|
| | Full Load | Reduced Load |
| Bellingham, Wash. | 94.17 | 34.81 |
| Seattle, Wash. | 533.25 | 272.59 |
| Spokane, Wash. | 141.89 | 71.96 |
| Yakima, Wash. | 175.17 | 59.73 |
| Astoria, Oregon | 26.92 | 16.87 |
| Bend, Oregon | 26.65 | 15.19 |
| Klamath Falls, Oregon | 33.12 | 19.60 |
| Medford, Oregon | 62.42 | 37.08 |
| Pendleton, Oregon | 51.35 | 21.11 |
| Portland, Oregon | 468.05 | 233.56 |
| Roseburg, Oregon | 37.37 | 17.85 |
| Las Vegas, Nevada | 116.74 | 82.04 |
| Reno, Nevada | 138.21 | 69.44 |
| Fresno, Calif. | 306.11 | 140.81 |
| Modesto, Calif. | 132.55 | 81.86 |
| Stockton, Calif. | 206.21 | 105.38 |
| Redding, Calif. | 87.94 | 30.88 |
| Sacramento, Calif. | 539.32 | 173.63 |
| San Jose, Calif. | 459.24 | 188.90 |
| Oakland 4M, Calif. | 848.22 | 288.14 |
| Oakland Fr., Calif. | 524.25 | 194.06 |
| Palo Alto, Calif. | 251.14 | 108.99 |
| San Francisco, Calif. | 375.40 | 198.45 |
| San Rafael, Calif. | 134.12 | 40.26 |
| Santa Rosa, Calif. | 386.08 | 107.64 |
| Bakersfield, Calif. | 167.84 | 93.56 |
| San Luis Obispo, Calif. | 109.15 | 45.59 |
| Compton, Calif. | 447.04 | 190.15 |
| Los Angeles, Calif. | 699.27 | 242.63 |
| El Monte, Calif. | 361.88 | 179.54 |
| Van Nuys, Calif. | 426.75 | 174.32 |
| Anaheim, Calif. | 158.45 | 73.78 |
| San Bernardino, Calif. | 576.81 | 202.72 |
| San Diego, Calif. | 424.33 | 248.71 |
| Total (orig. plus term.) | 9527.41 | 4061.54 |

along the final route chains, such as those from 1 to 3, have only a single choice of route.

The three-level network, shown in Fig. 3(b), allows a somewhat more complicated routing pattern. Calls from 1 to 2 in this network will attempt to reach nodes 2, 4, 6, 5 and 3 in that order, and all other routes will follow a similar pattern of hunting from low to high level in the distant region, and from high to low level in the home region. In no event can a call use more than one interregional trunk, and calls always travel up the hierarchy in the home region and down in the distant region. A fuller description of the process is given in Ref. 3. This restrictive routing

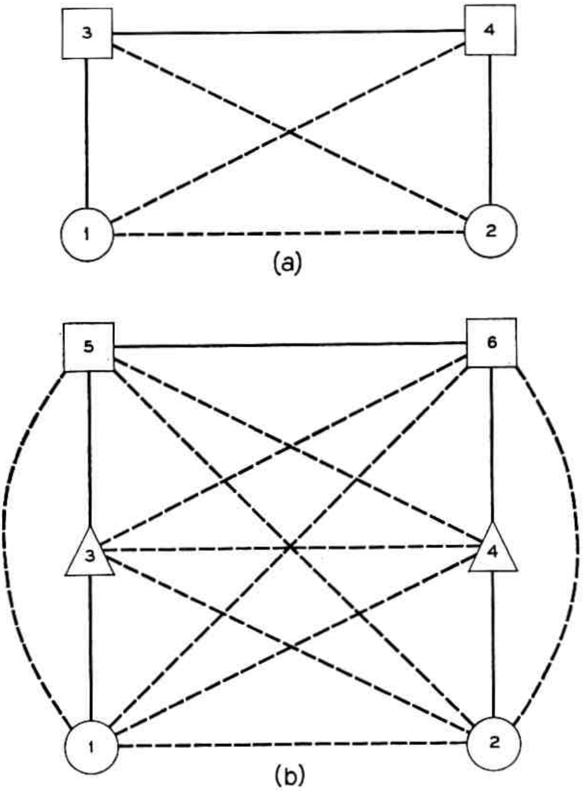


Fig. 3 — Organization of hierarchical networks.

pattern allows alternate routing to proceed without fear of "ring-around-the-rosie" or "shuttling" (which are types of looping routes), even though no information is carried with the call other than its destination code. The two-level hierarchical networks actually used contained six higher-level offices, located at Seattle, Portland, Sacramento, Oakland, San Bernardino, and Los Angeles. The three-level network took Portland, Sacramento, and San Bernardino as highest-level, or regional, centers; leaving Seattle, Oakland, and Los Angeles as middle-level, or sectional, centers. A sketch of the Washington-Oregon section of the full-load, two-level hierarchy is shown in Fig. 4.

The other network configuration examined for both load levels is the symmetrical network, in which alternate routes are selected approximately according to their total length. In all such networks studied, trunks are arbitrarily eliminated on links with less than 2 erlangs of directly offered traffic, and routing is then established using equipped links. Fig. 5 shows the trunk group layout for the Washington-Oregon section of the full-load symmetrical network. A basic restriction in all networks is that at most five outgoing choices are allowed from any node to any other, it being considered that further choices would lead to excessively circuitous routes. In addition, no route is allowed which is more than 1.5 times as long as the shortest nondirect route, or exceeds the shortest nondirect route by more than 2 links. These numbers were arrived at by trial and error and produced the most economical network for the full-load case, although they were not very critical in the determination of network cost or capability. Two symmetrical networks are studied in the full-load case, one which matches the blocking performance of the other networks at engineered loads, and one which has a higher blocking, as described below. Only one symmetrical network is used for the reduced-load model.

The method by which routes are selected is as follows. Initially, the shortest route between each two points is found. The route to the nearest neighbor node on this route is then listed as the first-choice route. The link from the originating node to the nearest neighbor node along the first-choice route is then made ineligible, and the shortest route again found. The link to the nearest neighbor node along this route is then

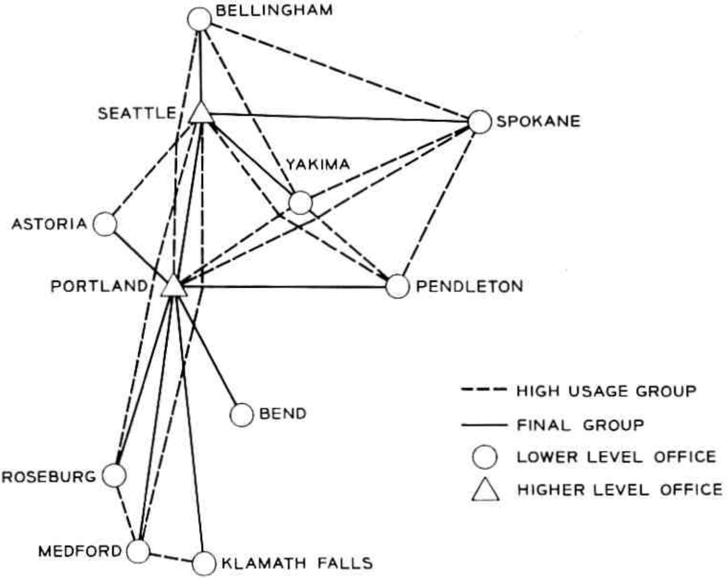


Fig. 4 — Full-load, two-level hierarchy — Washington-Oregon.

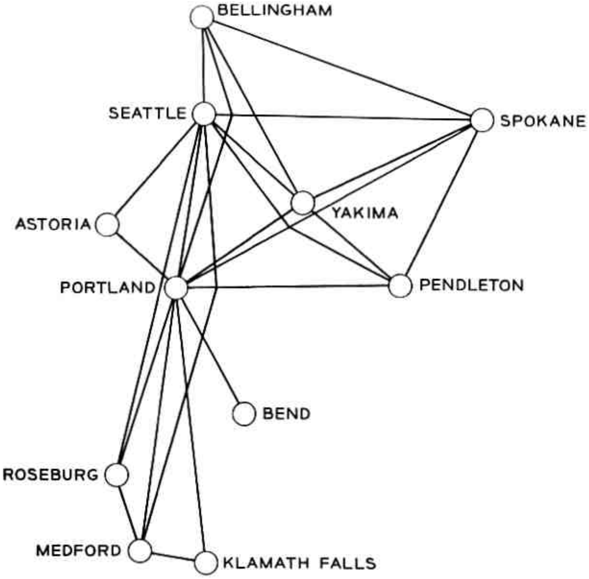


Fig. 5 — Full-load symmetrical network — Washington-Oregon.

denoted the second-choice route, and the distance and number of links calculated and compared with the first nondirect route. The entire procedure is repeated until no route falls within the distance ratio and link difference criteria, or five routes are selected, whichever occurs first. At this point the process is terminated and the routing table established. For example, in Fig. 5, to go from Yakima to Medford, the first-choice route is via Portland, the second is via Seattle, and the third is via Pendleton.

The third network configuration, considered in the full traffic case only, is the gateway network. This is essentially a two-level hierarchy with the interregional high-usage groups removed, as shown in Fig. 6 for Washington and Oregon. Traffic and trunks are therefore concentrated along the access routes to the gateway switching center and on the interregional finals. Although this kind of system clearly requires more trunks and trunk miles than a hierarchy to carry the same loads, it has been conjectured that savings in line and terminal equipment could be effected because of the large trunk cross sections involved. It also has been thought that this scheme might provide improved performance under shifting loads, which hypothesis is examined in this study. The gateway network studied assumed the gateway switches to be located at the same points as the higher-level offices in the two-level hierarchical networks.

In all of the above networks, stage-by-stage routing similar to that in the U.S. toll network is used. That is, once a call has reached a certain point in its path, its route selection is independent of its past history, and it is unable to back up and find another route out of a prior node. (This is not true if crankback is allowed, as will be discussed later.) In the symmetrical networks, the previous route is considered to the extent of preventing a call from returning to a node through which it has already been switched. In the hierarchy and gateway, this restriction is implicitly provided by the logic of the routing structure.

In sum, the networks examined are as follows:

(1) Full-load model

(a) two-level hierarchy
(b) three-level hierarchy
(c) symmetrical
(d) symmetrical with high blocking
(e) gateway.

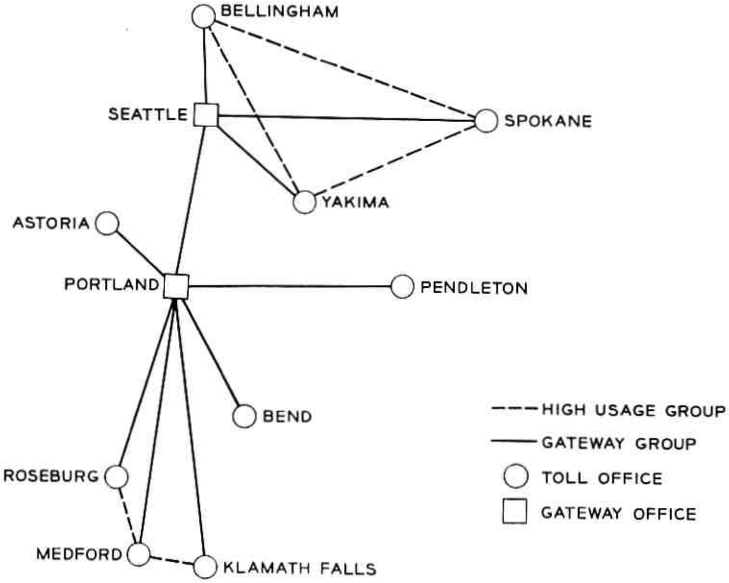


Fig. 6 — Gateway network — Washington-Oregon.

(2) Reduced-load model

  (a) two-level hierarchy
  (b) symmetrical.

### 3.4 *Engineering Procedures*

The size and complexity of the networks considered are such that manual engineering procedures or trial and error methods are not feasible. Accordingly, computer programs were prepared which established at least an initial network, which could then be adjusted if required by trial and error using the simulator. The objective for all networks was to attain an over-all average blocking probability of 0.01, with as small a dispersion of individual point-to-point probabilities as possible. This is a somewhat different criterion than the one normally used in existing hierarchical alternate routing systems, which specify the blocking probability observed on the final route, but it is closer in philosophy to local systems and others in which blocking probabilities produced by the system are the same to all customers.

The hierarchical networks were engineered with the aid of a computer program which essentially follows the procedure outlined in Ref. 4. Using this method, traffic is transferred from the direct to the alternate route when the direct route becomes so inefficient that the cost of adding a trunk to it is more than the cost of carrying the traffic on the alternate route. No account was taken of the nonrandomness of overflow traffic[5] or of the nonindependence of different links in the network. The errors resulting from these assumptions were not large and were corrected where required by trial and error using the simulator.

The process for engineering a symmetrical network is less well developed, and no method for designing an optimal, or even necessarily a very good network, exists. However, a program is in existence[6] which is capable of designing networks which will closely meet a desired blocking probability, using prespecified routes which are fully determined from origin to destination. It was necessary, in order to use this program, to convert the shortest route procedure described in Section 3.3 above to one which provides the full route rather than simply the order of hunt over the adjacent nodes. This resulted in networks which were engineered using a slightly different routing arrangement than the simulator actually used, and this, in conjunction with the basic assumptions implicit in the engineering program of random overflow traffic and independent links, led to blocking probabilities in the final network which were somewhat higher than desired. These were corrected for the purposes of comparing network configurations, but for certain studies of various methods of operating a symmetrical network, the networks with high blocking were retained.

### 3.5 *Load Changes*

Three patterns of load changes were used to measure the performance of the various networks under shifting load conditions. The two load changes examined in Ref. 1, uniform overload and overload of all traffic to and from a particular node, were avoided because of their limitations. The first case, uniform overload, represents a situation which is thought not to ordinarily occur in large real systems, and both load models are likely to obscure differences in behavior of competing networks, since the networks tend to be completely saturated or are limited by the specific overloaded nodes. Instead, three patterns of shifting loads, in which the total offered network load remained approximately unchanged, were used.

The first of these, called the "Christmas load," represents a type of shifting load normally seen in the U.S. on Christmas Day and on a few other special occasions. On these days, the normal long distance business traffic disappears and is replaced by a large volume of residential traffic. Typically, the increased traffic is of substantially longer haul than is the normal day traffic, so the phenomenon observed is that short-haul traffic decreases, but long-haul traffic increases. In order to represent this in the sample Pacific network, the network was broken down into four areas, consisting of Washington, Oregon, Northern California and Southern California. (Northern Nevada was included with Northern California and Southern Nevada with Southern California.) All intra-area traffic was reduced to 60 per cent of its normal value, and inter-area traffic was increased to from 150 to 275 per cent of its normal value, depending upon the distance. The total network load was 94 per cent of its normal value, as shown in Table II. Although these changes may appear extreme, they are not thought to be out of line with what actually occurs in the U.S. on Christmas and were applied to both full and reduced traffic models.

The second load change examined is not typical of any actual situation, but was designed to evaluate the effectiveness of the various networks in shifting load from an overloaded trunk group to a simultaneously underloaded one. In order to do this for the full-load network, all traffic items originating or terminating at the Oakland 4M machine, the largest office in the network (33 traffic items, total load 848 erlangs

TABLE II — CHRISTMAS LOAD CHANGES

| Traffic | % of Normal Day Busy Hour Load |
|---|---|
| Intraregional | 60 |
| Washington-Oregon | 150 |
| Washington-Northern Calif. | 210 |
| Washington-Southern Calif. | 275 |
| Oregon-Northern Calif. | 175 |
| Oregon-Southern Calif. | 230 |
| Northern Calif.-Southern Calif. | 150 |
| Network | 94 |

representing about 20 per cent of the network load), were either halved or doubled at random, although this was slightly modified so that the total network load remained at 99.5 per cent of its normal value. The normal and modified values of the Oakland loads are shown in Table III. In the reduced traffic model no single office had enough traffic to cause substantial changes in total network performance, so the halving and doubling were done at Seattle, Oakland and Los Angeles, which have total loads of 767 erlangs, representing about 40 per cent of the total network load. In this case the total network load increased by about 8 per cent. It is to be emphasized that this set of loads does not represent any expected realistic situation, but is a completely artificial test of the effectiveness of automatic rerouting under most favorable conditions.

The third load change examined was actually a series of load changes based on an assumed movement of the busy hour from north to south during a four-hour period. It was further assumed that, relative to the busy hour, an area's load was reduced 5 per cent in the adjacent hour, 10 per cent in the second hour, and 20 per cent in the third. Traffic between two areas (defined in the same way as for the Christmas loads) was taken to be the arithmetic mean of the levels of the terminal offices. That is, a traffic item between an area which is in its busy hour and one two hours distant is assumed to be reduced 5 per cent from its busy hour value.

Since the networks were engineered based on a single over-all load value, some normalization was done so that the over-all network load remained approximately constant. There were also some limitations in the simulation which prevented the desired changes from being reached exactly, but the final loads were quite close to the desired value. The sequence of changes is shown in Fig. 7. The first part of the line represents the basic engineered load, followed by the changes in each area's traffic as shown. The ordinate is a relative scale, so all loads are given as multiples of the basic value. The ramps between the hours were actually simulated as shown, but no measurements were taken during these periods. The inter-area traffic levels are not shown, but are arithmetic means of the levels of the terminal nodes, as described above.

This load change, which was applied to both full- and reduced-load networks, is designed to analyze a situation similar to that in the entire U.S.A., which has several time zones with a different busy hour in each. Although such differences are, of course, not actually observable in the network selected, which runs essentially north and south, we can, for the purposes of modeling, assume it runs from east to west, and is expanded in its dimensions. In this case, time zone changes like those postulated for this "busy hour load" would in fact be observed.

TABLE III — OAKLAND VARIATION LOADS (FULL-LOAD NETWORKS)

| Traffic between Oakland 4M and | Traffic Loads in Erlangs | |
|---|---|---|
| | Normal Loads | Changed Loads |
| Bellingham, Wash. | 3.58 | 7.16 |
| Seattle, Wash. | 26.64 | 13.32 |
| Spokane, Wash. | 3.31 | 6.62 |
| Yakima, Wash. | 2.39 | 1.20 |
| Astoria, Oregon | 0.23 | 0.46 |
| Bend, Oregon | 0.29 | 0.15 |
| Klamath Falls, Oregon | 1.08 | 2.16 |
| Medford, Oregon | 1.77 | 0.88 |
| Pendleton, Oregon | 0.50 | 1.00 |
| Portland, Oregon | 23.77 | 11.88 |
| Roseburg, Oregon | 0.50 | 1.00 |
| Las Vegas, Nevada | 3.78 | 1.89 |
| Reno, Nevada | 15.98 | 31.96 |
| Fresno, Calif. | 30.49 | 15.24 |
| Modesto, Calif. | 12.59 | 25.18 |
| Stockton, Calif. | 25.86 | 12.93 |
| Redding, Calif. | 8.68 | 17.36 |
| Sacramento, Calif. | 94.23 | 47.12 |
| San Jose, Calif. | 76.79 | 153.58 |
| Oakland Fr., Calif. | 122.95 | 61.47 |
| Palo Alto, Calif. | 45.47 | 90.94 |
| San Francisco, Calif. | 45.12 | 22.56 |
| San Rafael, Calif. | 26.18 | 52.36 |
| Santa Rosa, Calif. | 86.80 | 43.40 |
| Bakersfield, Calif. | 7.32 | 14.64 |
| San Luis Obispo, Calif. | 3.48 | 1.74 |
| Compton, Calif. | 33.57 | 67.14 |
| Los Angeles, Calif. | 60.73 | 30.36 |
| El Monte, Calif. | 26.74 | 53.48 |
| Van Nuys, Calif. | 28.77 | 14.38 |
| Anaheim, Calif. | 8.16 | 16.32 |
| San Bernardino, Calif. | 8.65 | 4.32 |
| San Diego, Calif. | 11.81 | 23.62 |
| Total | 848.21 | 847.82 |

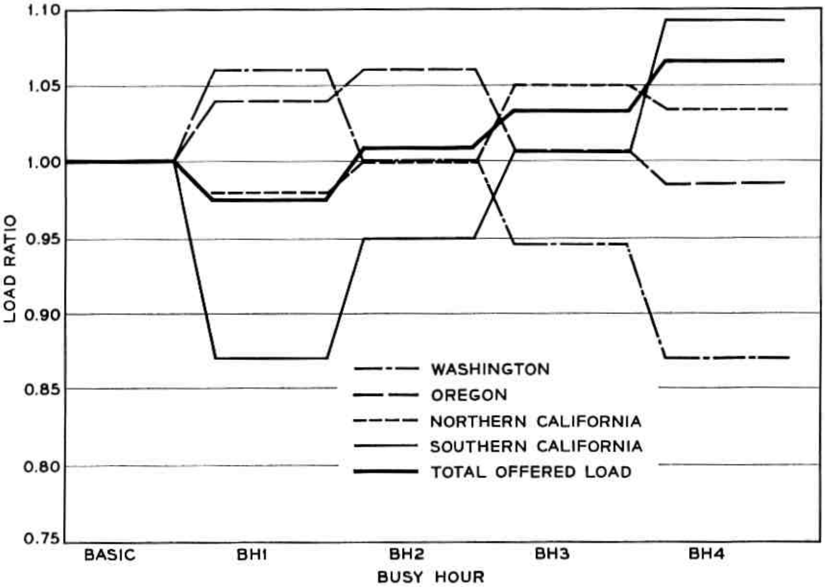


Fig. 7 — Busy hour load changes.

### 3.6 *Evaluation Criteria*

A communications network which must be engineered to meet a specific set of demands for service without being excessively costly, and which will then be subjected to demands for which it was never designed, is not easily evaluated by a single figure of merit, or even by a small number of parameters. The weight of overload performance versus engineered economy, performance under overload A as opposed to that under overload B, and service to traffic between points i and j as opposed to that provided between points k and m, provide ample opportunity for conflicting requirements. This is, of course, in addition to nontraffic considerations such as survivability, ease of engineering, administration and control, or ability to provide other services such as data and private line.

Nevertheless, in order to make a comparative evaluation of various network configurations, a set of criteria must be adopted which can be evaluated for each network under study and which will reflect the basic considerations of cost and service quality under all conditions.

The criteria which have been selected are four in number, two relating to cost and two relating to grade of service. They are:

(1) number of trunks required to provide the desired grade of service [of $P(0.01)$] at engineered load

(2) number of trunk miles required to provide the desired grade of service at engineered load

(3) the over-all weighted average blocking, $\bar{B}$, as defined in Section II above

(4) the dispersion of blocking $D_B$ (subsequently denoted simply $D$), as defined in Section II above.

Items (1) and (2) above can be provided with costs to derive approximate network costs, which will vary depending upon the cost of switching, terminal equipment and line facilities. This has been done for a few typical costs. The costs so derived are approximate because the trunk miles are defined as point-to-point airline miles, which is not the way actual facilities would normally be routed.

Items (3) and (4) are measures of service quality. $\bar{B}$ is by itself a measure of over-all network performance, and it is directly related to carried load. However, there may be severe distortions in the point-to-point blockings which would yield a low $\bar{B}$ but might still leave certain customers with extremely poor service. The inclusion of $D$ as a criterion will help to identify such a situation and ensure that network service is evaluated on a basis of balance as well as blocking level.

### 3.7 *Operating and Control Procedures*

The variations in operating procedures and the control methods employed all have the effect of changing the amount of alternate routing, normally also making different numbers of routes available to various point-to-point traffic items. Two control procedures were investigated for both the hierarchical and symmetrical networks. These were one-stage crankback and trunk reservation for first-routed traffic only (subsequently referred to simply as trunk reservation). One-stage crankback allows a call which has reached a point from which it is unable to proceed to back up one link along its previous route and attempt to complete via another route. This has been proposed as both a traffic improvement measure and as a means for allowing machine troubles to be circumvented without customer retrials. The investigation here relates, of course, only to its effect on the traffic capacity of the network. Trunk reservation allows only first-routed traffic to seize the last idle trunk in a group. Alternate routed calls can be served only if at least $m + 1$ trunks are idle, where $m$ is the number of trunks reserved. This procedure tends to maximize the number of calls which are carried on direct links at the expense of those carried on alternate routes. It also reduces group efficiencies somewhat, and the question is whether the reduction in circuitous routing is enough to compensate for this.

Finally, for symmetrical networks only, the maximum number of links per call was varied. In the case called "full routing," a maximum of five links was allowed for any call in the network. In the case called "limited routing," a maximum of only three or four links per call was allowed, depending upon the connectivity between the originating and terminating points of the call. This restriction, of course, reduces the average number of links per call, at the same time reducing the number of routes possible between any two points.

## IV. ANALYSIS OF NETWORK CONFIGURATIONS

### 4.1 *Facility Requirements*

It is difficult to arrive at an accurate measure of the cost differential between the various network configurations, since costs of trunk terminations, switching, and trunk lines vary from place to place and from network to network. However, it is expected that the relative costs of the various network configurations can be deduced from the number of trunks and the number of trunk miles by applying appropriate factors related to the distribution of trunk lengths and the types of switching and transmission equipment in general use in any given situation. If the unit costs of switching equipment, or of control features inherent in the routing plan, are significantly different for different networks, the magnitudes of these differences can be balanced against the differences in trunks and trunk miles to again deduce the total network relative costs. It should also be noted that the distances used for the trunk length calculations are based on airline mileage between originating and terminating points, which is ordinarily somewhat shorter than actual facility route mileage. This discrepancy can be corrected by introducing multiplying factors when determining network costs for any actual case.

Table IV shows the number of trunks and trunk miles required to provide the noted grade of service for each of the networks under consideration, both in absolute value and as per cent difference from the two-level hierarchy, which was arbitrarily selected as the standard. Although the blocking probabilities are not exactly the same for all networks due to inaccuracies in the engineering procedures and statistical fluctuations in the simulations, they are quite close.

The differences in facilities required for the various networks, with the exception of the gateway, are quite small, amounting to at most 4.1

TABLE IV — COMPARATIVE TRUNKING REQUIREMENTS

(a) Full-Load Networks

| Network | Trunk Miles | | Trunks | | $\bar{B}$ (Engineered) |
|---|---|---|---|---|---|
| | Actual (000) | % Diff. from 2-Level Hier. | Actual | % Diff. from 2-Level Hier. | |
| 2-level hier. | 1174 | 0 | 6659 | 0 | 0.007 |
| 3-level hier. | 1154 | −1.7 | 6679 | +0.3 | 0.008 |
| Symmetrical | 1129 | −3.8 | 6727 | +1.0 | 0.007 |
| Gateway | 1268 | +8.8 | 9236 | +38.6 | 0.010 |

(b) Reduced-Load Networks

| Network | Trunk Miles | | Trunks | | $\bar{B}$ (Engineered) |
|---|---|---|---|---|---|
| | Actual (00) | % Diff. from 2-Level Hier. | Actual | % Diff. from 2-Level Hier. | |
| 2-level hier. | 6047 | 0 | 3298 | 0 | 0.008 |
| Symmetrical | 5801 | −4.1 | 3256 | −1.3 | 0.008 |

per cent difference in trunk miles and 1.3 per cent difference in trunks between the symmetrical and hierarchical reduced-load networks.

The gateway network requires a much larger number of trunks and trunk miles than any of the others, reflecting the fact that many calls which in the other networks require only one link must use three in the gateway, and the fact that there is much excessive routing, or "back-haul" in traffic which is obliged to switch through gateways. In this case the resulting cost difference represents the savings in switching and line costs which would have to be achieved to offset the increased quantities of equipment required.

### 4.2 *Costs*

Table V gives the costs of the various networks, assuming a range of ratios of line to terminal costs which should include most actual situations. The differences between the hierarchical and symmetrical networks are quite small, as is that between the two- and three-level hierarchies, leading to a tentative conclusion that in these cases cost differential is not a primary reason for selection of one network over another. It must be remembered, however, that the hierarchy (two-level) was engineered using a known and proven economical procedure, while no such method is available for the symmetrical networks. Therefore, the hierarchies are probably close to optimal, while some additional economies might ultimately be realized for the symmetrical networks.

TABLE V — NETWORK COSTS

(a) Full-Load Networks

| Cost/Trunk Mile | Using $1500 Trunk Termination and Switching Costs | | | | | |
|---|---|---|---|---|---|---|
| | $10/Trunk Mile | | $50/Trunk Mile | | $100/Trunk Mile | |
| Network | Cost $(000,000) | % Diff. from 2-Level Hier. | Cost $(000,000) | % Diff. from 2-Level Hier. | Cost $(000,000) | % Diff. from 2-Level Hier. |
| 2-level hier. | 21.72 | — | 68.69 | — | 127.39 | — |
| 3-level hier. | 21.56 | −0.74 | 67.72 | −1.34 | 125.42 | −1.55 |
| Sym. | 21.38 | −1.57 | 66.54 | −3.13 | 122.99 | −3.45 |
| Gateway | 26.53 | +17.54 | 77.25 | +12.46 | 140.65 | +10.41 |
| (b) Reduced-Load Networks | | | | | | |
| 2-level hier. | 10.99 | — | 35.18 | — | 65.42 | — |
| Sym. | 10.69 | −2.73 | 33.89 | −3.67 | 62.89 | −3.86 |

Using the $50 per trunk-mile line cost figure, the data shown in Table V(a) indicate that there is about a 1.3 per cent savings in cost of transmission and switching facilities for a network of this size with full loads when a three-level rather than a two-level hierarchy is used, and another 1.8 per cent if a symmetrical network is considered. This, of course, does not include any differences in signaling and control equipment which might be required to implement a symmetrical network, nor can it take account of the nonoptimality of the network engineering procedure now in use.

The gateway network, as expected, costs about 12 per cent more than a hierarchy to carry the same traffic, assuming that trunk and terminal costs are the same for all networks. This difference will be somewhat mitigated by the fact that trunk and facility routes are more likely to be identical in the gateway than in the hierarchical configuration, and therefore the multiplier to convert from airline miles to facility miles may well be smaller. In addition, any savings in switching costs which can be effected because of the large volumes of traffic flowing through the gateway switches will of course work to the advantage of the gateway plan.

Table V(b) indicates that the reduced-load symmetrical network is about 3.7 per cent less expensive than the hierarchy. This is in agreement with earlier results[1] which indicated that lightly loaded networks show greater differences between configurations than do heavily loaded networks.

### 4.3 *Overload Performance*

Fig. 8 shows the over-all average blocking probability, $\bar{B}$, for four different full-load network configurations (two-level hierarchy, three-level hierarchy, symmetrical, and gateway), and for the load changes discussed in Section 3.5 above. The "base" load under the "BH Runs" heading represents the same average load as the "engineered" point. It is a shorter run, however, and any difference in blocking between the two points is due to statistical fluctuations. Only the points on the charts are meaningful, but lines have been drawn connecting them for visual clarity. Fig. 9 is a similar chart, showing the dispersion factor, $D$. Figs. 10 and 11 show the same factors for the reduced-load networks, where only the two-level hierarchy and symmetrical networks were examined.

Although there are apparently some differences in performance between the various networks under various load change conditions, it is clear from Figs. 8 and 9 that there is no single superior network configuration in terms of traffic capacity and performance under shifting loads

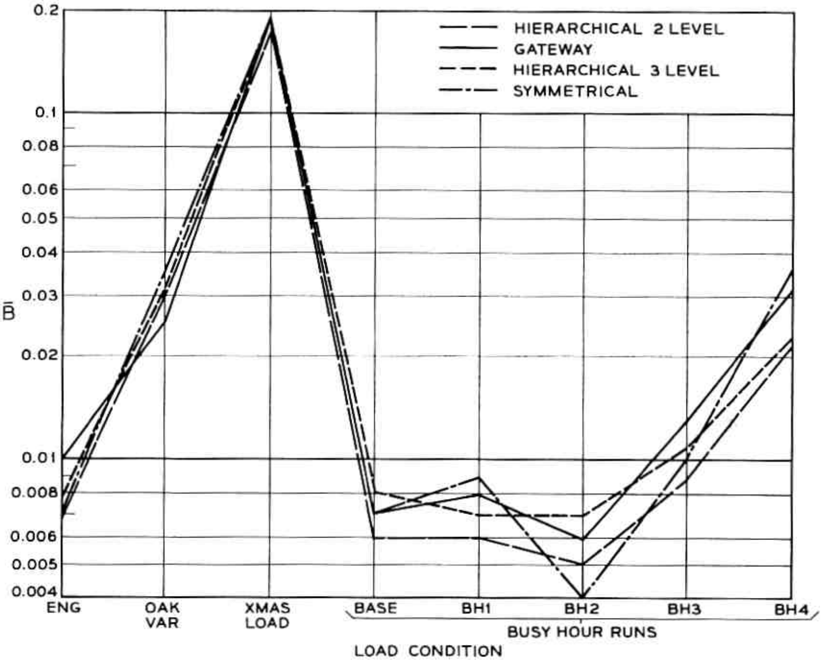


Fig. 8 — Over-all average blocking, $\bar{B}$: comparison of network configurations, full-load networks.

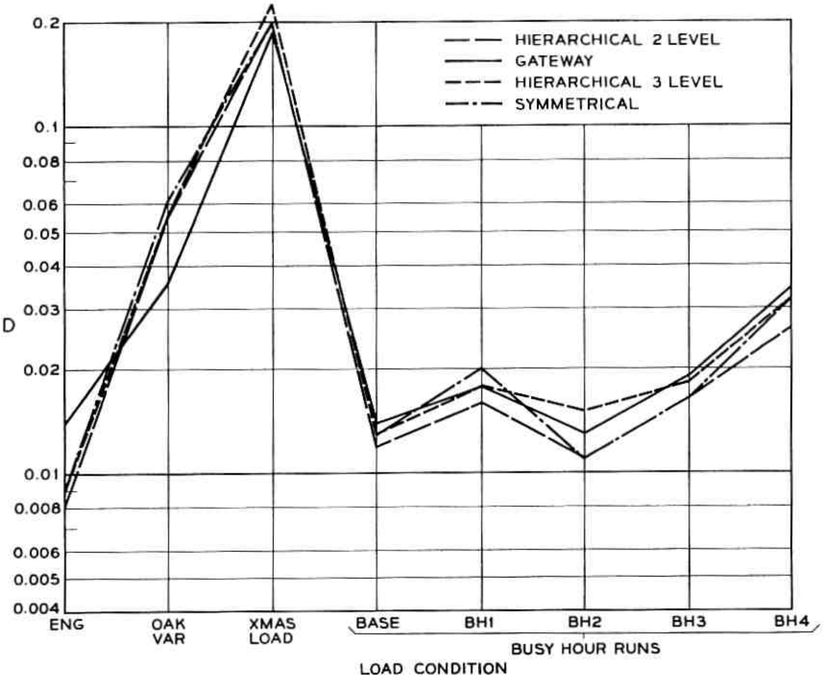


Fig. 9 — Dispersion of blocking, $D$: comparison of network configurations, full-load networks.

at full-load levels. Some small systematic differences are present, such as the fact that the two-level hierarchy appears to give slightly lower blocking than all other networks at all points except the Oakland variations case, where the gateway shows up best. This can, however, be a result of the initial engineered blocking level, which is slightly lower for the two-level hierarchy than for the three-level hierarchy or the gateway. This initial point does not so much denote a difference in performance under changed loads as it does the slight inaccuracies in engineering level, which are then reflected at every point on the chart. Although the simulation runs which produced these measurements used the identical set of calls for all networks at each load, the standard deviation of the results due to the finiteness of the simulation run is of the order of magnitude of the blocking probability at each point, and firm conclusions can be drawn only if a distinct superiority of one configuration over another manifests itself at almost all of the points considered. There are some such uniform results, but the differences are quite small, and may be offset by the differences in cost discussed above.

Figs. 10 and 11, on the other hand, show a small advantage for reduced-load hierarchical networks under all changed load conditions. In this case there is no initial error, and all evidence indicates that the hierarchy is slightly superior. It must be remembered, however, that the hierarchy costs somewhat more in this case, and this sensitivity to overloads may simply be the penalty paid for a more economical network at engineered loads.

The conclusion which must be reached from these results is that, for large networks with fairly high traffic densities, the performance of various alternate routing configurations in terms of traffic capacity under changing load conditions is quite similar. The reason for this is probably that the very density of traffic in these networks causes many of the trunk groups to be quite efficient, and the great bulk of the traffic is carried on the direct routes. Differences in alternate routing configuration, therefore, affect only a small proportion of the total traffic, with a correspondingly small effect on the network performance. In more lightly loaded networks, as has been observed, the differences are greater as

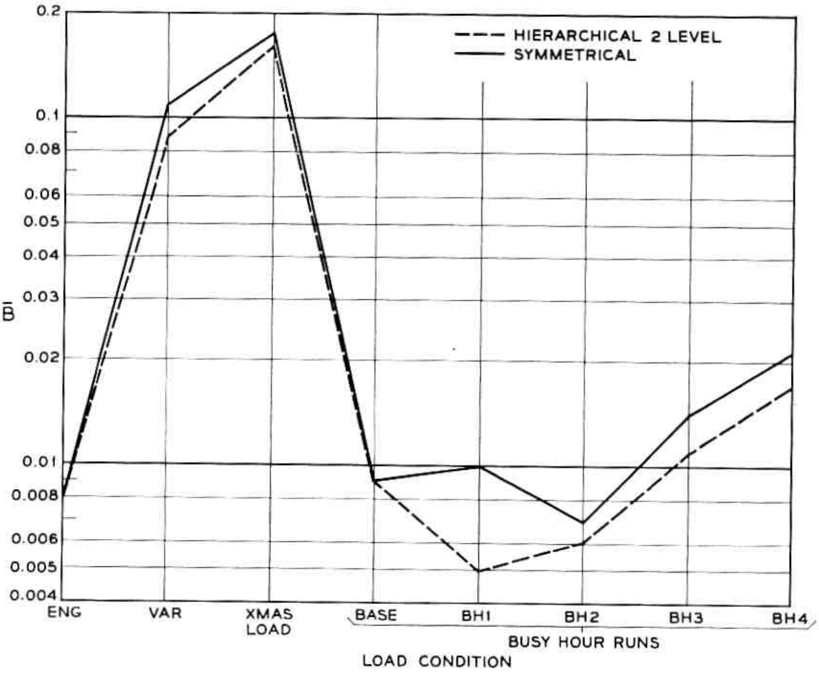


Fig. 10 — Over-all average blocking, $\bar{B}$: comparison of network configurations, reduced-load networks.

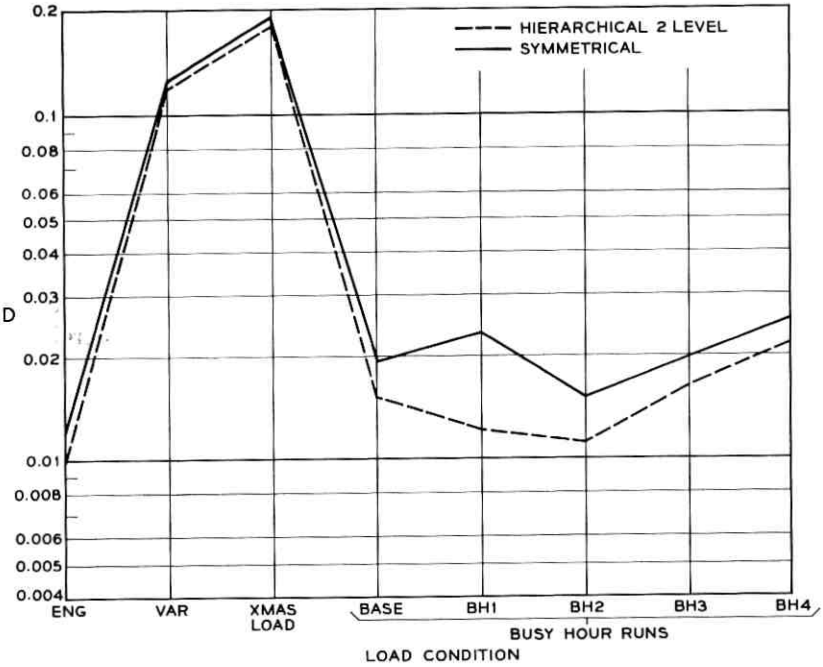


Fig. 11 — Dispersion of blocking, $D$: comparison of network configurations, reduced-load networks.

more of the traffic is alternate routed. Even in these cases, however, the differences are not large, and the comparison made here between symmetrical and hierarchical networks shows the slight superiority of one in cost to be offset by better performance of the other under shifting loads.

## V. ANALYSIS OF OPERATING AND CONTROL PROCEDURES

### 5.1 *Full versus Limited Routing*

As discussed earlier, symmetrical networks were operated in two ways. In the first of these, called "limited routing," a maximum of three or four links per call was allowed, depending upon the connectivity available to the traffic parcel. In the second, called "full routing," five links per call were allowed for all calls. Fig. 12 shows a comparison of the over-all average blocking for these two cases. It is clear from this figure that operation with limited routing is superior in traffic handling capacity. Although the differences at any point are still small, and the statistical variability of the results large, the fact that there is an advantage for the limited routing case for every point tested indicates that this is a real effect, and not merely a result of chance observation. Furthermore, since these two curves represent the same network in terms of trunk layout, there is no possibility of complicating or compensating factors due to cost differences or engineering errors. In fact, the difference in blocking probability under engineered loads in this case does not represent an engineering error, but instead an additional verification of the fact that operation with limited routing is superior. This result is further evidence of the fact that excessive alternate routing can cause service deterioration, even under light load conditions. (The routing used in the symmetrical networks discussed earlier was limited routing, chosen because it gave superior performance.)

The symmetrical network whose performance is plotted in Fig. 12 and in subsequent graphs is clearly not identical to that discussed previously, since the blocking probabilities at all points are somewhat higher.

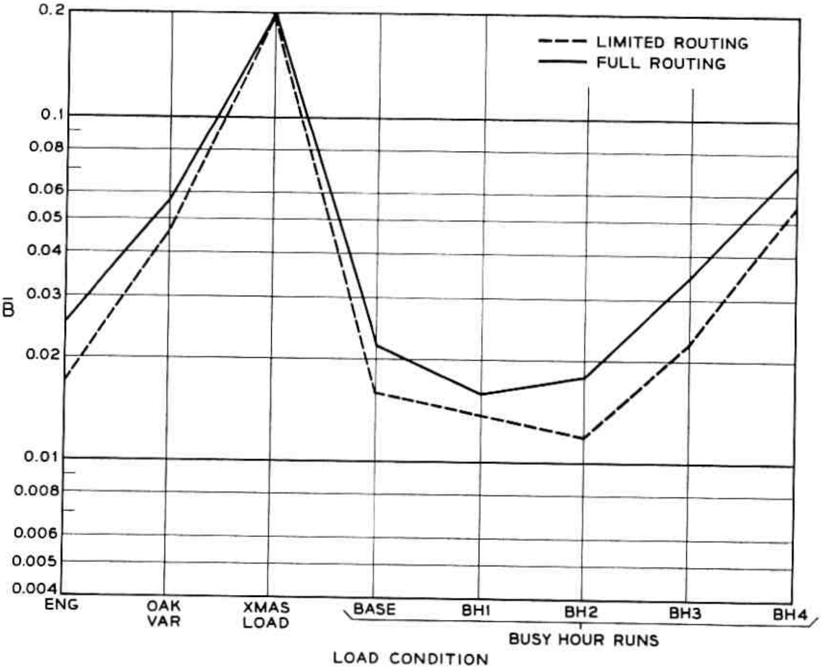


Fig. 12 — Over-all average blocking, $\bar{B}$: full vs limited routing; full-load, high-blocking, symmetrical network.

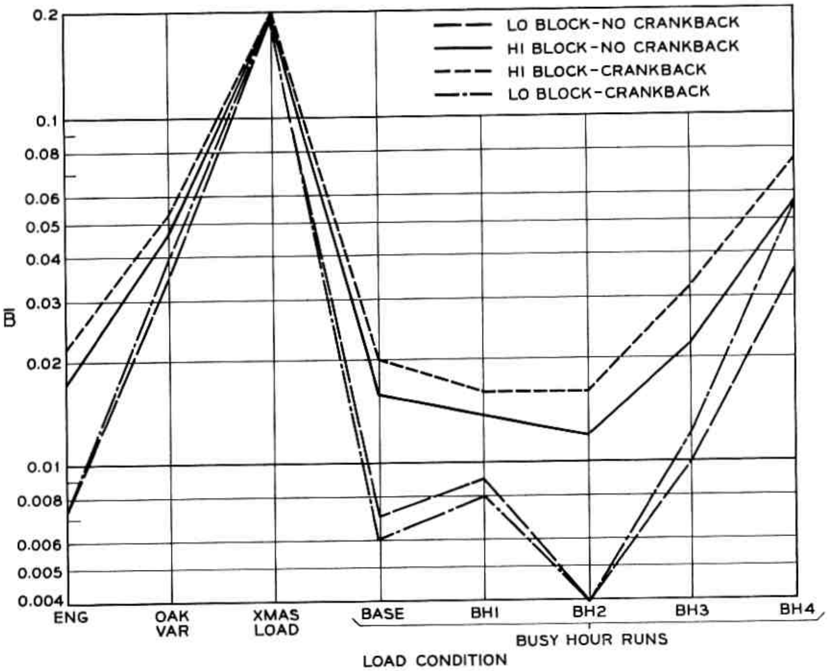


Fig. 13 — Over-all average blocking, $\bar{B}$: full-load symmetrical networks, effect of crankback.

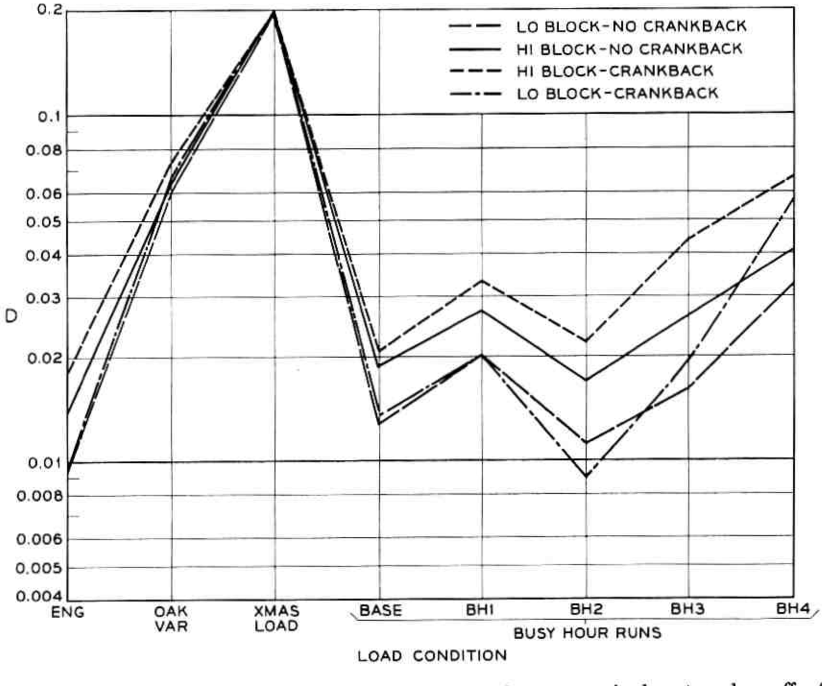


Fig. 14 — Dispersion of blocking, $D$: full-load symmetrical networks, effect of crankback.

This network, however, is the one which originally resulted from the engineering program, and it will be used for all studies concerned with differences in operating method using the same network. The earlier comparisons between network configurations on a basis of both cost and performance required that the blocking probability be approximately equal at engineered loads, and trial and error modifications were made to the symmetrical network to bring its blocking probability down to the proper level. The comparisons between different modes of operation of the same network should not be significantly affected by exact level of blocking at engineered loads and are expected to be valid for all networks of approximately the traffic densities considered. In the investigation of the effectiveness of crankback, however, symmetrical networks with both low and high over-all blocking probabilities were examined.

### 5.2 *Crankback*

Comparisons of networks operating with and without crankback were made for hierarchical and symmetrical networks using both the full traffic and reduced traffic models. In the hierarchical networks, no significant difference in behavior could be detected between the networks operated with and without crankback. This is because the structure of the hierarchical network is such that most of the blocking occurs on final route links, which are impossible to avoid even with the crankback option. For example, in Fig. 3(a) if a call from 1 to 2 is blocked at node 4, it may, with crankback, back up to node 1 and attempt to reach node 2. Even if this is possible, however, there is still a large probability of being blocked on route 3-2, and hence rearriving at 4 at some later time. In addition, those calls which do get through using the crankback option tend to use relatively long routes, causing later calls between other points to be blocked.

The over-all network blocking and dispersion of blocking for symmetrical networks with and without crankback are shown in Figs. 13 and 14 for full-load networks and in Figs. 15 and 16 for reduced-load networks. Figs. 13 and 14 have curves for both the symmetrical network as originally engineered (hi-block) and the corrected symmetrical network which was used for comparison with the hierarchy (lo-block). This was done so that any differences introduced by the general level of blocking would be apparent. The curves indicate that the level of blocking has, at most, marginal significance at these loads, and that crankback degrades the network performance at all but the lowest blocking levels, when it has virtually no effect. The networks here are operated with limited routing, but a similar test with full routing yields results which are substantially identical to those shown.

Figs. 15 and 16 show that crankback does offer a small advantage for less heavily loaded networks, although this advantage tends to disappear as the load increases, regardless of its distribution.

These results indicate that for large networks, operation with crankback at best offers a slight improvement in service when the service is good, and makes matters worse when the situation begins to deteriorate. An examination of the trunk occupancies and number of links per call shows that operation with crankback generally causes a larger number of links per call to be used on the average, with a higher over-all trunk occupancy. In effect, it therefore increases the amount of alternate routing allowed, and not always in the best way, so that degradation under overloads is a certainty. It therefore must be recommended that this device not be incorporated into large switching networks unless survivability, improved reliability, or other factors dictate it. If it is incorporated into a network for reliability or other purposes, means should be made available to disable it under overloads.

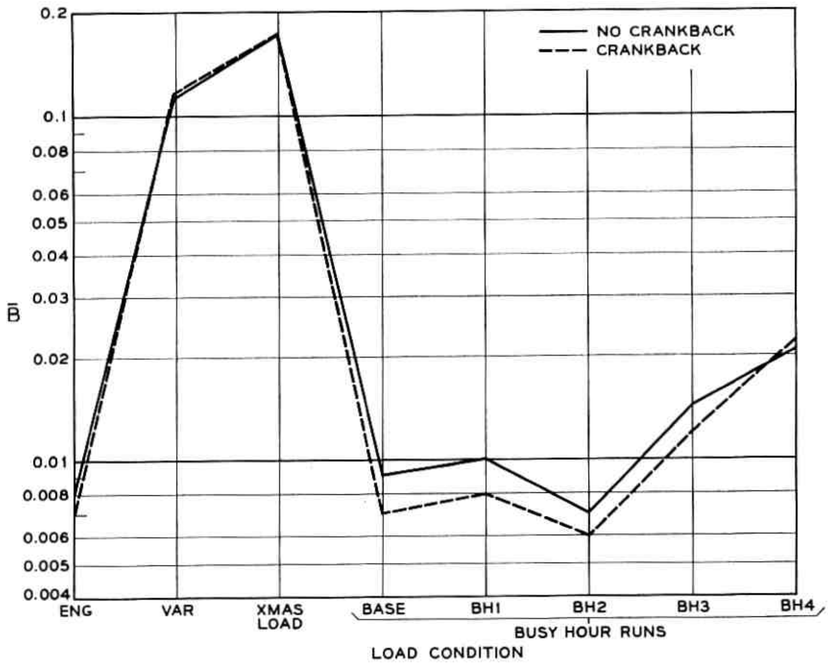


Fig. 15 — Over-all average blocking, $\bar{B}$: reduced-load symmetrical network, effect of crankback.

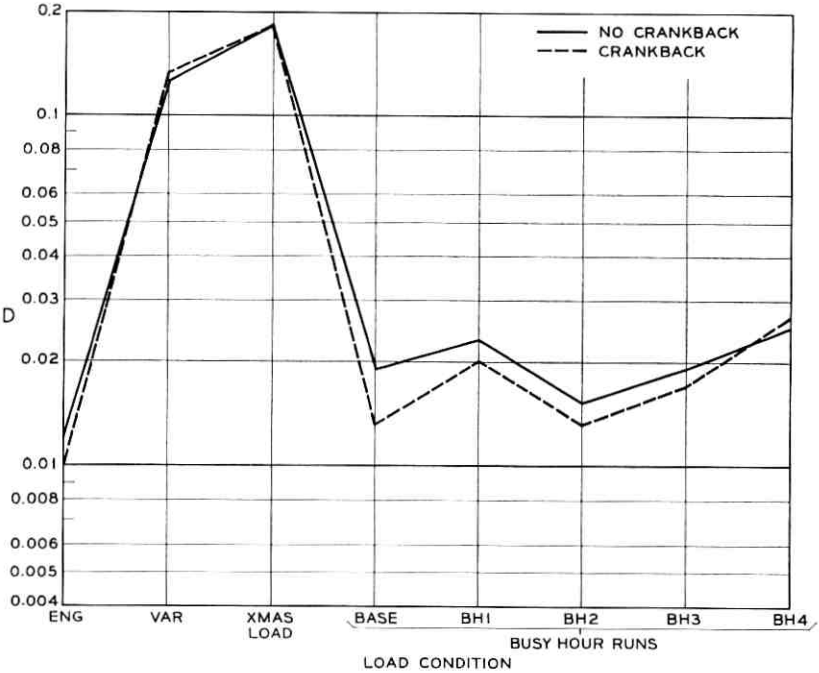


Fig. 16 — Dispersion of blocking, $D$: reduced-load symmetrical network, effect of crankback.

### 5.3 *Trunk Reservation for First-Routed Traffic*

Figs. 17 through 22 show the effect of trunk reservation for first-routed traffic on the blocking and dispersion of full- and reduced-load symmetrical and hierarchical networks. This measure, which reduces the amount of alternate routing on a selective basis, provides a uniform improvement in performance for all networks shown, although the improvement is more marked in the case of full-load than in reduced-load networks. The two-level hierarchies were not noticeably affected by the introduction of this measure.

In general, one trunk was reserved in each trunk group in the network, although two trunks were reserved on every group in some cases. It was generally found that reserving more trunks than noted in the charts had little additional effect upon the network performance. Figs. 17 and 18 show the effect of trunk reservation on symmetrical full traffic networks. It is interesting to note that the network with full routing has almost identical performance to the network with limited routing when trunk reservation is used. This is not illogical,

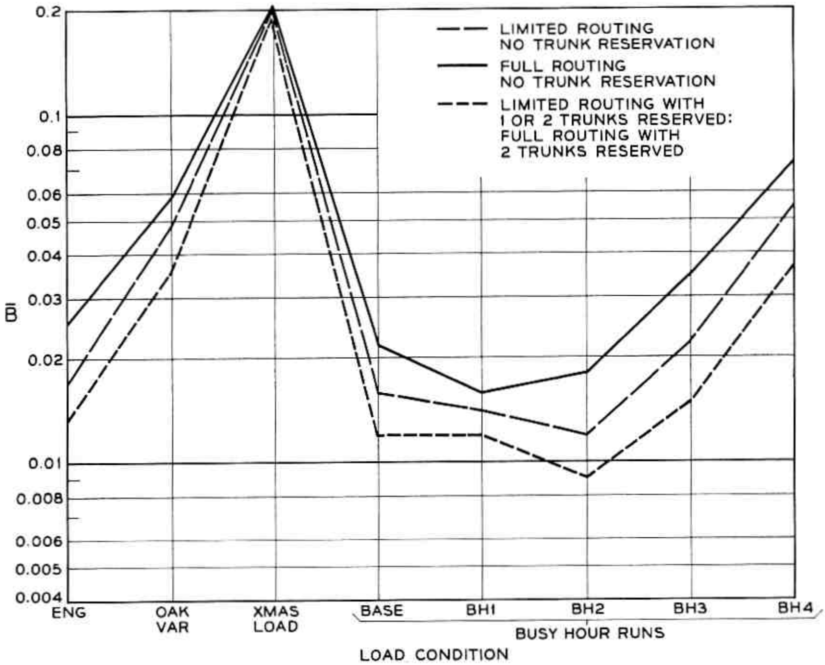


Fig. 17 — Over-all average blocking, $\bar{B}$: full-load, high-blocking symmetrical network, effect of trunk reservation.

since trunk reservation has a gross effect similar to that introduced by limiting the number of links per call.

Figs. 19 and 20 show the effect of trunk reservation on a three-level hierarchical network, and here we observe an improvement similar to that seen in the examination of symmetrical networks.

Figs. 21 and 22 show the blocking and dispersion for the reduced-traffic symmetrical network, in which the effect is similar but of lesser magnitude than that observed in the full-load networks.

It is quite likely that a selective application of trunk reservation to those groups which are large and have a large proportion of alternate routed traffic would be more effective than the across the board application used here. However, this study suffices to show that there is an advantage in the traffic handling capability of a network so equipped, and more detailed analysis will be required to determine the best number of trunks to be reserved in any given case.

Trunk reservation has essentially the opposite effect on the network as crankback; it reduces the amount of alternate routing during periods of momentary congestion, preventing calls from being completed using circuitous routes at such times. Subsequent calls are then not affected and the over-all network performance is improved.

One test was made using both trunk reservation and crankback, but the effect of trunk reservation appeared to dominate, and no difference was observed whether crankback was or was not used.

## VI. CONCLUSIONS

The first and most obvious conclusion to be drawn from the preceding results is that for networks with a high traffic density the selection of routing doctrine and control philosophy does not have any great effect upon the traffic handling capability of the trunking network. This fact is apparently due to the substantial trunk group size generally encountered in such networks, with the basic group efficiency sufficiently large to obviate any spectacular improvements due to clever routing or control schemes. Of course, these comments apply only to reasonable alterna-

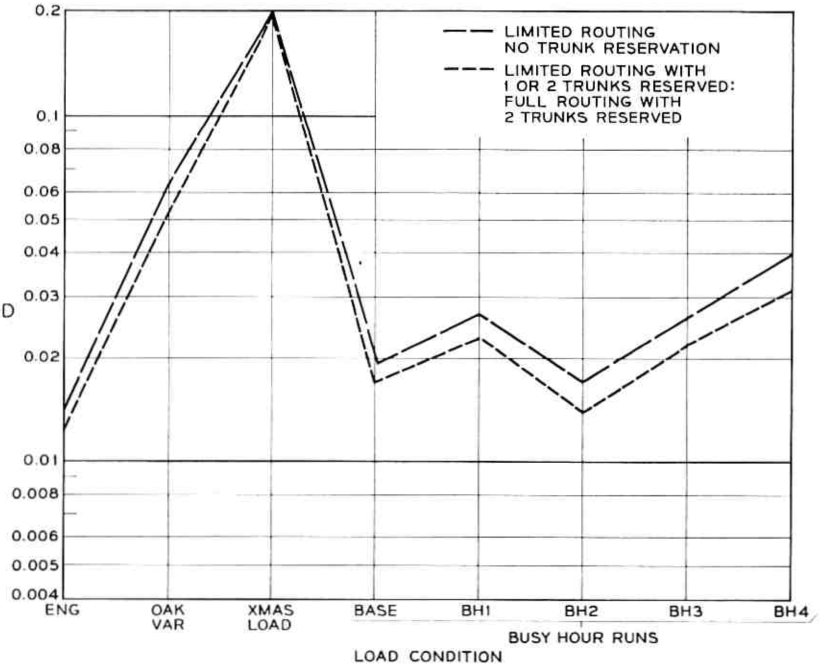


Fig. 18 — Dispersion of blocking, $D$: full-load, high-blocking symmetrical network, effect of trunk reservation.

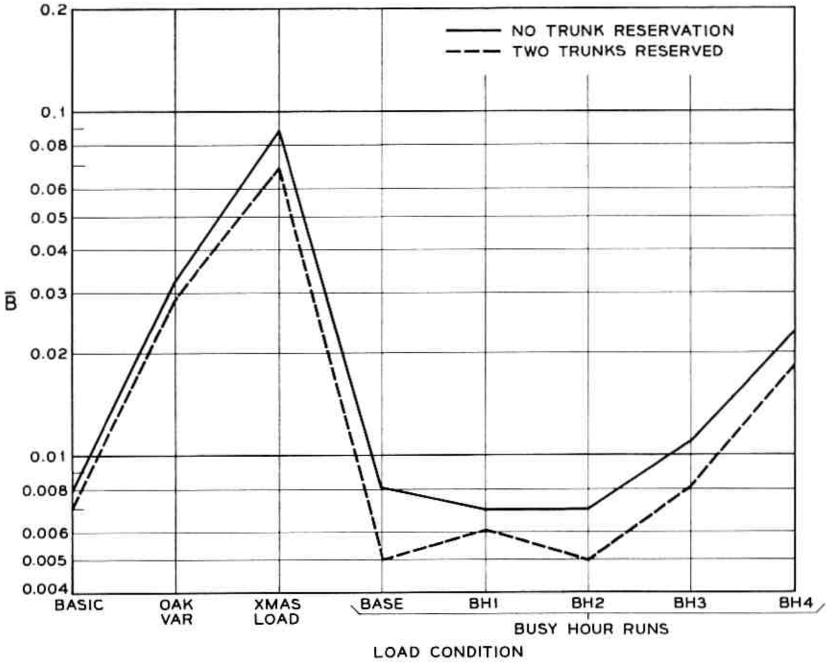


Fig. 19 — Over-all average blocking, $\bar{B}$: three-level hierarchical full-load network, effect of trunk reservation.

tives, such as those examined here. It is possible to develop a routing plan which would encourage circuitous routing at the expense of direct. Such a scheme would almost certainly show significantly poorer behavior than any of the networks investigated.

Planning for future networks should then be initially concerned with other factors, such as economics, survivability, flexibility and so forth, with a precise evaluation of traffic capacity to be determined after the fundamental design considerations are well formulated.

Having once accepted the basic idea that all differences are small in magnitude, we can nevertheless observe their direction, and, in the event that there are no other significant factors, decisions can be made on the basis of such small differences. A saving of one per cent in the toll trunk plant in the U.S.A. alone, for example, would amount to many millions of dollars, which is not insignificant in magnitude, even though it is a small fraction of the total network cost.

In the comparison of network configurations, the symmetrical networks have some cost advantages, particularly at lower load levels. This is to some extent offset by a tendency to deteriorate under overload slightly more rapidly than hierarchical or gateway networks. Furthermore, there is likely to be a not insignificant additional cost connected with the operation and control of such networks, and they are difficult to engineer and administer. They do have the advantage of improved survivability, however, since there is not so much concentration of facilities at regional switching centers.

The gateway network behaves well under overloads, but requires too high an initial cost to warrant its use with existing technology. If technological advances radically change the patterns of costs for such a network, then the gateway may be a suitable selection. The survivability aspects of these networks are particularly important, since sections of the network can be isolated by the destruction of a few critical points.

The hierarchical networks, which were the first alternate routing networks to be put into service, show a competitive initial cost and a reasonable reaction to shifting loads of all sorts. They are simple to engineer

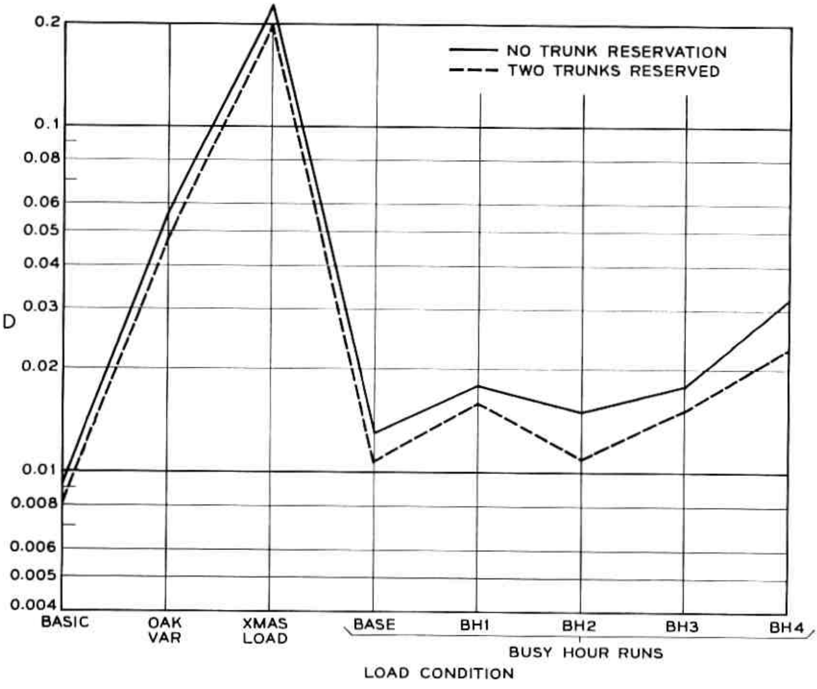


Fig. 20 — Dispersion of blocking, $D$: three-level hierarchical full-load network, effect of trunk reservation.

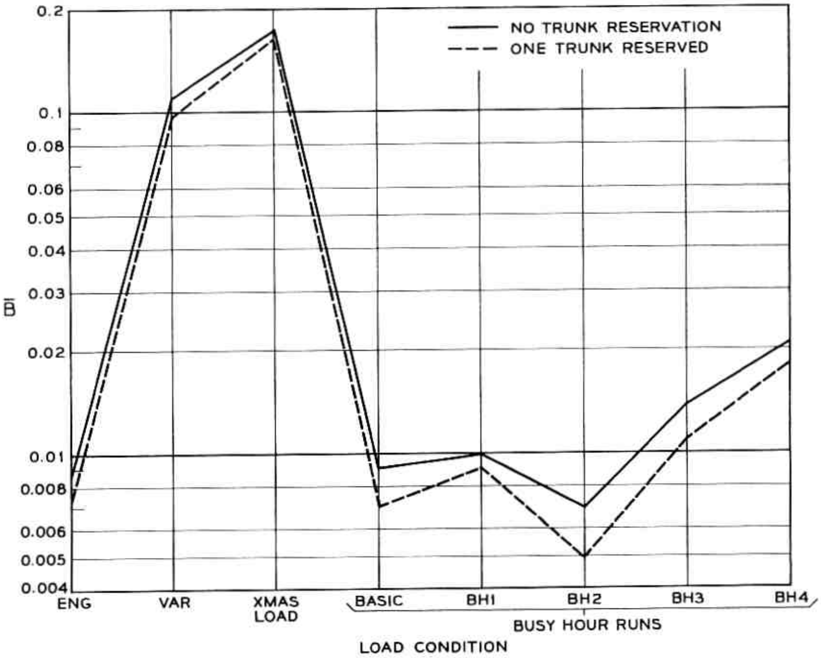


Fig. 21 — Over-all average blocking, $\bar{B}$: reduced-load symmetrical network, effect of trunk reservation.

and administer, and the logic associated with switching and routing control is relatively uncomplicated and economical. They pose an obvious survivability problem, since some traffic parcels have access to only a single route. This situation can be largely alleviated by dispersion of routes and liberal provision of high-usage groups.

In short, if a high-density communications network is desired, and concentration of traffic along backbone routes is allowable, then a hierarchical network is likely to be the best choice of network structure. As the traffic density declines, the symmetrical networks begin to show to advantage, and they are indispensable in some form if the survivability requirement rules out hierarchies. Symmetrical networks should, however, be implemented only in conjunction with an operating technique such as trunk reservation to maintain overload capability.

The investigations of control measures demonstrate conclusively that crankback is ineffective or harmful in all networks except perhaps those with extremely light traffic densities. It offers at most a small gain at engineered loads, and aggravates undesirable overload effects. There would therefore appear to be no reason for providing it other than the nontraffic one of improving the ability of a call to avoid an equipment malfunction. If it is used for this purpose it should be disabled under overload, when it shows the greatest traffic disadvantage.

Trunk reservation, on the other hand, almost always improves the traffic carrying capacity of networks, and is never harmful. It is an inexpensive measure to implement which is unquestionably worth using, and further studies of the strategy and extent of its use should be undertaken.

In sum, the basic factors relevant to the design of communications networks are:

(1) If there is a high density of traffic, and traffic concentration on backbone routes is allowed, then a hierarchical configuration probably should be selected, with the number of levels dependent upon the particular situation.

(2) If the traffic density is lower and/or the hierarchy is unacceptable

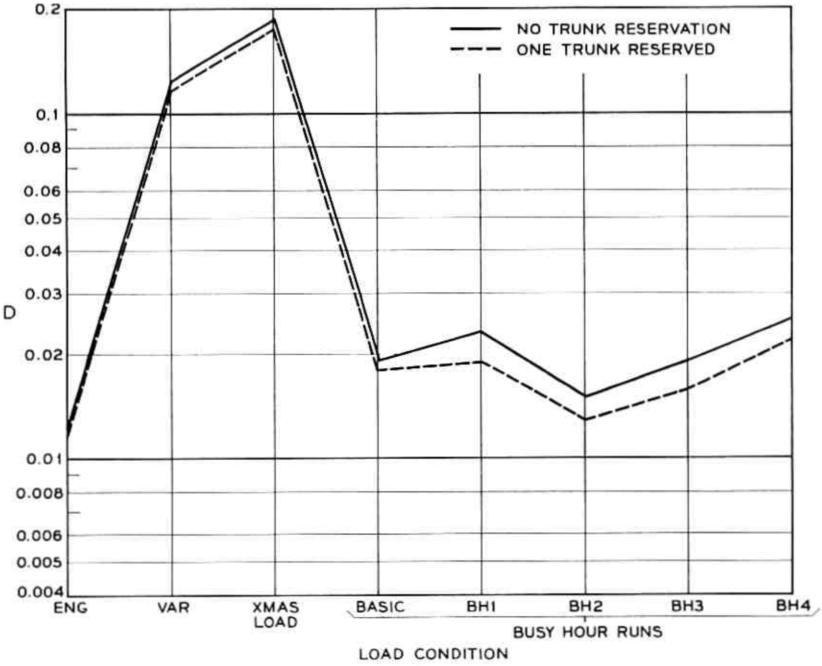


Fig. 22 — Dispersion of blocking, $D$: reduced-load symmetrical network, effect of trunk reservation.

for survivability reasons, then a symmetrical network may be more economical and can perform well if properly controlled.

(3) Crankback should not be used, except possibly as a means of alleviating the effects of equipment troubles. If used, its traffic disadvantages under overloads should be taken into account.

(4) Trunk reservation should be widely employed, since it is simple to implement and has noticeable traffic advantages under all load conditions with almost any network configuration.

Although these guidelines are, of course, qualitative in nature, this is necessary because of the large number of variables which exist in an actual network. Variations in traffic levels between and within networks, geographical distributions of switching offices and densities of traffic, equipment limitations and differing primary functions all lead to different constraints and weightings of various factors. It is the purpose of these studies to provide guides for the design of communications networks, with final choices dependent upon specific factors.

## VII. ACKNOWLEDGMENTS

This work could not have been done without the efforts of L. A. Gimpelson in organizing, and Mrs. E. E. Bailey and Misses G. C. Watling, S. A. Switch and A. Malone in programming, the simulation. Miss M. Lynch prepared most of the voluminous input data.

## REFERENCES

1. Weber, J. H., Some Traffic Characteristics of Communications Networks with Automatic Alternate Routing, B.S.T.J., **41,** Mar., 1962, pp. 769–796.
2. Gimpelson, L. A., and Weber, J. H., UNISIM—A Simulation Program for Communications Networks, Fall Joint Computer Conference, San Francisco, Oct., 1964.
3. Pilliod, J. J., Fundamental Plans for Toll Telephone Plant, AIEE Trans., **71,** pt. I, 1952, p. 257; also in B.S.T.J., **31,** July, 1952, p. 832.
4. Truitt, C. J., Traffic Engineering Techniques for Determining Trunk Requirements in Alternate Routing Trunk Networks, B.S.T.J., **33,** Mar., 1954, p. 277.
5. Wilkinson, R. I., Theories for Toll Traffic Engineering in the U.S.A., B.S.T.J., **35,** Mar., 1956, pp. 421–514.
6. Segal, M., Traffic Engineering of Communications Networks with a General Class of Routing Schemes, Fourth International Teletraffic Conference, London, July, 1964.

# An Experimental Study of Near-Field Cassegrainian Antennas*

By D. C. HOGG and R. A. SEMPLAK

(Manuscript received May 30, 1964)

*The near-field Cassegrainian antenna is a double-reflector system that employs, in its simplest form, confocal paraboloids. Unlike the standard Cassegrain which employs a hyperboloidal subreflector illuminated by a spherical wave, the near-field device is fed by a uniform phase front. Experimental data on noise performance, gain, and radiation patterns have been obtained at a frequency of 6 gc using two 16-foot paraboloids (focal length-to-diameter ratios of 0.375 and 0.25) in both standard and near-field configurations.*

*Using the shallow antenna, zenith noise temperatures of 10°K and 6°K were obtained for the standard and near-field systems, respectively; at an elevation angle of 10° the antenna temperatures were 50°K and 20°K. Using the deep secondary reflector, zenith noise temperatures of 4°K were obtained for both configurations; at 10° above the horizon, however, the standard Cassegrain has an antenna temperature of 30°K and the near-field device 13°K. In all cases, the antenna efficiencies are not far above 50 per cent. Discussion of noise produced by various methods of mounting subreflectors is included. Since noise produced by transmission lines and antenna environment is closely related to these experiments, it is discussed in detail in appendices.*

## I. INTRODUCTION

Large microwave antennas of high efficiency and low noise are desirable in radio astronomy, in tracking of space probes and in satellite communications. In all of these cases, convenient access to the associated electronic equipment is also a desirable feature. The horn reflector[1,2] is an antenna which provides this access and also admirably satisfies the electrical requirements. Nevertheless, it is of interest to

* Part of this material was presented to the URSI in Washington, D. C. (May, 1962).

examine other types of microwave antennas of more favorable ratio of geometrical aperture to total size with a view to improvement of their electrical performance toward that of the horn reflector.

The purpose of this study is twofold: to evaluate the near-field Cassegrain as a microwave antenna, and to compare its noise performance with that of other antennas. Actually, two 16-foot diameter paraboloids have been tested, one with an $f/D$ ratio of 0.375 and the other of 0.25. Measurements of antenna noise temperature, gain, and radiation patterns were made at a frequency of 6 gc using various feeding arrangements on both of these main reflectors.

Most paraboloids have relatively low aperture efficiencies and exhibit poor noise performance. For example, paraboloids fed by a horn at the focal point typically have intrinsic (back lobe) noise temperatures of 20 or 30 degrees Kelvin,[3,4] whereas the equivalent noise for the horn reflector is about 2°K.[5,6,7] This noise is due to thermal radiation from the environment of the antenna (mainly the ground) into the wide-angle or back lobes of the antenna; in what follows, it is designated by $T_b$.

Paraboloids fed by a source at the focus suffer from another deficiency: either the first circuit of the receiver must be mounted at the focal point (which is inconvenient), or a rather long transmission line (which results in a prohibitive increase in noise) must be provided. This undesirable feature is overcome by use of the Cassegrainian configuration[8,9] which, in the usual arrangement, has a point source feed at the apex of the main (secondary) reflector and a hyperboloidal (primary) subreflector near the focal plane. In this case, the receiving equipment may be situated at the apex of the secondary reflector, free space serving as the transmission medium to the subreflector. Often the location of equipment near the apex is restrictive; depending on the arrangement, it may or may not move with the main reflector. This type of feed is referred to here as the "standard" Cassegrain.

The near-field Cassegrain combines some of the useful properties of the horn reflector with those of the standard Cassegrain. Rather than a point-source feed at the apex of the main reflector, a plane-wave feed of the same dimension as the subreflector is used.* Of course, the subreflector blocks the field of the main aperture just as in the case of the standard Cassegrain configuration. The plane-wave feed used for the measurements to be discussed was a small horn-reflector antenna. This arrangement allows the electronic equipment to remain stationary while

* Experiments on an antenna of this type were described recently by Profera et al.[10] Some generalized antenna systems based on this concept are discussed by S. P. Morgan.[11]

the elevation angle of the antenna is changed, in much the same manner as with the horn-reflector antenna.

It should be mentioned that the near-field feeding system is suited only to antennas that are very large compared to the wavelength. In the model that has been tested here, where the antenna diameter is less than 100λ, this criterion is just met. However, it appears that the feed system is broadband, embracing all wavelengths shorter than that satisfying the criterion, and in this sense the near-field antenna is somewhat similar to the horn reflector. Methods for mode scanning[12] a horn-reflector antenna are equally applicable to a near-field Cassegrain.

In Section II, the geometry of the near-field Cassegrainian antenna is discussed; the fields produced by the near-field feed are also given there. Section III describes the equipment, siting, and the methods used for measurement of antenna noise temperature and gain. Sections IV and V contain the noise and gain measurements on the shallow and deep sixteen-foot paraboloids using various types of feeds; the effect of subreflector mounting structures on noise performance is also given in those sections. Measurement of noise due to loss in transmission lines is dealt with in Appendix A. In Appendix B, the back-lobe noise temperature ($T_b$) for an antenna in a given environment is discussed, and in Appendix C, a quality factor which governs the signal-to-noise ratio in antennas is proposed.

## II. THE NEAR-FIELD CASSEGRAINIAN ANTENNA

### 2.1 *Comparison of Standard and Near-Field Cassegrainian Antennas*

The standard and near-field Cassegrainian antennas are compared in the idealized sketches of Fig. 1. An extensive analysis of the standard Cassegrain antenna has been given[8] and it will not be discussed further here; however, it should be noted that radiation from the point-source feed tends to spill over the rim of the hyperboloid. It has been demonstrated recently[13] that suitable beam shaping of the source pattern can reduce this spill-over. The receiving equipment is stationary as the antenna changes elevation, a right-angle circular waveguide bend and rotating joint being provided in the transmission line (see Fig. 1a). A simple right-angle bend would not be used in systems employing circular polarization since, due to unequal phase velocities of orthogonal components, the circularity would be degraded. A simple bend was used in the measurements to be discussed since circular polarization was not involved.

The near-field Cassegrain, shown in Fig. 1(b), has a horn-reflector

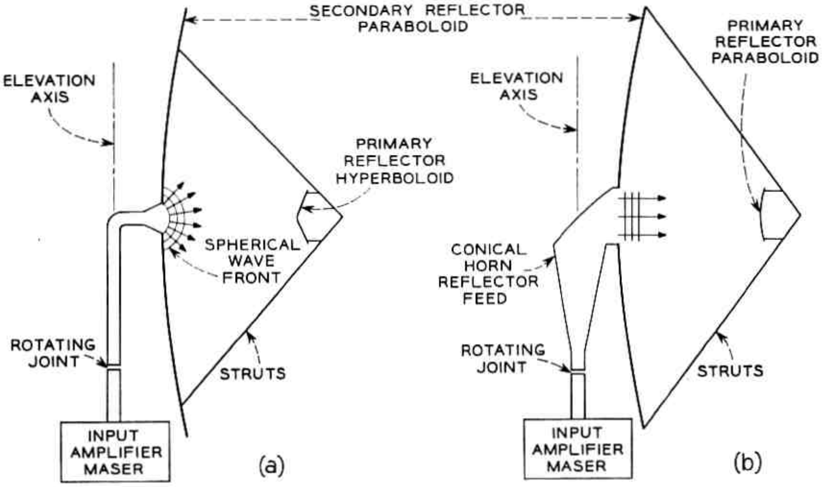


Fig. 1 — Idealized standard and near-field Cassegrainian antennas.

feed with an aperture of about the same diameter as that of the subreflector. To a geometrical optics approximation, the near field of this feed is collimated and of uniform phase.

### 2.2 *Geometry of Near-Field Cassegrain*

A simple derivation shows that the surface of the subreflector in the near-field configuration should be paraboloidal. Assume that the surface of the subreflector (see Fig. 2) is paraboloidal; it will then be sufficient to show that the path length of any ray from the plane wave in the feed aperture $EF$ to a reference plane in the secondary aperture is constant. Consider the ray of path length $\overline{AB} + \overline{BC} + \overline{CD}$. Equating this path to the length of the axial ray, one has

$$\overline{AB} + \overline{BC} + \overline{CD} = 2(f - f_1) + f = 3f - 2f_1 \tag{1}$$

where $f_1$ and $f$ are the focal lengths of the primary and secondary reflectors.* From Fig. 2, the line segments $\overline{AB}$, $\overline{BC}$, and $\overline{CD}$ are equal to $f - z_0$, $r_2 - r_1$, and $r_2 \cos\theta$ respectively. Equation (1) then becomes

$$f - z_0 + r_2 - r_1 + r_2 \cos\theta = 3f - 2f_1 \tag{2}$$

* Primary and secondary are used to designate the sub- and main reflectors respectively because the radiation patterns of the feed and main reflector are usually referred to as primary and secondary patterns.

where $z_0 = r_1 \cos \theta$, and from the equations of the paraboloids

$$r_1 = \frac{2f_1}{1 + \cos \theta} \quad \text{and} \quad r_2 = \frac{2f}{1 + \cos \theta}.$$

Making these substitutions, (2) becomes

$$f - \frac{2f_1(1 + \cos \theta)}{1 + \cos \theta} + \frac{2f(1 + \cos \theta)}{1 + \cos \theta} = 3f - 2f_1$$

which proves the equality.

As in the case of the standard Cassegrain, the degree of illumination on the surface of the secondary reflector of a near-field Cassegrain can be varied by using subreflectors of various focal lengths. Optimum illumination, as determined by geometrical optics, is achieved by using a subreflector of $f/D$ ratio identical to that of the secondary reflector.

Fig. 3 is an idealized sketch of the near field along the axis of the source aperture, the relative positions of subreflectors used in the experiments being indicated by arrows. Note that the subreflectors are

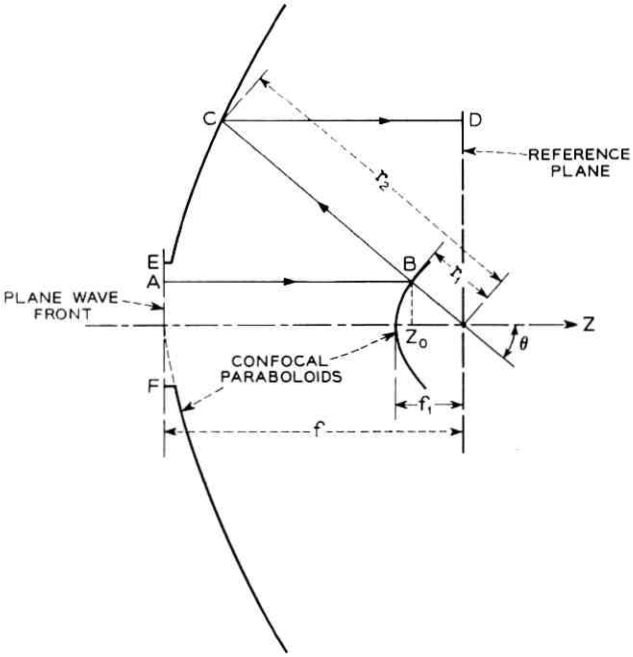


Fig. 2 — Geometry of near-field Cassegrainian antenna.

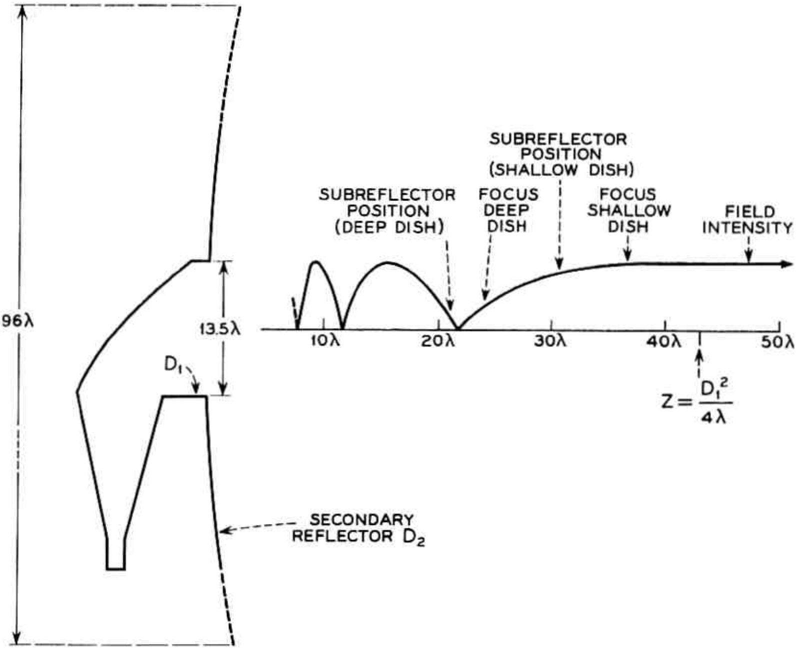

SUBREFLECTOR POSITION (SHALLOW DISH)
SUBREFLECTOR POSITION (DEEP DISH)
FOCUS DEEP DISH
FOCUS SHALLOW DISH
FIELD INTENSITY
96λ
13.5λ
$D_1$
10λ
20λ
30λ
40λ
50λ
$Z=\frac{D_1^2}{4\lambda}$
SECONDARY REFLECTOR $D_2$

Fig. 3 — Idealized near-field along the axis of the feed.

well out in the near-field. As mentioned in the introduction, the antenna size and wavelength used for these tests are far from optimum for the near-field type of feed. Preferably one would choose dimensions as large as possible with respect to wavelength (a high-gain antenna) in which case geometrical optics would hold more rigorously; an immediate consequence of this is that the subreflector positions (see Fig. 3) would be located where collimation and phase uniformity of the near-field are greatly improved.

### 2.3 *The Conical Horn-Reflector Feed*

A horn-reflector antenna[14] was used as the near-field feed for both the shallow and deep dishes. Theoretical and experimental studies of the far-field characteristics of this antenna have been published;[2] here, discussion is confined to its near-field characteristics.

The first 16-foot diameter secondary reflector used in the near-field Cassegrainian configuration had a focal length of 6 feet; therefore measurements of amplitude and phase of the field were made six feet in front of the aperture of the horn reflector, where the subreflector was to be mounted. The measurements were made in an anechoic chamber using a dipole probe. The data are plotted in Fig. 4 along with theoretical curves; these compare favorably in their general trend but not in detail. One notes that the phase departs from uniformity by about ±10° in some cases and that it is unsymmetrical with respect to the axis.

## III. EQUIPMENT AND METHODS OF MEASUREMENT

### 3.1 *Equipment*

Fig. 5 shows a complete antenna mounted on a motorized turntable carriage for azimuthal rotation. The cab to the left of the antenna is shielded; it houses the necessary equipment for measuring noise temperature, gain, and radiation patterns. The double A frame and cradle structure on which the secondary reflector is mounted is shown more clearly in Fig. 9 (p. 2691); it is a strong structural unit, no demand being made of the paraboloid for supplying rigidity. Elevation steering is provided by rotation of the cradle on bearings in the A frames. The transmission line, a circular waveguide of 2.8-inch diameter, is fed to the receiver in the cab through the bearing via a rotating joint.

The antenna is sited in a relatively flat, clear area; however, the site is ringed with trees which limit the horizon to an average elevation angle of 1.5 degrees.

The 6-gc maser receiver used for antenna noise temperature measurements has been discussed previously.[5,7]

### 3.2 *Method of Noise Measurements*

The setup used for measuring noise performance is shown in Fig. 6. The noise temperature at the input to the converter with the noise lamp off is given by

$$T_{ic} = G_m(T_a + T_m) \tag{3}$$

where $G_m$, $T_m$ are the maser gain and noise temperature, and $T_a = T_s + T_l + T_b$ is the antenna temperature. $T_s$ is the sky temperature observed by the main beam. $T_l$ is the noise temperature of the transmission line associated with the antenna.* $T_b$ is the noise contribution from the earth and sky through the wide angle side and back lobes of

* Here we neglect the actual attenuation due to transmission line loss; it is quite small.

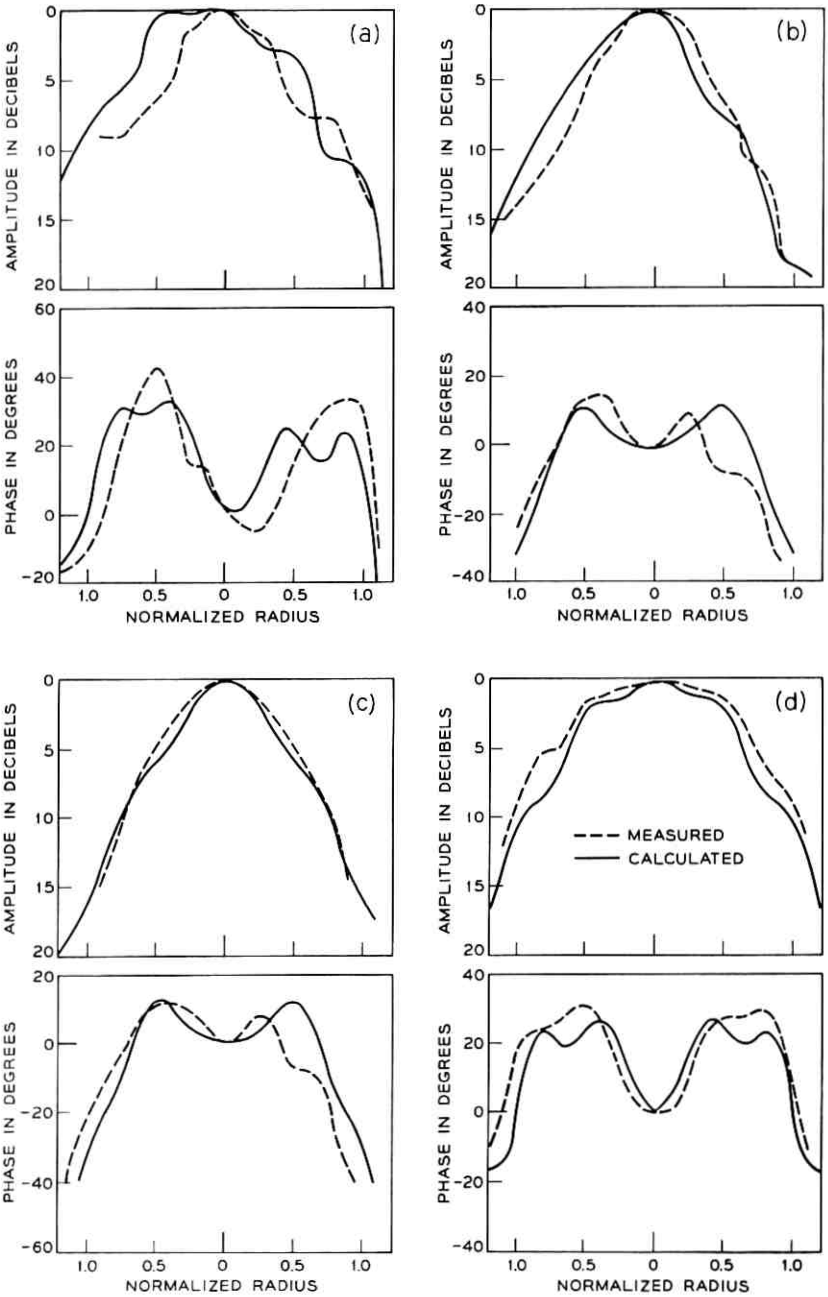


Fig. 4 — Near-field patterns of horn reflector ($Z$ = 6 ft.).

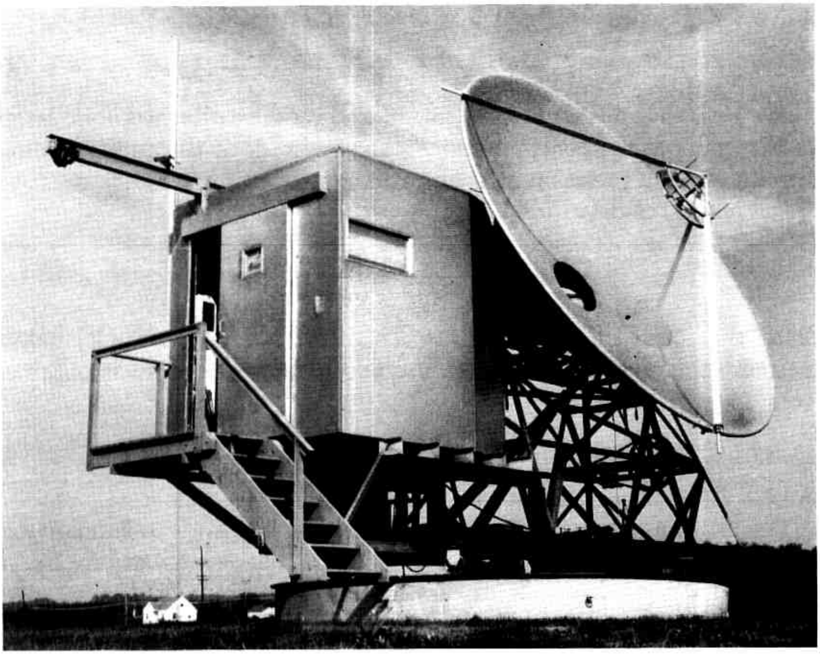


Fig. 5 — A 16-foot diameter paraboloid ($f/D = 0.375$) with near-field Cassegrainian feed.

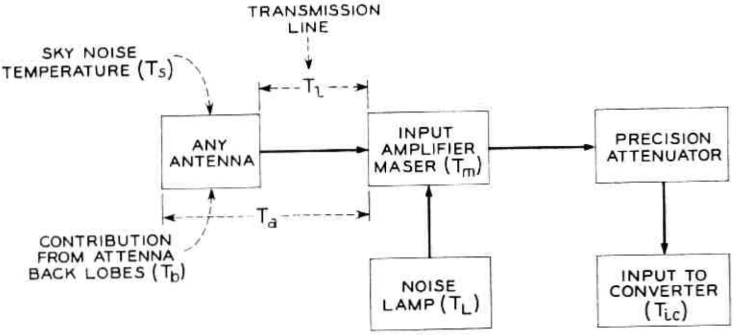

TRANSMISSION LINE
SKY NOISE TEMPERATURE ($T_s$)
$T_l$
ANY ANTENNA
INPUT AMPLIFIER MASER ($T_m$)
PRECISION ATTENUATOR
$T_a$
CONTRIBUTION FROM ATTENNA BACK LOBES ($T_b$)
NOISE LAMP ($T_L$)
INPUT TO CONVERTER ($T_{ic}$)

Fig. 6 — Noise contributors in the receiving system.

the radiation pattern. With the noise lamp fired, the noise input to the converter becomes

$$T_{ic} = G_m A(T_a + T_m + T_L) + (1 - A)T_0 \tag{4}$$

where $T_L$ is the additional noise introduced by the calibrated noise lamp, $T_0$ the ambient temperature and $A$ the reciprocal of the additional loss introduced by a precision attenuator to equalize (3) and (4).

Solving (3) and (4) for $T_a$,

$$T_a = \frac{AT_L}{1 - A} + \frac{T_0}{G_m} - T_m . \tag{5}$$

Since the terms on the right of (5) are determined by independent measurement, the noise temperature due to the back lobes is obtained from

$$T_b = T_a - T_s - T_l \tag{6}$$

provided $T_s$ and $T_l$ are known.

The sky temperature, $T_s$, for an atmosphere of given humidity is known from experience.* The transmission line contribution, $T_l$, is measured independently as discussed in Appendix A.

### 3.3 *Method of Gain and Radiation Pattern Measurement*

Gain measurements were made by comparing the power received by the antenna with that of a standard horn. A height run was made with this horn over the vertical extent of the main paraboloid for each gain measurement, the average of these data being taken as the reference value. Using the same equipment, azimuthal radiation patterns were obtained for the two principal polarizations.

## IV. MEASUREMENTS ON THE SHALLOW PARABOLOID ($f/D = 0.375$)

### 4.1 *Noise Measurements Using Various Feeds*

The 16-foot diameter spun-aluminum shallow paraboloid ($f/D = 0.375$) was first fed in a conventional manner using a cylindrical waveguide horn supported at the focal point by fiber glass struts, the feed waveguide running out from the apex. The radiation pattern of this

* In assigning values to $T_s$, the absolute water vapor density at the ground is determined from humidity and temperature measurements at the receiving site. Based on the particular value of water vapor density obtained, theoretical sky temperatures which have been verified previously[7] are calculated.

feed tapers to about −10 db at the rim of the paraboloid; this, in addition to an inverse distance effect of 3 db, results in a net illumination taper of 13 db.

Using the method for measuring noise discussed in Section 3.2, the two measurements on the curve of Fig. 7 labeled A were obtained* for zenith angles ($\theta$) of zero and 45°. In spite of the relatively strong taper, this feed produces a noise temperature of 26°K at both angles. By removing the fiber glass mounting struts and supporting the feed with fine guy ropes, the point labeled B was obtained, $T_b = 13.5$°K. Comparison of the zenith point on curve A with point B shows that the fiber glass struts contribute 12.5° of noise. Strut noise can be produced both by reflection of noise radiated from the ground into the antenna and by loss in the material comprising the strut; this point is elaborated upon later.

The second feeding arrangement measured on this paraboloid was a standard Cassegrain consisting of a precision hyperboloid (whose diameter could be changed from 30 to 24 inches by removal of an outer ring) fed by a horn of 3.5$\lambda$ diameter located at the apex of the main dish. The transmission line from horn to maser was about six feet of oversized circular guide including a right-angle bend.

With the hyperboloid mounted on fiber glass struts, noise measurements (at the zenith) produce $T_b = 24.5$°, as shown by point C in Fig. 7. The fiber glass struts were then covered with aluminum foil, essentially converting them to metal struts of the same geometry, and another measurement made, as indicated by the zenith value on curve D. At zenith the noise is now $T_b = 11$°K. Comparing this value with the 24.5° obtained using fiber glass struts, one sees that a decrease in noise of 13.5° has been effected. This result shows that most of the noise (at least for zenith orientation of the beam) is produced by loss in the dielectric; this conclusion seems to be valid because the decrease of 13.5° compares very favorably with the 12.5° decrease in $T_b$ obtained when the fiber glass struts were removed during the test using the waveguide feed. Measurements were also made versus zenith angle, as shown by the remainder of curve D in Fig. 7. As the horizon is approached, $T_b$ reaches values of the order 50°K. This noise is produced by spill-over beyond the rim of the hyperboloid and by reflection of noise from the earth by the sizable struts.

The last feeding arrangement to be discussed is the near-field Cassegrainian configuration. Fig. 5 shows a front view; mounted near the

* It may be well to mention here again that $T_b$ in Fig. 7 is the intrinsic antenna noise; sky and waveguide noise, etc. are not included.

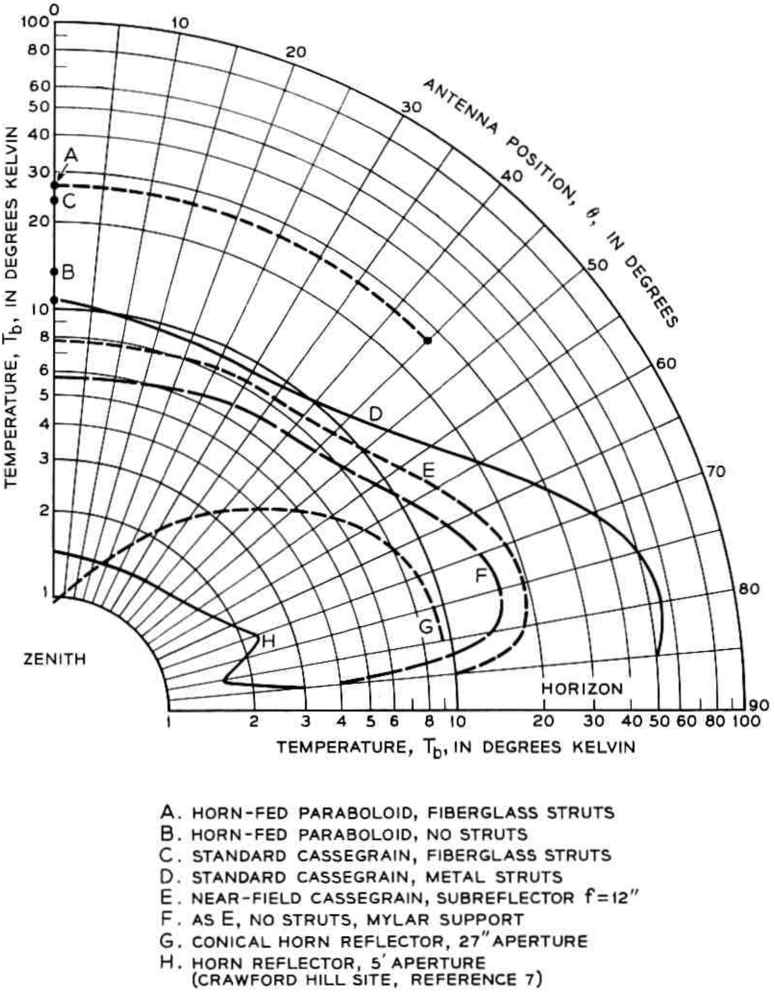


Fig. 7 — The 6-gc noise performance of 16-foot paraboloid ($f/D$ = 0.375).

focal plane is one of the several paraboloids used as primary reflectors. The mounting struts are of interest; they are metallic, light-weight, small, and extend to the rim of the secondary reflector, the latter being done so that the struts do not intercept direct radiation from the subreflector.

Noise data obtained for this feed using a primary reflector of 12-inch focal length are shown in Fig. 7 as curve E. $T_b$ at the zenith for this arrangement is less than 8°K. Equally important, however, $T_b$ is less than 20°K in the region of $\theta = 80°$, which is approximately the acquisition angle in a satellite communication system. At this angle the sky noise[7] is relatively high compared to the zenith value and it is desirable to have $T_b$ as low as possible.

By using a subreflector of longer focal length (14.5 inches), much less spill-over of the secondary reflector occurs and a zenith temperature of 4°K was achieved. For all elevation angles with this subreflector, $T_b$ was of the order of one-half that obtained with the subreflector of 12-inch focal length; of course, the secondary area illuminated is rather small and the gain is reduced by about 1.5 db; thus use of such subreflectors is of questionable value. A discussion of the signal-to-noise ratio in antennas is given in Appendix C.

To pursue the strut-noise effect further, a pressurized Mylar sheath support (similar to that shown in Fig. 9) with a wall thickness of 1.5 mils and inflated with nitrogen to a pressure of 0.15 psi was devised; it provides a remarkably rigid support. Nichrome guy wires from the subreflector to the rim of the secondary reflector are used for centering. Noise data for this arrangement (curve F in Fig. 7) show that there is an improvement of about 2°K (compared with curve E) for all elevation angles.

Curves G and H are included in Fig. 7 to serve as reference data. Curve G was obtained using the 27-inch horn-reflector feed (itself a low-noise antenna of respectable size with a far-field beamwidth of about 5°) mounted on the 16-foot paraboloid, struts and subreflector being removed. The arrangement amounts to a well shielded horn reflector with a large baffle (the secondary reflector); as indicated in curve G, $T_b$ is less than 1°K at zenith. The rather rapid increase to $T_b = 9°\text{K}$ at $\theta = 80°$ is attributed in part to the limited horizon* mentioned in Section 3.1. Curve H was obtained using a five-foot horn reflector at the Crawford Hill site, which has a clear horizon.[7]

### 4.2 *Radiation Pattern, Gain, and Impedance*

The idealized near field of the conical horn feed was shown in Fig. 1 as a plane wave perfectly collimated in the direction of the primary reflector. In reality this is not the case; as discussed in connection with Fig. 4, the phase varies as much as ±10° in places and is unsymmetrical with respect to the $Z$ axis, as is the amplitude. From an academic point

* The decrease in noise very near the horizon shown by the curves in Figs. 7 and 10 is also attributed to the environment; this effect was not observed using the Crawford Hill site (Refs. 5 and 7).

of view, a plane-wave source with a symmetrical aperture field (such as a horn-lens antenna) would be preferable as the feed, since the unsymmetrical phase produced by the horn reflector produces unsymmetrical radiation patterns in the near field. Evaluation of the field distribution in the aperture of the main dish is also a near-field diffraction problem. Preliminary calculations, using idealized fields, indicate that this distribution is toroidal; the secondary patterns produced therefore have relatively high immediate side lobes, similar to those of a heavily blocked aperture.

Fig. 8(a) shows the measured pattern for longitudinal (horizontal) polarization, using the subreflector of 12-inch focal length. Note that the side lobes are unsymmetrical and that the highest one is only some 13.5 db down from the main lobe. In Fig. 8(b), for transverse (vertical) polarization, the immediate side lobes are more than 20 db down. In both cases the half-power beamwidth is about 0.7°. For comparison, a calculated curve for constant amplitude distribution and 2 per cent aperture blocking is also shown in Fig. 8. The gain is 48.0 db for transverse and 47.4 db for longitudinal polarization. The calculated full area gain of the 16-foot dish is 50.1 db at 6.3 kmc; thus the average effective area is 2.4 db down or, in other words, the efficiency is 57.5 per cent.

The SWR for the above configuration is 1.17, equivalent to a return loss of 22 db. A slight improvement in impedance is obtained by remov-

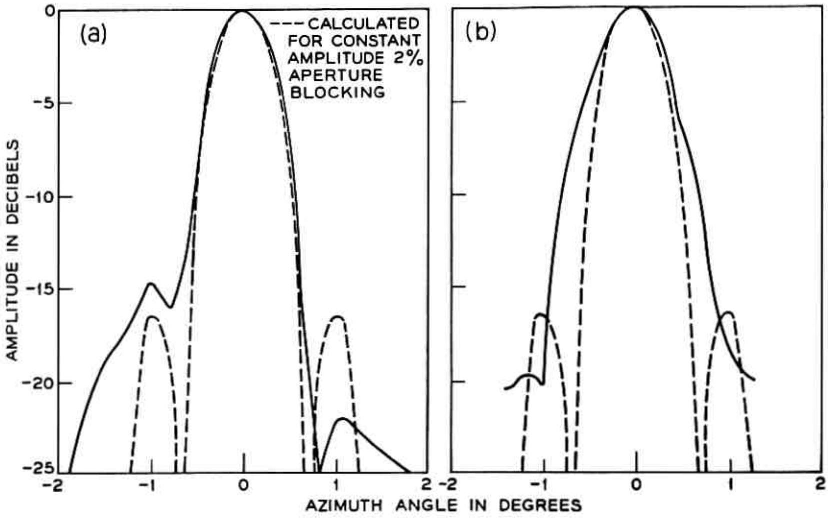


Fig. 8 — Radiation patterns of shallow paraboloid.

ing a circular area from the subreflector and illuminating the *concave* surface of the subreflector in the proper phase. This supplementary feeding arrangement (shown clearly in Fig. 5) resulted in a VSWR of 1.15 or a return loss of 23 db.

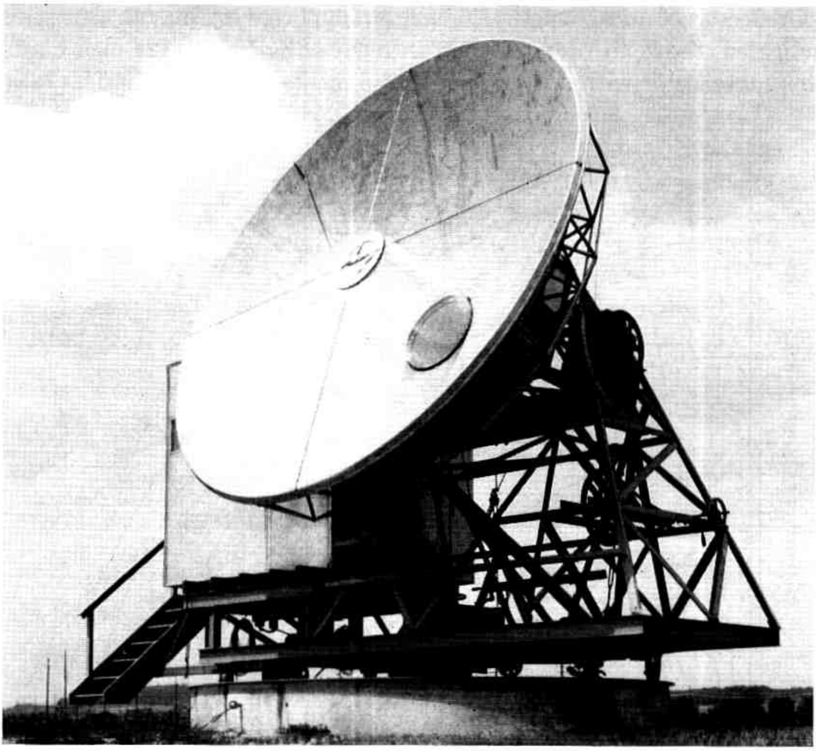


Fig. 9 — 16-foot paraboloid ($f/D = 0.25$) with near-field Cassegrainian feed. Note Mylar support for subreflector.

## V. MEASUREMENTS ON THE DEEP PARABOLOID ($f/D = 0.25$)

### 5.1 *Noise Measurements Using Various Feeds*

Performance was next examined using a deep paraboloid ($f/D = 0.25$) as secondary reflector; Fig. 9 shows a front view. This paraboloid (diameter 16 feet) was machined from urethane foam, a reflective surface of zinc being applied after machining.* Near the focal plane is one of the

* Precision in reflecting surfaces is an important factor in determining the wide angle lobes[15] and therefore the noise performance of antennas.

paraboloids used as a primary reflector, in this case supported by a Mylar sheath. An alternative support used four half-inch struts extending perpendicularly from the mounting ring surrounding the feed aperture to the subreflector.

Curve A of Fig. 10 shows data obtained using a 30-inch diameter subreflector of focal length 7.5 inches supported by metal struts; as indicated, the zenith temperature is about 4°K. Next, a 24-inch diameter subreflector with a focal length of 6 inches was used, the $f/D$ ratio

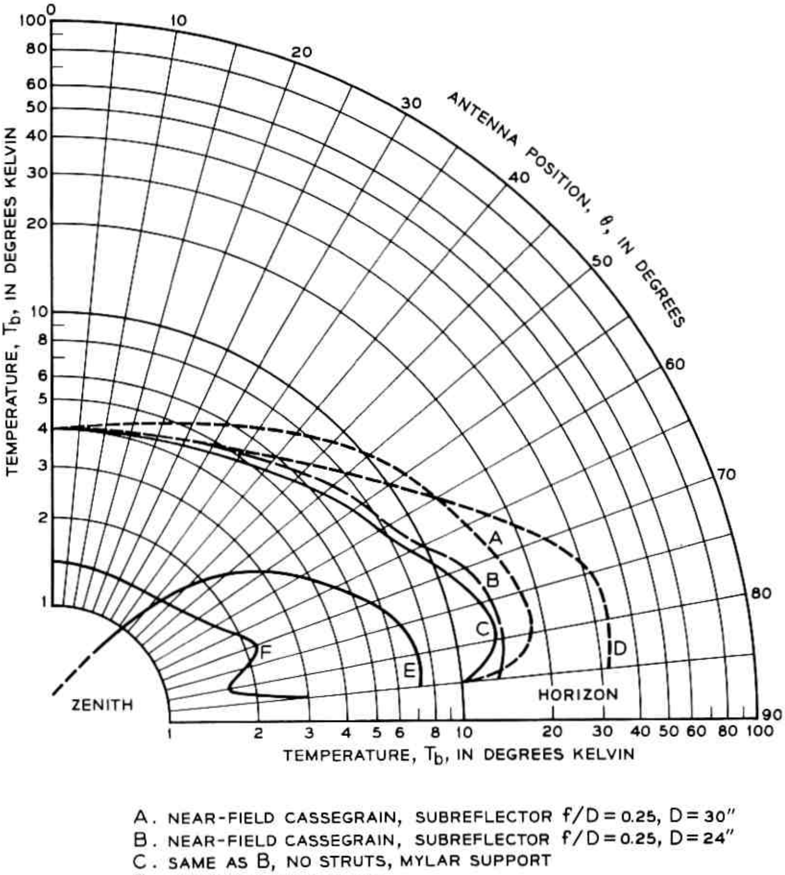


Fig. 10 — The 6-gc noise performance of 16-foot paraboloid ($f/D$ = 0.25).

being the same as in case A. As indicated by curve B, this feed arrangement displays very good characteristics, also achieving a zenith temperature of 4°K. Equally important, $T_b$ remains less than 10°K until a zenith angle of sixty-five degrees is reached. In the region of $\theta = 80°$, $T_b$ is less than 15°K.

Even though the struts supporting the subreflector were of small size, it was of interest to see what the change in $T_b$ would be if the struts were removed. Fig. 9 shows the nitrogen-filled Mylar sheath (previously discussed in Section 4.2) supporting the 24-inch diameter subreflector, and curve C in Fig. 10 shows the data obtained. There is no appreciable change in the zenith noise temperatures; however, there is improvement for angles approaching the horizon, indicating that the small struts were to some extent scattering the earth's radiation into the antenna.

A standard Cassegrain feed consisting of a precision 30-inch diameter hyperboloid machined from styrofoam and suitably surfaced with zinc and a 3.5λ diameter horn located at the apex of the secondary reflector was tested for comparison with the near-field feed; the noise measurements are shown as curve D in Fig. 10. At zenith, the noise temperature is 4°K, but spill-over effects quickly become apparent when the zenith angle exceeds sixty degrees. At $\theta = 80°$, $T_b$ has increased to about 30°K; this is attributed mainly to spill-over beyond the rim of the subreflector.

Reference curves E and F are included in Fig. 10; similar data were discussed in connection with Fig. 7.

### 5.2 *Radiation Pattern, Gain and Impedance*

Fig. 11 shows the azimuth patterns for the deep dish using the 24-inch diameter paraboloid as primary reflector. In Fig. 11(a) (longitudinal polarization) the immediate side lobes are unsymmetrical and the highest one is about 13.0 db down. Figure 11(b) is the secondary pattern for transverse polarization in the feed. In both cases, the 3-db beamwidth is about 0.7°.

During the measurements, the well known problem of properly illuminating a deep paraboloid became apparent; however, it was readily determined that deep reflectors may be illuminated more easily by Cassegrainian techniques than by focal point feeds. For example, using the near-field feed, the illumination at the rim of a 30-inch subreflector is down about 13 db; including a 6 db inverse distance attenuation, the resulting taper across the secondary reflector is approximately 20 db. The measured gain was 4 db down from full area. By reducing the diameter of the subreflector, the taper is also reduced, thereby in-

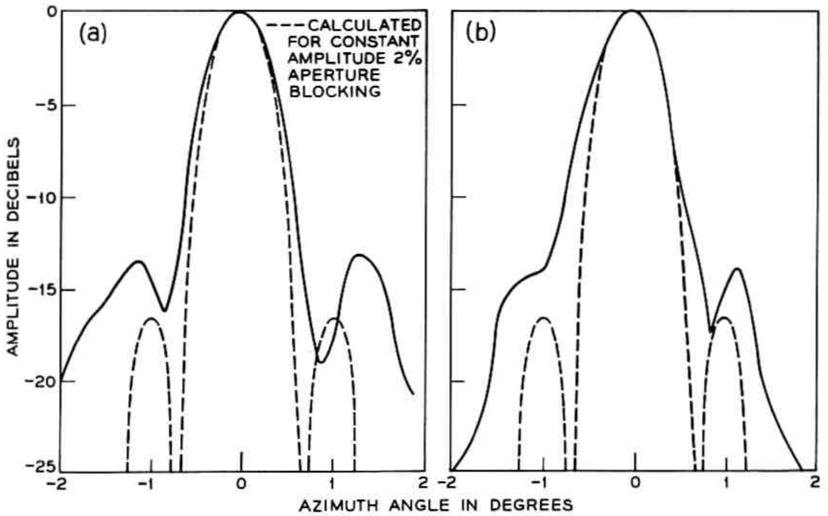


Fig. 11 — Radiation patterns of deep paraboloid.

creasing the gain (in this process, noise performance is often sacrificed for gain). A 24-inch subreflector was found to be a suitable compromise, measured gain for both polarizations being 47.3 db (2.8 db down from area). Somewhat surprisingly, as indicated by curve B of Fig. 10, there is no significant deterioration in noise performance.

Average SWR measurement for the configuration last mentioned is 1.11, which is equivalent to a return loss of 25 db, an improvement of 3 db over that of the shallow dish.

The gain of the standard Cassegrainian configuration was 47.4 db and 47.7 db for the vertical and horizontal polarizations, the average value being 0.24 db higher than that of the near-field device.

### VI. DISCUSSION

The measurements discussed here have been directed toward evaluation of the noise performance of several types of feeds for full paraboloidal reflectors. In particular, it is found that the near-field Cassegrainian feed, a device whose design is based on simple geometrical optics, performs exceptionally well, its low-noise characteristics being as good or better than those of the standard Cassegrainian feed over all angles of elevation. This result holds true for both shallow and deep secondary reflectors. The efficiency of the near-field Cassegrain is about 55 per cent, similar to that obtained using more conventional feeds; the radiation patterns are unsymmetrical due to lack of symmetry in the phase of the primary field.

Deterioration in noise performance due to the struts (or spars) used for supporting feed structures has been examined. Dielectric struts, such as those of fiber glass, have been found to introduce noise because of loss in the material. A pressurized membrane has been tested as a support for subreflectors; it appears satisfactory mechanically and minimizes degradation in electrical performance.

## VII. ACKNOWLEDGMENTS

We are indebted to R. Kompfner for suggestion of the compression cylinder as a subreflector support, to A. B. Crawford for his interest in these studies, to H. W. Anderson for assistance in mechanical design, to L. E. Hunt, Tingye Li, and Mrs. C. L. Beattie for help in measurement and calculation of the near field of the conical horn-reflector, and to J. S. Cook for constructive comment. H. A. Gorenflo provided valuable assistance throughout the construction and measurement phases of these experiments.

## APPENDIX A

### *Transmission Line Measurements*

### A.1 *Noise Measurements (Short Circuit)*

The noise that exists in a transmission line is caused by resistive losses in the line itself and by noise generators which may be at either or both ends of the line. Consider the infinite transmission line of Fig. 12(a), which for the moment is assumed to have a noise-free measuring device at terminal A. Let the thermodynamic temperature of the line be $T_l$, and $\alpha$ the power absorption coefficient. Since the line is of infinite length, the noise temperature measured at A is simply $T_a = T_l$. Divide the line into segments 1 and 2 at point $l$. The contribution to the noise at A by segment 2 is $T_l e^{-\alpha l}$ (since the line is infinite); therefore that contributed by segment 1 is $T_l(1 - e^{-\alpha l})$. A series expansion gives

$$T_l(1 - 1 + \alpha l - (\alpha^2 l^2/2) + \cdots)$$

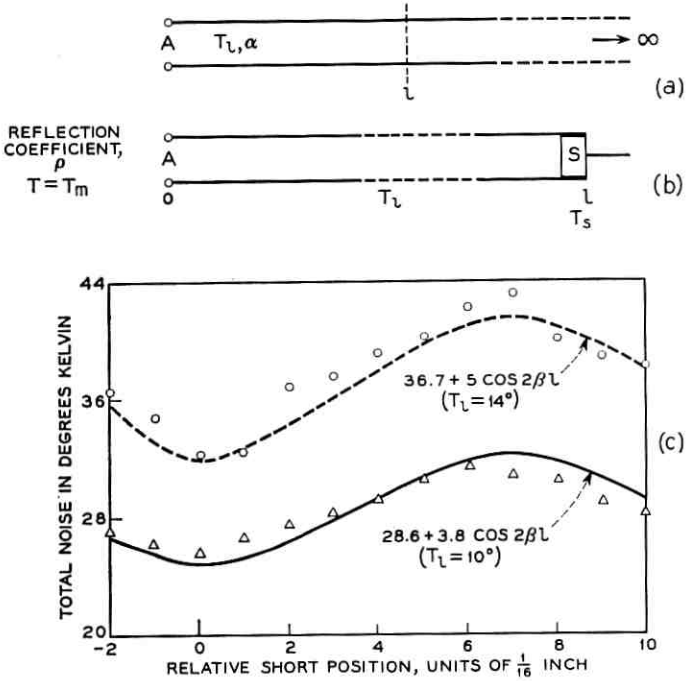


Fig. 12 — Noise measurements on shorted transmission line.

which for small $\alpha$ is approximately $T_l \alpha l$. Naturally, if the total loss, $\alpha l$, is known, the noise produced is obtained immediately.*

Fig. 12(b) shows a movable shorting piston, S, near the terminals A of the amplifier to be used for the noise measurements. The amplifier is not perfectly noise free, nor is it perfectly matched; therefore movement of the piston produces a cyclical variation in the noise at A.

Let the effective temperature of the maser amplifier at terminals A in Fig. 12(b) be designated $T_m$ ; this represents the intrinsic noise, which amounts to about 3°K. $T_l$ represents the effective temperature of the transmission line and $T_s$ that of the shorting piston, S. The voltage reflection coefficient at A (the input mismatch of the maser) is $\rho$. The

* This assumes that the thermodynamic temperature of the line is constant; if not, the effective noise temperature is given by

$$T = \int_0^l T(x)\alpha(x) \exp\left(-\int_0^x \alpha(x)\,dx\right) dx$$

where $T(x)$ is the temperature distribution along the line.

sources that give rise to noise $T$, traveling to the left (i.e., into the amplifier) at A are:

$T_{m_1}$ — intrinsic amplifier noise;
$T_{m_2}$ — amplifier noise traveling to the right, and reflected at S (uncorrelated with $T_{m_1}$);
$T_{l_1}$ — line noise initially traveling to right and reflected by S, also that reflected by $\rho$ at A and again by S;
$T_{l_2}$ — line noise initially traveling to left, also that reflected by $\rho$ and by S (uncorrelated with $T_{l_1}$);
$T_S$ — shorting piston noise, traveling to left, also that reflected by $\rho$ and by S.

Thus, to first order the noise entering the amplifier at A is

$$T = T_{m_1} + T_{m_2} + T_{l_1} + T_{l_2} + T_S .$$

If the attenuating effects of line loss are neglected (since they are only of the order 0.1 db), the above sum to first order becomes

$$T = T_{m_1} + (T_{m_2} + 2T_l + T_S)(1 + \rho^2) + 2(T_{m_2} + 2T_l + T_S)\rho \cos 2\beta l \tag{7}$$

where

$$2T_l = T_{l_1} + T_{l_2}$$

$\beta$ being the propagation constant of the line.

Measured data (noise versus short position) shown as crosses in Fig. 12(c) were taken using a short circuit in round waveguide. Also shown (as a solid curve) are data calculated using the following constants in (7): $T_{m_1} = 3°$, $T_{m_2} = 3°$, $T_l = 10°$, $T_S = 2.5°$ (short circuit with a standing wave ratio of 250) and $\rho = 0.075$ (22-db return loss), the last two being measured values; these result in

$$T = 28.6 + 3.8 \cos 2\beta l. \tag{8}$$

Now let an additional length of transmission line be added to $l$ such that the total length is $l_1$, and let the short $S$ be moved to the end of this line. The data, shown as dots in Fig. 12(c), were obtained when approximately 6 feet of 2.8-inch diameter line* were added. The dashed curve is a plot of (7) with the following constants:

$$T_{m_1} = T_{m_2} = 3°, \qquad T_{l_1} = 14°, \qquad T_S = 2.5°, \qquad \rho = 0.075,$$

$$T_1 = 36.7 + 5.0 \cos 2\beta l. \tag{9}$$

* The line actually comprised 45 inches of straight guide and a right-angle bend of 15-inch radius used in the standard Cassegrain feed.

If (8) is subtracted from (9), one has

$$2(T_{l_1} - T_l)(1 + \rho^2) + 4(T_{l_1} - T_l)\rho \cos 2\beta l = 7.9 + 1.2 \cos 2\beta l. \tag{10}$$

From the first terms of both sides of (10), one obtains $2(T_{l_1} - T_l) = 7.9$, or an increased noise temperature $T_{l_1} - T_l = 3.95°$ due to the additional length of line. From the second terms one obtains $4(T_{l_1} - T_l)\rho = 1.2$ or $T_{l_1} - T_l = 4°$. The accuracy of the latter value is very dependent upon an accurate value for $\rho$ (measured as 0.075), whereas the first value, 3.95° (which is really the difference between the average values of the plots in Fig. 12(c), is good to order $1 + \rho^2$. The value 3.95°K is equivalent to a loss of 0.052 db in the additional length of line.

### A.2 *SWR Measurements (Short Circuit)*

Using the same shorting piston and transmission line as above but adding an additional 60-inch length of waveguide (diameter 2.8 inches), the measured VSWR was 85, which is equivalent to a loss of 0.0093 db/foot.

The loss per unit length derived from the noise measurement just discussed is 0.0092 db/foot, indicating close agreement between the two methods.

## APPENDIX B

### *The Antenna Noise — $T_b$*

Since typical antenna patterns have significant levels in the side and back lobes, it is necessary to consider the effects of noise due to thermal radiation from the environment into the antenna.* This effective noise temperature is designated by $T_b$.

Consider first the ideal radiation pattern shown in Fig. 13(a): it has a very narrow beam of width $\alpha$, the gain $G$ being constant over the angle $\alpha$; it has no back lobes. The antenna is assumed to be lossless and to be mounted height $h$ above the ground. Beamed at various angles $\theta$ with respect to zenith, this antenna sees the true brightness temperature, $T(\theta)$, due to various noise sources. For $0 < \theta < \pi/2$, $T(\theta) = T_s(\theta)$ is the sky temperature. For $\pi/2 < \theta < \pi$, the brightness temperature is due to both sky and earth, as shown in Fig. 13(b), since the sky noise from angle $\pi - \theta$ is reflected at point P according to the reflection

* This effect has been discussed recently in Ref. 16.

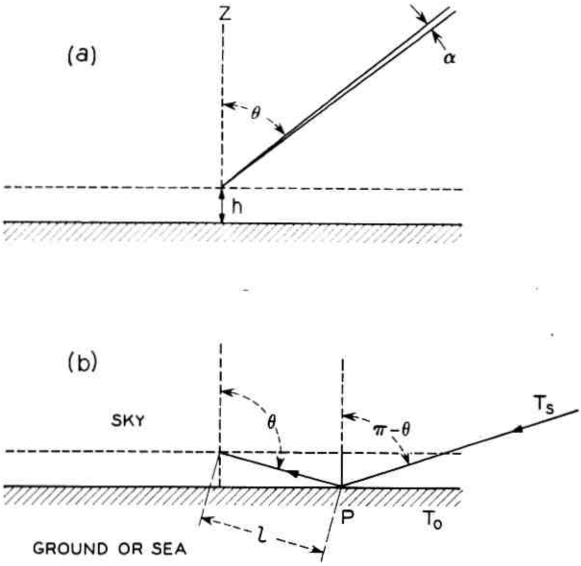


Fig. 13 — Ideal antenna and its environment.

coefficient of the earth, $r(\theta)$, at P, whereas the earth generates noise that enters the antenna directly dependent upon the coefficient $a(\theta)$, which represents absorption at P. Noise due to loss in the atmosphere along path length $l$ also contributes to the brightness temperature. Since $r(\theta) + a(\theta) = 1$, the brightness temperature for $\pi/2 < \theta < \pi$ is

$$T(\theta) = r(\theta)T_s(\pi - \theta) + [1 - r(\theta)]T_0 + T_{s_l}(\theta - \pi/2) \qquad (11)$$

where $T_0$, the temperature of the ground, is assumed to be 300°K.

The term $T_{s_l}(\theta - \pi/2)$ of (11) represents noise due to the path in the atmosphere between the antenna and the point P. Compared with other noise sources, it is found to be negligible, and therefore has been disregarded in what follows.

The reflection coefficient $r(\theta)$ is highly dependent upon the environment and to some extent on polarization; it usually varies with time, being a function of the ground conditions over vegetated areas and the wave conditions over water. Using representative data at 10 cm wavelength for the reflection coefficient,[17] the sky temperature, and (11), one can estimate the brightness temperature distribution for all angles $\theta$, as shown in Fig. 14. Curve B is for smooth sea water and curve A for a perfectly reflecting mirror (which images the sky noise), whereas the

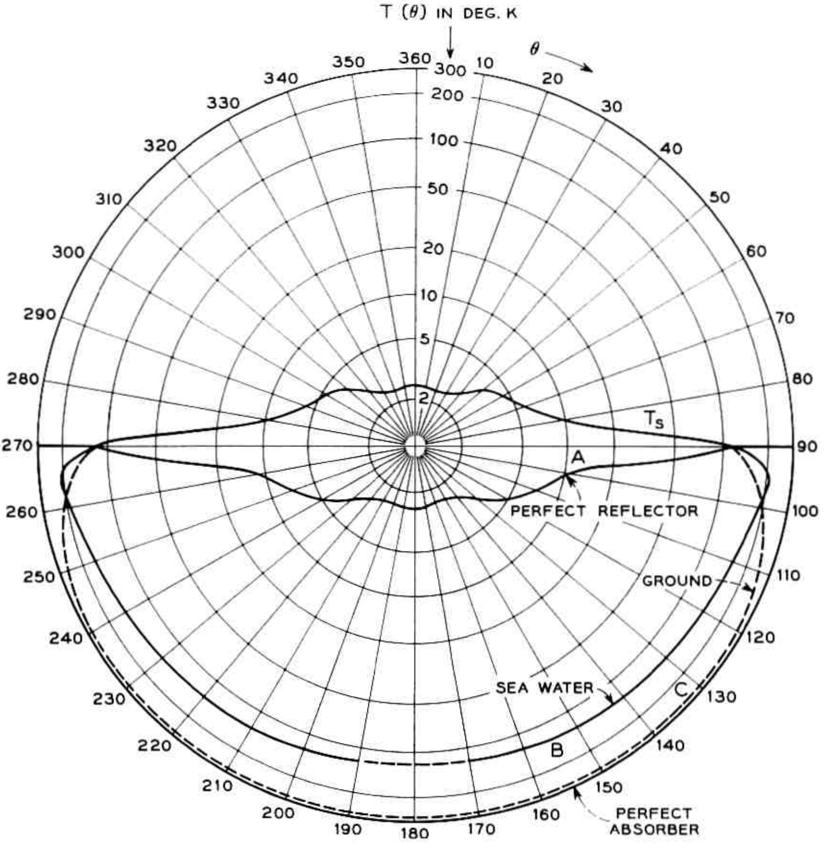


Fig. 14 — Brightness temperature distributions for middle cm-wave band.

poorly reflecting ground environment, curve C, approximates a perfect absorber.

An actual antenna has a finite radiation pattern $G(\theta)$; assuming it to be symmetrical about the main axis of the antenna beam, the equation relating antenna temperature to radiation pattern and brightness temperature is

$$T_s + T_b = \frac{1}{2} \int_0^{\pi} G(\theta) T(\theta) \sin \theta \, d\theta \tag{12}$$

for the antenna beamed vertically. If the antenna is beamed at angle $\theta'$ with respect to the zenith, (12) becomes

$$T_s(\theta') + T_b(\theta') = \frac{1}{2} \int_0^{\pi} G(\theta - \theta') T(\theta) \sin \theta \, d\theta. \tag{13}$$

As a simple application of (12), consider an isotropic antenna surrounded by a noise-free sky and a perfectly absorbing earth. In this case $G(\theta) = 1$ and $T(\theta) = T_0$ for $\pi/2 < \theta < \pi$; thus

$$T_s + T_b = \frac{1}{2}\int_{\pi/2}^{\pi} T_0 \sin\theta \, d\theta = \frac{T_0}{2} = 150°\text{K}.$$

An idealized radiation pattern for a microwave antenna is shown in Fig. 15, where

$$G(\theta) = G_0, \qquad 0 < \theta < \alpha/2$$

and

$$G(\theta) = G_b, \qquad \alpha/2 < \theta < \pi,$$

$G_b$ being the average gain in the side and back lobes. Again assuming a noise-free sky and perfectly absorbing earth,

$$T_{b_2} = \frac{1}{2}\int_{\pi/2}^{\pi} G_b T_0 \sin\theta \, d\theta = \frac{T_0 G_b}{2} = 150 G_b°\text{K}.$$

Thus, for example, if $G_b = 0.1$ (10 db below isotropic), $T_{b_2} = 15°\text{K}$.

Using the idealized antenna pattern of Fig. 15 and the data of Fig. 14, let us now integrate numerically according to (12). The noise contribution from the main beam (the so-called sky noise) is

$$T_s = \frac{1}{2}\int_0^{\alpha/2} G_0 \, T_s(\theta) \sin\theta \, d\theta \approx 2.5°$$

which is readily taken from Fig. 14.

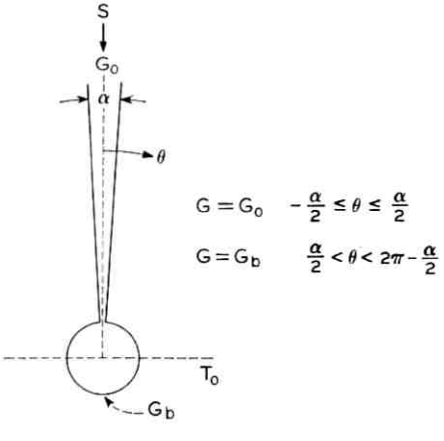


Fig. 15 — Idealized antenna pattern.

Using $G_b = 0.1$, the contribution due to sky noise in the far side lobes $(\alpha/2 < \theta < \pi/2)$ is

$$T_{b_1} = \frac{G_b}{2} \int_{\alpha/2}^{\pi/2} T_s(\theta) \sin \theta \, d\theta = 0.7°.$$

From the region $\pi/2 < \theta < \pi$ (ground, etc.), the contribution is

$$T_{b_2} = \frac{G_b}{2} \int_{\pi/2}^{\pi} T(\theta) \sin \theta \, d\theta$$

which amounts to 0.7° for the antenna above a perfect reflector, 7.6° above sea water and 15° above a perfectly absorbing earth.

Thus for an antenna with far side and back lobes 10 db below an isotropic radiator, the total antenna noise due to atmosphere and environment for zenith orientation of the beam is

$$T_a = T_s + T_{b_1} + T_{b_2}$$

which amount to 3.9° (perfect reflector)
10.8° (calm sea water)
18.2° (ground with vegetation which approximates a perfect absorber).

$T_b = T_{b_1} + T_{b_2}$ for the above conditions is
1.4° (perfect reflector)
8.3° (calm sea water)
15.7° (ground with vegetation)

obtained simply by subtracting the sky noise (2.5°K) from the previous numbers.

## APPENDIX C

### *The Signal-To-Noise Ratio and Quality Factor of an Antenna*

For the idealized antenna pattern of Fig. 15, the received power at the terminals of the antenna oriented toward a white noise signal source is

$$P_S = SAB = SG_0(\lambda^2/4\pi)B$$

where $S$ is the incident signal flux, $B$ the bandwidth, and $A$ the effective area of the antenna. The total noise in the antenna is $P_N = kT_aB$, $k$ being Boltzmann's constant and $T_a = T_s + T_b + T_l$. The contribution $T_s$ is the sky noise in the main beam; it is essentially independent of the gain $G_0$ for high-gain antennas. $T_b$ and $T_l$ are the effective noise temperatures due to back lobes and line losses. The signal-to-noise ratio for the antenna is therefore

$$\frac{P_S}{P_N} = \frac{S\lambda^2}{4\pi k}\frac{G_0}{T_a} = \frac{S\lambda^2}{4\pi k}\frac{G_0}{(T_s + T_b + T_l)}. \tag{14}$$

This assumes that the noise figure of the receiving amplifier is negligible. Of course, the receiver noise and the antenna impedance must both be considered in calculating the system noise.

Of the terms contributing to $T_a$ in (14), $T_s$ is unavoidable and only $T_b$ and $T_l$ can be attributed to deficiencies in the antenna. We can define a quality factor for the antenna in the following way: set $T_s$ equal to zero and multiply numerator and denominator of (14) by $T_0$; then

$$\frac{P_S}{P_N} = \frac{S\lambda^2}{4\pi k T_0}\frac{G_0 T_0}{(T_b + T_l)} = \frac{S\lambda^2}{4\pi k T_0} Q \tag{15}$$

where $Q = G_0T_0/(T_b + T_l)$ is the quality factor. Examples of typical values of $Q$ are:

(1) An isotropic antenna completely surrounded by a perfect absorber at $T_0 = 300°$ ($T_l = 0$, no line losses), $Q = 1$.

(2) An isotropic antenna surrounded by a perfectly absorbing earth and noise-free sky, (no line losses), $Q = 2$.

(3) The antenna above ground as discussed in Appendix B with far side and back lobes 10 db below an isotropic radiator, (where $T_b \approx 15°$), $Q = 20G_0$.

(4) The near-field Cassegrain as discussed in Section V

$$(G_0 = 5.4 \times 10^4, \qquad T_l + T_b = 4° + 4° = 8°), \qquad Q = 2 \times 10^6.$$

(5) A horn-reflector antenna with the same aperture area and transmission line loss as in (4) above

$$(G_0 = 7 \times 10^4, \qquad T_l + T_b = 4° + 1° = 5°), \qquad Q = 4.2 \times 10^6.$$

Similarly, one can define a quality factor for the total receiving system as

$$Q_T = \frac{G_0 T_0}{T_R + T_a}$$

where $T_R$ represents all noise associated with the receiver proper, and $T_a$, all noise associated with the antenna.

REFERENCES

1. Crawford, A. B., Hogg, D. C., and Hunt, L. E., A Horn-Reflector Antenna for Space Communication, B.S.T.J., **40**, July, 1961, p. 1095.

2. Hines, J. N., Li, Tingye, and Turrin, R. H., The Electrical Characteristics of the Conical Horn-Reflector Antenna, B.S.T.J., **42,** July, 1963, p. 1187.
3. Giordmaine, J. A., Alsop, L. E., Mayer, C. H., and Townes, C. H., A Maser Amplifier for Radio Astronomy at X-Band, Proc. IRE, **47,** No. 6, 1959, p. 1064.
4. Cook, J. J., Gross, L. G., Bair, M. E., and Terhume, R. W., A Low-Noise X-Band Radiometer Using Maser, Proc. IRE, **49,** No. 4, 1961, p. 768.
5. DeGrasse, R. W., Hogg, D. C., Ohm, E. A., and Scovil, H. E. D., Ultra-Low-Noise Receiving System for Satellite or Space Communications, Proc. Nat. Electr. Conf., **15,** 1959, p. 370.
6. Ohm, E. A., Receiving System — Project Echo, B.S.T.J., **40,** July, 1961, p. 1065.
7. Hogg, D. C., and Semplak, R. A., The Effect of Rain and Water Vapor on Sky Noise at Centimeter Wavelengths, B.S.T.J., **40,** Sept., 1961, p. 1331.
8. Hannan, P. W., Microwave Antennas Derived from the Cassegrain Telescope, IRE Trans. on Antennas and Propagation, **AP-9,** March, 1961, p. 140.
9. Potter, P. D., Aperture Efficiency of Large Paraboloidal Antennas as a Function of their Feed System Radiation Characteristics, IEEE Trans. on Antennas and Propagation, **AP-11,** May, 1963.
10. Profera, C. E., and Sciambi, A. F., A High Efficiency Low-Noise Antenna Feed System, PTGAP International Symposium, Boulder, Colorado, July, 1963.
11. Morgan, S. P., Some Examples of Generalized Cassegrainian and Gregorian Antennas, to be published in Nov., 1964, IEEE Trans. on Antennas and Propagation.
12. Cook, J. S., and Lowell, R., The Autotrack System, B.S.T.J., **42,** July, 1963, p. 1283.
13. Potter, P. D., Unique Feed System Improved Space Antennas, Electronics, **25,** June 22, 1962, p. 36.
14. Schelkunoff, S. A., and Friis, H. T., *Antennas Theory and Practice*, J. Wiley & Sons, 1952, Ch. 18.
15. Dragone, C., and Hogg, D. C., Wide-Angle Radiation Due to Rough Reflectors, B.S.T.J., **42,** Sept., 1963, p. 2285.
16. Croom, D. L., Naturally Occurring Thermal Radiation in the Range 1–10 Gcs, Proc. IEE, **3,** No. 5, May, 1964, p. 967.
17. Kerr, D. E., *Propagation of Short Waves*, McGraw-Hill Book Co., 1951, ch. 5.

# A Two-Gyro, Gravity-Gradient Satellite Attitude Control System

By J. A. LEWIS and E. E. ZAJAC

(Manuscript received April 24, 1964)

*This article gives the results of an analytical and numerical study of a two-gyro, gravity-oriented communications satellite. The principal purpose of the study was to uncover and solve the analytical problems arising in the design of passive gravity-gradient attitude control systems. Although the study was directed at satellite orientation, it is felt that many of the techniques developed have general use in the investigation of dynamical systems.*

*We consider both small and large motions about the desired earth-pointing orientation. In the small-motion study, the goal is simultaneous optimization of the transient response and the forced response to perturbations caused by orbital eccentricity, magnetic torques, solar torques, thermal rod bending, and micrometeorite impact. In the large-motion study, we enumerate all possible equilibrium positions of the satellite and then consider initial despin after injection into orbit, inversion of the satellite from one stable equilibrium position to another by switching of gyro bias torques, and the decay of transient motions resulting from large initial angular rates.*

*As a specific numerical example, we have treated a 300-lb satellite in a 6000-nm orbit, stabilized by a 60-ft extensible rod with a 20-lb tip mass, and by two single-degree-of-freedom gyros, each with an angular momentum of $10^6$ cgs units. Without a detailed discussion of hardware, it is concluded that such a system, having a total weight of 50 to 75 pounds including power supply, will provide a settling time for small disturbances of less than one orbit and will hold the antenna pointing error within a few degrees.*

TABLE OF CONTENTS

## I. INTRODUCTION

It has been known for over two hundred years that the variation in the gravitational field over the length of an earth satellite generates torques which tend to keep the axis of minimum inertia of the satellite pointing toward the earth. In particular, this mechanism keeps one face of the moon earth-pointing.

Such gravity-gradient orientation of communications satellites is very attractive because the simplicity of the effect leads to the possibility of simple attitude control and hence high reliability and long life. On the other hand, the tiny size of the gravity-gradient torques means formidable mechanization problems, and although Pierce suggested its use as early as 1955,[1] gravity-gradient stabilization has been widely held to be impractical.

However, several recent analytical and hardware studies have resulted in proposals for practical, gravity-gradient controlled satellites. All the proposed schemes work on the same principle. Steady-state perturbations, due, for example, to magnetic and solar torques, are kept within tolerable limits by making the satellite inertia sufficiently large, usually with some sort of extensible rod-tip mass combination. Damping of transient perturbations is provided by connecting the satellite through a dissipative joint to an "anchor," that is, to some object that will allow energy dissipation by virtue of relative motion between itself and the satellite proper. The anchor may be one or more gyros, as in the schemes discussed by Ogletree, et al.,[2,3,4] by Burt,[5] and by Scott;[6] a second rigid body, either hinged to the satellite, as proposed by Kamm,[7] by Paul, West, Yu, et al.;[8,9] or a second rigid body at the end of a compliant dumbbell as discussed by Paul,[10] by Newton,[11] and by Fischell and Mobley;[12] or a second, fluid body, as considered by Lewis.[13]

In this article we examine a gravity-gradient system anchored by two gyros. A schematic of the system is shown in Fig. 1, where also is indicated the standard nomenclature for axes: the pitch axis is normal to the orbit plane, the yaw axis is along the local vertical, and the roll axis is along the orbital track. Each gyro rotor is contained in a gimbal can (not shown in the schematic), mounted on bearings, and immersed in a fluid bath. Thus, fluid shear produces the required energy dissipation. The

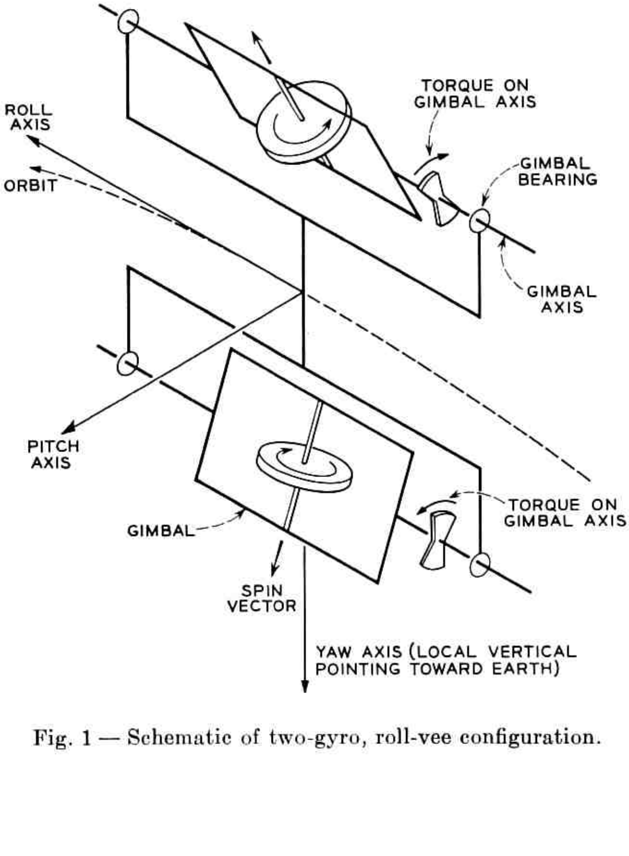


Fig. 1 — Schematic of two-gyro, roll-vee configuration.

gyros are single-axis gyros: that is, the spin vectors are constrained by the gimbal bearings to lie in a single plane within the satellite. In the position shown, this is the pitch-yaw plane.

Because of the small physical dimensions of the gyro "anchor," this system has the virtue that the dissipative joints can be sealed within the satellite; the joints are not exposed to the space environment when the satellite is operating. Also, the required inertia augmentation is particularly simple: only a single extended rod-tip mass.

A simple explanation of how single-axis gyros damp out an arbitrary motion can be given in terms of the rate or torque-seeking property of gyros. A torque applied to a gyro will cause it to precess. By conservation of angular momentum, the precession will try to line up the gyro spin vector with the applied torque or angular rate vector.

Bearing this in mind, assume that the satellite is in orbit in its earth-pointing orientation. It then is rotating at the rate of one revolution per orbit about the pitch axis. If the gyro gimbals were free, this pitch rate would cause the spin vectors to align themselves in the direction of the pitch axis. However, in order to obtain three-axis damping, the spin vectors are held in a vee position by equal and opposite constant torques (see Fig. 1), applied to the gimbals.

Now, if the satellite is disturbed about the pitch axis, both gyros seek the disturbance, resulting in a scissoring motion of the gimbals relative to the satellite, damping out the pitch disturbance. A yaw disturbance causes an in-phase motion of the gyros and again energy is dissipated. Since the gyro spin vectors are constrained to move in the pitch-yaw plane, they are constrained from moving toward a disturbance about the roll axis. However, the roll and yaw motions are coupled. Hence in this case the gyros again try to line up with the yaw axis. Thus three-axis damping is obtained.

Our work continues a study carried out by the Instrumentation Laboratory[2,3] of the Massachusetts Institute of Technology, under the sponsorship of Bell Telephone Laboratories, in which the particular two-gyro configuration studied here was shown to be the most promising of several possible gyro-anchored systems. Our primary objective, however, was not to design a specific attitude control system, which in any case would have to be integrated with the design of a specific satellite, but rather to develop general guiding principles and analytical and numerical techniques useful in such a design problem. Thus, we consider only the broad hardware questions that affect the analysis — for example, the design of extensible rods necessary to augment satellite inertias — but we do not go into the detail of specific gyro hardware, as would be required in a complete design.

The organization of the article is as follows. Section II, for the general reader, summarizes the results of our study in some detail in nonmathematical terms. Following some general remarks about inertia levels, applicable to all gravity-gradient systems, we more fully describe the two-gyro system studied. We then summarize the system's small-angle performance, stressing, in particular, the performance obtained when the inertia of the satellite is augmented by the erection of a single rod. Next we discuss the effects of the main small-angle perturbations: orbital eccentricity, magnetic torques, solar radiation pressure, micrometeorite bombardment and thermal rod bending. Finally, we consider large-angle motions, starting with initial despin upon orbital injection by a combination of rod erection and uncaging of the gyros. In the discussions of large-angle motions, we indicate that there may exist equilibrium positions far removed from the desired, earth-pointing position; we also show how these may be avoided.

Gravity-gradient systems are bistable: that is, associated with a stable, earth-pointing orientation is a second, equally stable orientation obtained by a 180° rotation about the pitch axis. In the concluding section of Section II we describe how the satellite can be flipped from one stable orientation to the other by means of a torque pulse applied to the gimbals.

The results pertaining to the two-gyro system given in Section II serve as an outline of the analysis required for the design of any gravity-gradient attitude control system. They also serve as an introduction to the theory in Sections III and IV. In these parts we present several results and methods that we feel apply generally to the design of many-parameter, linear dynamical systems (see Section III) and to large-angle motions of a satellite (see Section IV).

Specifically, in Section III we develop various bounds on system settling time, and then show how series expansions in terms of system parameters can be used to explore the behavior of a linear system as a function of its parameters. We next describe a computer program based on the Routh criteria, which allows very rapid computation of system response as a function of system parameters. By these means, we are able to survey system behavior over the entire range of six system parameters.

In Section IV, we develop the equations of large-angle motion, including the case of variable inertia, occurring during rod erection. Here we stress the superiority of direction cosines or Euler parameters as compared to Euler angles in satellite kinematics, both from the point of view of computing speed and of ease in visualizing satellite motions. We then give the analysis of equilibrium positions, despin, and flipping or inversion.

## II. SYSTEM DESCRIPTION AND SUMMARY OF RESULTS

### 2.1 *Gravity-Gradient Attitude Control Systems*

All gravity-gradient systems have one feature in common, namely the low magnitude of the gravity-gradient restoring torque, of the order of $I\Omega^2$, where $I$ is a typical satellite moment of inertia and $\Omega$ is the orbital rate. The level of this torque is the main factor determining the steady-state response to constant and periodic disturbing torques. In particular, in the case of a typical communications satellite at an altitude of 6000 nm, the magnitude of the torque exerted by the geomagnetic field on the residual magnetic moment of the satellite is such that the satellite inertia must be increased by a factor of about forty to reduce the steady-state response to an acceptable level.

The low level of the gravity-gradient restoring torque also implies low system natural frequencies, of the order of the orbital rate $\Omega$. Corresponding to this low natural frequency is a minimum $1/e$ settling time of the order of a fraction of an orbit. Zajac[14] has shown that all the systems mentioned above have pitch settling times no less than about one tenth of an orbit. This, of course, is a lower bound on minimum settling time for three-axis motion.

Based on these simple considerations, we would expect that all well designed gravity-gradient attitude control systems would have about the same transient and steady-state performance, that they would all have settling times of a fraction of an orbit, and that they would all require some form of inertia augmentation to obtain acceptable steady-state response. Thus the choice of a particular gravity-gradient attitude control system should be based mainly on ease of mechanization and long-time reliability, rather than system performance.

In the present case, at least, requirements on the large-angle performance of the system (despin, satellite inversion, etc.) preclude choosing the system parameters to give minimum settling time, although the settling time is not greatly increased by meeting the other requirements. It is likely that such a compromise would be necessary for optimum overall performance of any gravity-gradient system, so that the minimum settling time is of academic interest only. Of more importance is the variation of system performance with variation in system parameters. We have thus taken the view that a broad survey of performance as a function of system parameters is of more interest than an optimization based on a single measure of system performance, e.g., settling time.

In the following sections we describe the configuration and performance of the two-gyro system in detail. The interested reader may find the corresponding theoretical analyses in Sections III and IV.

### 2.2 *System Description*

Fig. 2 shows the more important features of the typical single-axis gyro indicated schematically in Fig. 1. The basic element of the gyro is a rotor which spins rapidly about the spin axis and generates a certain angular momentum vector.

The spinning rotor element is enclosed in a sealed gimbal can, mounted on bearings so that it can rotate about a single axis, the gimbal or output axis. A fluid-filled gap between gimbal can and gyro case provides damping as the gimbal rotates.

In the system considered, the two gyros have their gimbal axes along the satellite roll axis. The gyro spin axes are disposed in a vee configuration around the satellite pitch axis, which is also the axis about which the satellite rotates to remain aligned with the local vertical as it traverses its orbit. To distinguish this arrangement from other possible two-gyro configurations,[2-6] it will be called a "roll-vee" configuration.

In the vee arrangement, torques must be supplied constantly to change the direction of the gyro angular momentum vectors, as the satellite traverses its orbit. These torques, constant in magnitude and exerted about the gimbal axes, are provided by a constant electrical signal into electromechanical torquers on the gimbal axes.

It is also possible to inject a signal into the torquers on ground com-

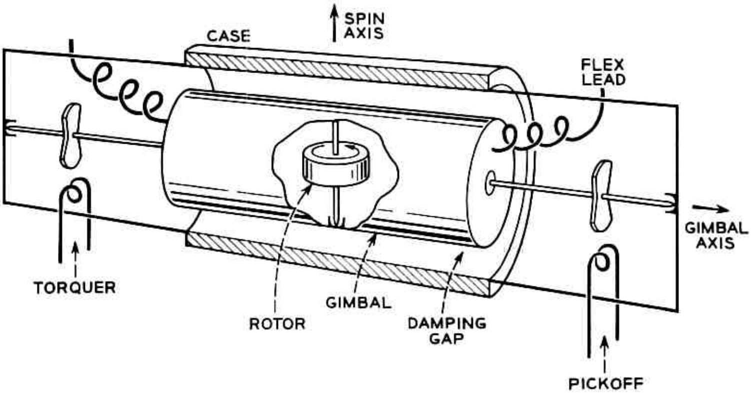


Fig. 2 — Single-axis gyro.

mand. This can be used to invert the satellite if it should get into an undesirable equilibrium position. This possibility is discussed in the sequel.

In order to spin the gyro motor, current must be brought into the gimbal can. This is done by means of highly compliant flex leads. In the present application, the flex-lead spring constants, exerting a small restoring torque around the gimbal axis, can be neglected. However, for a typical communications satellite without inertia augmentation, the flex-lead torques can be of the same order as the gravity-gradient torques.

In any case the gimbal excursions must be limited by suitably placed stops. The location and nature of these stops is an important design consideration. In the first place, undesired equilibrium positions, with the gyro gimbals against the stops, may occur if the stop positions are not carefully chosen. In the second place, large tumbling rates may force the gimbals against the stops, where they are capable of only limited relative motion, depending on the stop elasticity. In both cases the available damping may be greatly reduced. The equilibrium positions may be dealt with analytically, while the large motion may be studied numerically with the stops simulated by hardening springs.

### 2.2.1 *Weight and Power Requirements*

For the attitude control of a typical communications satellite in a 6000-nm altitude orbit, we require two single-axis gyros, each with a rotor angular momentum of about $10^6$ cgs units, weighing about 10 pounds and requiring from 7 to 10 watts power to drive the rotor motor. In addition we require some sort of inertia augmentation which we shall assume is supplied by a single extensible 60-foot rod of the STEM (self-storing tubular extensible member) type, designed and developed by DeHavilland Aircraft of Canada, Ltd., and described in detail in Ref. 8, together with a 20-pound tip mass, which also serves as the tape storage drum. We then have the attitude control system weight breakdown given below:

| | |
|---|---|
| 2 $10^6$ cm-gm-sec gyros | 20 lbs |
| 1 tip mass | 20 lbs |
| 1 extensible rod | 4 lbs |
| gyro power supply<br>(2 lbs of solar cells/watt) | 40 lbs |
| total | 84 lbs |

We have assumed that the satellite proper is a four-foot diameter sphere, weighing 300 pounds, with a moment of inertia of 20 slug-ft$^2$.

It is believed that the above estimates are quite conservative and subject to considerable reduction. The power is used to maintain the gyro rotor speed constant mainly against bearing drag. In a zero-g environment the bearing drag might be substantially reduced. In any case there is probably a trade-off between gyro life, requiring heavily lubricated bearings, and minimum rotor power, requiring light lubrication.

The rod length is chosen to increase the satellite pitch and roll inertias from 20 to 2000 slug-ft². This inertia augmentation sufficiently despins the satellite from currently estimated injection rates of 0.5-1.0 rpm to cause capture by the gravity-gradient field. The required inertia augmentation varies roughly linearly with initial injection rates (see Section 2.4.1). With a sufficiently small injection rate, the augmented inertia could be reduced to the 700 slug-ft² level required to counter magnetic torques (see Section 2.3.4.1). Such a reduction in inertia would mean smaller gyros, and, again, less power.

### 2.3 *Small-Angle Performance*

In order to study the small-angle transient and steady-state response of the roll-vee gyro attitude control system, extensive tables giving decay rates, response to orbital eccentricity, and response to periodic torques at zero, one, and two times orbital frequency Ω as functions of the system parameters were produced by an IBM 7090 computer in a running time of 0.04 hour by a procedure outlined in Section III. Figs. 3 through 15 summarize this broad survey. For each pair of inertia ratios, $B/A$, $C/A$,

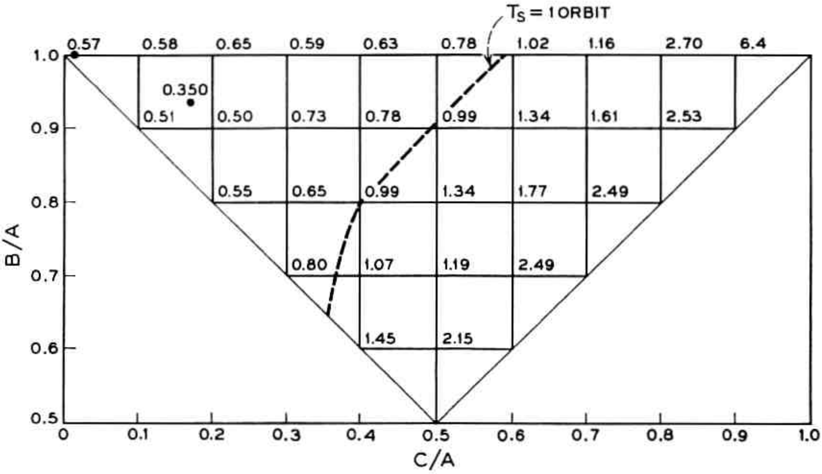


Fig. 3 — Asymptotic settling time in orbits (reduction of $1/e$).

where $A$,$B$,$C$, are the satellite pitch, roll, and yaw moments of inertia, satisfying the inequalities

$$A \geqq B \geqq C, \qquad B + C \geqq A,$$

values of gyro parameters were chosen from these tables to minimize the asymptotic settling time, i.e., the time in which the most lightly damped mode of motion is reduced by $1/e$. Fig. 3 shows the corresponding settling times, while Figs. 4 and 5 give the gyro dimensionless parameters

$$h = (H/A\Omega) \cos \alpha, \qquad h' = (H/C_D) \cos \alpha,$$

where $\alpha$ is the vee half-opening angle, $H$ the gyro angular momentum, and $C_D$ the gyro damping constant for both gyros. Since the small roll-yaw motion depends only on $H$ in the form $H \cos \alpha$, the above is a convenient choice of parameterization. In all cases, except those indicated, the best value of $\alpha$ was 60°, at least over the relatively coarse grid of $\Delta h = 0.25$, $\Delta h' = 0.25$ and $\Delta\alpha = 20°$ used in the tables.

Figs. 6 through 15 give the steady-state response to an orbit eccentricity $\epsilon = 0.01$ and to periodic torques of amplitude 0.01 $A\Omega^2$ for the same values of gyro parameters. Note that the eccentricity response when $\epsilon = 0.01$ is of the order of 1° over the entire range of inertia ratios, having a maximum value of less than 3°. Both the pitch offset, due to a constant pitch torque, and the roll amplitude, due to a periodic roll torque

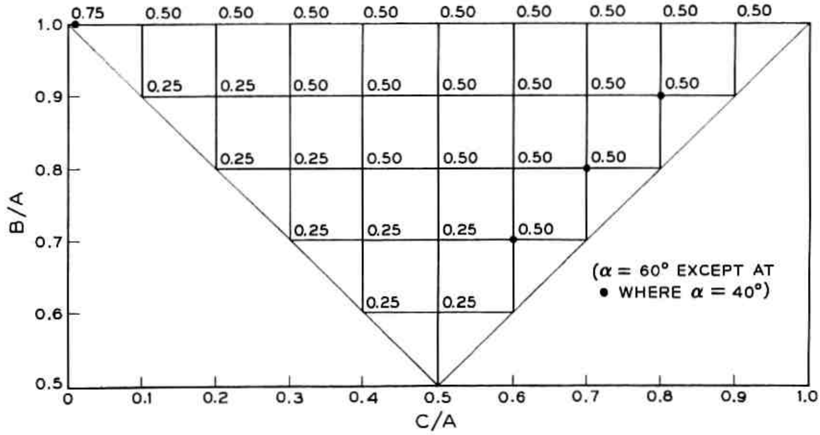


Fig. 4 — Gyro parameter $h = H \cos\alpha/A\Omega$.

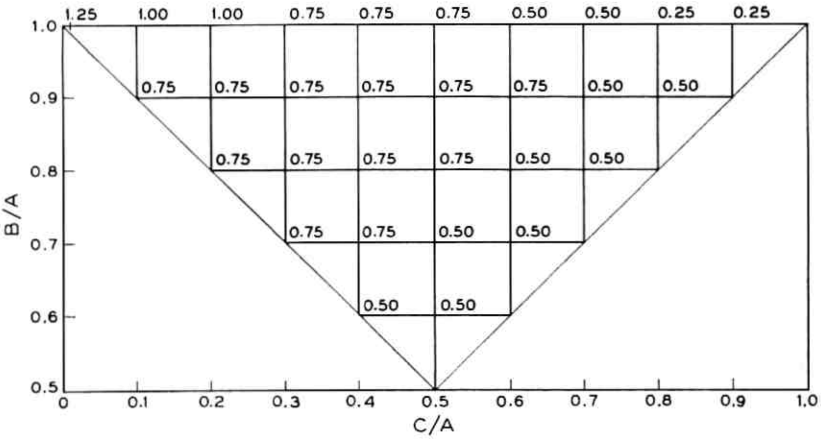


Fig. 5 — Gyro parameter $\boldsymbol{h}' = H \cos \alpha / C_D$ .

at orbital frequency, depend only on the satellite inertias, being given in radians by the simple relations

$$|\varphi_x|_0 = M/3(B - C)\Omega^2, \qquad |\varphi_y|_1 = M/3(A - C)\Omega^2,$$

for a torque of amplitude $M$. Similarly, for torques at frequency $\omega \gg \Omega$, the pitch, roll, and yaw amplitudes tend to the values $M/A\omega^2$, $M/B\omega^2$, $M/C\omega^2$, again independent of the gyro parameters.

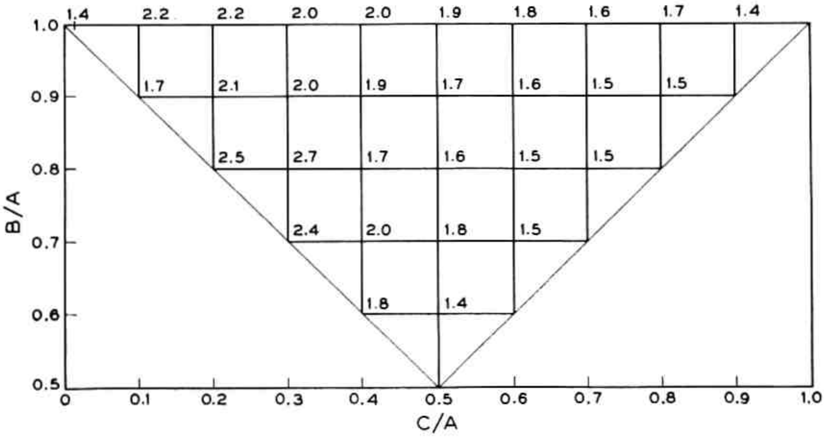


Fig. 6 — Pitch amplitude (degrees) for eccentricity $\epsilon = 0.01$ at orbital frequency.

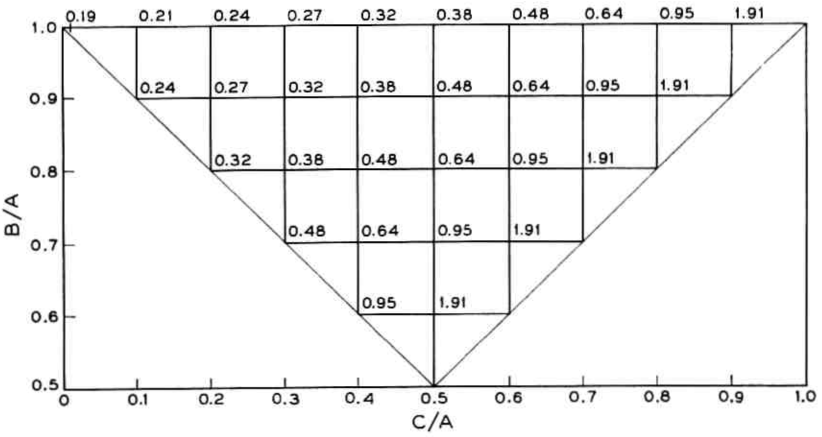


Fig. 7 — Pitch offset (degrees) for constant pitch torque $0.01A\Omega^2$.

### 2.3.1 *The Minimum Settling Time*

These plots do not show the values of inertia ratios and gyro parameters which yield the smallest settling time. A search over a finer grid of parameter values gives a minimum value of settling time of 0.332 orbits, attained for $B/A = 0.925$, $C/A = 0.175$, $h = 0.260$, $h' = 0.688$, $\alpha = 64°$. To attain this value, a slightly negative gimbal spring $K = -0.15\ H\Omega \cos \alpha$ must be used. A negative spring constant may be realized by a simple feedback circuit between gimbal pickoff and gimbal torquer. This

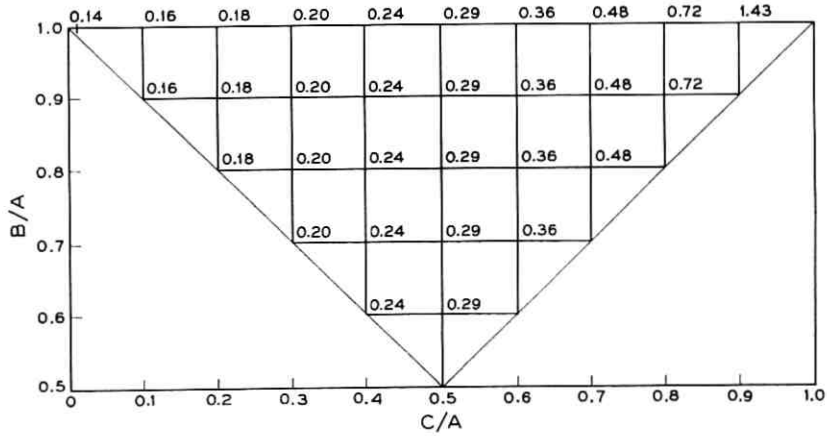


Fig. 8 — Roll offset (degrees) for constant roll torque $0.01A\Omega^2$.

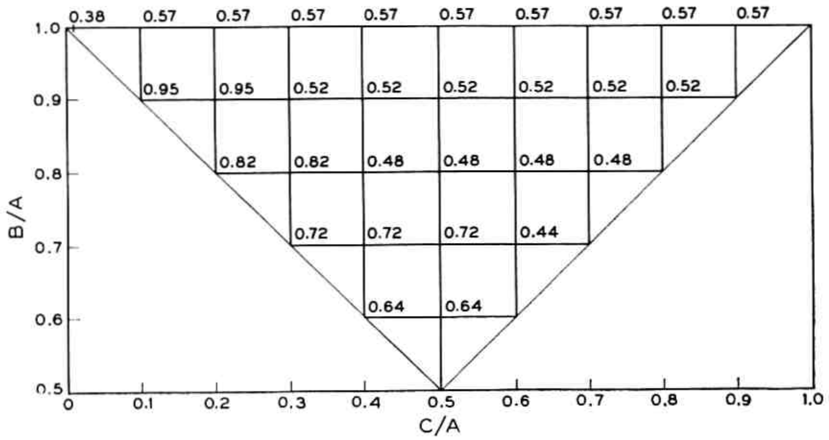


Fig. 9 — Yaw offset (degrees) for constant yaw torque $0.01A\Omega^2$.

minimum settling time is useful as a lower bound, but of more practical interest is the broad range of system parameters over which settling times of less than one orbit can be obtained.

### 2.3.2 *The Spindle*

The figures also do not give performance values for a "dumbbell" or "spindle," i.e., a body for which $A = B \gg C$. This case is of particular

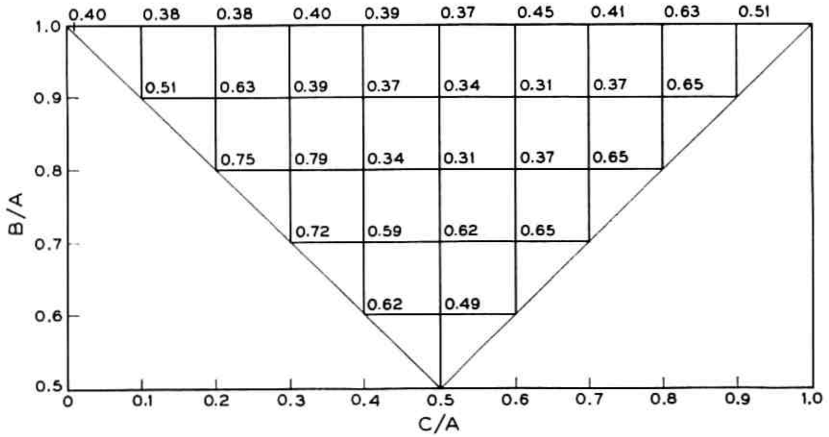


Fig. 10 — Pitch amplitude (degrees) for pitch torque amplitude $0.01A\Omega^2$ at orbital frequency.

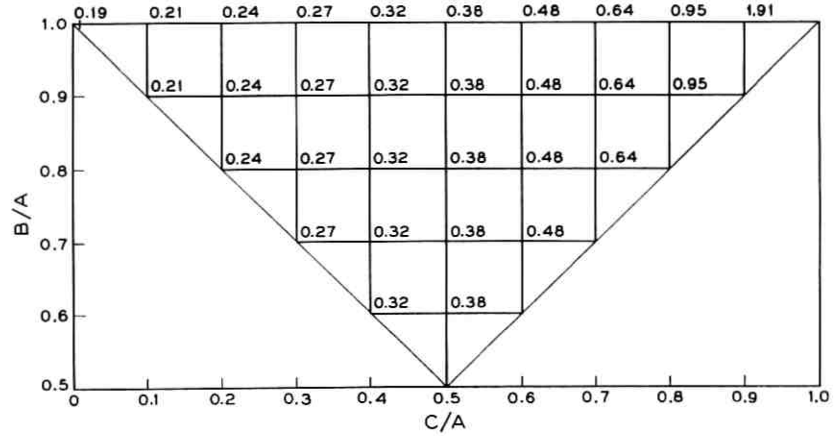


Fig. 11 — Roll amplitude (degrees) for roll torque amplitude $0.01A\Omega^2$ at orbital frequency.

interest, since it may be realized by the erection of a single rod-tip mass combination. To describe the spindle, as well as to give a sample of the tables made by the computer, we reproduce the computer output for $B/A = 1$, $C/A = 0.01$ in Table I. Since the IBM printer has only a limited range of symbols, the following replacements were used:

$$\text{BB} = b = B/A, \qquad \text{CC} = c = C/A,$$

$$\text{KAPPA} = \kappa = 1 + [K/(H\Omega \cos \alpha)],$$

where $K$ is the gimbal spring constant, so that $\kappa = 1$ means zero gimbal spring constant,

$$\text{HH} = h, \ \text{HP} = h',$$

$$\text{ALPHA} = \alpha.$$

The remaining quantities give the transient and steady-state responses. In particular, P0, P1, P2, R0, R1, R2, Y0, Y1, Y2 are the pitch, roll, and yaw amplitudes in degrees for pitch, roll, and yaw torques of amplitude $0.01\ A\Omega^2$ at zero, one, and two times orbital frequency. Note that P0 and R1 are constant, since they depend only on $b$ and $c$, while R0,Y0 are fixed for fixed values of $h$. The quantity E is the pitch amplitude in degrees at orbital frequency for an orbit eccentricity $\epsilon = 0.01$. Finally the quantities labeled "QUINT" and "C" give the real parts of smallest magnitude of the characteristic roots of the roll-yaw quintic and the pitch cubic in terms of the orbital rate $\Omega$. The smallest of these values

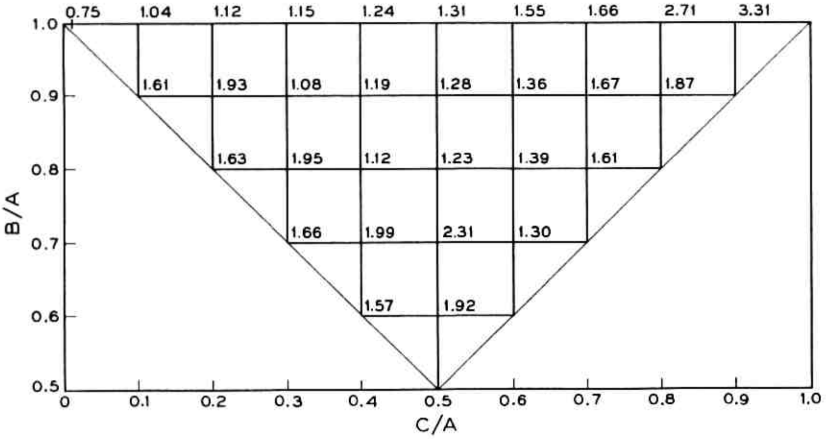


Fig. 12 — Yaw amplitude (degrees) for yaw torque amplitude $0.01A\Omega^2$ at orbital frequency.

$D$ (say) determines the asymptotic settling time $T_s = 1/(2\pi D)$. Inspection of the table reveals that, for $h = 0.750$, $h' = 1.25$, $\alpha = 40°$, we have the smallest settling time, for QUINT $= -0.340(2), -0.657(2)$, i.e., two roots with real parts $-0.340$ and two roots with real part $-0.657$, and one negative real root (not listed) of larger magnitude. Similarly, in this case $C = -0.279(2)$, i.e., two roots of the cubic with real part $-0.279$ and one unlisted negative real root of larger magnitude. The asymptotic

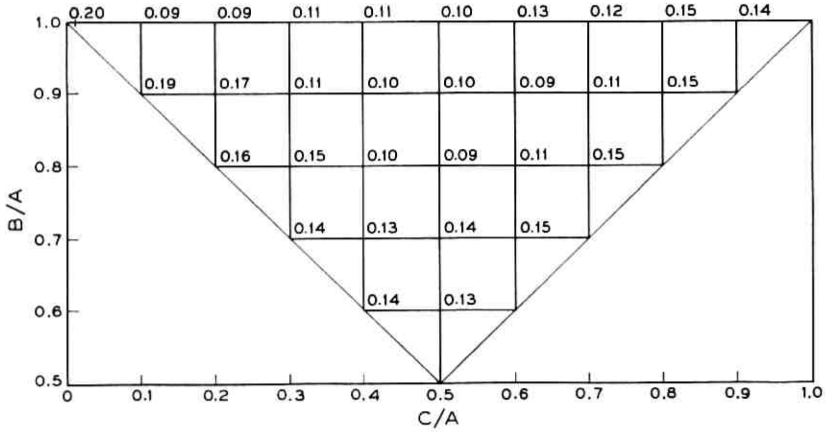


Fig. 13 — Pitch amplitude (degrees) for pitch torque amplitude $0.01A\Omega^2$ at twice orbital frequency.

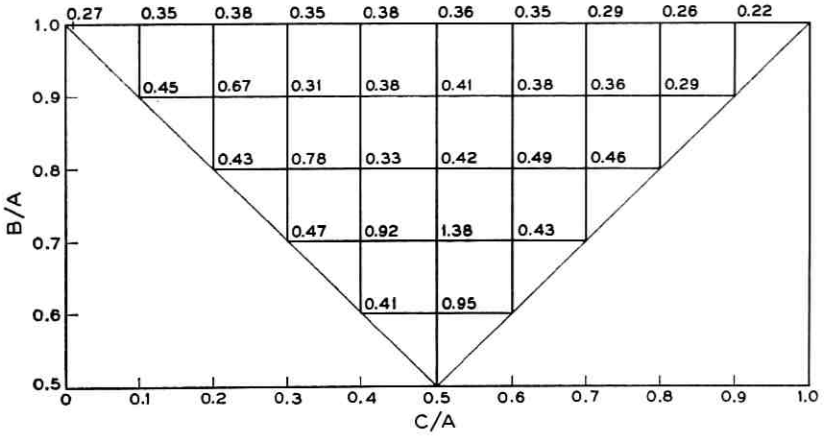


Fig. 14 — Roll amplitude (degrees) for roll torque amplitude $0.01A\Omega^2$ at twice orbital frequency.

settling time is then given by $1/[(2\pi)\ (0.279)] = 0.57$ orbits. Despite the coarseness of the table, this is very close to the minimum value of 0.50 orbits for a spindle, attained for $h = 0.77$, $h' = 1.29$, $\alpha = 38°$. This minimum value may be calculated by an asymptotic expansion in the large quantity $h/c = (H \cos \alpha)/C\Omega$.

Let us now attempt a specific "design." This design must be regarded

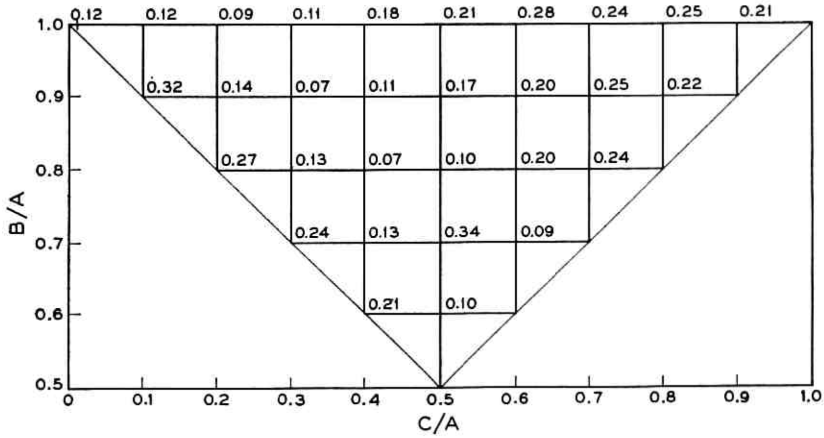


Fig. 15 — Yaw amplitude (degrees) for yaw torque amplitude $0.01A\Omega^2$ at twice orbital frequency.

as illustrative, rather than definitive, since a real design must take into account the fine details of gyro hardware as well as requirements imposed by the use of the satellite in an actual communications system. For example, it is not at all clear what limits on maximum settling time would be imposed by system requirements. We have tentatively set this maximum settling time at one orbit.

### 2.3.3 *Transient Response for a Spindle.*

Fig. 16 shows the asymptotic settling time in orbits as a function of the dimensionless gyro angular momentum $H/A\Omega$ for $\alpha = 40°$, $h' = 1.25$ and for $\alpha = 60°$, $h' = 1.00$. The former gives a minimum settling time very near the optimum value for a spindle for $H/A\Omega$ nearly unity, but varies more rapidly with $H/A\Omega$ than does the other system. Also, we are particularly interested in large values of $H/A\Omega$—i.e., $H/A\Omega > 2$—since we propose to use the gyros as inertia wheels in the initial despin of the satellite after injection into orbit. In this case the second system gives a considerably smaller settling time (0.85 orbits, compared with 1 orbit, at $H/A\Omega = 2$). We may actually increase $H/A\Omega$ to about 2.4 in this case and stay within the maximum settling time of 1 orbit. Undoubtedly, by trimming the values of $\alpha$ and $h'$, we may increase $H/A\Omega$ even more, but, since this is intended to be an illustrative design, we do not consider these questions further here; instead, we simply take as our "design" $\alpha = 60°$, $h' = h = 1.00$ $(H/A\Omega = 2.00)$. In the illustrative examples of the sequel, these parameter values will be assumed. From the table, they yield

$$\begin{aligned} \text{QUINT} &= -0.189(2),\ -1.318(2), \\ C &= -0.190(2). \end{aligned}$$

Given the real parts of 4 roots of the roll-yaw quintic and 2 roots of the pitch cubic, it is a simple matter to calculate all the characteristic roots completely, especially for a spindle. In the present case we find solutions of the form

$$e^{-0.19\Omega t} \begin{matrix} \sin \\ \cos \end{matrix} (0.64\Omega t), \qquad e^{-6.62\Omega t},$$

for the pitch motion, and

$$e^{-0.19\Omega t} \begin{matrix} \sin \\ \cos \end{matrix} (1.40\Omega t), \qquad e^{-1.32\Omega t} \begin{matrix} \sin \\ \cos \end{matrix} (0.53\Omega t), \qquad e^{-200\Omega t},$$

TABLE I — COMPUTER OUTPUT FOR SPINDLE SHAPE ($A = B \gg C$)

BB = 1.000, CC = 0.010, KAPPA = 1.000

```
              HH=0.250                                    HH=0.500                                    HH=0.750
      P0= 0.19 R0= 0.14 Y0 =1.15                  P0= 0.19 R0= 0.14 Y0= 0.57                  P0= 0.19 R0= 0.14 Y0= 0.38
   HP=0.500,QUINT=-0.113(2),  -0.576(1)        HP=0.500,QUINT=-0.118(2),  -0.474(1)        HP=0.500,QUINT=-0.101(2), -0.410(1)
ALPHA R1= 0.19,R2= 0.46,Y1= 1.48,Y2= 0.35     R1= 0.19,R2= 0.23,Y1= 0.82,Y2= 0.18         R1= 0.19,R2= 0.16,Y1= 0.61,Y2= 0.12
 20.0 P1= 0.29, P2= 0.55,E= 0.59,C=-0.015(2)   P1= 0.29, P2= 0.54,E= 0.61,C=-0.030(2)      P1= 0.30, P2= 0.52,E= 0.62,C=-0.450(2)
 40.0     0.30      0.49     0.65   -0.080(2)      0.31      0.42     0.73   -0.160(2)         0.32      0.35     0.81    -0.239(2)
 60.0     0.32      0.29     0.92   -0.337(2)      0.31      0.17     1.26   -0.637(2)         0.27      0.12     1.43    -0.687(2)
 80.0     0.09      0.04     1.35   -0.170(2)      0.04      0.02     1.25   -0.088(2)         0.03      0.01     1.22    -0.059(2)
   HP=0.750,QUINT=-0.150(2),  -0.863(2)        HP=0.750,QUINT=-0.184(2),  -1.170(2)        HP=0.750,QUINT=-0.161(2), -0.804(1)
ALPHA R1= 0.19,R2= 0.56,Y1= 1.62,Y2= 0.35     R1= 0.19,R2= 0.28,Y1= 0.88,Y2= 0.18         R1= 0.19,R2= 0.19,Y1= 0.65,Y2= 0.12
 20.0 P1= 0.29, P2= 0.54,E= 0.60,C=-0.020(2)   P1= 0.30, P2= 0.52,E= 0.63,C=-0.041(2)      P1= 0.30, P2= 0.49,E= 0.65,C=-0.061(2)
 40.0     0.31      0.44     0.71   -0.109(2)      0.33      0.35     0.85   -0.212(2)         0.34      0.28     1.01    -0.307(2)
 60.0     0.36      0.22     1.22   -0.408(2)      0.34      0.12     1.71   -0.456(2)         0.26      0.08     1.78    -0.336(2)
 80.0     0.07      0.02     1.37   -0.110(2)      0.03      0.01     1.25   -0.058(2)         0.02      0.01     1.22    -0.039(2)
   HP=1.000,QUINT=-0.159(2),  -0.582(2)        HP=1.000,QUINT=-0.250(2),  -0.750(2)        HP=1.000,QUINT=-0.236(2), -1.017(2)
ALPHA R1= 0.19,R2= 0.69,Y1= 1.79,Y2= 0.35     R1= 0.19,R2= 0.34,Y1= 0.96,Y2= 0.18         R1= 0.19,R2= 0.23,Y1= 0.69,Y2= 0.12
 20.0 P1= 0.30, P2= 0.53,E= 0.61,C=-0.024(2)   P1= 0.30, P2= 0.50,E= 0.64,C=-0.049(2)      P1= 0.31, P2= 0.47,E= 0.68,C=-0.072(2)
 40.0     0.32      0.40     0.76   -0.124(2)      0.35      0.29     0.97   -0.231(2)         0.37      0.23     1.21    -0.312(2)
 60.0     0.40      0.17     1.52   -0.364(2)      0.36      0.10     2.13   -0.309(2)         0.25      0.07     2.00    -0.236(2)
 80.0     0.06      0.02     1.38   -0.082(2)      0.03      0.01     1.26   -0.044(2)         0.02      0.01     1.22    -0.029(2)
   HP=1.250,QUINT=-0.149(2),  -0.437(2)        HP=1.250,QUINT=-0.277(2),  -0.516(2)        HP=1.250,QUINT=-0.340(2), -0.657(2)
ALPHA R1= 0.19,R2= 0.82,Y1= 1.99,Y2= 0.35     R1= 0.19,R2= 0.41,Y1= 1.06,Y2= 0.18         R1= 0.19,R2= 0.27,Y1= 0.75,Y2= 0.12
 20.0 P1= 0.30, P2= 0.52,E= 0.62,C=-0.027(2)   P1= 0.30, P2= 0.48,E= 0.66,C=-0.052(2)      P1= 0.31, P2= 0.44,E= 0.70,C=-0.076(2)
 40.0     0.32      0.37     0.80   -0.130(2)      0.36      0.26     1.07   -0.226(2)         0.40      0.20     1.39    -0.279(2)
 60.0     0.45      0.15     1.83   -0.297(2)      0.39      0.08     2.48   -0.237(2)         0.25      0.06     2.14    -0.184(2)
 80.0     0.05      0.02     1.38   -0.066(2)      0.02      0.01     1.26   -0.035(2)         0.02      0.01     1.22    -0.023(2)
```

Table I — *continued*

| | HH = 1.000 | | | | HH = 1.250 | | | | HH = 1.500 | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | P0= 0.19 R0= 0.14 Y0= 0.29 | | | | P0= 0.19 R0= 0.14 Y0= 0.23 | | | | P0= 0.19 R0= 0.14 Y0= 0.19 | | | |
| | HP=0.500, QUINT=−0.085(2), −0.365(1) | | | | HP=0.500, QUINT=−0.071(2), −0.329(1) | | | | HP=0.500, QUINT=−0.060(2), −0.301(1) | | | |
| ALPHA | R1= 0.19, R2= 0.12, Y1= 0.50, Y2= 0.09 | | | | R1= 0.19, R2= 0.09, Y1= 0.44, Y2= 0.08 | | | | R1= 0.19, R2= 0.08, Y1= 0.40, Y2= 0.06 | | | |
| 20.0 | P1= 0.30, | P2= 0.51, | E= 0.63, | C=−0.061(2) | P1= 0.30, | P2= 0.50, | E= 0.65, | C=−0.076(2) | P1= 0.30, | P2= 0.48, | E= 0.66, | C=−0.091(2) |
| 40.0 | 0.32 | 0.30 | 0.90 | −0.317(2) | 0.32 | 0.26 | 0.99 | −0.393(2) | 0.32 | 0.23 | 1.08 | −0.467(2) |
| 60.0 | 0.23 | 0.09 | 1.48 | −0.465(2) | 0.19 | 0.08 | 1.47 | −0.361(2) | 0.16 | 0.06 | 1.45 | −0.298(2) |
| 80.0 | 0.02 | 0.01 | 1.20 | −0.045(2) | 0.02 | 0.01 | 1.19 | −0.036(2) | 0.01 | 0.01 | 1.18 | −0.030(2) |
| | HP=0.750, QUINT=−0.133(2), −0.645(1) | | | | HP=0.750, QUINT=−0.110(2), −0.556(1) | | | | HP=0.750, QUINT=−0.091(2), −0.494(1) | | | |
| ALPHA | R1= 0.19, R2= 0.14, Y1= 0.53, Y2= 0.09 | | | | R1= 0.19, R2= 0.11, Y1= 0.46, Y2= 0.08 | | | | R1= 0.19, R2= 0.10, Y1= 0.41, Y2= 0.06 | | | |
| 20.0 | P1= 0.30, | P2= 0.47, | E= 0.67, | C=−0.082(2) | P1= 0.31, | P2= 0.45, | E= 0.70, | C=−0.102(2) | P1= 0.31, | P2= 0.43, | E= 0.72, | C=−0.122(2) |
| 40.0 | 0.36 | 0.23 | 1.17 | −0.389(2) | 0.36 | 0.19 | 1.33 | −0.453(2) | 0.36 | 0.17 | 1.47 | −0.489(2) |
| 60.0 | 0.20 | 0.06 | 1.70 | −0.264(2) | 0.16 | 0.05 | 1.61 | −0.218(2) | 0.13 | 0.04 | 1.54 | −0.186(2) |
| 80.0 | 0.02 | 0.01 | 1.20 | −0.029(2) | 0.01 | 0.01 | 1.19 | −0.023(2) | 0.01 | 0.00 | 1.18 | −0.020(2) |
| | HP=1.000, QUINT=−0.189(2), −1.318(2) | | | | HP=1.000, QUINT=−0.152(2), −0.990(1) | | | | HP=1.000, QUINT=−0.125(2), −0.786(1) | | | |
| ALPHA | R1= 0.19, R2= 0.17, Y1= 0.56, Y2= 0.09 | | | | R1= 0.19, R2= 0.14, Y1= 0.48, Y2= 0.08 | | | | R1= 0.19, R2= 0.11, Y1= 0.43, Y2= 0.06 | | | |
| 20.0 | P1= 0.31, | P2= 0.44, | E= 0.71, | C=−0.095(2) | P1= 0.32, | P2= 0.41, | E= 0.74, | C=−0.118(2) | P1= 0.32, | P2= 0.38, | E= 0.78, | C=−0.140(2) |
| 40.0 | 0.40 | 0.18 | 1.46 | −0.359(2) | 0.41 | 0.15 | 1.70 | −0.371(2) | 0.41 | 0.13 | 1.90 | −0.361(2) |
| 60.0 | 0.18 | 0.05 | 1.81 | −0.190(2) | 0.14 | 0.04 | 1.67 | −0.159(2) | 0.11 | 0.03 | 1.58 | −0.137(2) |
| 80.0 | 0.01 | 0.00 | 1.20 | −0.022(2) | 0.01 | 0.00 | 1.19 | −0.018(2) | 0.01 | 0.00 | 1.18 | −0.015(2) |
| | HP=1.250, QUINT=−0.266(2), −0.933(2) | | | | HP=1.250, QUINT=−0.202(2), −1.199(2) | | | | HP=1.250, QUINT=−0.161(2), −1.441(2) | | | |
| ALPHA | R1= 0.19, R2= 0.20, Y1= 0.60, Y2= 0.09 | | | | R1= 0.19, R2= 0.16, Y1= 0.51, Y2= 0.08 | | | | R1= 0.19, R2= 0.14, Y1= 0.45, Y2= 0.06 | | | |
| 20.0 | P1= 0.32, | P2= 0.41, | E= 0.74, | C=−0.101(2) | P1= 0.32, | P2= 0.38, | E= 0.78, | C=−0.123(2) | P1= 0.33, | P2= 0.35, | E= 0.83, | C=−0.144(2) |
| 40.0 | 0.44 | 0.16 | 1.74 | −0.296(2) | 0.46 | 0.13 | 2.07 | −0.291(2) | 0.46 | 0.11 | 2.33 | −0.278(2) |
| 60.0 | 0.17 | 0.04 | 1.87 | −0.150(2) | 0.13 | 0.03 | 1.70 | −0.126(2) | 0.10 | 0.03 | 1.60 | −0.108(2) |
| 80.0 | 0.01 | 0.00 | 1.20 | −0.017(2) | 0.01 | 0.00 | 1.19 | −0.014(2) | 0.01 | 0.00 | 1.18 | −0.011(2) |

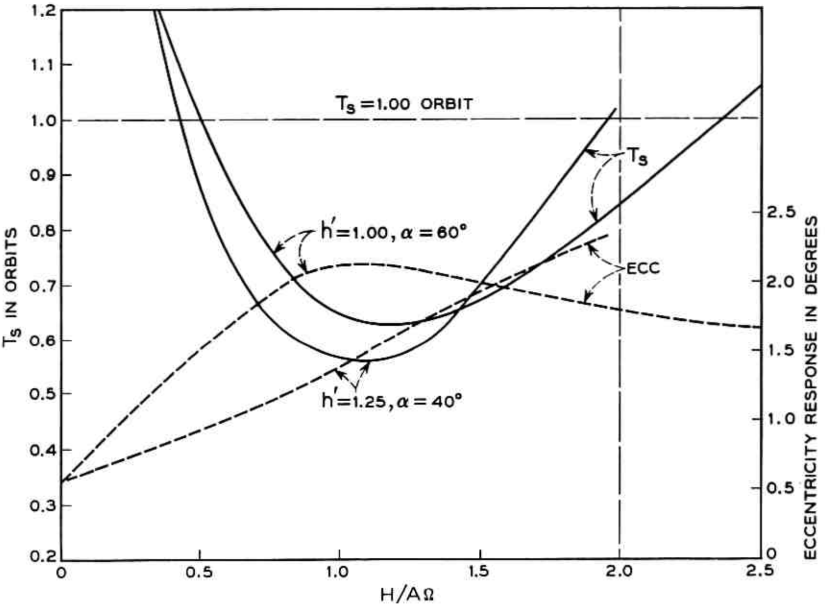


Fig. 16 — Asymptotic settling time in orbits versus gyro angular momentum.

for the roll-yaw motion. Note that the period of the oscillatory solutions is comparable to the orbital period, as we would expect, and that both motions have rapidly damped exponential solutions. The latter feature, typical of spindle-shaped bodies, causes difficulties in numerical integration of the differential equations, both for small and large motion, for it implies that derivatives may be very much larger than the dependent variable itself.

2.3.3.1 *Micrometeorite Impact.* One source of transient disturbance is the angular momentum imparted by micrometeorite impact. It was estimated in Ref. 8 that for a satellite of comparable inertia level, impacts producing offsets greater than 5° would occur every two years and impacts large enough to tumble the satellite every 23 years, on the average. A more recent study[24] of the present two-gyro system indicates similar times if Whipple's 1958 micrometeorite data are used. For Whipple's 1963 data, the corresponding 5° and tumbling times are 40 years and 1000 years. From a systems point of view, the low frequency of occurrence of these disturbances suggests that a settling time of 1 orbit is quite adequate.

### 2.3.4 *Steady-State Response*

Disturbances producing constant or periodic pointing errors may be classified as either kinematic or dynamic. The response to the former, of which orbit eccentricity is a typical example, is essentially independent of satellite inertia; the response to the latter type of disturbance may be reduced simply by increasing the level of satellite inertia to a suitable value. The kinematic response limits the minimum attainable pointing error; the dynamic response to given disturbing torques sets the level of satellite inertia.

In the case of the spindle, the table yields an eccentricity response amplitude at orbital frequency of $E = 1.81°$ for an orbit eccentricity $\epsilon = 0.01$. We also find the steady-state response amplitudes due to torque amplitudes of $0.01A\Omega^2$ given by

$$\begin{array}{lll} \mathrm{P0} = 0.19°, & \mathrm{P1} = 0.18°, & \mathrm{P2} = 0.05°, \\ \mathrm{R0} = 0.14°, & \mathrm{R1} = 0.19°, & \mathrm{R2} = 0.17°, \\ \mathrm{Y0} = 0.29°, & \mathrm{Y1} = 0.56°, & \mathrm{Y2} = 0.09°. \end{array}$$

We are particularly interested in the pitch offset P0, due to a constant pitch torque, and the roll amplitude R1, due to a roll torque at orbital frequency. Both of these are independent of gyro parameters and equal in the case of a spindle. Together with a given disturbing torque, they serve to set the satellite inertia level.

2.3.4.1 *Magnetic Torque — Satellite Inertia Level.* In the case of a communications satellite, one of the principal disturbing torques is the torque exerted by the geomagnetic field on the residual magnetic moment of the satellite. It has been estimated in Ref. 8 that this torque might be as large as $5 \times 10^{-6}$ ft-lb for a satellite like the Telstar satellite at an altitude of 6000 nm. At this altitude $\Omega = 2.73 \times 10^{-4}$ rad/sec ($\sim$1 rad/hr). Because of the steady rotation of the earth-pointing satellite, this torque does not have a constant pitch component, but it will have a roll component at orbital frequency. Thus R1 is the response amplitude of interest. To make R1 equal to the eccentricity response of 1.81° requires a satellite pitch moment of inertia $A$ such that $A\Omega^2$ is ten times the above torque, yielding $A = 670$ slug-ft$^2$. Since a typical moment of inertia for a satellite somewhat larger than the Telstar satellite is 20 slug-ft$^2$, this calculation indicates that some sort of inertia augmentation is required. We shall assume that the satellite proper has equal moments of inertia $A_0 = B_0 = C_0 = 20$ slug-ft$^2$ and that the pitch and roll moments of inertia are increased to 2000 slug-ft$^2$ by the erection of a single 60-ft extensible rod and a tip mass of 20 pounds. As indicated in Section 2.2.1, we make the inertia somewhat larger than the minimum required to counter magnetic torque in order that rod erection may be used for satellite despin.

2.3.4.2 *Solar Radiation Pressure.* When rods are erected to augment the satellite inertia, the perturbing torque due to incident solar radiation is in general increased, but not at the same rate as the inertia, if a thin rod and a dense tip mass are used. To give some idea of the order of magnitude of this torque, let us consider a 2-foot radius, 300-pound satellite, joined to a 0.2-foot radius, 20-pound tip mass by a 58-foot-long, 0.04-foot diameter rod. This yields a maximum solar torque of about $2 \times 10^{-6}$ ft-lb ($0.0075\ A\Omega^2$ for $A = 2000$ slug-ft$^2$) around the center of mass of the system, which lies about 4 feet from the center of the satellite proper. This low torque is the result of a partial balance between the resultant force on the satellite and the resultant force on the rod, both yielding torques of the order of $5 \times 10^{-6}$ ft-lb ($0.019\ A\Omega^2$). Even using this figure, the deflection due to solar pressure will be no larger than that due to magnetic torque. Thus such a satellite need not be especially designed to balance out solar torques.

2.3.4.3 *Bending Due to Solar Heating.* In Ref. 8 the bending of an extensible rod of the STEM type, due to differential solar heating, was analyzed. Further unpublished work by P. Hrycak and by J. G. Engstrom at Bell Telephone Laboratories leads to the formula

$$d/L = (L/4r)\pi\alpha T_0/[\kappa + 4 + 16\beta/3],$$

for the deflection $d$ of a rod of length $L$, radius $r$, and expansion coefficient $\alpha$, where

$$T_0 = (a_0 S/\pi e_0 \sigma)^{\frac{1}{4}},$$

$$\kappa = \pi k h T_0/r^2 a_0 S, \qquad \beta = e_i/e_0\,,$$

with $a_0$ and $e_0$ the rod external absorptivity and emissivity, $e_i$ the rod internal emissivity, $\sigma$ the Stefan-Boltzmann constant, $S$ the flux of solar radiation through unit area in unit time, $k$ the rod thermal conductivity, and $h$ the rod wall thickness. The dimensionless quantity $\kappa$ gives the ratio between heat transferred by conduction and by radiation. Typical values of the above quantities are:

$$\begin{array}{lll}
L = 60 \text{ ft}, & r = 0.02 \text{ ft}, & h = 2 \times 10^{-4} \text{ ft}, \\
S = 442 \text{ Btu/ft}^2\text{-hr}, & \sigma = 171 \times 10^{-11} \text{ Btu/ft}^2\text{-hr-}(^\circ\text{R})^4, & \\
k = 65 \text{ Btu/ft-hr-}^\circ\text{F}, & \alpha = 10^{-5}/^\circ\text{F}, & \\
a_0 = 0.67, & c_0 = e_i = 0.33, &
\end{array}$$

the last five values being appropriate for beryllium copper. In this case $T_0 = 635°\text{R}$, $\kappa = 219$, $\beta = 1$, and $d/L = 0.0656$. Note that in this case $\kappa \gg 1$, so that conduction effects dominate. Unless $\beta \gg 1$, i.e., the outside of the rod is highly reflecting and the inside "black," we may use the simpler formula

$$d/L = (L/4r)(\alpha r^2 a_0 S/kh),$$

obtained by neglecting the remaining terms in the denominator of the previous formula in comparison with $\kappa$. In the present case this yields $d/L = 0.0684$, compared with the more exact value of 0.0656.

The above displacement $d$ is the displacement of the tip mass at the end of the rod and hence produces a corresponding rotation of the yaw principal axis of inertia through an angle of order $d/L = 0.0656 = 3.7°$ and an antenna pointing error of the same size. Note that this angle increases linearly with $L$, so that thermal bending sets an upper limit on the length of a given type of rod which may be used for inertia augmentation. In observations of the Applied Physics Laboratory 1963 22A satellite,[12] thermal bending manifested itself apparently as a high-frequency oscillation of the satellite's attitude, attributed to the rapid heating of an extensible rod on passage from shadow into sunlight.

### 2.4 *Large-Angle Motion*

We shall discuss various large-angle motions of the gravity-oriented, gyro-stabilized satellite in order of their occurrence. First we consider the injection, despin, and capture of the satellite in orbit and the equilibrium positions into which it may settle. Next we discuss the use of the gyros to flip the satellite in case it settles into the inverted equilibrium position. Finally we report the results of computer studies of various large motions.

#### 2.4.1 *Satellite Despin*

We assume that the satellite is injected into a nearly circular, 6000-nm orbit with an initial spin rate of less than 1 rpm around an arbitrary axis. After injection, erection of a single 60-foot rod with a 20-pound tip mass then increases the moment of inertia around axes normal to the rod from 20 slug-ft$^2$ to 2000 slug-ft$^2$ and decreases the spin rate around these axes by a corresponding factor, e.g., from 250 rpo (revolutions per orbit) to 2 rpo. The component of spin around the rod axis is, of course, unaffected by rod erection. This component of spin is removed by uncaging the gyros from their nominal equilibrium position, in which they have a zero net component of angular momentum around the rod axis (the body yaw axis) and allowing them to precess toward the spin rate vector. The net change of yaw angular momentum due to this precession is of the order of the gyro angular momentum $H = 2\,A\Omega$, where $A$ is the final moment of inertia (2000 slug-ft$^2$) around the body pitch and roll axes normal to the rod, and the angular momentum due to the initial spin around the rod axis is $250\,C\Omega = 250\,A_0\Omega = 2.5\,A\Omega$, of the same order of magnitude. Note that this latter despin is in proportion to the difference of angular momenta, rather than their ratio, so that we might expect difficulties with the small differences of large numbers, leaving us with a sizeable angular velocity around the body yaw axis. However, the yaw component of angular velocity rapidly settles out; it is the yaw angular momentum, rather than the yaw angular velocity, which is of importance.

This is shown in Figs. 17–18 where, for the design of Section 2.3.2,

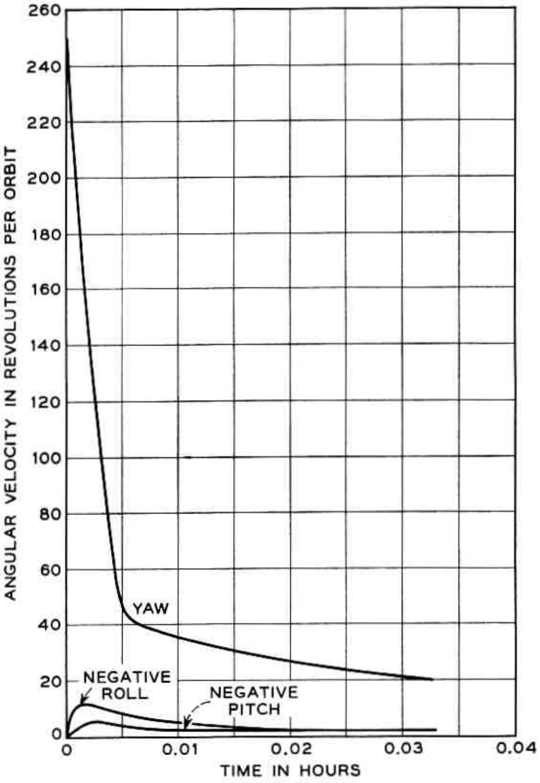


Fig. 17 — Despin during boom erection.

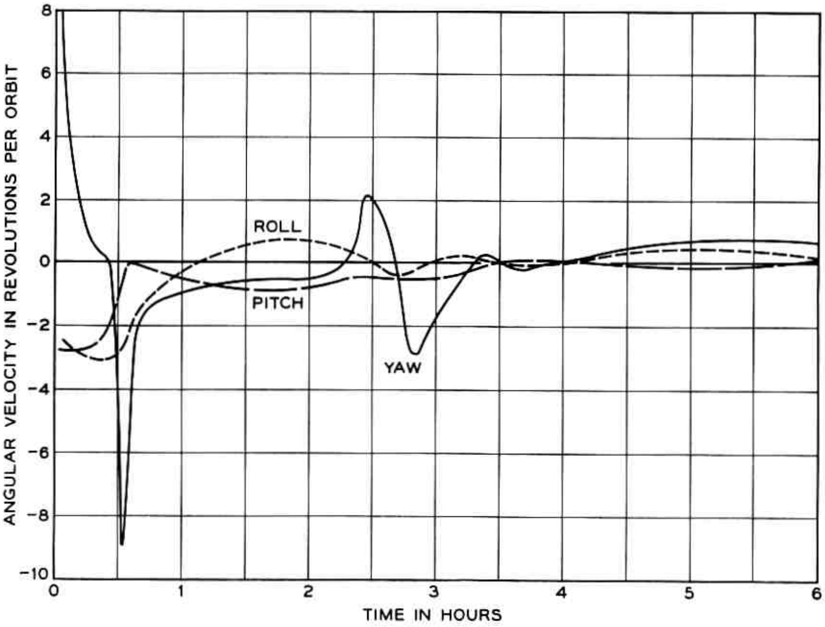


Fig. 18 — Despin to capture following boom erection.

yaw, pitch, and roll rate, obtained by a digital computer, are plotted against time. At $t = 0$, the satellite was assumed injected into the desired orientation, with gyros at the null position, and with a yaw rate of 250 rpo (approximately $\frac{2}{3}$ rpm). The elapsed time of Fig. 17 is two minutes, corresponding to boom erection. In this time, the yaw rate decreases to 20 rpo, while pitch and roll first peak at $-12$ rpo and $-6$ rpo respectively and then decay to $-2.5$ rpo at the end of boom erection. Subsequently, as shown in Fig. 18, all three rates decrease to less than 1 rpo at the end of 1 orbit.

### 2.4.2 *Equilibrium Positions*

Four equilibrium positions, in which the satellite is stationary with respect to the rotating local vertical, may be found by inspection. Two of these, shown schematically in Fig. 19, are the stable roll-vee positions with the gyro angular momentum vectors making a symmetrical vee with the orbit pitch axis (normal to the orbit plane), the gyro gimbal axes along the orbit roll axis (tangent to the orbit track), and the rod along the local vertical. The satellite antenna in this case is either directed toward the earth or away from the earth. We discuss the inversion

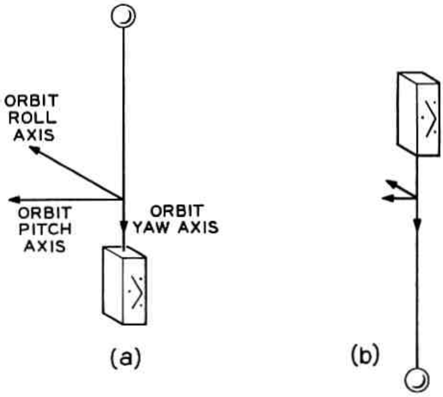


Fig. 19 — Equilibrium positions, stable roll-vee.

of the satellite from the latter position in the sequel. Two other equilibrium positions are yaw-vee positions, Fig. 20, with the gimbal axes along the orbit yaw axis, i.e., the local vertical, and the gyro angular momentum vectors again making a symmetrical vee with the orbit pitch axis, and the rod along the orbit roll axis. These two positions are unstable, however, just as they would be without the gyros.

Other equilibrium positions occur because of the presence of the gyro gimbal stops. Suppose, for example, that the satellite is rotated around the local vertical through 180° from its normal operating position. The gyro gimbal torquers which normally hold the gyro vee open against the 1 rpo steady precession of the satellite in orbit, now act with the precession to force the gyro gimbals against stops located near the body yaw axis. The resulting symmetrical reverse vee configuration (see Fig. 21a) is a possible satellite equilibrium position. Although the satellite antenna is still directed toward the earth in this position, it is an undesirable equilibrium position, because, when the gyro gimbals are against stops, their damping capability is severely reduced. This reversed equilibrium position can be made unstable by moving the gimbal stops in

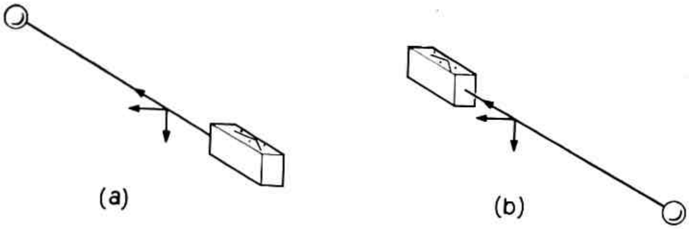


Fig. 20 — Equilibrium positions, unstable yaw-vee.

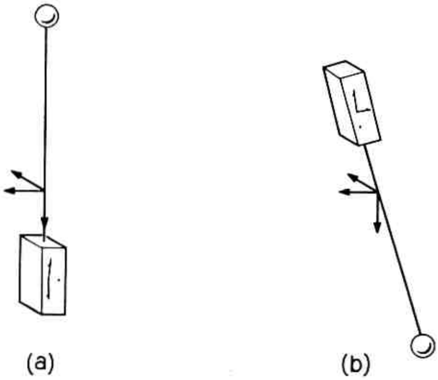


Fig. 21 — Equilibrium positions, reverse-vee and skewed.

from the yaw axis and by choosing the gyro angular momentum to have a suitable value, as discussed in Section IV. The inversion of the position shown in Fig. 21(a) and the corresponding reversed yaw-vee positions are unstable as before.

Finally we note the possibility of skewed equilibrium positions, in which the body principal axes do not coincide with the orbit axes and both gyro gimbals are against stops (see Fig. 21b). Examples are discussed in Section 4.3. Such unsymmetrical equilibrium positions may be easily eliminated by appropriate choices of stop positions and gyro angular momentum, but their occurrence suggests the necessity of a thorough investigation of equilibrium positions for any attitude control system, especially one in which constraints due to stops are present. The investigation of equilibrium positions also may serve as a guide in singling out lightly damped modes of large motion.

### 2.4.3 *Satellite Inversion*

As we have already noted, the satellite may be captured, after injection into orbit, in inverted position with its antenna pointing away from the earth. With sufficiently large gyros it may be flipped from this position by changing the net gyro angular momentum by means of a simple signal injected by ground command into the gyro gimbal torquers. We simply reverse the polarity of the bias signals into the gyro torquers for a preset short time interval. The resulting change in angular momentum is just enough to cause the satellite to tumble, so that it is captured again in its normal operating position. Fig. 22 shows the result of such an inversion procedure, where the polarity is switched for $\frac{1}{2}$ orbit. Here,

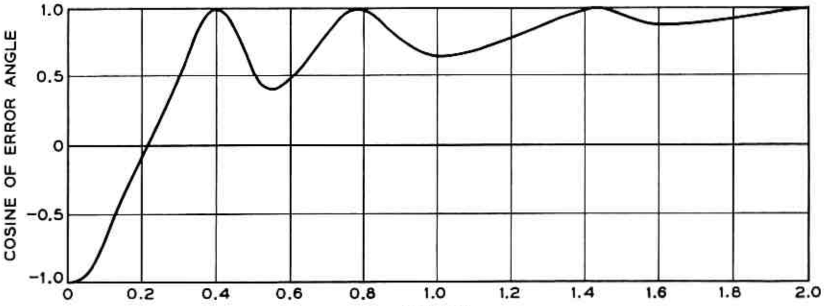


Fig. 22 — Error angle versus time during satellite inversion. Torquer polarities interchanged for 1/2 orbit.

we have plotted the cosine of the error angle, i.e., the angle between the body yaw axis and the local vertical.

### 2.4.4 *Computer Runs for Large-Angle Motions*

Since only a limited amount of energy may be imparted by initial displacement of the satellite, computer studies were directed at the effects of high initial angular velocities. In Figs. 23–26 are shown some sample results of computer runs for the response of the satellite design of Section 2.3.2 to high initial rates, applied to the satellite in the stable roll-vee orientation. These may be regarded as responses to micrometeorite impacts, or as representative of initial transients following inadequate despin.

To save space, we again plot, as a function of time, only the cosine of the error angle between the yaw axis and the local vertical. However, the orientation of the satellite and of the gyro spin vectors are shown every half orbit in computer-made perspective drawings of a rectangular parallelepiped representing the satellite. The view is along the orbital track in the rotating, earth-pointing reference frame, so that the local vertical and normal to the orbital plane are in the plane of the paper. Plus signs are placed on the faces of the parallelepiped to avoid optical illusions. The gyro stops are indicated by dots. The reader may find more details about these drawings, as well as a description of computer-made movies showing large motions of the two-gyro satellite, in Ref. 15.

As is seen from Figs. 23–26, rates of the order of 4 rpo about pitch and roll damp out in about 10 orbits, whereas yaw rates of even 100 rpo settle out in about 5 orbits. In roll and yaw, the settling time and motion are similar if negative rather than positive rates are applied. The responses to positive and negative pitch rates are, however, different in character. A high positive pitch rate collapses the gyros toward the pitch axes, and a slowly decaying, essentially single-axis spin ensues. A high negative pitch rate opens up the gyros and drives them against the yaw stops. This sends the satellite into a complicated tumbling which eventually settles out.

If a micrometeorite of linear momentum $m$ strikes the satellite at a lever arm $L$ from a principal axis with moment of inertia $I$, the angular velocity $\omega$ imparted around that axis will be $\omega = mL/I$. This velocity varies directly with $L$ and inversely with $I$. For the design of Section 2.3.2, the yaw and pitch or roll lever arms are in the ratio 2/60, while the inertias are in the ratio 2000/20. A micrometeorite which imparts a pitch or roll rate of $4\Omega$ will impart a yaw rate of $(2/60)\cdot(2000/20)\cdot 4\Omega = 13.3\Omega$. Therefore we see from Figs. 23–26 that the two-gyro spindle satel-

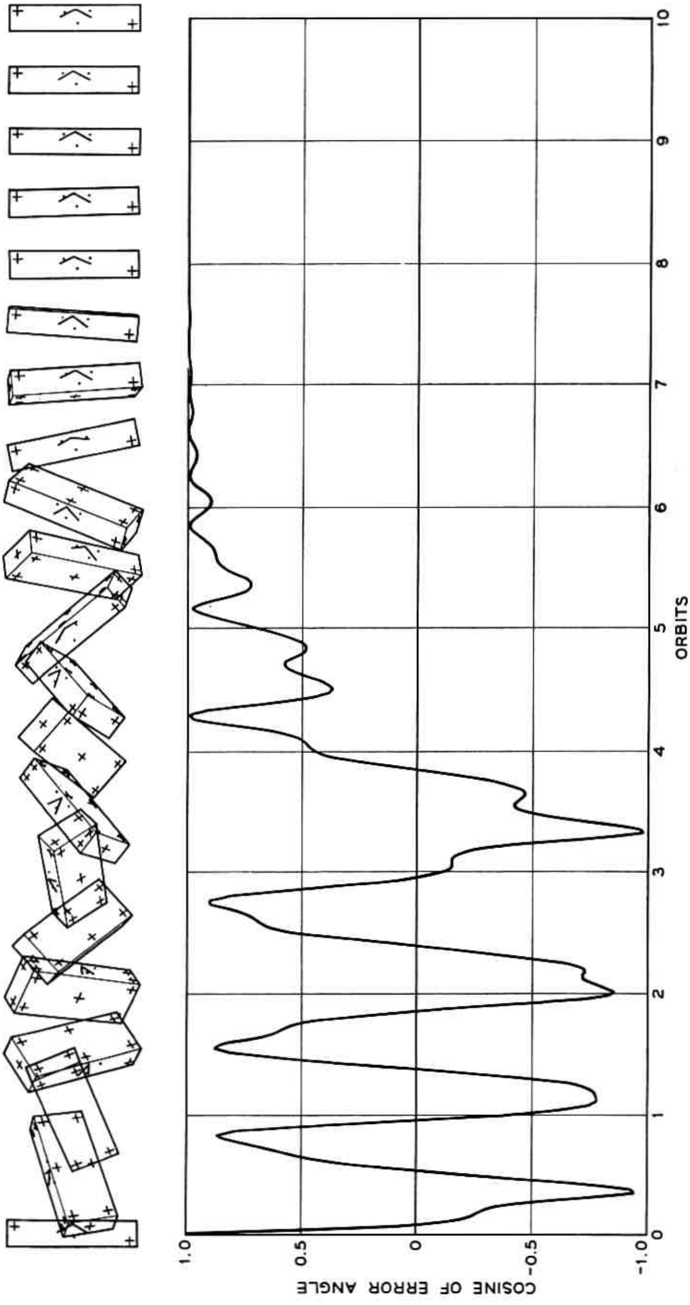


Fig. 23 — Response to +5 rpo roll rate.

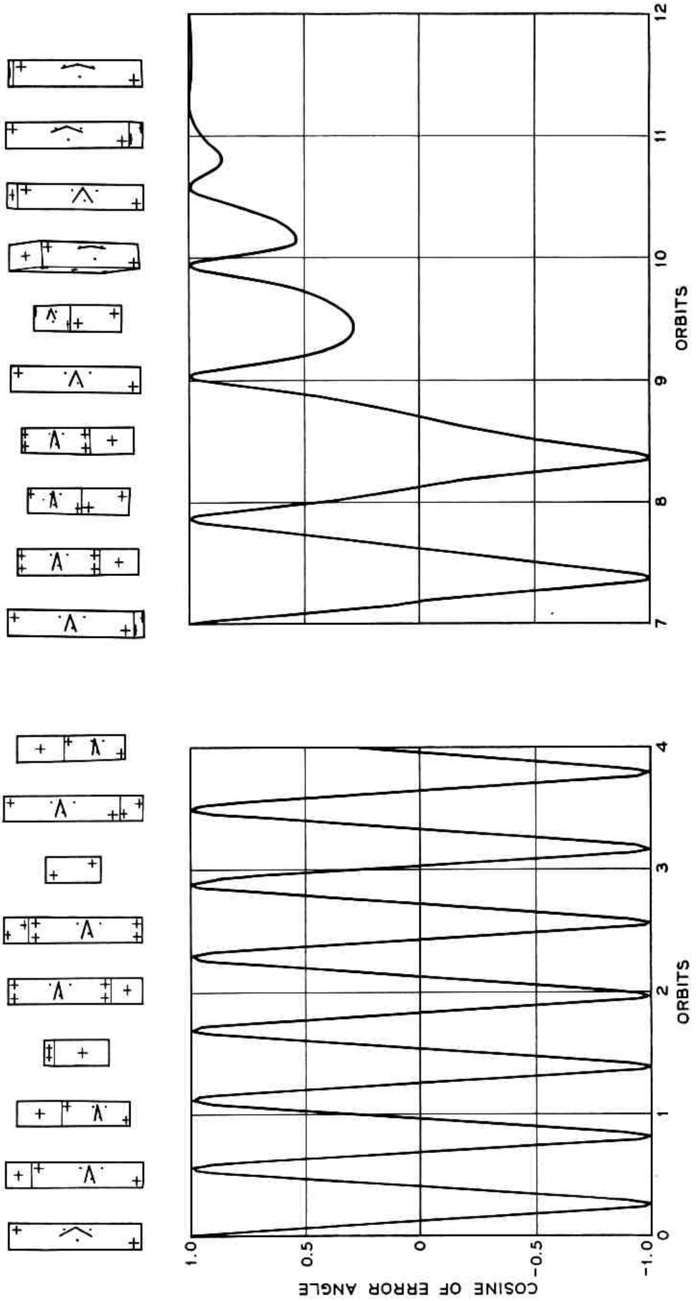


Fig. 24 — Response to +4 rpo pitch rate.

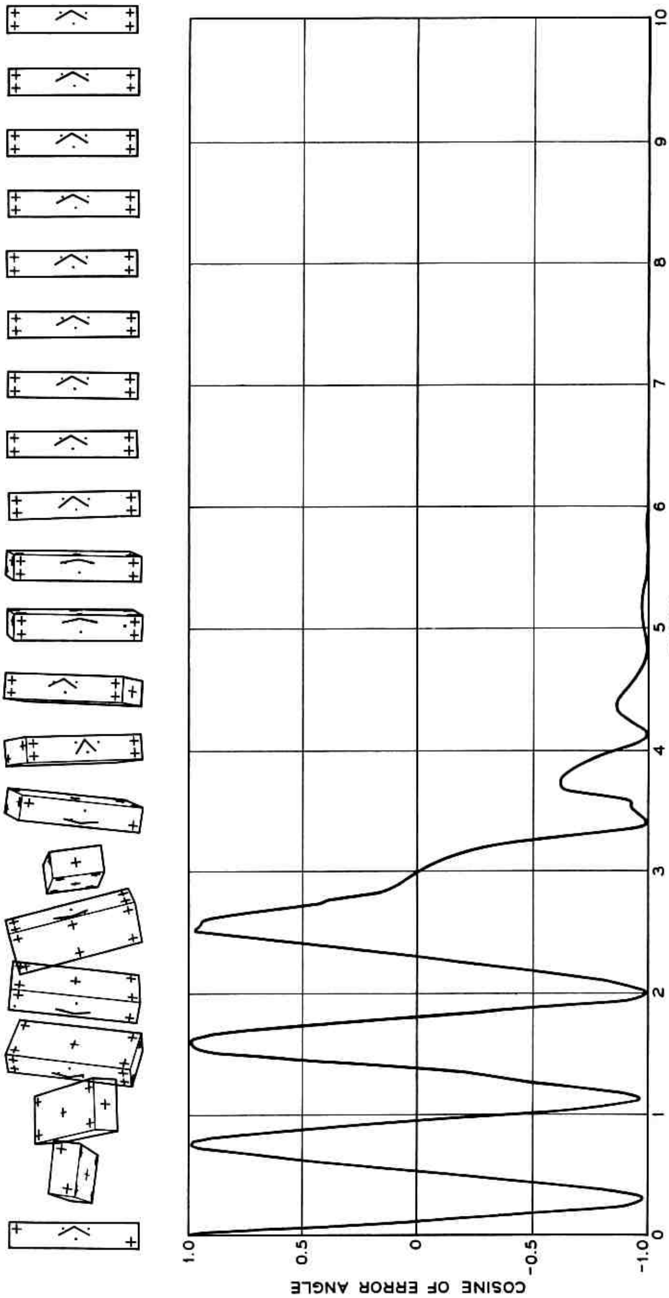


Fig. 25 — Response to −4 rpo pitch rate.

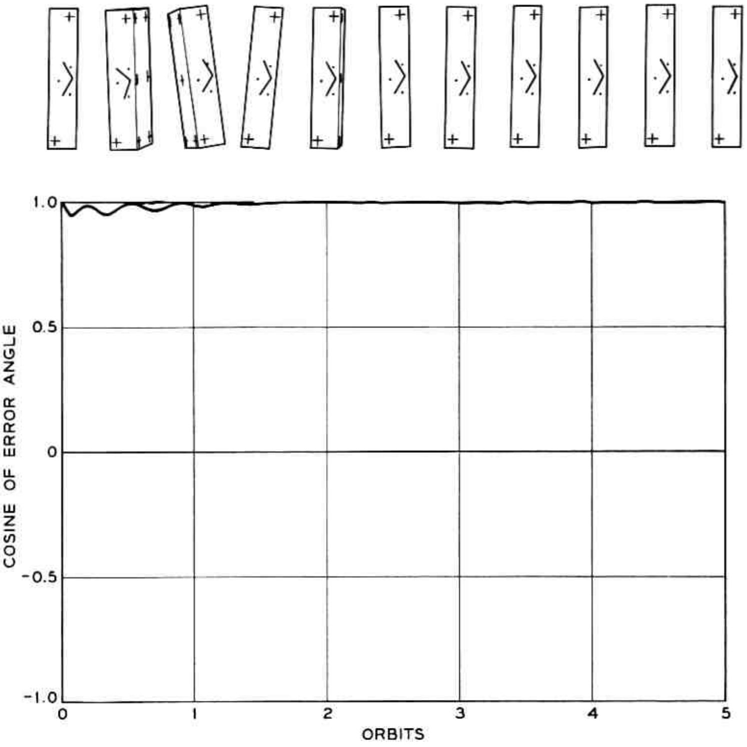


Fig. 26 — Response to 100 rpo yaw rate.

lite is considerably more resistive to micrometeorite impact about yaw than about pitch and roll.

It is well known that gravity-gradient satellites will tumble if placed in a sufficiently eccentric orbit. Computer experiments showed that for the design of Section 2.3.2 this occurred at an eccentricity of about 0.2. Computer results for $\epsilon = 0.225$ are shown in Fig. 27.

## III. SMALL-ANGLE MOTION

### 3.1 *Satellite Configuration*

To settle the vexing questions of nomenclature and sign convention once for all, we commence with a brief description of the quantities characterizing a gravity-oriented satellite moving in a circular orbit (including the effect of small eccentricity later on) at the orbital angular

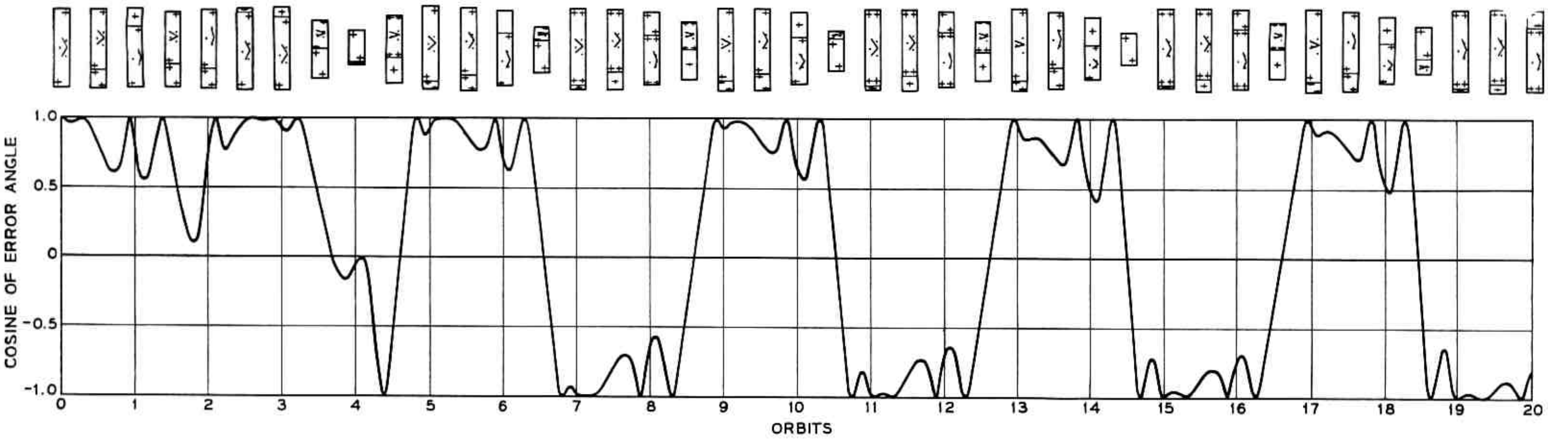


Fig. 27 — Tumbling for orbital eccentricity of 0.225.

velocity $\Omega$. A convenient number to remember, to fix the magnitude of $\Omega$, is that an orbit at an altitude of about 6800 statute miles corresponds to an orbital rate $\Omega = 1$ radian/hr and an orbital period of $2\pi$ or about 6 hours, 15 minutes.

For our purposes the satellite is described by its principal moments $A \geqq B \geqq C$ about principal axes $x',y',z'$, along which the principal unit vectors $\mathbf{i}',\mathbf{j}',\mathbf{k}'$ lie. These principal axes form the body pitch, roll, and yaw axes, respectively.

When the satellite is in a position of stable equilibrium, the $x',y',z'$ body axes coincide with orbit pitch, roll, and yaw axes $x,y,z$ as in Fig. 19, with corresponding unit vectors $\mathbf{i},\mathbf{j},\mathbf{k}$, normal to the orbital plane, along the orbit track, and along the local vertical toward the center of the earth. These orbit reference axes rotate at the orbital rate $\Omega$ (1 rpo) about the orbit pitch axis. It should be noted that, although a spindle-shaped body, formed by the extension of a single rod and tip mass, is shown, for which $B \approx A$ and $C \ll B$, the small-angle analysis which follows covers the whole range of inertias, given by the inequalities

$$A > B > C,$$

required for stability, and

$$B + C > A,$$

imposed by rigid-body geometry.

For small perturbations from equilibrium the satellite orientation is specified by the small pitch, roll, and yaw angles $\varphi_x$, $\varphi_y$, $\varphi_z$, through which the body axes $x',y',z'$ are rotated from the orbit axes $x,y,z$, as in Fig. 28. The corresponding satellite angular velocity vector $\boldsymbol{\omega}_s$, with respect to inertial space, is given by

$$\boldsymbol{\omega}_s = \mathbf{i}(\Omega + \dot{\varphi}_x) + \mathbf{j}\dot{\varphi}_y + \mathbf{k}\dot{\varphi}_z ,$$

with respect to orbit axes, or

$$\boldsymbol{\omega}_s = \mathbf{i}'(\Omega + \dot{\varphi}_x) + \mathbf{j}'(-\Omega\varphi_z + \dot{\varphi}_y) + \mathbf{k}'(\Omega\varphi_y + \dot{\varphi}_z),$$

with respect to body axes.

### 3.2 *Roll-Vee System Equations for Small Motion*

By neglecting second-order terms in the dynamical equations of Section IV, as indicated in Section 4.2.2, we obtain the satellite equations of motion

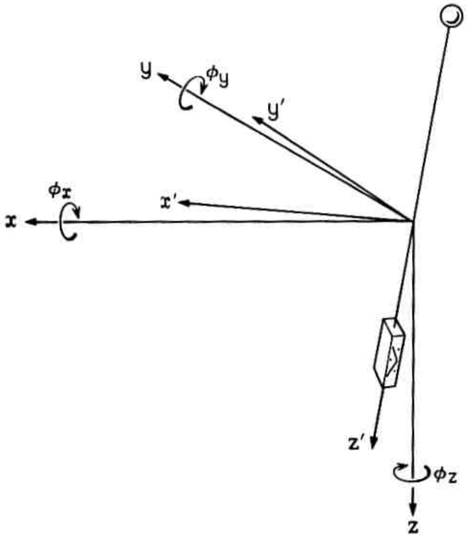


Fig. 28 — Definition of small pitch, roll, and yaw angles.

$$A\ddot{\varphi}_x + 3(B - C)\Omega^2\varphi_x + 2H_s\dot{\psi}_g = 0,$$

$$\begin{aligned} B\ddot{\varphi}_y &+ [4(A - C)\Omega^2 + 2H_c\Omega]\varphi_y \\ &+ [(A - B - C)\Omega + 2H_c]\dot{\varphi}_z + 2H_c\Omega\varphi_g = 0, \qquad (1) \\ C\ddot{\varphi}_z &+ [(A - B)\Omega^2 + 2H_c\Omega]\varphi_z \\ &- [(A - B - C)\Omega + 2H_c]\dot{\varphi}_y - 2H_c\dot{\varphi}_g = 0, \end{aligned}$$

for the satellite pitch, roll, and yaw angles $\varphi_x$, $\varphi_y$, $\varphi_z$. The sum $\varphi_g = \frac{1}{2}(\varphi_{g1} + \varphi_{g2})$ and difference $\psi_g = \frac{1}{2}(\varphi_{g1} - \varphi_{g2})$ of the two gimbal angles satisfy the equations

$$C_D\dot{\psi}_g + (K + H_c\Omega)\psi_g - H_s\dot{\varphi}_x = 0,$$

$$C_D\dot{\varphi}_g + (K + H_c\Omega)\varphi_g + H_c\Omega\varphi_y + H_c\dot{\varphi}_z = 0.$$

Here $H_c = H \cos \alpha$ and $H_s = H \sin \alpha$. This is an eighth-order linear system of equations for $\varphi_x$, $\varphi_y$, $\varphi_z$, $\psi_g$, $\varphi_g$, which splits immediately into a cubic pitch system for $\varphi_x$ and $\psi_g$ and a quintic roll-yaw system for $\varphi_y$, $\varphi_z$, $\varphi_g$, since the pitch motion depends only on the out-of-phase, or "scissoring," motion of the gyro gimbals, given by the difference angle $\psi_g$, and the roll-yaw motion depends only on the in-phase gimbal motion, given by the sum angle $\varphi_g$.

These equations can be reduced to dimensionless form by setting

$$p = (1/\Omega)d/dt, \qquad b = B/A, \qquad c = C/A, \qquad h = (H/A\Omega) \cos \alpha,$$
$$h' = (H/C_D) \cos \alpha, \qquad \kappa = 1 + [K/(H\Omega \cos \alpha)],$$

yielding the two sets of equations:

*Pitch:*

$$[p^2 + 3(b - c)]\varphi_x + (2h \tan \alpha)p\psi_g = 0,$$
$$-(h' \tan \alpha)p\varphi_x + (p + \kappa h')\psi_g = 0,$$

*Roll-Yaw:*

$$[bp^2 + 4(1 - c) + 2h]\varphi_y + (1 - b - c + 2h)p\varphi_z + 2h\varphi_g = 0,$$
$$-(1 - b - c + 2h)p\varphi_y + [cp^2 + (1 - b) + 2h]\varphi_z - 2hp\varphi_g = 0,$$
$$h'\varphi_y + h'p\varphi_z + (p + \kappa h')\varphi_g = 0.$$

If we insert appropriate terms on the right-hand sides of these equations to include the effect of given initial conditions and perturbing torques, we may regard the above systems as the Laplace transforms of the original set of differential equations, with transform variable $p$. The solution is then found by solving this set of linear, algebraic equations for $\varphi_x$, $\psi_g$, etc., now interpreted as Laplace transforms, and calculating the residues at the poles of these functions of $p$. The transient response is entirely determined by the location of these poles and by the specific initial conditions. The steady-state response to a periodic perturbing torque at frequency $N\Omega$ may be determined by inserting constant right-hand sides, in general complex, setting $p = iN$, and solving for the amplitudes $|\varphi_x|$, $|\psi_g|$, etc.

### 3.3 *Transient Response*

For given initial conditions and zero perturbing torques, the transforms are rational functions of $p$, with the characteristic pitch and roll-yaw polynomials as denominators, given by

*Pitch Cubic:*

$$f_3(p) = p^3 + c_1p^2 + c_2p + c_3, \tag{2}$$

*Roll-Yaw Quintic:*

$$f_5(p) = p^5 + a_1p^4 + a_2p^3 + a_3p^2 + a_4p + a_5, \tag{3}$$

where

$$c_1 = h'(\kappa + 2h \tan^2 \alpha), \qquad c_2 = 3(b - c), \qquad c_3 = 3(b - c)\kappa h',$$

and

$$a_1 = \kappa h' + \frac{2hh'}{c},$$

$$a_2 = \frac{1 - b + 2h}{c} + \frac{4(1 - c) + 2h}{b} + \frac{(1 - b - c + 2h)^2}{bc},$$

$$a_3 = \kappa h' a_2 + \frac{2hh'(2 + 2b - 3c - 2h)}{bc},$$

$$a_4 = \frac{(1 - b + 2h)(4 - 4c + 2h)}{bc},$$

$$a_5 = \kappa h' a_4 - \frac{2hh'(1 - b + 2h)}{bc}.$$

For stability, it is necessary that all the roots of the above polynomials have negative real parts. In particular, the magnitude of the real part of the root nearest the imaginary axis determines the rate of decay of the most lightly damped mode of motion. If this real part is $-D$, we can define a $1/e$ asymptotic settling time $T_s = 1/2\pi D$ (in orbit periods) and use $T_s$ as a measure of transient response, particularly suited for use with a digital computer. In Section 3.6 we discuss the determination of $D$ as a function of $b$, $c$, $\alpha$, $h$, $h'$, and $\kappa$. Once it is reduced to a suitably small value by some choice of system parameters, the short-time transient response can be determined by solution of the differential equations with specific initial conditions and the system parameters readjusted, if necessary. Actually systems chosen on the basis of minimum asymptotic settling time seem to have quite adequate short-time, as well as steady-state, response.

### 3.4 *Steady-State Response*

The steady-state response to periodic perturbing torques at frequency $N\Omega$, determined as previously outlined, is given in various cases by the following relations:

*Pitch amplitude for pitch torque $A\Omega^2$:*

$$|\varphi_x| = (N^2 + \kappa^2 h'^2)^{\frac{1}{2}}/|f_3(iN)|, \tag{4}$$

*Roll amplitude for roll torque $A\Omega^2$:*

$$|\varphi_y| = \left[\left(\frac{1-b+2h}{c}\kappa h' - a_1N^2\right)^2 + \left(\frac{1-b+2h}{c}N - N^3\right)^2\right]^{\frac{1}{2}} \bigg/ |f_5(iN)|, \tag{5}$$

*Yaw amplitude for yaw torque $A\Omega^2$:*

$$|\varphi_z| = \left[\left(\frac{4-4c+2h}{b}\kappa h' - \frac{2hh'}{b} - \kappa h'N^2\right)^2 + \left(\frac{4-4c+2h}{b}N - N^3\right)^2\right]^{\frac{1}{2}} \bigg/ c\,|f_5(iN)|, \tag{6}$$

*Roll amplitude for yaw torque $A\Omega^2$ and yaw amplitude for roll torque $A\Omega^2$:*

$$|\varphi_y| = |\varphi_z| = N\left[\left(\frac{1-b+2h}{c}\kappa h' - a_1\right)^2 + \left(\frac{1-b-c+2h}{c}N\right)^2\right]^{\frac{1}{2}} \bigg/ b\,|f_5(iN)|, \tag{7}$$

where, by (2) and (3),

$$|f_3(iN)|^2 = (c_3 - c_1N^2)^2 + (c_2N - N^3)^2,$$

$$|f_5(iN)|^2 = (a_5 - a_3N^2 + a_1N^4)^2 + (a_4N - a_2N^3 + N^5)^2.$$

In particular, a constant pitch torque $A\Omega^2(N = 0)$ gives the constant pitch offset

$$|\varphi_x|_0 = 1/[3(b - c)],$$

while, for a roll torque of amplitude $A\Omega^2$ at the orbital frequency $(N = 1)$

$$|\varphi_y|_1 = 1/[3(1 - c)].$$

These amplitudes, independent of the gyro parameters, limit the minimum permissible satellite inertias for given perturbing torques.

Finally, an elliptic orbit of small eccentricity $\epsilon$ induces forced pitch vibrations at the orbital frequency $\Omega$ with amplitude

$$|\varphi_x|_\epsilon = 2\epsilon(1 + c_1^2)^{\frac{1}{2}}/[(c_3 - c_1)^2 + (c_2 - 1)^2]^{\frac{1}{2}}. \tag{8}$$

By straightforward differentiation, it is easily shown that the eccentricity response $|\varphi_x|_\epsilon$ has a single maximum as a function of the gyro opening angle $\alpha$. For $\alpha$ larger than the value at which the maximum is attained, the eccentricity response decreases monotonically, approaching $2\epsilon$ as $\alpha$ approaches 90°.

### 3.5 *Bounds on the Asymptotic Damping Rate D*

As mentioned in Section 3.3, a convenient single measure of transient response is the parameter $D$, the distance from the imaginary axis of the right-most root of the characteristic equations. One would like to know $D$ as a function of the system parameters, $D = D(b,c,\alpha,h,h',\kappa)$. In general, this function is impossible to determine analytically and must be computed numerically. In order to limit such computations to ranges of the parameters $b,c,\alpha,h,h',\kappa$ that give reasonable values of $D$, it is convenient to have bounds on $D$.

One set of bounds is given by the following theorem:[16]

*If the coefficients $q_0$, $q_1$, $\cdots$, of a polynomial $P(p)$ are positive,*

$$P(p) = q_0p^n + q_1p^{n-1} + \cdots + q_n$$

*then D is bounded by*

$$D^{k-l} \leqq \frac{q_k \Big/ \binom{n}{k}}{q_l \Big/ \binom{n}{l}} \qquad k > l, \qquad \begin{array}{l} k = 1, 2, \cdots, n \\ l = 0, 1, \cdots. \end{array}$$

We note that in both $f_3(p)$ and $f_5(p)$, in (2) and (3), $q_0 = 1$. Hence by the above theorem with $l = 0$, if the system parameters are such that any of the coefficients $c_1$, $c_2$, $c_3$, $a_1$, $\cdots$, $a_5$ is small, then $D$ will be small.

Likewise, if any coefficient in $f_3(p)$ or $f_5(p)$ is large compared to a subsequent coefficient, then the theorem tells us that $D$ will again be small.

We note also that $b - c < 1$, so that the theorem applied to $c_2$ gives, in pitch,

$$D^2 \leqq b - c \leqq 1,$$

i.e., the asymptotic settling time $T_s$ in pitch for the roll-vee system is bounded by $T_s \geqq 1/2\pi = 0.159$ orbit. (This is slightly larger than the corresponding bound $T_s \geqq 5^{\frac{1}{4}}/2\sqrt{3}\pi = 0.137$ orbit obtained in Ref. 17 for a two-body satellite.)

From these bounds we conclude immediately that at best $D$ can be of order unity, and to get a $D$ of this order of magnitude the coefficients and ratios of coefficients in $f_3(p)$ and $f_5(p)$ must be at least of order unity.

Another useful bound is obtained by shifting the origin in the $p$ plane to $p = -D$ and applying the *Routh criteria* (see Section 3.7), to the shifted $f_3(p)$ polynomial. It is then found that one of the terms in the Routh array is

$$r = -2Dx^2 + [8D^2 + 3(b - c)]x - 8D^3 - 6(b - c)D - 3(b - c)\kappa h',$$

where

$$x = \kappa h' + 2hh' \tan^2 \alpha.$$

To have all roots to the left of the line Re $p = -D$, $r$ must be positive. But it is easily verified that $r > 0$ only if

$$D \leqq \frac{3}{8} \frac{(b - c)}{\kappa h'},$$

which gives an additional bound on $D$.

### 3.6 *Determination of D by Series Expansions*

When the coefficients in $f_3(p)$ or $f_5(p)$ are either large or small, $D$ can sometimes be expanded in a power series around a known root. This again restricts the parameter ranges over which $D$ must be determined numerically. For example, suppose $h'$ is small. The roots of $f_3(p)$ at $h' = 0$ are $p = 0$, $p = \pm i\sqrt{3(b - c)}$. However, it is well known[18] that each of the three branches of the triple-valued function $p = p(h')$ is analytic in $h'$. Expanding around $h' = 0$, say for $p(0) = i\sqrt{3(b - c)}$, we have

$$\begin{aligned} p &= i\sqrt{3(b - c)} + h' \left(\frac{dp}{dh'}\right)_{h'=0} + \cdots \\ &= i\sqrt{3(b - c)} - h'h \tan^2 \alpha + \cdots, \end{aligned}$$

with similar expressions easily obtained for the other two branches of $p = p(h')$.

A particular case of interest is that of a spindle-shaped body. In this case, $c \to 0$, $b \to 1$, and the coefficients $a_1, \cdots, a_5$ of $f_5(p)$ all become large. One can then consider the equation $cf_5(p)$, in which the leading coefficient is small. However at $c = 0$, this equation is singular because it is reduced in degree from a quintic to a quartic. The quartic, with coefficients $ca_1, ca_2, \cdots, ca_5$, gives only four of the limiting roots as $c \to 0$. The fifth limiting root is however easily found by setting $p = \sigma/c$, yielding $f_5^*(\sigma) = c^5 f_5(\sigma/c)$. Application of the expansion theorem to $f_5^*(\sigma)$ then yields the fifth limiting root $p \to -2hh'/c$ as $c \to 0$.

This root gives a highly damped mode, and has a real part far in the left half-plane. The roots of interest in finding $D$ are thus those of the limiting quartic:

$$f_4(p) = p^4 + (\beta_1/h')p^3 + \beta_2 p^2 + (\beta_3/h')p + \beta_4 = 0,$$

where

$$\beta_1 = 2h + 1, \qquad \beta_2 = \kappa(2h + 1) + 4 - 2h,$$
$$\beta_2 = 2h + 4, \qquad \beta_4 = \kappa(2h + 4) - 2h.$$

The limiting quartic $f_4(p)$ is a function of the three parameters $\kappa,h,h'$, whereas, in this limiting case, the cubic $f_3(p)$ is a function of $\kappa$, $h$, $h'$, and $\alpha$. It turns out that $D = D(h,h',\kappa,\alpha)$ can be obtained graphically. Further, $D_m$, the maximum $D$ for all possible $h,h',\kappa,\alpha$, can be found and has the value $D_m = 0.317$, attained at the values

$$0.77 < h < 0.78, \qquad h' = 1.29, \qquad \kappa = 0.92, \qquad \alpha = 38°.$$

(The value $D_m$ corresponds to an asymptotic settling time

$$T_s = 1/2\pi D_m = 0.502$$

orbits.) However, the description of the graphical technique and the derivation of $D_m$ are too lengthy for inclusion here.

### 3.7 *Computation of the Over-All Small-Angle Response*

The over-all small-angle performance of the satellite attitude control system is characterized by its steady-state response to constant and periodic disturbances (solar torques, magnetic torques, orbital eccentricity) and by its transient response to sporadic disturbances (initial injection, micrometeorite impact). In proper design, one wants to diminish the response to all disturbances to below a suitably small level.

The steady disturbances have their main components at zero, orbital, and twice orbital frequency. As indicated earlier, their amplitudes may be diminished by inertia augmentation with extensible rods. Fortunately, it is easy to write down the formulas, (4)–(8), for satellite response to steady disturbances, and also easy to program these formulas for digital computation.

The computation of the transient response is not so straightforward, even in terms of the single measure $D$. An interesting theoretical problem is to find the maximum $D$ as a function of all system parameters. Gradient or steepest descent methods, which first come to mind for the solution to this problem, seem to be difficult to apply, since the maximum $D$ usually occurs in the neighborhood of multiple roots where the function $D = D(b,c,h,h',\kappa,\alpha)$ is singular.

However, although this is a theoretically interesting problem, its solution is not of great practical importance, as indicated in Section 2.1. It is more important to have a cheap method of computing $D$. A method that we have found useful involves the Routh criteria as follows:

Write the polynomial $f(p)$ as

$$f(p) = a_0 p^n + a_1 p^{n-1} + \cdots + a_m p^{n-m} + \cdots + a_n = 0,$$

and form the Routh array

$$\begin{array}{l} a_0\,,\ a_2\,,\ a_4\,,\ \cdots, \\ b_0\,,\ b_2\,,\ b_4\,,\ \cdots, \\ c_0\,,\ c_2\,,\ c_4\,,\ \cdots, \\ d_0\,,\ d_2\,,\ d_4\,,\ \cdots, \end{array}$$

where

$$b_0 = a_1\,,\ b_2 = a_3\,,\ \cdots,\ b_{2i} = a_{2i+1}\,,\ \cdots,$$

and

$$\begin{aligned} c_{2i} &= a_{2i+2} - (b_{2i+2}a_0/b_0), \\ d_{2i} &= b_{2i+2} - (c_{2i+2}b_0/c_0), \quad \text{etc.,} \qquad i = 0, 1, 2, \cdots. \end{aligned}$$

Then the number of sign changes in the sequence $a_0$, $b_0$, $c_0$, $d_0$, etc. (providing no term is zero) is the number of roots in the right-half plane. Because of its recurrence structure, this scheme is easily programmed on a digital computer.

To determine the real parts of the roots of $f(p)$, one applies the scheme to a succession of translated half-planes as follows. If $p = -D + \zeta$, then

$$\begin{aligned} & f(-D + \zeta) \\ &\quad = f(-D) + f'(-D)\zeta + \frac{f''(-D)\zeta^2}{2!} + \cdots + \frac{f^{(n)}(-D)\zeta^n}{n!} = 0 \\ &\quad = q_0\zeta^n + q_1\zeta^{n-1} + \cdots + q_n = 0, \end{aligned}$$

where it is easily verified that

$$q_0 = \frac{f^{(n)}(-D)}{n!} = a_0\,,$$

$$q_1 = \frac{f^{(n-1)}(-D)}{(n-1)!} = -na_0D + a_1\,,$$

$$q_2 = \frac{f^{(n-2)}(-D)}{(n-2)!} = \binom{n}{2} a_0 D^2 - \binom{n-1}{1} a_1 D + a_2 ,$$

$$\cdots$$

$$q_k = \frac{f^{(n-k)}(-D)}{(n-k)!}$$

$$= \binom{n}{k} a_0(-D)^k + \binom{n-1}{k-1} a_1(-D)^{k-1} + \cdots + a_k .$$

The Routh array applied to the coefficients $q_0$, $q_1$, $\cdots$, $q_n$ then indicates the number of roots to the right of the line Re $p = -D$ (Re $\zeta = 0$). In order to locate the real parts of the roots to an arbitrary degree of accuracy, one applies this array on a sequence of nested intervals. For example, start with some large $D = D^*$ such that the Routh array applied on Re $p = -D^*$ indicates roots to its right. Take as the initial interval $-D^* <$ Re $p < 0$. In a stable system there will be roots between the right boundary (Re $p = 0$), and the left boundary (Re $p = -D^*$). Next apply the Routh criteria on Re $p = -D^*/2$. There are two possibilites: (a) if there *are no* roots to the right of Re $p = -D^*/2$, make this the new right boundary; the interval $-D^* <$ Re $p < -D^*/2$ now has the same properties as the initial interval, (b) if there *are* roots to the right of Re $p = -D^*/2$, make this line the *left* boundary of the new interval $-D^*/2 <$ Re $p < 0$, which again has the same properties as the initial interval. By applying this process $n$ times, one ends up with an interval of width $D^*/2^n$, which contains roots but has no roots to its right. The accuracy of the location of the real parts of the roots closest to the imaginary axis can be set by prescribing the width of the final interval. Since the widths of the successive intervals go down as $1/2^n$, the process converges rapidly.

After the real parts of the roots closest to the imaginary axis are found within some interval of desired width, say Interval 1, the same procedure can be used to find the next closest roots to the imaginary axis. One starts again at some sufficiently large $D^*$, such that some roots fall to the right of Re $p = -D^*$ and to the left of Re $p = -D_{L1}$, the left boundary of Interval 1. One makes these the left and right boundaries respectively, of an initial Interval 2, and applies the nested interval iteration again. The right boundary of Interval 2 in each iteration is characterized by having $m$ roots to its right, where $m$ is the number of roots contained in Interval 1.

The starting value $D^*$ can be chosen in various ways. If one is interested only in the roots closest to the imaginary axis, he can pick $D^*$ as

$$D^* = \min\left[\frac{a_m \Big/ \binom{n}{m}}{a_{m-1} \Big/ \binom{n}{m-1}}\right],$$

for then (see Section 3.5) there will be at least one root to the right of $\operatorname{Re} p = -D^*$. If it is desired to find the real parts of all the roots, $D^*$ can be chosen as

$$D^* = \max\left(\frac{a_m}{a_{m-1}}\right),$$

since it is well known[19] that this value of $D^*$ is a bound on the modulus of the maximum root and hence all the roots will be to the right of $\operatorname{Re} p = -D^*$.

We remark that this procedure may be easily extended to a method for finding both the real and imaginary parts of the roots of a real polynomial. It is only necessary to use well-known relations between the imaginary parts of the roots and certain members of the Routh array.

The above scheme goes rapidly on the IBM 7090 computer. For example, if the widths of Interval 1, Interval 2, etc. are set at 0.005, the running time is about 1000 cases a minute to find the real parts of all the roots of both the quintic, $f_5(p)$, and the cubic, $f_3(p)$. Tables calculated by this process were used in making the parameter survey whose results are summarized in Section 2.3.

## IV. LARGE-ANGLE MOTION

### 4.1 *Introduction*

The large-angle motion of the satellite is of course governed by nonlinear differential equations, which in general must be integrated numerically. Nevertheless, a few analytical and intuitive insights are available. These are pointed out in the sections which follow.

We begin with a discussion of the pertinent dynamical and kinematic equations, including the effect of variable inertia, due to rod erection. Then we enumerate the equilibrium positions of the satellite, in which it is at rest with respect to the orbiting reference frame in a circular orbit, and show that certain restrictions must be placed on gyro angular momentum to eliminate undesired positions. This is followed by a discussion of satellite despin by the erection of a single rod and tip mass. Finally we show how the satellite may be inverted by ground command to the gyro torquers.

### 4.2 *Large-Angle Equations of Motion*

In the following, we make an explicit distinction between dynamical equations, valid in any coordinate system when written in proper vector form, and kinematic relations between various specific coordinate systems. This allows us to introduce a minimum number of different coordinate systems and to avoid a good deal of irrelevant algebraic complexity.

#### 4.2.1 *Dynamical Equations*

The rate of change of angular momentum **L** about the satellite center of mass, with respect to a reference frame rotating at the satellite angular velocity $\boldsymbol{\omega}_s$, is governed by the equation

$$\dot{\mathbf{L}} + \boldsymbol{\omega}_s \times \mathbf{L} = \mathbf{M}, \tag{9}$$

where **M** is the resultant torque around the center of mass, the sum of the gravity-gradient torque $\mathbf{M}_G$, the total gyro precession torque $\mathbf{M}_H$, and the external disturbing torque $\mathbf{M}_E$. For a rigid body

$$\mathbf{L} = \mathbf{I}\cdot\boldsymbol{\omega}_s \tag{10}$$

where **I** is the inertia dyadic, given in terms of the principal moments of inertia $A > B > C$ and corresponding principal vectors $\mathbf{i}',\mathbf{j}',\mathbf{k}'$, by

$$\mathbf{I} = A\mathbf{i}'\mathbf{i}' + B\mathbf{j}'\mathbf{j}' + C\mathbf{k}'\mathbf{k}'. \tag{11}$$

If $\boldsymbol{\omega}$ is the satellite angular velocity relative to orbit reference axes,

$$\boldsymbol{\omega}_s = \mathbf{i}\dot{\psi} + \boldsymbol{\omega}, \tag{12}$$

where **i** is a unit vector normal to the orbit plane and $\psi(t)$ is the polar angle of the satellite center of mass, measured from orbit perigee in earth-centered coordinates and satisfying the orbit equation

$$\dot{\psi} = \Omega(1 + \epsilon \cos\psi)^2/(1 - \epsilon^2)^{\frac{3}{2}}, \tag{13}$$

where $\epsilon$ is the orbit eccentricity and $\Omega = 2\pi/T_0$, $T_0$ being the orbit period. The gravity-gradient torque $\mathbf{M}_G$ is given by

$$\mathbf{M}_G = 3\Omega^2(1 + \epsilon\cos\psi)^3[\mathbf{k} \times (\mathbf{I}\cdot\mathbf{k})]/(1 - \epsilon^2)^3, \tag{14}$$

where **k** is a unit vector directed along the local vertical toward the center of the earth. Here and in the following, we consider only what Beletskii[20] calls the "restricted problem" for which the motion of the center of mass is given by (13) and is unaffected by the motion around the center of mass. Finally the resultant gyro torque for a two-gyro, roll-vee configuration is

$$\mathbf{M}_H = \mathbf{H}_1 \times \boldsymbol{\omega}_{g_1} + \mathbf{H}_2 \times \boldsymbol{\omega}_{g_2} \,, \tag{15}$$

where $\mathbf{H}_i$'s are the gyro angular momenta, of fixed magnitude $H$, and the $\boldsymbol{\omega}_{g_i}$'s are the gyro gimbal angular velocities. In terms of the gimbal angles $\varphi_{g_i}$ and the nominal roll-vee half-opening angle $\alpha$, we have

$$\omega_{g_i} = \boldsymbol{\omega}_s + \mathbf{j}'\dot{\varphi}_{g_i} \,, \tag{16}$$

$$\mathbf{H}_1 = H[\mathbf{i}' \cos (\alpha - \varphi_{g_1}) + \mathbf{k}' \sin (\alpha - \varphi_{g_1})] \tag{17}$$

$$\mathbf{H}_2 = H[\mathbf{i}' \cos (\alpha + \varphi_{g_2}) - \mathbf{k}' \sin (\alpha + \varphi_{g_2})]. \tag{18}$$

The set of dynamical equations is completed by the gimbal equations of motion. If the gyro gimbals are not against stops, these are

$$C_D\dot{\varphi}_{g_i} + K\varphi_{g_i} = M_{g_i} + \mathbf{j}' \cdot (\mathbf{H}_i \times \boldsymbol{\omega}_{g_i}), \tag{19}$$

where $C_D$ is the gyro damping constant, $K$ the gimbal spring constant, and the constant bias torques $M_{g_i}$ are given by $M_{g_2} = -M_{g_1} = H\Omega \sin \alpha$. When the gimbals are against stops, the reaction torques from the stops on the gimbals must be added to (19).

### 4.2.2 *Kinematic Relations*

The orientation of the satellite body axes $x',y',z'$, or the corresponding unit vectors $\mathbf{i}',\mathbf{j}',\mathbf{k}'$, with respect to the orbit axes $x,y,z$ or corresponding unit vectors $\mathbf{i},\mathbf{j},\mathbf{k}$,* may be specified in a number of ways. In classical dynamics, Euler angles have been traditionally used. They specify a rigid body's orientation with a minimum set of three numbers, and, in some of the soluble problems of rigid body dynamics, lead to straightforward analytical manipulations.

From a computing point of view, Euler angles, however, have three serious disadvantages: (1) they involve trigonometric functions, which are expensive to compute, (2) they are singular when the nutation angle is zero, and (3) they are difficult to use in the visualization of complicated motions. We have chosen to use the so-called Euler parameters, rather than the Euler angles. A set of variables, perhaps even more suitable for the matrix algebra typical of modern computer programming, might be the direction cosines $\alpha,\beta,\gamma$, etc., satisfying the relations

* See Section 3.1.

$$\begin{aligned}\mathbf{i} &= \mathbf{i}'\alpha + \mathbf{j}'\alpha' + \mathbf{k}'\alpha'',\\ \mathbf{j} &= \mathbf{i}'\beta + \mathbf{j}'\beta' + \mathbf{k}'\beta'',\\ \mathbf{k} &= \mathbf{i}'\gamma + \mathbf{j}'\gamma' + \mathbf{k}'\gamma''.\end{aligned} \tag{20}$$

These $\alpha$'s should not be confused with the nominal vee opening angle, which we shall distinguish from the direction cosines, whenever they are used together, with a subscript. By using identities of the form

$$\dot{\mathbf{k}} + \boldsymbol{\omega} \times \mathbf{k} = 0, \tag{21}$$

satisfied by **i**, **j**, and **k**, we may obtain 9 equations giving the rates of change of the direction cosines in terms of the direction cosines and the components of $\boldsymbol{\omega}$. We would then have 15 equations, including 3 satellite equations of angular motion, 1 equation of motion for the satellite's mass center, 2 gimbal equations of motion, and 9 equations for the direction cosine rates, yielding the 3 components of $\boldsymbol{\omega}$, the satellite polar angle $\psi$, the 2 gimbal angles, and the 9 direction cosines. The identities

$$\begin{aligned}\alpha^2 + \beta^2 + \gamma^2 = \alpha^2 + \alpha'^2 + \alpha''^2 &= 1, \quad \text{etc.},\\ \alpha\beta + \alpha'\beta' + \alpha''\beta'' = \alpha\alpha' + \beta\beta' + \gamma\gamma' &= 0, \quad \text{etc.},\end{aligned}$$

which must be satisfied initially, would then serve as checks on the numerical solution. Incidentally, it should be noted that the cosine of the antenna pointing error angle is given by the direction cosine $\gamma''$, between the local vertical and the body yaw axis.

We shall use the direction cosines to study equilibrium positions, but Euler parameters in the study of general satellite motion, since they are simply related to the deflection angles for small motion. If we assume that the $(x',y',z')$-axes are formed by rotation of the $(x,y,z)$-axes through the angle $\theta$ around an axis with direction cosines $m_x$, $m_y$, $m_z$, the Euler parameters $\xi_x$, $\xi_y$, $\xi_z$, $\chi$ are defined by the relations

$$(\xi_x, \xi_y, \xi_z) = (m_x, m_y, m_z) \sin(\theta/2), \qquad \chi = \cos(\theta/2).$$

We now have[21]

$$\begin{aligned}\mathbf{i} &= \mathbf{i}'(\xi_x^2 - \xi_y^2 - \xi_z^2 + \chi^2) + 2\mathbf{j}'(\xi_x\xi_y - \chi\xi_z) + 2\mathbf{k}'(\xi_x\xi_z + \chi\xi_y),\\ \mathbf{j} &= 2\mathbf{i}'(\xi_x\xi_y + \chi\xi_z) + \mathbf{j}'(-\xi_x^2 + \xi_y^2 - \xi_z^2 + \chi^2)\\ &\quad + 2\mathbf{k}'(\xi_y\xi_z - \chi\xi_x),\\ \mathbf{k} &= 2\mathbf{i}'(\xi_{.}\xi_y - \chi\xi_y) + 2\mathbf{j}'(\xi_y\xi_z + \chi\xi_z)\\ &\quad + \mathbf{k}'(-\xi_x^2 - \xi_y^2 + \xi_z^2 + \chi^2),\end{aligned} \tag{22}$$

giving the direction cosines, and

$$\begin{aligned}
2\dot{\xi}_x &= \chi\omega_{x'} + \xi_y\omega_{z'} - \xi_z\omega_{y'} , \\
2\dot{\xi}_y &= \chi\omega_{y'} + \xi_z\omega_{x'} - \xi_x\omega_{z'} , \\
2\dot{\xi}_z &= \chi\omega_{z'} + \xi_x\omega_{y'} - \xi_y\omega_{x'} , \\
-2\dot{\chi} &= \xi_x\omega_{x'} + \xi_y\omega_{y'} + \xi_z\omega_{z'} ,
\end{aligned} \tag{23}$$

completing a set of 10 equations for the three components of angular velocity $\boldsymbol{\omega}$, two gimbal angles, four Euler parameters, and the polar angle $\psi$. Just as in the case of the direction cosines, the single identity

$$\xi_x^2 + \xi_y^2 + \xi_z^2 + \chi^2 = 1$$

serves as a check on the numerical solution. Also note that, for small rotation $\theta$, $2\xi_x \sim \varphi_x$ , $2\xi_y \sim \varphi_y$ , $2\xi_z \sim \varphi_z$ , $\chi \sim 1$, $\omega_{x'} \sim \dot{\varphi}_x$ , $\omega_{y'} \sim \dot{\varphi}_y$ , $\omega_{z'} \sim \dot{\varphi}_z$ , and

$$\begin{aligned}
\mathbf{i} &\sim \mathbf{i}' - \mathbf{j}'\varphi_z + \mathbf{k}'\varphi_y , \\
\mathbf{j} &\sim \mathbf{i}'\varphi_z + \mathbf{j}' - \mathbf{k}'\varphi_x , \\
\mathbf{k} &\sim -\mathbf{i}'\varphi_y + \mathbf{j}'\varphi_x + \mathbf{k}',
\end{aligned}$$

where $\varphi_x$ , $\varphi_y$ , $\varphi_z$ are the small pitch, roll, and yaw angles. If these relations are inserted into the dynamical equations and second-order terms neglected, the linear equations for the small motion, (1), are obtained.

In coding the differential equations for the digital computer, it was found convenient to define cross-product and dot-product subroutines:

$$\begin{aligned}
\mathbf{A} \times \mathbf{B} &= (-A_3B_2 + A_2B_3 , A_3B_1 - A_1B_3 , -A_2B_1 + A_1B_2), \\
\mathbf{A}\cdot\mathbf{B} &= A_1B_1 + A_2B_2 + A_3B_3 .
\end{aligned}$$

This allowed the coding to follow closely the vector form of (9)–(11), which was useful from the standpoint of both coding simplicity and debugging.

For nondumbbell satellite shapes, say $b = 0.9$, $c = 0.5$, the five-point predictor-corrector with $\frac{2}{3}$ rule as given by Hamming (Ref. 22, Chapter 15) was used. However, in the spindle case, $b = 1.0$, $c = 0.01$, the differential equations become singular because the small number $c$ multiplies a derivative, and the five-point, $\frac{2}{3}$ rule scheme was found to be very slow. Following a suggestion of R. W. Hamming, a simple three-point predictor-corrector scheme (Ref. 22, p. 186) was then tried. It turned out to be three to four times faster than the five-point scheme and to give about the same accuracy.

In the computer runs, the gyro stops were simulated by hardening springs. For example, for $\varphi_{g_1} > \beta$, the normal spring restoring torque of $K\varphi_{g_1}$ was replaced by

$$K\varphi_{g_1} + B(\varphi_{g_1} - \beta) + \frac{C}{\theta - \varphi_{g_1}},$$

where $B$, $C$, $\beta$, and $\theta$ are constants, to simulate the pitch stop of the first gyro. The same expression but with different constants was used for the yaw stop. Specifically, in all the computer runs for the spindle-shaped satellite reported here, $K = 0$ and the pitch-stop values for Gyro 1 were

$$B = 50A\Omega^2, \qquad C = 0.01A\Omega^2, \qquad \beta = 58°, \qquad \theta = 60°.$$

For the yaw stop, $\beta$ and $\theta$ above were replaced by $\beta = -20°$ and $\theta = -30°$. Corresponding, symmetrical stop constants were used for Gyro 2.

### 4.2.3 *The Rate of Change of Energy*

For a circular orbit we can easily obtain a useful expression for the rate of change of kinetic and potential energy, relative to orbit axes. We take the scalar product of the satellite equation of motion (9) with the relative angular velocity $\boldsymbol{\omega}$ and combine it with the two gyro equations, (11), multiplied by $\dot{\varphi}_{g_i}$. After some routine algebra, we obtain the relation

$$(d/dt)(T + V + G) = -C_D(\dot{\varphi}_{g_1}{}^2 + \dot{\varphi}_{g_2}{}^2),$$

where the relative kinetic energy

$$T = \tfrac{1}{2}\boldsymbol{\omega}\cdot\mathbf{I}\cdot\boldsymbol{\omega},$$

the potential energy

$$V = \tfrac{1}{2}\Omega^2(3\mathbf{k}\cdot\mathbf{I}\cdot\mathbf{k} - \mathbf{i}\cdot\mathbf{I}\cdot\mathbf{i}),$$

and the gyro energy

$$G = G_1 + G_2,$$

with

$$G_i = \tfrac{1}{2}K\varphi_{g_i}{}^2 + M_{g_i}\varphi_{g_i} - \Omega\mathbf{i}\cdot\mathbf{H}_i.$$

This expression is useful in the estimation of various quantities, in particular the velocity required to tumble the satellite, and conditions necessary for capture.[23]

### 4.3 *Satellite Equilibrium Positions*

When the satellite moves in a circular orbit in such a way as to remain stationary with respect to the local vertical, its motion satisfies the equilibrium equation

$$\Omega^2 \mathbf{i} \times (\mathbf{I} \cdot \mathbf{i}) = 3\Omega^2 \mathbf{k} \times (\mathbf{I} \cdot \mathbf{k}) + \Omega \mathbf{H} \times \mathbf{i},$$

where **H** is the resultant gyro angular momentum. For a symmetrical satellite

$$\mathbf{I} = A(\mathbf{i}'\mathbf{i}' + \mathbf{j}'\mathbf{j}') + C\mathbf{k}'\mathbf{k}'$$

and the above equation yields the relations

$$(A - C)\beta''\gamma'' = 0$$

$$\mathbf{H} \cdot \mathbf{j} = H_y = (A - C)\Omega\alpha''\beta'', \tag{24}$$

$$\mathbf{H} \cdot \mathbf{k} = H_z = 4(A - C)\Omega\alpha''\gamma'', \tag{25}$$

with $\alpha'' = \mathbf{i} \cdot \mathbf{k}'$, $\beta'' = \mathbf{j} \cdot \mathbf{k}'$, $\gamma'' = \mathbf{k} \cdot \mathbf{k}'$, as defined by the table of direction cosines, (20). Thus we have the following general result:

I. *The equilibrium positions of any symmetrical, gravity-oriented, gyro-stabilized body in a circular orbit must be such that the principal axis of least inertia and the resultant angular momentum are perpendicular either to the orbit roll axis* ($\beta'' = H_y = 0$) *or to the orbit yaw axis* ($\gamma'' = H_z = 0$).

In the case of a roll gyro system, with all gimbal axes parallel to the body roll axis, the resultant gyro angular momentum must have the form

$$\mathbf{H} = \mathbf{i}' H_{x'} + \mathbf{k}' H_{z'},$$

so that

$$H_y = \beta H_{x'} + \beta'' H_{z'},$$

$$H_z = \gamma H_{x'} + \gamma'' H_{z'}.$$

Thus, $\beta'' = H_y = 0$ implies $\beta H_{x'} = 0$, and $\gamma'' = H_z = 0$ implies $\gamma H_{x'} = 0$. If we now assume that the motion of the gyro gimbals is restricted by stops along the body yaw axis, so that $H_{x'} > 0$, an assumption appropriate for the case of the two-gyro roll-vee, we have the following result:

II. *The equilibrium positions of any symmetrical, gravity-oriented, roll-gyro-stabilized body in a circular orbit must be such that the body roll axis is either parallel to the orbit roll axis* ($\beta' = \pm 1$) *or parallel to the orbit yaw axis* ($\gamma' = \pm 1$).

So far we have not made use of the gyro gimbal equilibrium equations. If the gimbals are not against stops, from (19) for the roll-vee, these take the form

$$\Omega \mathbf{j}' \cdot (\mathbf{H}_1 \times \mathbf{i}) - H\Omega \sin \alpha_0 - K\varphi_{g_1} = 0,$$

$$\Omega \mathbf{j}' \cdot (\mathbf{H}_2 \times \mathbf{i}) + H\Omega \sin \alpha_0 - K\varphi_{g_2} = 0,$$

where we now denote the nominal gimbal angle by $\alpha_0$ to avoid confusion with the direction cosines. If the flex-lead constraint is negligible, i.e., $K = 0$, these two equations imply that

$$\gamma' H_y - \beta' H_z = 0$$

which in either case in II ($\beta' = \pm 1$, $H_y = 0$ or $\gamma' = \pm 1$, $H_z = 0$) implies that $H_y = H_z = 0$, so that no torque is exerted on the satellite by the gyros. We then have:

III. *The equilibrium positions of a symmetrical, gravity-oriented, free roll-vee-gyro-stabilized body must be such that the resultant gyro torque vanishes* ($\boldsymbol{i} \times \boldsymbol{H} = 0$) *and either the body pitch, roll, and yaw axes are parallel to the orbit pitch, roll, and yaw axes* ($\boldsymbol{i}' = \pm \boldsymbol{i}$, $\boldsymbol{j}' = \pm \boldsymbol{j}$, $\boldsymbol{k}' = \pm \boldsymbol{k}$) *or the body pitch, roll, and yaw axes are parallel to the orbit pitch, yaw, and roll axes, respectively* ($\boldsymbol{i}' = \pm \boldsymbol{i}$, $\boldsymbol{j}' = \pm \boldsymbol{k}$, $\boldsymbol{k}' = \pm \boldsymbol{j}$).

The signs of course must be chosen so that the above represents a proper rotation. Note that the above applies to any roll gyro system for which the resultant torque around the body roll axis exerted by the gyros on the satellite vanishes. If $i' = +i$, the second set of equilibrium positions gives Fig. 20, with the gyro gimbal axes along the orbit yaw axis in a yaw-vee configuration. Small pitching motion around these equilibrium positions is governed by a characteristic equation of the same form as that for the roll-vee, (1), except that the coefficient $3(B - C)/A > 0$ is replaced by $3(C - B)/A < 0$. Thus these equilibrium positions are unstable.

The equilibrium position $\mathbf{i}' = \mathbf{i}$, $\mathbf{j}' = \mathbf{j}$, $\mathbf{k}' = \mathbf{k}$ of the first set is shown in Fig. 19(a). It corresponds to the normal operating position with the body yaw axis, on which the antenna is situated, directed toward the earth. The inverted position (see Fig. 19b) $\mathbf{i}' = \mathbf{i}$, $\mathbf{j}' = -\mathbf{j}$, $\mathbf{k}' = -\mathbf{k}$ is also stable, since it merely corresponds to an interchange of the two gyros. This bistability is characteristic of gravity-oriented bodies and a gravity-oriented communications satellite must either use two antennas, with associated switching, or incorporate some means of flipping the satellite in response to ground command. The latter possibility is discussed in some detail in the sequel.

The reversed roll-vee equilibrium positions Fig. 21(a), with $\mathbf{i}' = -\mathbf{i}$, remain to be investigated. The corresponding yaw-vee positions are still unstable. If the gyro gimbals were completely free, satellite precession and the bias torques, now acting together, would rotate the gyro gimbals from the reversed roll-vee until they formed a normal roll-vee around the orbit pitch axis. But with gimbal stops, making the angles $\pm\alpha^*$ with the body pitch axis, the gimbals rotate until the stop reaction torque and the bias torque sustain the 1 rpo steady precession of the satellite in orbit. The stability of this reversed roll-vee position can be investigated by using the characteristic equations for the normal roll-vee, with $H \cos \alpha_0$ replaced by $-H \cos \alpha^*$ and a large spring constant $K^*$, introduced to take the stop compliance into account. In particular the coefficient

$$a_4 = [(A - B)\Omega + 2H \cos \alpha_0][4(A - C)\Omega + 2H \cos \alpha_0]/BC\Omega^2,$$

in the roll-yaw characteristic equation, is replaced by

$$a_4^* = [(A - B)\Omega - 2H \cos \alpha^*][4(A - C)\Omega - 2H \cos \alpha^*]/BC\Omega^2.$$

If

$$(A - B)\Omega < 2H \cos \alpha^* < 4(A - C)\Omega,$$

this is negative and the equilibrium position is unstable.

The instability of the reversed roll-vee when $\alpha^*$ satisfies the above inequalities is shown in Fig. 29. In this case, the system parameters are the same as those of the sample design in Section 2.3.2 with $\alpha^* = 80°$. Initially the gyros are against the stops and the satellite has rates of 0.05 rpo about all three axes. It is seen that the satellite turns around the yaw axis and settles down to rest in the desired orientation in less than five orbits.

When the gyro gimbals are against stops, the gyros exert a torque on the body and in general there are other, skewed equilibrium positions. To investigate these positions without getting involved in the details of stop compliances, etc., which depend on the specific gyros used, we consider only two idealized cases, the first with stops along the positive and negative body yaw axes but with no stops along the body-pitch axis, and the second with stops along the pitch axis as well as along the yaw axis.

In both cases the gyro spin axes may be back-to-back along the yaw axis, but this is a case of zero net gimbal torque already treated and is easily eliminated by moving the gimbal stops in slightly. In the first case both spin axes may lie along the body yaw axis against stops, so that

$$\mathbf{H}_1 = \mathbf{H}_2 = \mathbf{k}'H, \qquad \mathbf{H} = 2\mathbf{k}'H$$

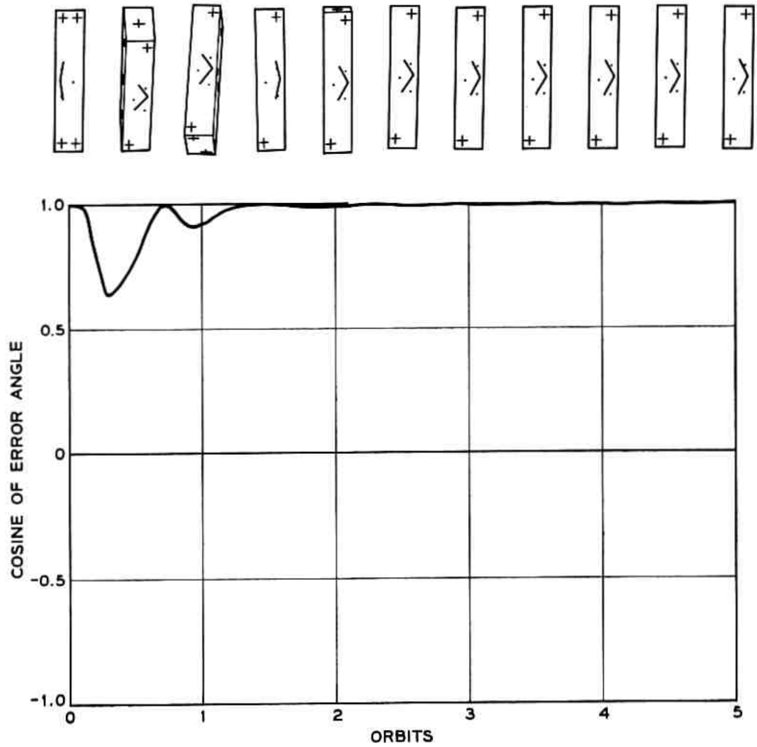


Fig. 29 — Instability of reversed roll-vee when gyro stops are suitably disposed.

From (24) and (25) we again have two cases to consider: (a) $\beta'' = H_y = 0$, and (b) $\gamma'' = H_z = 0$. In case (a) we have

$$H_z = 2\gamma'' H = 4(A - C)\Omega\alpha''\gamma''.$$

The subcase $\beta'' = H_y = 0; \gamma'' = 0$ is easily shown to be unstable, so there remains only the position given by

$$\alpha'' = H/[2(A - C)\Omega].$$

Unless $H > 2(A - C)\Omega$, this yields an equilibrium position which can be maintained by the stop reaction torques. These torques are of course one-sided, since the stop can only "push" and not "pull." This undesirable skew equilibrium position can be eliminated by making

$$H > 2(A - C)\Omega.$$

A similar position for case (b) ($\gamma'' = H_z = 0$) can be eliminated by satisfying the less restrictive condition $H > (A - C)\Omega/2$.

The corresponding situation with yaw and pitch stops finds one gyro against the yaw stop and the other against the pitch stop (see Fig. 21b). For example, suppose

$$\mathbf{H}_1 = \mathbf{k}'H, \qquad \mathbf{H}_2 = \mathbf{i}'H, \qquad \mathbf{H} = H(\mathbf{i}' + \mathbf{k}').$$

Now in case (a), $\beta'' = H_y = 0$, we have

$$H_z = H(\gamma + \gamma'') = 4(A - C)\Omega\alpha''\gamma'',$$

or, in terms of the angle $\theta$ between the $x$ and $x'$ axes,

$$\sin 2\theta = -[H/2^{\frac{1}{2}}(A - C)\Omega] \sin [\theta + (\pi/4)].$$

The two roots of this equation in the interval $-\pi/4 < \theta < \pi$ are excluded, because they require stops which "pull" on the gimbals. On the other hand, the two roots in the interval $\pi < \theta < 3\pi/2$ yield possible equilibrium positions. These roots exist only if $H < 2^{\frac{1}{2}}(A - C)\Omega$. Again the case (b) $\gamma'' = H_z = 0$, yields no equilibrium positions of this type under the less restrictive condition $H > 2^{-\frac{1}{2}}(A - C)\Omega$. Since an increase in gyro angular momentum tends to degrade the transient performance of the system, we shall assume in the following that the gyro gimbals are limited in excursion by both yaw and pitch stops, so that only the restriction $H > 2^{\frac{1}{2}}(A - C)\Omega$ need be satisfied.

In the case of an unsymmetric satellite, a similar but more complicated analysis of the equilibrium positions can be carried out.

### 4.4 *Rod Extension and Satellite Despin*

We have already indicated the necessity of augmenting the satellite inertia to increase the gravity-gradient restoring torques to required levels. If this inertia augmentation is done after injection into orbit, it also reduces the satellite angular velocity to a level where the gravity-gradient torques may become effective in aligning the satellite with the local vertical. One method of inertia augmentation is the extension of so-called STEM rods described in detail in Ref. 8. These are beryllium copper tapes which form straight, tubular rods when unwound from a drum. If they are used together with dense tip masses, the satellite inertia may be increased by several hundredfold without a proportional increase in solar torque. In the following sections we first consider the effect a variable inertia has on the general form of the satellite equation of motion and then discuss satellite despin using a single extensible rod in combination with two gyros.

### 4.4.1 *Equations of Motion for Variable Satellite Inertia*

We may derive all of the dynamical equations for the motion of a gravity-oriented body by integration of the general equations of motion for a continuous medium. In fact this is perhaps the most direct way of calculating the gravity-gradient torque, which is due to the variable gravitational body force acting on each mass element of the body. The resulting equation of motion, (9), applies to rigid and flexible bodies alike, provided that the angular momentum **L** is calculated correctly. **L** is given in general by the integral

$$\mathbf{L} = \int_B \mathbf{r} \times (\partial \mathbf{r}/\partial t + \boldsymbol{\omega}_s \times \mathbf{r})\,dm, \tag{26}$$

where **r** is the radius vector from the center of mass of the body $B$ to the mass element $dm$. For a rigid body, **r** differs from its initial value $\mathbf{r}_0$ only by a rotation and $\partial \mathbf{r}/\partial t = 0$, yielding the usual form

$$\mathbf{L} = \int_B \mathbf{r} \times (\boldsymbol{\omega}_s \times \mathbf{r})\,dm = \mathbf{I}\cdot\boldsymbol{\omega}_s\,,$$

but in general **r** depends both on $\mathbf{r}_0$ and $t$, so that

$$\mathbf{L} = \mathbf{I}\cdot\boldsymbol{\omega}_s + \int_B \mathbf{r} \times (\partial \mathbf{r}/\partial t)\,dm,$$

where the inertia dyadic **I** depends on $t$.

Let us now consider the extension of a single massless rod with tip mass $m_a$. If $\mathbf{a}(t)$ is the radius vector from the center of mass of the satellite proper to the tip mass, (26) yields

$$\mathbf{L} = \mathbf{I}\cdot\boldsymbol{\omega}_s + m\mathbf{a} \times \dot{\mathbf{a}},$$

$$\mathbf{I} \times \boldsymbol{\omega}_s = \mathbf{I}_0 \times \boldsymbol{\omega}_s + m\mathbf{a} \times (\boldsymbol{\omega}_s \times \mathbf{a}),$$

$$m = m_a m_s/(m_a + m_s),$$

for satellite mass $m_s$, and the inertia dyadic for the satellite around its center of mass

$$\mathbf{I}_0 = A_0\mathbf{i}'\mathbf{i}' + B_0\mathbf{j}'\mathbf{j}' + C_0\mathbf{k}'\mathbf{k}'.$$

When the rod is erected parallel to itself, as would normally be the case, $\mathbf{a} \times \dot{\mathbf{a}} = 0$ and the effect of rod extension is entirely taken into account by the time-dependent inertia **I**. If the rod is extended along the axis of minimum moment of inertia, $\mathbf{a} = \mathbf{k}'a(t)$ and

$$\mathbf{I} = (A_0 + ma^2)\mathbf{i}'\mathbf{i}' + (B_0 + ma^2)\mathbf{j}'\mathbf{j}' + C_0\mathbf{k}'\mathbf{k}'.$$

### 4.4.2 *Satellite Despin*

When the moments of inertia of a torque-free spinning body are increased by a factor of $N$, conservation of angular momentum requires that the angular velocity of the body decrease by a factor of $1/N$. On the other hand, if the spinning body contains a spinning rotor, an increase in the angular momentum of the rotor produces a corresponding decrease in the angular momentum of the body and hence a reduction in the angular velocity of the body. Both elements exist in the gravity-oriented, gyro-stabilized satellite. Inertia augmentation is required to obtain gravity-gradient torques of the proper level; rotation of the gyro gimbals provides a change in angular momentum.

A single extensible rod-tip mass combination provides adequate gravity-gradient torques, if erected along the satellite yaw axis. However, erection of such a rod reduces only pitch and roll injection angular velocities; the yaw component is unaffected. This may be removed by using the gyros as reaction wheels.

Suppose the gyros are caged at their null position during the rod erection phase. Then, neglecting gravity-gradient and external torques during the short erection time, we have a freely spinning body. If the initial components of angular momentum are all of the order of magnitude $A_0N\Omega$, where $A_0$ is the moment of inertia about all three axes, we may reduce the pitch and roll components of angular velocity to order $\Omega$ (1 rpo), with respect to inertial space, by extending a single rod that increases the pitch and roll moments of inertia from $A_0$ to $A = NA_0$. The yaw angular velocity remains equal to $N\Omega$.

The yaw angular momentum, $A_0N\Omega$, however, is of the order $A\Omega$, the same order of magnitude as the angular momentum $H$ of each gyro. The gyros then are large enough to absorb the residual angular momentum. If the gyro gimbals are now released, the spin axes will tend to line up with the residual angular velocity around the yaw axis. One gyro spin axis rotates until constrained by the yaw axis gimbal stop; the other rotates until constrained by the pitch axis gimbal stop. There is thus a net change in yaw gyro angular momentum of order $H$ and, furthermore, because of the rate-seeking property of the gyro, it always occurs in the correct sense to reduce the satellite angular momentum.

This qualitative argument has been supported by computer runs for given parameter values and initial rates. In particular one may determine the gyro size needed to despin from a given initial yaw rate. Actually it is not necessary to separate the erection and uncaging phases of injection, so long as the gyros are not uncaged before erection.

The result of such a computer run was given in Figs. 17–18. In one orbit the angular rates are reduced to less than 1 $\Omega$ and capture occurs.

### 4.5 *Satellite Inversion*

As we have already noted, gravity-oriented bodies are bistable, i.e., they are in stable equilibrium with the axis of least inertia, on which the antenna would be mounted, both directed toward the earth and away from the earth. In this section we discuss a method of flipping the satellite by means of a simple ground command injected into the gyro gimbal torquers.

When the satellite is in either of the above stable equilibrium positions, its total angular momentum is $A\Omega + 2H \cos \alpha_0$ around the pitch axis. If we could somehow rotate the two gyro gimbals instantaneously, so that both spin axes pointed along the pitch axis, the total angular momentum would become $A(\Omega + \omega) + 2H$, where $\omega$ is the pitch angular velocity with respect to the orbit frame. Since the gimbal rotation is assumed instantaneous, the total angular momentum is conserved, i.e.,

$$A\Omega + 2H \cos \alpha_0 = A(\Omega + \omega) + 2H,$$

or

$$\omega = -2H(1 - \cos \alpha_0)/A.$$

Thereafter, the single-axis, pitching motion is governed by an equation of the form

$$A\ddot{\varphi} + 3(B - C)\Omega^2 \sin \varphi \cos \varphi = 0,$$

where $\varphi$ is the pitch angle around the orbit pitch axis. A first integral of this equation is

$$A\dot{\varphi}^2 + 3(B - C)\Omega^2 \sin^2 \varphi = A\omega^2,$$

since $\varphi(0) = 0$, $\dot{\varphi}(0) = \omega$. In order that $\dot{\varphi}$ be one-signed, i.e., in order that tumbling occur, we must have

$$A\omega^2 > 3(B - C)\Omega^2$$

or

$$(H/A\Omega)^2 > 3(B - C)/4A(1 - \cos \alpha_0)^2. \tag{27}$$

For a gyro angular momentum satisfying this condition we can excite a tumbling motion by collapsing the gyro spin axes toward the pitch axis. Actually, since we only want to rotate the satellite through a half revolution, the gyro angular momentum need barely exceed this minimum value. Furthermore, we may collapse the gyro spin axes toward the pitch axis by simply reversing the bias torques applied to the gimbals for a suitable length of time. For a spindle satellite with $(B - C)/A = 0.99$, computer runs (see Fig. 22) show that the satellite may be inverted by applying this reversed bias for about a half orbit. For a satellite with $(B - C)/A = 0.4$ it turns out that it is only necessary to remove the bias torques for a fraction of an orbit. For any satellite a suitable combination of bias torque and time can always be found to flip the satellite into its desired operating position, providing the relation (27) is satisfied.

We remark that bias torques could also be used to rotate the gyro gimbals against the yaw stops. A similar, single-axis argument then gives an expression like (27). However, with the gyros back-to-back against yaw stops, the satellite has negligible yaw stiffness, and is vulnerable to yaw disturbances. This possibility of inversion was therefore not pursued.

## V. ACKNOWLEDGMENTS

The authors would like to acknowledge the contributions of Dr. J. E. DeLisle, B. M. Hildebrant, and E. G. Ogletree, of the MIT Instrumentation Laboratory, not only in the initial study but also in subsequent consultations. They would also like to thank J. Sautter and Mrs. C. Kimme for assistance with computer programming.

## APPENDIX

### *List of Symbols*

| | |
|---|---|
| $A,B,C$ | principal moments of inertia |
| $b = B/A,\ c = C/A$ | dimensionless principal inertias |
| $C_D$ | gyro damping constant |
| $D = 1/2\pi T_s$ | damping rate, inversely proportional to settling time |
| $\mathbf{H}_1\ ,\ \mathbf{H}_2$ | angular momentum vectors of gyros 1 and 2 |
| $H = \mid \mathbf{H}_1 \mid = \mid \mathbf{H}_2 \mid$ | magnitude of gyro angular momentum |
| $\mathbf{H} = \mathbf{H}_1 + \mathbf{H}_2$ | resultant gyro angular momentum vector |
| $H_x\ ,\ H_y\ ,\ H_z$ | orbital pitch, roll, and yaw components of total gyro momentum |

| | |
|---|---|
| $H_{x'}, H_{y'}, H_{z'}$ | body pitch, roll, and yaw components of total gyro momentum |
| $h = (H/A\Omega) \cos \alpha$ | dimensionless gyro angular momentum |
| $h' = (H/C_D) \cos \alpha$ | dimensionless damping parameter |
| $\mathbf{I}$ | inertia dyadic |
| $\mathbf{i},\mathbf{j},\mathbf{k}$ | unit vectors along orbital pitch, roll, and yaw axes |
| $\mathbf{i}',\mathbf{j}',\mathbf{k}'$ | unit vectors along body pitch, roll, and yaw axes |
| $K$ | gimbal spring constant |
| $\mathbf{L}$ | satellite angular momentum vector |
| $\mathbf{M}_G$ | gravity-gradient torque |
| $\mathbf{M}_H$ | resultant gyro torque |
| $m_s, m_a$ | satellite mass, tip mass |
| $m = m_a m_s/(m_a + m_s)$ | effective mass of satellite and rod assembly |
| $p = \dfrac{1}{\Omega}\dfrac{d}{dt}$ | differentiation operator or transform parameter |
| $T_s$ | asymptotic $1/e$ settling time in orbits |
| $x,y,z$ | coordinate axes in orbital pitch, roll, and yaw directions |
| $x',y',z'$ | coordinate axes in body pitch, roll, and yaw directions |
| $\alpha$ | vee half-opening angle |
| $\alpha,\alpha',\alpha''$ | direction cosines between $\mathbf{i}$ and $\mathbf{i}',\mathbf{j}',\mathbf{k}'$ vectors |
| $\beta,\beta',\beta''$ | direction cosines between $\mathbf{j}$ and $\mathbf{i}',\mathbf{j}',\mathbf{k}'$ vectors |
| $\gamma,\gamma',\gamma''$ | direction cosines between $\mathbf{k}$ and $\mathbf{i}',\mathbf{j}',\mathbf{k}'$ vectors |
| $\epsilon$ | orbital eccentricity |
| $\kappa = 1 + [K/(H\Omega \cos \alpha)]$ | dimensionless gimbal spring parameter |
| $\xi_x, \xi_y, \xi_z, \chi$ | Euler parameters |
| $\varphi_x, \varphi_y, \varphi_z$ | small pitch, roll, and yaw angles |
| $\varphi_g = \frac{1}{2}(\varphi_{g_1} + \varphi_{g_2})$ | gyro sum angle |
| $\psi$ | polar angle of satellite center of mass |
| $\psi_g = \frac{1}{2}(\varphi_{g_1} - \varphi_{g_2})$ | gyro difference angle |
| $\Omega$ | orbital rate |
| $\boldsymbol{\omega}$ | satellite angular velocity referred to orbital axes |
| $\boldsymbol{\omega}_s$ | satellite angular velocity referred to inertial space |

REFERENCES

1. Pierce, J. R., Orbital Radio Relays, Jet Propulsion, **25**, 1955, pp. 153–157.
2. Ogletree, E. G., Sklar, S. J., and Mangan, J., Satellite Attitude Control

Study, Part I, Report No. R-308, Instrumentation Laboratory, MIT, July, 1961.

3. Ogletree, E. G., Mangan, J. G., Petranic, T. D., Sklar, S. J., Merkle, R. F., and Hutchinson, R. C., Satellite Attitude Control Study, Part II, Report No. R-308, Instrumentation Laboratory, MIT, Feb., 1962.
4. DeLisle, J. E., Ogletree, E. G., and Hildebrant, B. M., Applications of Gyrostabilizers to Satellite Attitude Control, *Progress in Astronautics and Aeronautics*, Vol. 13, Academic Press, 1964, pp. 149–175.
5. Burt, E. G. C., On the Attitude Control of Earth Satellites, Eighth Anglo-American Aeronautical Conference, London, Sept., 1961.
6. Scott, E. D., Control Moment Gyro Gravity Stabilization, *Progress in Astronautics and Aeronautics*, Vol. 13, Academic Press, 1964, pp. 103–149.
7. Kamm, L. J., Vertistat: An Improved Satellite Orientation Device, ARS Journal, **32,** 1962, pp. 911–913.
8. 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.
9. 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.
10. Paul, B., Planar Librations of an Extensible Dumbbell Satellite, AIAA Journal, **1,** Feb., 1963, pp. 411–418.
11. Newton, R. R., Damping of a Gravitationally Stabilized Satellite, AIAA Journal, **2,** Jan., 1964, pp. 20–25.
12. Fischell, R. E., and Mobley, F. F., A System for Passive Gravity-Gradient Stabilization of Earth Satellites, *Progress in Astronautics and Aeronautics*, Vol. 13, Academic Press, 1964, pp. 37–71.
13. Lewis, J. A., Viscous Damping of Gravitationally Stabilized Satellites, Proc. Fourth U.S. Nat. Congr. Appl. Mech., Berkeley, Calif., June, 1962.
14. Zajac, E. E., Limits on the Damping of Two-Body Gravitationally Oriented Satellites, AIAA Journal, **1,** Feb., 1963, pp. 498–499.
15. Zajac, E. E., Computer-Made Perspective Movies as a Scientific and Communication Tool, Comm. ACM, **7,** March, 1964, pp. 169–170.
16. Zajac, E. E., Bounds on the Decay Rate of Damped Linear Systems, Quart. Appl. Math., **20,** Jan., 1963, pp. 383–384.
17. Zajac, E. E., Damping of a Gravitationally Oriented Two-Body Satellite, ARS Journal, **32,** Dec., 1962, pp. 1871–1875.
18. Knopp, K., *Theory of Functions*, Vol. II, Dover, N. Y., 1947, p. 121.
19. Uspensky, J. V., *Theory of Equations*, McGraw-Hill, N. Y., 1948, p. 75.
20. Beletskii, V. V., The Librations of a Satellite, in *Artificial Earth Satellites*, Vol. 3 (translated from the Russian), Plenum Press, N. Y., 1961, pp. 18–45.
21. Whittaker, E. T., *Analytical Dynamics of Particles and Rigid Bodies*, Cambridge Univ. Press, 1961, p. 8.
22. Hamming, R. W., *Numerical Methods for Scientists and Engineers*, McGraw-Hill, 1962.
23. Zajac, E. E., Capture Problem in Gravitational Attitude Control of Satellites, ARS Journal, **31,** Oct., 1961, pp. 1464–1466.
24. Morgan, S. P., and Yu, E. Y., Expected Frequencies of Meteoritic Disturbances of Gravitationally Oriented Satellites, to appear.

# ERRATA

**Design of Armorless Ocean Cable,** M. W. Bowker, W. G. Nutt and R. M. Riley, B.S.T.J. **43,** July, 1964, pp. 1185–1208.

Equation on p. 1197 should read:

$$\alpha = k_1\sqrt{f}\left(\frac{1}{d\sqrt{\sigma_i}} + \frac{1}{D\sqrt{\sigma_o}}\right)\frac{\sqrt{\epsilon}}{\log D/d} + k_2 f F_p \sqrt{\epsilon}.$$

**Analysis of a Tubular Gas Lens,** D. Marcuse and S. E. Miller, B.S.T.J., **43,** July, 1964, pp. 1759–1782.

Equation (25), p. 1778, should read:

$$f = 0.596\,\frac{a^2}{z}\,\frac{T_0}{\theta_0(n_0 - 1)}\,.$$

# ERRATA

**Design of Armorless Ocean Cable,** M. W. Bowker, W. G. Nutt and R. M. Riley, B.S.T.J. **43,** July, 1964, pp. 1185–1208.

Equation on p. 1197 should read:

$$\alpha = k_1\sqrt{f}\left(\frac{1}{d\sqrt{\sigma_i}} + \frac{1}{D\sqrt{\sigma_o}}\right)\frac{\sqrt{\epsilon}}{\log D/d} + k_2 f F_p \sqrt{\epsilon}.$$

**Analysis of a Tubular Gas Lens,** D. Marcuse and S. E. Miller, B.S.T.J., **43,** July, 1964, pp. 1759–1782.

Equation (25), p. 1778, should read:

$$f = 0.596\,\frac{a^2}{z}\,\frac{T_0}{\theta_0(n_0 - 1)}\,.$$

Study, Part I, Report No. R-308, Instrumentation Laboratory, MIT, July, 1961.

3. Ogletree, E. G., Mangan, J. G., Petranic, T. D., Sklar, S. J., Merkle, R. F., and Hutchinson, R. C., Satellite Attitude Control Study, Part II, Report No. R-308, Instrumentation Laboratory, MIT, Feb., 1962.
4. DeLisle, J. E., Ogletree, E. G., and Hildebrant, B. M., Applications of Gyrostabilizers to Satellite Attitude Control, *Progress in Astronautics and Aeronautics*, Vol. 13, Academic Press, 1964, pp. 149–175.
5. Burt, E. G. C., On the Attitude Control of Earth Satellites, Eighth Anglo-American Aeronautical Conference, London, Sept., 1961.
6. Scott, E. D., Control Moment Gyro Gravity Stabilization, *Progress in Astronautics and Aeronautics*, Vol. 13, Academic Press, 1964, pp. 103–149.
7. Kamm, L. J., Vertistat: An Improved Satellite Orientation Device, ARS Journal, **32,** 1962, pp. 911–913.
8. 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.
9. Fletcher, H. J., Rongved, L., and Yu, E. Y., Dynamics Analysis of a Two-Body Gravitationally Oriented Satellite, B.S.T.J., **42,** Sept., **1**963, pp. 2239–2266.
10. Paul, B., Planar Librations of an Extensible Dumbbell Satellite, AIAA Journal, **1,** Feb., 1963, pp. 411–418.
11. Newton, R. R., Damping of a Gravitationally Stabilized Satellite, AIAA Journal, **2,** Jan., 1964, pp. 20–25.
12. Fischell, R. E., and Mobley, F. F., A System for Passive Gravity-Gradient Stabilization of Earth Satellites, *Progress in Astronautics and Aeronautics*, Vol. 13, Academic Press, 1964, pp. 37–71.
13. Lewis, J. A., Viscous Damping of Gravitationally Stabilized Satellites, Proc. Fourth U.S. Nat. Congr. Appl. Mech., Berkeley, Calif., June, 1962.
14. Zajac, E. E., Limits on the Damping of Two-Body Gravitationally Oriented Satellites, AIAA Journal, **1,** Feb., 1963, pp. 498–499.
15. Zajac, E. E., Computer-Made Perspective Movies as a Scientific and Communication Tool, Comm. ACM, **7,** March, 1964, pp. 169–170.
16. Zajac, E. E., Bounds on the Decay Rate of Damped Linear Systems, Quart. Appl. Math., **20,** Jan., 1963, pp. 383–384.
17. Zajac, E. E., Damping of a Gravitationally Oriented Two-Body Satellite, ARS Journal, **32,** Dec., 1962, pp. 1871–1875.
18. Knopp, K., *Theory of Functions*, Vol. II, Dover, N. Y., 1947, p. 121.
19. Uspensky, J. V., *Theory of Equations*, McGraw-Hill, N. Y., 1948, p. 75.
20. Beletskii, V. V., The Librations of a Satellite, in *Artificial Earth Satellites*, Vol. 3 (translated from the Russian), Plenum Press, N. Y., 1961, pp. 18–45.
21. Whittaker, E. T., *Analytical Dynamics of Particles and Rigid Bodies*, Cambridge Univ. Press, 1961, p. 8.
22. Hamming, R. W., *Numerical Methods for Scientists and Engineers*, McGraw-Hill, 1962.
23. Zajac, E. E., Capture Problem in Gravitational Attitude Control of Satellites, ARS Journal, **31,** Oct., 1961, pp. 1464–1466.
24. Morgan, S. P., and Yu, E. Y., Expected Frequencies of Meteoritic Disturbances of Gravitationally Oriented Satellites, to appear.

# Optimum Reception of Binary Gaussian Signals

By T. T. KADOTA

(Manuscript received May 4, 1964)

*The problem of optimum reception of binary Gaussian signals is to specify, in terms of the received waveform, a scheme for deciding between two alternative covariance functions with minimum error probability. Although a considerable literature already exists on the problem, an optimum decision scheme has yet to appear which is both mathematically rigorous and convenient for physical application. In the context of a general treatment of the problem, this article presents such a solution. The optimum decision scheme obtained consists in comparing, with a predetermined threshold $k$, a quadratic form (of function space) in the received waveform $x(t)$, namely,*

$$\text{choose} \quad r_0(s,t) \quad \textit{if} \quad \iint x(s)h(s,t)x(t)\, ds\, dt < k,$$

$$\text{choose} \quad r_1(s,t) \quad \textit{if} \quad \iint x(s)h(s,t)x(t)\, ds\, dt \geqq k,$$

*where $r_0(s,t)$ and $r_1(s,t)$ are the covariance functions while $h(s,t)$ is given as a solution of the integral equation,*

$$\iint r_0(s,u)h(u,v)r_1(v,t)\, du\, dv = r_1(s,t) - r_0(s,t).$$

*This may be regarded as a generalization of the "correlation detection" in the case of binary sure signals in noise.*

*Section I defines the problem, reviews the literature, and, together with certain pertinent remarks, summarizes principal results. A detailed mathematical treatment follows in Section II and the Appendices.*

## I. INTRODUCTION AND SUMMARY

### 1.1 *Definition and Nature of Problem*

The problem of optimum reception of binary Gaussian signals arises as a mathematical idealization of a common communication problem.

Consider a radio communication link containing a random medium. The transmitter sends one of two possible signals with known frequency rates (a priori probabilities), and the receiver decides which one of the two has been transmitted. Even if the transmitted signals are deterministic, the observable waveforms at the receiver appear to be random owing to effects of the random medium and the ever-present thermal noise at the receiver. The task of the so-called optimum (or ideal) receiver is to decide, upon observation of the received waveform for a finite time, which one of the two signals has been transmitted in such a way as to minimize the so-called probability of error. Thus, the problem of optimum reception amounts to specifying in terms of the received waveform such an optimum decision scheme for given a priori probabilities.

It is assumed that the values of the received waveforms at arbitrary instants of time during the observation interval, say $0 \leqq t \leqq 1$, are jointly Gaussian distributed with means zero and a covariance matrix which is determined by either one of two known covariance functions, depending upon which one of the two signals is transmitted. Then, the above problem may be stated as one of testing simple hypotheses as follows: Suppose there are two ensembles of real functions of time $t$, $0 \leqq t \leqq 1$, which are statistically characterized as being Gaussian distributed with identically vanishing mean functions and two distinct covariance functions. A sample (function) $x(t)$ is drawn either from the first ensemble with probability $\alpha$ (the null hypothesis: $H_0$) or from the second with probability $1 - \alpha$ (the alternative hypothesis: $H_1$). Determine a "critical region" $\Lambda_\alpha$ (a subset of a space of real functions $\Omega$) for rejecting $H_0$ (or accepting $H_1$) if $x(t)$ belongs to $\Lambda_\alpha$ and accepting $H_0$ if $x(t)$ does not, in such a way that the associated error probability,

$$P_e(\Lambda_\alpha) = \alpha P_0(\Lambda_\alpha) + (1 - \alpha)P_1(\Omega - \Lambda_\alpha), \tag{1}$$

is no greater than $P_e(\Lambda)$ for an arbitrary $\Lambda \subset \Omega$; where $P_0$ and $P_1$ are two Gaussian (probability) measures defined on (measurable) subsets of $\Omega$ by the two zero mean functions and two covariance functions. Thus, the problem of optimum reception amounts to dividing the function space into two parts in such a way that the weighted probabilities on them specified by (1) are minimum among all possible divisions.

There are two features worth noting in this formulation. One is the lack of uniqueness of the optimum division as a consequence of adopting the minimum error probability as the optimality criterion. Namely, it is immaterial whether a certain set $N$ (of functions) with both probabilities zero, i.e., $P_0(N) = 0 = P_1(N)$, should be a part of $\Lambda_\alpha$ or $\Omega -$

$\Lambda_\alpha$, since it does not contribute to the error probability $P_e$. Thus, those sets upon which $P_0$ and $P_1$ vanish can effectively be ignored. The other feature is a stipulation that the division be specified in terms of the general sample (function), namely, the general element $\omega$ of the function space $\Omega$, so that each sample (a received waveform) can be classified as a member of $\Lambda_\alpha$ or $\Omega - \Lambda_\alpha$. From the probability theoretical point of view, these features dictate specification of the division (or the decision scheme) to be made in terms of the "almost all sample functions" (or "almost surely," "with probability one," etc.) proposition. While this offers flexibility in one sense, it presents a restriction in another. For example, anticipating the forthcoming results, if the division of $\Omega$ is made by means of a certain $\omega$ function on $\Omega$, this function can be arbitrary or even undefined on the sets of $\omega$ upon which $P_0$ and $P_1$ vanish. Yet, if the function is defined as a certain limit (or, obtained by a limit operation), then the sense of convergence must be at least "for almost all sample functions," but not "in quadratic mean (in the mean)," "in probability," and "in distribution," which are in general weaker.

The problem of optimum reception of binary Gaussian signals may be regarded as a generalization of an almost classical problem in communication theory, namely, optimum detection of binary *sure* signals in Gaussian noise. It is well known that such detection consists in comparing, with a preassigned threshold, the correlation integral of the received waveform and a certain function determined by the two signals and noise characteristics. More precisely, let $\{x_t, 0 \leqq t \leqq 1\}$ be a Gaussian process whose covariance function is $r(s,t)$, $0 \leqq s,t \leqq 1$, continuous and positive-definite, and whose mean function is either $m_0(t)$ or $m_1(t)$, both continuous, corresponding to the two sure signals. Denote the sample function of the process by $x(t)$ and the threshold by $c > 0$. Then Grenander[1] shows that if the integral equation

$$\int_0^1 r(s,t)g(s)\,ds = m_1(t) - m_0(t) \tag{2}$$

has a square-integrable solution, the optimum decision scheme under the Neyman-Pearson criterion is the following:

$$\begin{aligned} &\text{choose} \quad m_0(t) \quad \text{if} \quad \int_0^1 x(t)g(t)\,dt < c, \\ &\text{choose} \quad m_1(t) \quad \text{if} \quad \int_0^1 x(t)g(t)\,dt \geqq c. \end{aligned} \tag{3}$$

Suppose the two sure signals in the above problem are replaced by two *stochastic* (Gaussian) signals and the additive noise is included in these signals so that the decision between two sure signals becomes now the decision between two Gaussian signals. Furthermore, suppose the optimality criterion is changed from the Neyman-Pearson's to the error-probability minimization. Then, the problem becomes optimum reception of Gaussian signals under the minimum error-probability criterion. More precisely, let $\{x_t, 0 \leqq t \leqq 1\}$ be a Gaussian process whose mean function is identically zero and whose covariance function is either $r_0(s,t)$ or $r_1(s,t)$, continuous and positive-definite, with the accompanying a priori probabilities $\alpha$ and $1 - \alpha$ respectively. Then what are the counterparts of (2) and (3)? That is, under what conditions can the optimum decision scheme be specified in terms of a correlation integral involving the sample function, and what is the decision scheme itself?

### 1.2 *Review of Literature*

Despite momentous foundations laid by Grenander in 1950, little progress was made toward rigorous solution of the above problem during the succeeding decade, due primarily to restrictions of the mathematical scope to elementary probability theory. The majority of the work is characterized by two features: (*i*) use (and misuse) of the classical method of likelihood ratio and (*ii*) attempts to specify the decision scheme in terms of some integrals involving the sample function. In order to use the classical method, however, the continuous (parameter) process must first be "represented" by a (finite) sequence of random variables. Thus Middleton[2] and Price [3] sample $\{x_t, 0 \leqq t \leqq 1\}$ to obtain the representing sequence $x_{t_1}, \cdots, x_{t_n}$ and form their likelihood ratio $l_n$:

$$l_n(x_{t_1}, \cdots, x_{t_n}) = | R_0^{(n)}(R_1^{(n)})^{-1} |^{\frac{1}{2}} \exp\left\{\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} [(R_0^{(n)})^{-1} - (R_1^{(n)})^{-1}]_{ij} x_{t_i} x_{t_j}\right\}, \tag{4}$$

where $R_0^{(n)}$ and $R_1^{(n)}$ are two alternative covariance matrices of $x_{t_1}, \cdots, x_{t_n}$ given respectively by $(R_0^{(n)})_{ij} = r_0(t_i, t_j)$ and $(R_1^{(n)})_{ij} = r_1(t_i, t_j)$; $i, j = 1, \cdots, n$. Then, as $n \to \infty$ and each sampling interval becomes infinitesimal, the decision scheme is specified in terms of the limits of the exponent and the factor before the exponential in (4), provided these limits exist. Middleton argues on a formal basis that the exponent of (4) becomes an integral

$$\int_0^1 x_t y_t \, dt,$$

where the new process $\{y_t, 0 \leqq t \leqq 1\}$ is given as one of the solutions of a pair of certain simultaneous "stochastic integral equations." Price also formally argues that the exponent converges to an integral

$$\int_0^1 \int_0^1 x_s[g_1(s,t) - g_0(s,t)]x_t \, ds \, dt,$$

where $g_0$ and $g_1$ are given as solutions of a certain pair of ordinary integral equations.

Davis,[4] Bello[5] and Turin,[6] on the other hand, make orthonormal expansions of the process and use the Fourier coefficients as the representing sequence. However, the formulation of Davis and Bello is based upon a ratio of probability density functions of *two* sequences of Fourier coefficients corresponding to two separate orthonormal expansions, which is *not* a likelihood ratio; while the fundamental notion in Turin's formulation is "probability density functions of processes," which are unbounded functions in general.

One difficulty common among all the papers is the total absence of convergence proofs for series of random variables. As mentioned in Section 1.1, the sense of convergence must be "for almost all sample functions." Yet, for example, it is not clear on what ground the exponent of (4) should converge for almost all sample functions to those stochastic integrals, nor is the existence of the integrals themselves shown.

The other common difficulty, of a more fundamental nature, is the lack of optimality proofs. Considering the process as an ensemble of "well behaved" functions of time, it is intuitively plausible that such an ensemble should be "adequately" described by the distributions of the "infinitely densely" sampled values of the member functions or by the distributions of the Fourier coefficients of some orthogonal expansions in $\mathfrak{L}_2$ (the space of square-integrable functions). Namely, the continuous (parameter) process should somehow be "representable" by a sequence (infinite in general) of random variables. However, the optimality of the resultant decision scheme should in general be affected by selection of the representing sequence. Obviously, there are innumerable ways of sampling the process, resulting in innumerable decision schemes. Similarly, there are as many sequences of Fourier coefficients to represent the process as orthonormal bases of $\mathfrak{L}_2$. Yet, should all the representing sequences eventually yield the decision schemes with the same error probability, the minimum? If not, which sequences are the best representations? Even if the best sequence is chosen, on what grounds will the error probability remain minimum in the limit as $n \to \infty$, since, after all, the classical method is valid only for a finite $n$?

Note that there is no a priori need for the use of either likelihood ratios or representations, so long as the proposed decision scheme is shown to have the minimum error probability. In fact, Slepian[7] shows interesting special examples (of the singular case) where minimality of the error probability is explicitly proved. From a different point of view, Parzen recently restores Grenander's basic formulation, where what is called the Radon-Nikodym derivative plays the role of the likelihood ratio in the classical theory, and puts the sampling method on a more rigorous basis.

### 1.3 *Summary of Main Results and Remarks*

Solution of the problem of optimum reception stated in Section 1.1 rests on the following two fundamental (measure theoretical) facts:

(a) If $P_0$ and $P_1$ are two Gaussian (probability) measures, they must be either (*i*) "equivalent," i.e., $P_0 \equiv P_1$, or (*ii*) "orthogonal" (or "singular"), i.e., $P_0 \perp P_1$.

(*b*) If $P_0$ and $P_1$ are equivalent, there exists a certain nonnegative random variable $f(\omega)$, called the Radon-Nikodym derivative of $\mathrm{P}_1$ with respect to $P_0$, and a set of $\omega$ points in $\Omega$ such that $f(\omega) \geqq \alpha/(1-\alpha)$ can be taken as the desired critical region, denoted by $\Lambda_\alpha$ in Section 1.1. On the other hand, if $\mathrm{P}_0$ and $P_1$ are orthogonal, there exists a set $H$ of $\omega$ points in $\Omega$ such that $P_0(H) = 0$ and $P_1(H) = 1$, and the critical region can be taken to be such a set $H$. In short, the following set $S_\alpha$ serves as the critical region:

$$S_\alpha = \begin{cases} \{f(\omega) \geqq \alpha/(1-\alpha)\} & \text{if} \quad P_0 \equiv P_1, \\ H & \text{if} \quad P_0 \perp P_1. \end{cases} \tag{5}$$

Thus, the problem of determining the critical region now becomes the problem of finding such a random variable and a set $H$.

Next, through the use of theory of martingales, the following facts can be established:

For almost all sample functions,

(*i*) if (and only if)

$$\lim_{n\to\infty} \operatorname{tr}\,[(R_0^{(n)})^{-1}\,R_1^{(n)} - 2I - R_0^{(n)}\,(R_1^{(n)})^{-1}] < \infty, \tag{6}$$

then

$$\lim_{n\to\infty} l_n(x_{t_1}, \cdots, x_{t_n}) = f(\omega) \quad \text{under both hypotheses;*} \tag{7}$$

(*ii*) if (and only if) (6) is not satisfied, then

$$\lim_{n\to\infty} l_n(x_{t_1}, \cdots, x_{t_n}) = \begin{cases} 0 & \text{under null hypothesis,} \\ \infty & \text{under alternative hypothesis,} \end{cases} \tag{8}$$

provided that the sequence $\{t_i\}$ is dense in the interval $0 \leqq t \leqq 1$, where "tr" stands for "trace" and the likelihood ratio $l_n$, together with $R_0^{(n)}$ and $R_1^{(n)}$, is previously defined in (4).†

Examination of (7) and (8) in conjunction with (5) immediately leads to the conclusion that, irrespective of the hypotheses,

$$S_\alpha = \{\lim_{n\to\infty} l_n(x_{t_1}, \cdots, x_{t_n}) \geqq \alpha/(1 - \alpha)\}. \tag{9}$$

Thus, if $x(t_1), \cdots, x(t_n)$ are the values of the sample function (the received waveform) $x(t)$, $0 \leqq t \leqq 1$, sampled arbitrarily but with the restriction that each sampling interval becomes infinitesimal as $n \to \infty$, then the optimum decision scheme becomes the following:

$$\begin{aligned} &\text{choose} \quad r_0(s,t) \quad \text{if} \quad \lim_{n\to\infty} l_n[x(t_1), \cdots, x(t_n)] < \alpha/(1 - \alpha), \\ &\text{choose} \quad r_1(s,t) \quad \text{if} \quad \lim_{n\to\infty} l_n[x(t_1), \cdots, x(t_n)] \geqq \alpha/(1 - \alpha). \end{aligned} \tag{10}$$

Furthermore, according to (*i*), if the given covariance functions $r_0(s,t)$ and $r_1(s,t)$ are such that (6) is satisfied by their covariance matrices $R_0^{(n)}$ and $R_1^{(n)}$ obtained through sampling, then, regardless of whether $r_0(s,t)$ or $r_1(s,t)$ is the true covariance function, the above limit is finite for almost all sample functions, and the error probability associated with the decision scheme (10) is minimum. According to (*ii*), on the other hand, if $r_0(s,t)$ and $r_1(s,t)$ are such that $R_0^{(n)}$ and $R_1^{(n)}$ do not satisfy (6), then for almost all sample functions the limit vanishes if $r_0(s,t)$ is true, while the limit diverges if $r_1(s,t)$ is true; and, independent of the given a priori probabilities, the associated error probability simply vanishes, resulting in the case of "perfect reception."

* Recall that the null hypothesis is the hypothesis that $r_0(s,t)$ is the true covariance function of the process while the alternative is the hypothesis that $r_1(s,t)$ is the true covariance function.

† (6) and (7) are also found in Parzen.[8]

It should be noted, first of all, that the sequence of sampled values is not used to represent the continuous process but to obtain the crucial random variable $f$ and set $H$ through formation of the likelihood ratio. Secondly, under the assumption of the covariance functions being continuous, it can be proved that, regardless of the sampling manner, the limit of the likelihood ratio satisfies either (7) or (8), thus yielding the same error probability, so long as each sampling interval becomes infinitesimal as $n \to \infty$.* Lastly, negation of condition (6) can be regarded as a necessary and sufficient condition for perfect reception.

Having obtained the optimum decision scheme (10), the question of possible simplification naturally arises next. Examination of the form of the likelihood ratio (4) suggests that, if the limits of the exponent and the factor before the exponential exist separately, decision scheme (10) may be rewritten in terms of these limits. Such an attempt already appears in the literature, as mentioned in Section 1.2. However, the crucial mathematical consideration hinges upon the condition under which such a procedure can be justified. Here, the following condition is shown to be necessary and sufficient:

$$\begin{aligned} &\lim_{n\to\infty} \operatorname{tr}\,[(R_0^{(n)})^{-1}\,R_1^{(n)} - I] < \infty \\ &\lim_{n\to\infty} \operatorname{tr}\,[R_0^{(n)}\,(R_1^{(n)})^{-1} - I] < \infty. \end{aligned} \tag{11}$$

Note that this condition implies (6), as it should, and excludes the case of perfect reception. In fact, condition (11) states not only that the sum of two traces converge as condition (6) requires, but also that the two traces converge individually. In conclusion: If condition (11) is satisfied, then there exist a positive constant $\beta$ and a random variable $\theta$ such that

$$\beta = \lim_{n\to\infty} |\,R_0^{(n)}\,(R_1^{(n)})^{-1}\,|, \tag{12}$$

$$\theta = \lim_{n\to\infty} \sum_{i=1}^{n} \sum_{j=1}^{n} [(R_0^{(n)})^{-1} - (R_1^{(n)})^{-1}]_{ij}\, x_{t_i} x_{t_j} \tag{13}$$

for almost all sample functions under both hypotheses; and the optimum decision scheme (10) is reduced to the following:

$$\begin{aligned} &\text{choose} \quad r_0(s,t) \quad \text{if} \quad \theta(x) < \log\,(1/\beta)[\alpha/(1-\alpha)]^2, \\ &\text{choose} \quad r_1(s,t) \quad \text{if} \quad \theta(x) \geqq \log\,(1/\beta)[\alpha/(1-\alpha)]^2, \end{aligned} \tag{14}$$

* This does not imply that two different decision schemes yield the same decision for every sample function; rather, a set $N$ of sample functions, for which two decisions differ, give no contribution to the error probability, i.e., $P_0(N) = 0 = P_1(N)$.

where $\theta(x)$ is the value of $\theta$ for the sample function $x(t)$, which is obtained by simply replacing $x_{t_i}$ and $x_{t_j}$ in (13) by $x(t_i)$ and $x(t_j)$.

Although the above decision scheme is certainly a step toward simplification compared with (10), it is still inconvenient, if not unfeasible, for physical application, since it requires limit operations for each received waveform. Yet, so long as the likelihood ratio is formed in terms of the sampled values, elimination of the limit operation appears to be impossible. Recall, however, the problem of optimum detection of sure signals in noise mentioned in Section 1.1. There, the likelihood ratio is formed in terms of the Fourier coefficients of the so-called Karhunen-Loéve expansion of the process instead, thus resulting in the decision scheme specified in terms of an *integral* in place of an *infinite series*, as shown by (3). Needless to say, in the present problem where there are two covariance functions instead of one, additional mathematical complications should be inevitable. Nevertheless, an optimum decision scheme which is essentially comparable to (3) can be obtained, as will now be shown.

Let $\lambda_1 \geqq \lambda_2 \geqq \cdots$ and $\psi_1(t), \psi_2(t), \cdots$ be the eigenvalues and the orthonormal eigenfunctions associated with the covariance function $r_0(s,t)$, and, similarly, let $\mu_1 \geqq \mu_2 \geqq \cdots$ and $\varphi_1(t), \varphi_2(t), \cdots$ be those associated with $r_1(s,t)$. Then, it can be shown that, under the assumption of $r_0(s,t)$ and $r_1(s,t)$ being continuous and positive-definite, the integrals

$$\xi_i = \int_0^1 x_t \psi_i(t)\, dt, \qquad i = 1, 2, \cdots, \tag{15}$$

exist for almost all sample functions under both hypotheses, and are Gaussian distributed with means zero. Furthermore, the covariance matrix determining the joint distribution of $\xi_1, \cdots, \xi_n$ is given by either

$$\begin{aligned} (Q_0^{(n)})_{ij} &= \lambda_i \delta_{ij}, \quad \text{or} \\ (Q_1^{(n)})_{ij} &= a_{ij} = \sum_{k=1}^{\infty} \mu_k u_{ki} u_{kj}, \qquad u_{ij} = \int_0^1 \varphi_i(t)\psi_j(t)\, dt, \end{aligned} \tag{16}$$

depending upon which one of $r_0(s,t)$ and $r_1(s,t)$ is the true covariance function of the process.

Thus the likelihood ratio of $\xi_1, \cdots, \xi_n$ becomes

$$\hat{l}_n = | Q_0^{(n)} (Q_1^{(n)})^{-1} |^{\frac{1}{2}} \exp \left\{ \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} [(Q_0^{(n)})^{-1} - (Q_1^{(n)})^{-1}]_{ij} \xi_i \xi_j \right\}, \tag{17}$$

which corresponds to (4). It turns out that, under the previous assumption on the covariance functions, there is a complete parallel between the two formulations, one based upon $x_{t_1}, \cdots, x_{t_n}$ and the other based upon $\xi_1, \cdots, \xi_n$. Thus, for almost all sample functions,

($i$) if (and only if)

$$\lim_{n\to\infty} \operatorname{tr} \left[(Q_0^{(n)})^{-1} Q_1^{(n)} - 2I - Q_0^{(n)} (Q_1^{(n)})^{-1}\right] < \infty, \tag{18}$$

then

$$\lim_{n\to\infty} \hat{l}_n(\xi_1, \cdots, \xi_n) = f(\omega) \quad \text{under both hypotheses;} \tag{19}$$

($ii$) if (and only if) (18) is not satisfied, then

$$\lim_{n\to\infty} \hat{l}_n(\xi_1, \cdots, \xi_n) = \begin{cases} 0 & \text{under null hypothesis,} \\ \infty & \text{under alternative hypothesis.} \end{cases} \tag{20}$$

Then, the optimum decision scheme corresponding to (10) becomes:

$$\begin{aligned} &\text{choose} \quad r_0(s,t) \quad \text{if} \quad \lim_{n\to\infty} \hat{l}_n[\xi_1(x), \cdots, \xi_n(x)] < \alpha/(1-\alpha), \\ &\text{choose} \quad r_1(s,t) \quad \text{if} \quad \lim_{n\to\infty} \hat{l}_n[\xi_1(x), \cdots, \xi_n(x)] \geqq \alpha/(1-\alpha), \end{aligned} \tag{21}$$

where $\xi_i(x)$, $i = 1, \cdots, n$, are the values of the random variables $\xi_i$ for the sample function $x(t)$, namely,

$$\xi_i(x) = \int_0^1 x(t)\psi_i(t)\, dt.$$

Again, note first the role of $\{\xi_i\}$, which is not a representing sequence of the process but a means for obtaining the crucial random variable $f$ and set $H$ by forming the likelihood ratio. Secondly, it can be shown that, under the assumption of the two covariance functions being continuous and positive-definite, $\{\varphi_i(t)\}$ can be used in place of $\{\psi_i(t)\}$ to form $\{\xi_i\}$, but *not* any orthonormal basis of $\mathfrak{L}_2$. Lastly, as before, negation of (18) can be interpreted as a necessary and sufficient condition for perfect reception. Completing the parallel, if (and only if)

$$\begin{aligned} &\lim_{n\to\infty} \operatorname{tr} \left[(Q_0^{(n)})^{-1} Q_1^{(n)} - I\right] < \infty, \\ &\lim_{n\to\infty} \operatorname{tr} \left[Q_0^{(n)} (Q_1^{(n)})^{-1} - I\right] < \infty, \end{aligned} \tag{22}$$

then there exist $\hat{\beta}$ and $\hat{\theta}$ such that

$$\hat{\beta} = \lim_{n\to\infty} | Q_0^{(n)} (Q_1^{(n)})^{-1} |, \tag{23}$$

$$\hat{\theta} = \lim_{n\to\infty} \sum_{i=1}^{n} \sum_{j=1}^{n} [(Q_0^{(n)})^{-1} - (Q_1^{(n)})^{-1}]_{ij}\, \xi_i \xi_j \tag{24}$$

for almost all sample functions under both hypotheses; and decision scheme (21) is reduced to the following:

$$\begin{aligned} &\text{choose} \quad r_0(s,t) \quad \text{if} \quad \hat{\theta}(x) < \log (1/\hat{\beta})[\alpha/(1-\alpha)]^2, \\ &\text{choose} \quad r_1(s,t) \quad \text{if} \quad \hat{\theta}(x) \geqq \log (1/\hat{\beta})[\alpha/(1-\alpha)]^2. \end{aligned} \tag{25}$$

Returning to the original goal of eliminating the limit operation, examination of (24) immediately suggests the possibility of rewriting $\hat{\theta}$ as a quadratic form in $x_t$. That is, if one defines

$$h^{(n)}(s,t) = \sum_{i=1}^{n} \sum_{j=1}^{n} h_{ij}^{(n)} \psi_i(s) \psi_j(t), \tag{26}$$

where

$$h_{ij}^{(n)} = [(Q_0^{(n)})^{-1} - (Q_1^{(n)})^{-1}]_{ij},$$

then, from (15),

$$\hat{\theta} = \lim_{n\to\infty} \int_0^1 \int_0^1 x_s h^{(n)}(s,t) x_t \, ds\, dt, \tag{27}$$

and $h_{ij}^{(n)}$; $i,j = 1, \cdots, n$, can be given as a solution of the matrix equation

$$Q_0^{(n)}\ (h_{ij}^{(n)})\ Q_1^{(n)} = Q_1^{(n)} - Q_0^{(n)},$$

or, more directly, $h^{(n)}(s,t)$ can be given as a solution of the integral equation

$$\int_0^1 \int_0^1 r_0^{(n)}(s,u) h^{(n)}(u,v) r_1^{(n)}(v,t)\, du\, dv = r_1^{(n)}(s,t) - r_0^{(n)}(s,t), \tag{28}$$

where

$$r_0^{(n)}(s,t) = \sum_{i=1}^{n} \lambda_i \psi_i(s) \psi_i(t), \quad r_1^{(n)}(s,t) = \sum_{i=1}^{n} \sum_{j=1}^{n} a_{ij} \psi_i(s) \psi_j(t). \tag{29}$$

Then, the following conjecture should be imminent:

$$\hat{\theta} = \int_0^1 \int_0^1 x_s h(s,t) x_t \, ds\, dt, \tag{30}$$

where $h(s,t)$ is a solution of

$$\int_0^1 \int_0^1 r_0(s,u)h(u,v)r_1(v,t)\,du\,dv = r_1(s,t) - r_0(s,t), \tag{31}$$

which are formally the limits of (27) and (28) respectively. The essential part of the above conjecture can be shown to be correct. That is, if (31) has a solution $h(s,t)$ such that

$$\begin{gathered}\int_0^1 \int_0^1 h^2(s,t)\,ds\,dt < \infty, \quad \text{then} \\ \int_0^1 \int_0^1 x_s h(s,t) x_t\,ds\,dt = \lim_{n\to\infty} \sum_{i=1}^{n} \sum_{j=1}^{n} h_{ij}^{(n)} \xi_i \xi_j\end{gathered} \tag{32}$$

for almost all sample functions under both hypotheses, provided that, for all $i, j = 1, 2, \cdots$,

$$a_{ii} < 1, \qquad a_{ii} > \sum_{j=1}^{\infty}{}' \mid a_{ij} \mid, \qquad \frac{\left| \dfrac{a_{ij}}{\lambda_i} - \delta_{ij} \right|}{1 - \sum_{k=1}^{\infty} \mid \delta_{jk} - a_{jk} \mid} \leqq K, \tag{33}$$

where $K$ is a positive constant independent of $i$ and $j$.
Then the optimum decision scheme (25) is immediately reduced to the following desired form:

$$\begin{aligned}&\text{choose} \quad r_0(s,t) \quad \text{if} \quad \int_0^1 \int_0^1 x(s)h(s,t)x(t)\,ds\,dt < \log \frac{1}{\hat{\beta}} \left( \frac{\alpha}{1-\alpha} \right)^2, \\ &\text{choose} \quad r_1(s,t) \quad \text{if} \quad \int_0^1 \int_0^1 x(s)h(s,t)x(t)\,ds\,dt \geqq \log \frac{1}{\hat{\beta}} \left( \frac{\alpha}{1-\alpha} \right)^2.\end{aligned} \tag{34}$$

Difficulty of the proof lies mainly in the fact that, as $n$ increases, the coefficients $h_{ij}^{(n)}$ themselves vary with $n$ as well as the number of the terms of the sum, yet $h^{(n)}(s,t)$ must approach $h(s,t)$ in such a way that

$$\lim_{n\to\infty} \int_0^1 \int_0^1 x_s h^{(n)}(s,t) x_t\,ds\,dt = \int_0^1 \int_0^1 x_s h(s,t) x_t\,ds\,dt$$

for almost all sample functions under both hypotheses. This accounts for need of the auxiliary conditions (33). The first condition is not a restriction in physical application since

$$\sum_{i=1}^{\infty} a_{ii} = \sum_{j=1}^{\infty} \mu_i = \int_0^1 r_1(t,t)\,dt$$

is the average energy of the waveform in the interval $0 \leqq t \leqq 1$, which can always be normalized to assure $a_{ii} < 1$. Although the remaining two conditions are restrictive, current knowledge of infinite systems of equations does not seem to allow their removal. Thus this calls for a future investigation of the degree of restriction imposed by them in physicial application.

As anticipated, there is an apparent correspondence between the classical case of sure signals in noise mentioned in Section 1.1 and the present case of stochastic signals, namely, between (2)–(3) and (31)–(34), except for the fact that the constituent functions in the latter case are functions of two variables instead of one. As the integral of the decision scheme (3) has a simple physical interpretation (the output of a linear filter with $g(t)$ as its impulse response), so does the integral in (34). Namely, it is the output of a quadratic filter whose impulse response is $h(s,t)$. The advantage of this scheme over the others — namely, (10, (14), (21) and (25) — is obvious. Given two covariance functions, the impulse response of the filter is uniquely determined by the integral equation (31) if a solution exists, and decision is made by comparing, with a preassigned threshold, the appropriately sampled output of the filter with the received waveform as its input, instead of having to perform the limit operation for each received waveform.

Finally, it should be remarked that the optimum decision scheme above differs *formally* from those previously obtained by others.* A further, and more significant, distinction lies in the assured optimality of this scheme, inherent in its derivation, while the optimality of the others has yet to be proved separately.†

## II. MATHEMATICAL THEORY

### 2.1 *Gaussian Processes*

Let $\{x_t, t \in T\}$ be a real Gaussian process with a parameter set $T = [0,1]$ and a finite dimensional distribution function $F_{t_1, \cdots, t_n}$, which is determined by given mean function and covariance function where $t_1, \cdots, t_n$ are an arbitrary finite subset of $T$. It is assumed that the mean function is identically zero on $T$ while the covariance function is positive-definite and continuous on $T \times T$. In the present problem it is desirable to have an explicit representation of the given process $\{x_t, t \in T\}$ on a function space.‡

Let $\Omega$ be a space of real-valued functions of $t \in T$. Let $x_s(\omega)$ be the

* Although their work is briefly reviewed in Section 1.2, their decision schemes are not stated explicitly in this paper.

† This excludes Parzen's[8] case where the decision scheme is essentially (10).

‡ The next paragraph follows closely Example 2.3 in Supplement, Doob,[9] pp. 609–610.

$\omega$ function with the value $\zeta(s)$ if $\omega$ is the function $\zeta(\cdot)$, so that $x_s(\omega) = \zeta(s)$. If the $t$ function $\omega$ has values $\zeta(t_1), \cdots, \zeta(t_n)$ at $t_1, \cdots, t_n$, the condition

$$\zeta(t_1) \leqq \rho_1, \cdots, \zeta(t_n) \leqq \rho_n$$

defines an $\omega$ set, which is denoted by

$$\{x_{t_i}(\omega) \leqq \rho_i, \qquad i = 1, \cdots, n\} \tag{35}$$

where $\rho_1, \cdots, \rho_n$ are arbitrary real numbers. Next, let $\mathfrak{F}$ be the class of all $\omega$ sets obtained in this way for arbitrary $n$, $t_1, \cdots, t_n$, and let $\mathcal{B}_T$ be the Borel field generated by $\mathfrak{F}$, and lastly let $P$ be a probability measure defined on the sets of $\mathcal{B}_T$ whose value is given by

$$P\{x_{t_i}(\omega) \leqq \rho_i, i = 1, \cdots, n\} = F_{t_1,\cdots,t_n}(\rho_1, \cdots, \rho_n). \tag{36}$$

Then, $\{x_t(\omega), t \in T\}$ is a representation of the given process $\{x_t, t \in T\}$ on the function space $\Omega$, and $(\Omega, \mathcal{B}_T, P)$ is the explicit probability measure space for the representation.*

(Remark) By virtue of the choice of representation space, the general elements of the space $\Omega$ coincide with the general sample functions of the process $\{x_t(\omega), t \in T\}$. Thus, the phrases, "almost everywhere (or almost surely)" and "for almost all sample functions," have the same meaning.

The assumption of continuous covariance function has the following significant consequences:

(*i*) $\{x_t(\omega), t \in T\}$ has an equivalent (with respect to $P$) separable and measurable process on the same $\omega$ space.† Hence, so long as the almost-everywhere valid properties of a given process are of interest, as in the case of this paper, the given process may as well be taken to be separable and measurable. Therefore, the Gaussian process $\{x_t(\omega), t \in T\}$ is henceforth assumed to be separable and measurable.

(*ii*) $\{x_t(\omega), t \in T\}$ is sample (Lebesgue) square-integrable on $T$ almost everywhere with respect to $P$.‡

This immediately implies that a Lebesgue integral

* Symbolic distinction between the given process and its representation on the function space is made by explicitly writing the argument $\omega$ for the latter.

† Note continuity of the covariance function of a process is equivalent to continuity in quadratic mean of the process (Loéve,[10] p. 470), and hence it implies continuity in probability of the process. Then, according to Theorem 2.6 in Doob,[9] pp. 61–62, there exists an equivalent separable and measurable process on the same space.

‡ See Loéve,[10] pp. 520–521.

$$\xi(\omega) = \int_T x_t(\omega)\psi(t)\,dt$$

exists almost everywhere, in which $\psi(t)$ is any continuous function on $T$. Furthermore, since the sample Lebesgue integral of a process coincides almost everywhere with the Riemann integral in quadratic mean criterion,* and also the Riemann integral in quadratic mean criterion of a Gaussian process is a Gaussian (random) variable,† $\xi(\omega)$ is a Gaussian variable.

### 2.2 *Formulation of Problem*

Let $F_{0;t_1,\cdots,t_n}$ and $F_{1;t_1,\cdots,t_n}$ be two alternative Gaussian finite dimensional distribution functions of a real separable and measurable process $\{x_t(\omega), t \in T\}$, whose mean functions are identically zero and whose covariance functions, denoted respectively by $r_0(s,t)$ and $r_1(s,t)$, are positive-definite and continuous on $T \times T$. Let $P_0$ and $P_1$ be the Gaussian probability measures defined respectively by $F_{0;t_1,\cdots,t_n}$ and $F_{1;t_1,\cdots,t_n}$ on the Borel field $\mathcal{B}_T$ of subsets of $\Omega$ as defined previously. It is well known that $P_0$ and $P_1$ are either equivalent, $P_0 \equiv P_1$, or orthogonal, $P_0 \perp P_1$.‡

Define a set function $P_e$ by

$$P_e(\Lambda) = \alpha P_0(\Lambda) + (1 - \alpha)P_1(\Omega - \Lambda), \qquad \Lambda \in \mathcal{B}_T, \quad (37)$$

where $\alpha$ is a constant, $0 < \alpha < 1$.§ Let $\Lambda_\alpha \in \mathcal{B}_T$ be such a set that

$$P_e(\Lambda_\alpha) \leqq P_e(\Lambda) \qquad \text{for all } \Lambda \in \mathcal{B}_T. \quad (38)$$

Then, the problem of interest is to specify such a set $\Lambda_\alpha$ in terms of $x_t(\omega)$.‖

Now, if $P_0 \equiv P_1$, let $f(\omega)$ be a Radon-Nikodym derivative of $P_1$ with respect to $P_0$; while, if $P_0 \perp P_1$, let $H \in \mathcal{B}_T$ be a set such that $P_0(H) = 0$ and $P_1(H) = 1$. Then, it can be shown that the following

* Henceforth, the "sample Lebesgue integral of a process" will simply be called the "integral of a process," unless otherwise specified. A definition of Riemann integral in quadratic mean criterion is in Loéve,[10] pp. 471–474.

† See Loéve,[10] p. 485.

‡ See Hajek.[11,12]

§ $P_e$ is the so-called error probability. Although $0 \leqq P_e \leqq 1$ for all $\Lambda \in \mathcal{B}_T$, $P_e$ is not a probability measure, and its full meaning is given in Section 1.1.

‖ Equivalence between this problem and that of "optimum reception of binary Gaussian processes" is discussed in detail in Section I.

set $S_\alpha$ satisfies condition (38):*

$$\begin{aligned} &\text{if } P_0 \equiv P_1, \qquad S_\alpha = \{f(\omega) \geqq \alpha/(1-\alpha)\}, \\ &\text{if } P_0 \perp P_1, \qquad S_\alpha = H. \end{aligned} \tag{39}$$

Thus the above stated problem is reduced to that of finding

($i$), if $P_0 \equiv P_1$, a function of $x_t(\omega)$ equal to $f(\omega)$ almost everywhere with respect to $P_0$ and $P_1$, and

($ii$), if $P_0 \perp P_1$, some such set $H$ expressible in terms of $x_t(\omega)$.

## 2.3 *Solutions — I*

### 2.3.1 *General Solution*

Let $\{\tau_k\}$ be a sequence of points in $T = [0,1]$, which is dense in $T$. Let $\mathcal{B}_n$ be a Borel field generated by a class of $\omega$ sets of the form

$$\{x_{\tau_i}(\omega) \leqq \rho_i, \qquad i = 1, \cdots, n\}, \tag{40}$$

and let $\mathcal{B}_\infty$ be the minimal Borel field containing $\bigcup_{n=1}^{\infty} \mathcal{B}_n$. Obviously,

$$\mathcal{B}_1 \subset \mathcal{B}_2 \subset \cdots \subset \mathcal{B}_\infty \subset \mathcal{B}_T. \tag{41}$$

Then, since $\{x_t(\omega), t \in T\}$ is a separable process, continuous in probability (with respect to $P_0$ and $P_1$), and the sequence $\{\tau_k\}$ is dense in $T$, it follows that, for an arbitrary set $\Lambda \in \mathcal{B}_T$, there exists a set $\Lambda' \in \mathcal{B}_\infty$ such that

$$P_0(\Lambda\Delta\Lambda') = 0 = P_1(\Lambda\Delta\Lambda'). \tag{42}$$†

Now, through the use of the covariance functions $r_0(s,t)$ and $r_1(s,t)$ and the fact that the mean functions are identically zero, the density functions $p_0$ and $p_1$ of the random variables $x_{\tau_i}(\omega)$, $i = 1, \cdots, n$, corresponding to $P_0$ and $P_1$ respectively, are obtained as follows:

$$\begin{aligned} p_m(\nu_1, \cdots, \nu_n) &= (2\pi)^{-(n/2)} \mid R_m^{(n)} \mid^{-\frac{1}{2}} \\ &\times \exp\left\{-\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} [(R_m^{(n)})^{-1}]_{ij}\,\nu_i\nu_j\right\}, \qquad m = 0, 1, \end{aligned} \tag{43}$$

where the $\tau_i$, $i = 1, \cdots, n$, are a finite subset of $\{\tau_k\}$, and $R_m^{(n)}$, $m = 0$, 1, are $n \times n$ symmetric, positive-definite matrices defined by

$$(R_m^{(n)})_{ij} = r_m(\tau_i, \tau_j); \qquad m = 0, 1; \qquad i,j = 1, \cdots, n. \tag{44}$$

* See Appendix A. The first assertion of (39) follows from Corollary 1 in this appendix, while the second assertion is self-evident.

† See Doob,[9] pp. 51–55; in particular, Theorem 2.2 ($i$).

Then define a random variable $l_n(\omega)$ by

$$l_n(\omega) = \frac{p_1[x_{\tau_1}(\omega), \cdots, x_{\tau_n}(\omega)]}{p_0[x_{\tau_1}(\omega), \cdots, x_{\tau_n}(\omega)]} = | R_0^{(n)}(R_1^{(n)})^{-1} |^{\frac{1}{2}} \times \exp\left\{\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} [(R_0^{(n)})^{-1} - (R_1^{(n)})^{-1}]_{ij} x_{\tau_i}(\omega) x_{\tau_j}(\omega)\right\}. \tag{45}$$

Note that

$$l_n(\omega) \geqq 0 \quad \text{for all } n. \tag{46}$$

Furthermore, since $R_m^{(n)}$, $m = 0, 1$, are positive-definite, $p_1 = 0$ whenever $p_0 = 0$ and vice versa. Then, it can be shown that the processes $\{l_n(\omega), n \geqq 1\}$ and $\{1/l_n(\omega), n \geqq 1\}$ are martingales with respect to $P_0$ and $P_1$ respectively.*

(*i*) $P_0 \equiv P_1$: Let $E_0\{f(\omega) | \mathcal{B}_n\}$, $n = 1, 2, \cdots$, be a conditional expectation of $f(\omega)$, given $\mathcal{B}_n$, with respect to $P_0$. Namely,

$$\int_\Lambda E_0\{f(\omega) \mid \mathcal{B}_n\}\, dP_0 = \int_\Lambda f(\omega)\, dP_0 \quad \text{for any } \Lambda \in \mathcal{B}_n .$$

Then,

$$l_n(\omega) = E_0\{f(\omega) | \mathcal{B}_n\}, \qquad \text{a.e. } (P_0),\dagger \tag{47}$$

and, from (41)‡

$$\lim_{n\to\infty} E_0\{f(\omega) | \mathcal{B}_n\} = E_0\{f(\omega) | \mathcal{B}_\infty\}, \qquad \text{a.e. } (P_0). \tag{48}$$

Yet, from the definition of $E_0\{f(\omega) | \mathcal{B}_\infty\}$ and (42),

$$E_0\{f(\omega) | \mathcal{B}_\infty\} = f(\omega), \qquad \text{a.e. } (P_0). \tag{49}$$

Hence,

$$\lim_{n\to\infty} l_n(\omega) = f(\omega), \qquad \text{a.e. } (P_0). \tag{50}$$

Since $P_0 \equiv P_1$, the above implies

$$\lim_{n\to\infty} l_n(\omega) = f(\omega), \qquad \text{a.e. } (P_1). \tag{51}$$

Thus, the desired function, which is equal to $f(\omega)$, a.e. $(P_0, P_1)$, is

* See Doob,[9] pp. 91–93.

† "a.e. $(P_m)$," $m = 0, 1$, is used as a shorthand notation of "almost everywhere with respect to $P_m$." Similarly, "a.e. $(P_0, P_1)$" will be used to denote "almost everywhere with respect to both $P_0$ and $P_1$."

‡ See Doob,[9] p. 331.

$l_\infty(\omega)$, which is defined by

$$l_\infty(\omega) = \lim_{n\to\infty} l_n(\omega). \tag{52}$$

(*ii*) $P_0 \perp P_1$ : From (46), $\lim_{n\to\infty} l_n(\omega) < \infty$, a.e. $(P_0)$.* In fact, it can be shown that

$$\lim_{n\to\infty} l_n(\omega) = 0, \qquad \text{a.e. } (P_0).\dagger \tag{53}$$

By using the same argument, it follows that

$$\lim_{n\to\infty} [1/l_n(\omega)] = 0, \qquad \text{a.e. } (P_1). \tag{54}$$

Hence, for an arbitrary constant $c > 0$,

$$P_0\{\lim_{n\to\infty} l_n(\omega) \geqq c\} = 0, \qquad P_1\{\lim_{n\to\infty} l_n(\omega) \geqq c\} = 1.$$

Thus, the desired set $H$, with $P_0(H) = 0$ and $P_1(H) = 1$, is

$$H = \{\lim_{n\to\infty} l_n(\omega) \geqq \alpha/(1 - \alpha)\}. \tag{55}$$

In summary, upon combination of (52) and (55) in conjunction with (39), the desired set $S_\alpha$ is

$$S_\alpha = \{\lim_{n\to\infty} l_n(\omega) \geqq \alpha/(1 - \alpha)\}, \tag{56}$$

irrespective of whether $P_0 \equiv P_1$ or $P_0 \perp P_1$.

### 2.3.2 *Special Solutions and Summary*

Under certain restrictive conditions, the set $S_\alpha$ can be specified in terms of well defined functions of $x_t(\omega)$. It is the purpose of this subsection to obtain such specifications as well as the accompanying conditions in terms of the given covariance functions $r_0(s,t)$ and $r_1(s,t)$.

(*i*) If $P_0 \equiv P_1$, it has already been shown that

$$S_\alpha = \{l_\infty(\omega) \geqq \alpha/(1 - \alpha)\}.$$

Thus, it is of interest to obtain a condition for $P_0 \equiv P_1$.‡

Define

$$\eta_n(\omega) = [l_n(\omega) - 1] \log l_n(\omega), \qquad n = 1, 2, \cdots. \tag{57}$$

* See Doob,[9] p. 319; Theorem 4.1 (*i*).

† See Doob,[9] pp. 345–346.

‡ Such conditions are already available (e.g., Parzen,[8] Shepp[13]). For more detail, see Yaglom.[14]

Then, since $(\rho - 1) \log \rho$, $\rho > 0$, is a real, continuous and convex function of $\rho$ and $E_0\{|(l_n(\omega) - 1) \log l_n(\omega)|\} < \infty$, $n = 1, 2, \cdots$; $\{\eta_n(\omega), n \geqq 1\}$ constitute a semi-martingale (with respect to $P_0$).* Hence, $E_0\{\eta_n(\omega)\}$, $n = 1, 2, \cdots$, forms a monotone nondecreasing sequence

$$E_0\{\eta_1(\omega)\} \leqq E_0\{\eta_2(\omega)\} \leqq \cdots, \qquad (58)†$$

which must either converge or diverge. Then, according to (53),

if $P_0 \perp P_1$, then

$$\lim_{n\to\infty} E_0\{\eta_n(\omega)\} = \infty. \qquad (59)$$

Hence, since $P_0$ and $P_1$ can be either equivalent or orthogonal, it follows that

$P_0 \equiv P_1$, if

$$\lim_{n\to\infty} E_0\{\eta_n(\omega)\} < \infty. \qquad (60)$$

It can be shown that the converse of (59) is also true,‡ i.e.,

$$\text{if} \quad \lim_{n\to\infty} E_0\{\eta_n(\omega)\} = \infty, \qquad \text{then } P_0 \perp P_1. \qquad (61)$$

This implies that the condition of (60) is also necessary. Thus, through substitution of (45) into (57) and application of (43) for expectation calculation,§

$P_0 \equiv P_1$, if and only if

$$\lim_{n\to\infty} \text{tr}\, [(R_0^{(n)})^{-1} R_1^{(n)} - 2I + R_0^{(n)} (R_1^{(n)})^{-1}] < \infty. \qquad (62)\|$$

where $R_0^{(n)}$ and $R_1^{(n)}$ are defined in terms of $r_0(s,t)$ and $r_1(s,t)$ by (44).

(*ii*) Examination of (45), (50) and (51) indicates that, in addition to condition (62), if

$$\lim_{n\to\infty} | R_0^{(n)} (R_1^{(n)})^{-1} | = \beta, \qquad 0 < \beta < \infty, \qquad (63)$$

then

* See Doob,[9] pp. 295–296, Theorem 1.1 (iii). "$E_0$" denotes expectation with respect to $P_0$, namely, an integration over $\Omega$ with respect to $P_0$.

† See Doob,[9] p. 324, Theorem 4.1s.

‡ See Hajek;[12] in particular, Lemma 2.1.

§ For this calculation, use the following equality: $E_0\{\eta_n(\omega)\} = E_1\{\log l_n(\omega)\} - E_0\{\log l_n(\omega)\}$, $n = 1, 2, \cdots$.

‖ "tr" denotes "trace," and $I$ is the $n \times n$ identity matrix.

$$\lim_{n\to\infty} \sum_{i=1}^{n} \sum_{j=1}^{n} [(R_0^{(n)})^{-1} - (R_1^{(n)})^{-1}]_{ij}\, x_{\tau_i}(\omega) x_{\tau_j}(\omega) < \infty, \qquad \text{a.e. } (P_0, P_1). \tag{64}$$

Thus, by defining $\theta(\omega)$ as the above limit, i.e.,

$$\theta(\omega) = \lim_{n\to\infty} \sum_{i=1}^{n} \sum_{j=1}^{n} [(R_0^{(n)})^{-1} - (R_1^{(n)})^{-1}]_{ij} x_{\tau_i}(\omega) x_{\tau_j}(\omega), \tag{65}$$

the set $S_\alpha \in \mathcal{B}_T$ can be specified as follows:

$$S_\alpha = \{\theta(\omega) \geqq \log (1/\beta)(\alpha/(1-\alpha))^2\}. \tag{66}$$

It will now be shown that two conditions (62) and (63), required for the above specification of $S_\alpha$, are equivalent to the following pair of conditions:

$$\lim_{n\to\infty} \operatorname{tr} [(R_0^{(n)})^{-1} R_1^{(n)} - I] < \infty,$$

and (67)

$$\lim_{n\to\infty} \operatorname{tr} [R_0^{(n)} (R_1^{(n)})^{-1} - I] < \infty.$$

Define

$$\begin{aligned} \zeta_n(\omega) &= -\log l_n(\omega), \\ \zeta_n'(\omega) &= l_n(\omega) \log l_n(\omega) \qquad n = 1, 2, \cdots . \end{aligned} \tag{68}$$

Thus,

$$\eta_n(\omega) = \zeta_n'(\omega) + \zeta_n(\omega), \qquad n = 1, 2, \cdots . \tag{69}$$

Again, just as in the case of $\eta_n(\omega)$, both $\{\zeta_n(\omega), n \geqq 1\}$ and $\{\zeta_n'(\omega), n \geqq 1\}$ are semi-martingales with respect to $P_0$, and

$$\begin{aligned} E_0\{\zeta_1(\omega)\} &\leqq E_0\{\zeta_2(\omega)\} \leqq \cdots, \\ E_0\{\zeta_1'(\omega)\} &\leqq E_0\{\zeta_2'(\omega)\} \leqq \cdots . \end{aligned} \tag{70}$$

Furthermore, from (53),

$$\text{if} \quad P_0 \perp P_1, \qquad \text{then} \quad \lim_{n\to\infty} E_0\{\zeta_n(\omega)\} = \infty. \tag{71}$$

However, from (69) and (70), divergence of $E_0\{\zeta_n(\omega)\}$ implies that of $E_0\{\eta_n(\omega)\}$. Hence, according to (61), the converse of (71) holds. Then, again from (70) and the equivalence-or-orthogonality dichotomy of $P_0$ and $P_1$,

$$P_0 \equiv P_1 \quad \text{if and only if} \quad \lim_{n\to\infty} E_0\{\zeta_n(\omega)\} < \infty. \tag{72}$$

Thus, upon substitution of (45) into (68) and application of (43) for expectation calculation, an alternative necessary and sufficient condition for $P_0 \equiv P_1$ is obtained as follows:

$$\lim_{n\to\infty} \{\log |(R_0^{(n)})^{-1} R_1^{(n)}| + \operatorname{tr} [R_0^{(n)} (R_1^{(n)})^{-1} - I]\} < \infty. \tag{73}$$

Now, under the condition (63), the above condition implies that

$$\lim_{n\to\infty} \operatorname{tr} [R_0^{(n)} (R_1^{(n)})^{-1} - I] < \infty. \tag{74}$$

Then, upon combination of conditions (62) and (74), condition (67) immediately follows.

The result of this section may be summarized as follows:

(*i*) In general,

$$S_\alpha = \{\lim_{n\to\infty} l_n(\omega) \geqq \alpha/(1-\alpha)\},$$

where $l_n(\omega)$ is defined by (45).

(*ii*) If $P_0 \equiv P_1$, which is true if and only if

$$\lim_{n\to\infty} \operatorname{tr} [(R_0^{(n)})^{-1} R_1^{(n)} - 2I + R_0^{(n)} (R_1^{(n)})^{-1}] < \infty,$$

then $\lim_{n\to\infty} l_n(\omega) = f(\omega)$, a.e. $(P_0, P_1)$; thus by defining $l_\infty(\omega) = \lim_{n\to\infty} l_n(\omega)$,

$$S_\alpha = \{l_\infty(\omega) \geqq \alpha/(1-\alpha)\}.$$

(*iii*) if

$$\lim_{n\to\infty} \operatorname{tr} [(R_0^{(n)})^{-1} R_1^{(n)} - I] < \infty,$$

$$\lim_{n\to\infty} \operatorname{tr} [R_0^{(n)} (R_1^{(n)})^{-1} - I] < \infty, \quad .$$

then

$$S_\alpha = \{\theta(\omega) \geqq \log (1/\beta)(\alpha/(1-\alpha))^2\}$$

where $\theta(\omega)$ and $\beta$ are defined by (65) and (63) respectively.

### 2.4 *Solutions — II*

#### 2.4.1 *General and Special Solutions*

Let $\lambda_1 \geqq \lambda_2 \geqq \cdots$ and $\psi_1(t), \psi_2(t), \cdots$ be the eigenvalues and the corresponding orthonormal eigenfunctions associated with the covariance function $r_0(s,t)$.* Similarly, let $\mu_1 \geqq \mu_2 \geqq \cdots$ and $\varphi_1(t), \varphi_2(t), \cdots$ be such eigenvalues and eigenfunctions associated with $r_1(s,t)$. Then, according to the discussion in 2.1 (*ii*), continuity on $T$ of each $\psi_i(t)$ implies that the integrals

$$\xi_i(\omega) = \int_T x_t(\omega)\psi_i(t)\,dt, \qquad i = 1, 2, \cdots, \tag{75}$$

exist a.e. $(P_0, P_1)$, and are Gaussian random variables. In fact, it can be shown that the density functions $\hat{p}_0$ and $\hat{p}_1$ of $\xi_1(\omega), \cdots, \xi_n(\omega)$ corresponding to $P_0$ and $P_1$ are given by†

$$\begin{aligned}\hat{p}_m(\nu_1, \cdots, \nu_n) = {} & (2\pi)^{-(n/2)} \mid Q_m^{(n)} \mid^{-\frac{1}{2}} \\ & \exp\left\{-\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}[(Q_m^{(n)})^{-1}]_{ij}\,\nu_i\nu_j\right\}, \qquad m = 0, 1,\end{aligned} \tag{76}$$

where $Q_m^{(n)}$, $m = 0, 1$, are $n \times n$ symmetric and positive-definite matrices defined by

$$(Q_0^{(n)})_{ij} = \lambda_i\delta_{ij}, \qquad (Q_1^{(n)})_{ij} = \sum_{k=1}^{\infty}\mu_k u_{ki} u_{kj}, \tag{77}$$

where

$$u_{ij} = \int_T \varphi_i(t)\psi_j(t)\,dt. \tag{78}$$

Let $\hat{\mathfrak{B}}_n$ be a Borel field generated by a class of $\omega$ sets of the form

$$\{\xi_i(\omega) \leqq \rho_i, i = 1, \cdots, n\}, \tag{79}$$

and let $\hat{\mathfrak{B}}_\infty$ be the minimal Borel field containing $\bigcup_{n=1}^{\infty} \hat{\mathfrak{B}}_n$. Obviously,

$$\hat{\mathfrak{B}}_1 \subset \hat{\mathfrak{B}}_2 \subset \cdots \subset \hat{\mathfrak{B}}_\infty \subset \hat{\mathfrak{B}}_T. \tag{80}$$

It can be shown that, for an arbitrary $\Lambda \in \mathfrak{B}_T$, there exists some $\hat{\Lambda} \in \hat{\mathfrak{B}}_\infty$ such that

$$P_0(\Lambda\Delta\hat{\Lambda}) = 0. \tag{81}‡$$

Now define a random variable $\hat{l}_n(\omega)$ by

* More precise definitions of these eigenvalues and eigenfunctions are given in Appendix B.

† See Appendix C.

‡ See Appendix D.

$$\begin{aligned}\hat{l}_n(\omega) &= \frac{\hat{p}_1[\xi_1(\omega), \cdots, \xi_n(\omega)]}{\hat{p}_0[\xi_1(\omega), \cdots, \xi_n(\omega)]} \\ &= | Q_0^{(n)} (Q_1^{(n)})^{-1} |^{\frac{1}{2}} \exp \left\{ \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} [(Q_0^{(n)})^{-1} \right. \\ &\quad \left. - (Q_1^{(n)})^{-1}]_{ij} \xi_i(\omega) \xi_j(\omega) \right\}, \end{aligned} \tag{82}$$

where (76) is substituted for the second equality. Again, note that $\hat{l}_n(\omega)$ is nonnegative for all $n$ and also the fact that $\hat{p}_1 = 0$ whenever $\hat{p}_0 = 0$ and vice versa since $Q_m^{(n)}$, $m = 0, 1$, are positive-definite. Thus, again the processes $\{\hat{l}_n(\omega), n \geqq 1\}$ and $\{1/\hat{l}_n(\omega), n \geqq 1\}$ are martingales with respect to $P_0$ and $P_1$ respectively.

By following step-by-step the same procedure as the one in the preceding section,* the following results are obtained:

(*i*) In general,

$$S_\alpha = \{\lim_{n \to \infty} \hat{l}_n(\omega) \geqq \alpha/(1 - \alpha)\}. \tag{83}$$

(*ii*) If $P_0 \equiv P_1$, which is true if and only if

$$\lim_{n \to \infty} \operatorname{tr} [(Q_0^{(n)})^{-1} Q_1^{(n)} - 2I + Q_0^{(n)} (Q_1^{(n)})^{-1}] < \infty, \tag{84}$$

then

$$\lim_{n \to \infty} \hat{l}_n(\omega) = f(\omega), \qquad \text{a.e. } (P_0, P_1); \tag{85}$$

thus by defining

$$\hat{l}_\infty(\omega) = \lim_{n \to \infty} \hat{l}_n(\omega), \tag{86}$$

$$S_\alpha = \{\hat{l}_\infty(\omega) \geqq \alpha/(1 - \alpha)\}. \tag{87}$$

(*iii*) If

$$\begin{aligned} &\lim_{n \to \infty} \operatorname{tr} [(Q_0^{(n)})^{-1} Q_1^{(n)} - I] < \infty, \\ &\lim_{n \to \infty} \operatorname{tr} [Q_0^{(n)} (Q_1^{(n)})^{-1} - I] < \infty, \end{aligned} \tag{88}$$

then there exists a constant $\hat{\beta}$, $0 < \hat{\beta} < \infty$, such that

* In effect, it amounts to replacing $\mathfrak{B}_n$ and $l_n(\omega)$, $n = 1, 2, \cdots n$, by $\hat{\mathfrak{B}}$ and $\hat{l}_n(\omega)$ respectively.

$$\lim_{n\to\infty} | Q_0^{(n)} (Q_1^{(n)})^{-1} | = \hat{\beta}; \tag{89}$$

and, from (85) and (82), it follows that

$$\lim_{n\to\infty} \sum_{i=1}^{n} \sum_{j=1}^{n} [(Q_0^{(n)})^{-1} - Q(_1{}^{(n)})^{-1}]_{ij}\xi_i(\omega)\xi_j(\omega) < \infty, \quad \text{a.e. } (P_0 , P_1); \tag{90}$$

thus, by defining $\hat{\theta}(\omega)$ as the above limit,

$$S_\alpha = \{\hat{\theta}(\omega) \geqq \log (1/\hat{\beta})(\alpha/(1 - \alpha))^2\}. \tag{91}$$

### 2.4.2 *Integral Expression for $\hat{\theta}(\omega)$*

For the purpose of physical application, it is desirable to express the random variable $\hat{\theta}(\omega)$ as a simpler function of $x_t(\omega)$, in particular, without involving limit operation. Examination of the definition of $\hat{\theta}(\omega)$, i.e.,

$$\hat{\theta}(\omega) = \lim_{n\to\infty} \sum_{i=1}^{n} \sum_{j=1}^{n} [(Q_0^{(n)})^{-1} - (Q_1^{(n)})^{-1}]_{ij}\xi_i(\omega)\xi_j(\omega), \tag{92}$$

indicates that $\hat{\theta}(\omega)$ might be expressible as a quadratic form in $x_t(\omega)$, i.e.,

$$\int_T\int_T x_s(\omega)h(s,t)x_t(\omega)\, ds\, dt$$

if such a square-integrable function $h(s,t)$, $(s,t) \in T \times T$, exists and can be determined uniquely. It is the purpose of this subsection to make the above statement more definite and precise.

Define an $n \times n$ symmetric matrix $H^{(n)}$ by

$$H^{(n)} = (Q_0^{(n)})^{-1} - (Q_1^{(n)})^{-1}.$$

Then,

$$Q_0^{(n)}H^{(n)}Q_1^{(n)} = Q_1^{(n)} - Q_0^{(n)},$$

or, through (77), the equation for the $i$-$j$th element becomes

$$\sum_{k=1}^{n} \lambda_i(H^{(n)})_{ik}(Q_1^{(n)})_{kj} = (Q_1^{(n)}{}_{ij} - \lambda_i\delta_{ij}; \qquad i, j = 1, \cdots , n.$$

In other words, every $i$th row of $H^{(n)}$ satisfies the following system of equations:*

$$\sum_{k=1}^{n} a_{jk}h_{ik}^{(n)} = b_j(i) \qquad j = 1, \cdots , n,$$

* Note that the solution is unique, since the matrix $(a_{ij})$ is positive-definite.

where

$$\begin{aligned} a_{ij} &= \sum_{k=1}^{\infty} \mu_k u_{ki} u_{kj}, \qquad i, j = 1, 2, \cdots, \\ b_j(i) &= (a_{ij}/\lambda_i) - \delta_{ij}, \end{aligned} \tag{93}*$$

or its standard form

$$h_{ij}^{(n)} = \sum_{k=1}^{n} c_{jk} h_{ik}^{(n)} + b_j(i), \qquad j = 1, \cdots, n, \tag{94}$$

where

$$c_{ij} = \delta_{ij} - a_{ij}.$$

Now, for each $i = 1, 2, \cdots$, consider the following infinite system of equations:

$$h_{ij} = \sum_{k=1}^{\infty} c_{jk} h_{ik} + b_j(i), \qquad j = 1, 2, \cdots. \tag{95}$$

According to the theory of infinite systems of equations,† if (95) has a solution $(h_{i1}, h_{i2}, \cdots)$ for each $i = 1, 2, \cdots$, such that

$$\sum_{i=1}^{\infty} \sum_{j=1}^{\infty} h_{ij}^2 < \infty, \tag{96}$$

then $(h_{i1}, h_{i2}, \cdots)$ is unique and

$$h_{ij} = \lim_{n \to \infty} h_{ij}^{(n)}, \qquad j = 1, 2, \cdots, \tag{97}$$

for each $i = 1, 2, \cdots$, where $(h_{i1}^{(n)}, \cdots, h_{in}^{(n)})$, $i = 1, \cdots, n$, is the solution of (94); provided that (95) satisfies the following conditions: for each $i = 1, 2, \cdots$,

$$\sum_{j=1}^{\infty} | c_{ij} | < 1, \tag{98}$$

and there exists a constant $K_i > 0$, independent of $j$, such that

$$| b_j(i) | \leqq K_i \left( 1 - \sum_{k=1}^{\infty} | c_{jk} | \right), \qquad j = 1, 2, \cdots. \tag{99}$$

On the other hand, if $(h_{i1}, h_{i2}, \cdots)$ is a solution of (95) for each $i = 1, 2, \cdots$, satisfying (96), then the following integral equation

$$\int_T \int_T r_0(s,u) h(u,v) r_1(v,t) \, du \, dv = r_1(s,t) - r_0(s,t) \tag{100}$$

* Note that $(Q_1^{(n)})_{ij} = a_{ij}$; $i, j = 1, \cdots, n$.

† See Kantorovich and Krylov,[15] pp. 20–33.

has a square-integrable solution $h(s,t)$,

$$\int_T \int_T h^2(s,t)\, ds\, dt < \infty, \tag{101}$$

such that

$$h(s,t) = \sum_{i=1}^{\infty} \sum_{j=1}^{\infty} h_{ij} \psi_i(s) \psi_j(t), \qquad \text{in the mean.}^* \tag{102}$$

Conversely, if $\tilde{h}(s,t)$ is a square-integrable solution of (100), then (95) has a unique solution $(\tilde{h}i_1, \tilde{h}_{i2}, \cdots)$ for each $i = 1, 2, \cdots$, satisfying $\sum_{i=1}^{\infty} \sum_{j=1}^{\infty} \tilde{h}_{ij}^2 < \infty$, such that

$$\tilde{h}_{ij} = \int_T \int_T \psi_i(s) \tilde{h}(s,t) \psi_j(t)\, ds\, dt.\dagger \tag{103}$$

Now, extend the definition of $h_{ij}^{(n)}$, $i = 1, 2, \cdots, n$,‡ by adding

$$h_{ij}^{(n)} = 0; \qquad i, j = n + 1, n + 2, \cdots. \tag{104}$$

Then, (90) and (92) can be rewritten as

$$\lim_{n \to \infty} \sum_{i=1}^{\infty} \sum_{j=1}^{\infty} h_{ij}^{(n)}\, \xi_i(\omega) \xi_j(\omega) < \infty, \qquad \text{a.e. } (P_0, P_1), \tag{105}$$

and

$$\hat{\theta}(\omega) = \lim_{n = \infty} \sum_{i=1}^{\infty} \sum_{j=1}^{\infty} h_{ij}^{(n)}\, \xi_i(\omega) \xi_j(\omega). \tag{106}$$

According to the theory of coordinate and projective limits in sequence spaces,§ (97) and (105) imply that

$$\hat{\theta}(\omega) = \sum_{i=1}^{\infty} \sum_{j=1}^{\infty} h_{ij} \xi_i(\omega) \xi_j(\omega), \qquad \text{a.e. } (P_0, P_1), \tag{107}$$

since

$$\sum_{i=1}^{\infty} \sum_{j=1}^{\infty} \xi_i^2(\omega) \xi_j^2(\omega) < \infty, \qquad \text{a.e. } (P_0, P_1). \tag{108}$$

On the other hand, from (102) and square-integrability, a.e. $(P_0, P_1)$,

* See Appendix E.1.

† See Appendix E.2.

‡ Namely, $(h_{i1}, \cdots, h_{in})$ is the solution of (94) for each $i = 1, \cdots, n$.

§ See Cooke,[16] pp. 282–289; in particular, Theorem (10.3, II), extended to the case of double sequences.

of $\{x_t(\omega),\ t \in T\}$,

$$\int_T\int_T x_s(\omega)h(s,t)x_t(\omega)\ ds\ dt = \sum_{i=1}^{\infty}\sum_{j=1}^{\infty} h_{ij}\xi_i(\omega)\xi_j(\omega), \qquad \text{a.e. } (P_0, P_1). \tag{109}$$

Hence,

$$\hat{\theta}(\omega) = \int_T\int_T x_j(\omega)h(s,t)x_t(\omega)\ ds\ dt, \qquad \text{a.e. } (P_0, P_1). \tag{110}$$

### 2.4.3 *Discussion and Summary*

Recall that, in order to specify the set $S_\alpha \in \mathcal{B}_T$ for given $\alpha$ as (91), it is sufficient to assume (88), which assures existence of $\hat{\theta}(\omega)$ and $\hat{\beta}$ defined by (89) and (92) respectively. Moreover, in order to express $\hat{\theta}(\omega)$ as (110), it requires the additional assumptions that (*i*) the integral equation (100) have a square-integrable solution and (*ii*) the conditions (98) and (99) be satisfied.

It can be shown, however, under the assumptions (*i*) and the following:

$$a_{ii} < 1, \qquad i = 1, 2, \cdots, \tag{111}$$

the conditions (*ii*) and (88) can be replaced by the following:

$$a_{ii} > \sum_{j=1}^{\infty}{}' \mid a_{ij} \mid, \qquad i = 1, 2, \cdots, \tag*{(112)*}$$

and that there exists a constant $K > 0$, independent of $i, j = 1, 2, \cdots$, such that

$$\mid (a_{ij}/\lambda_i) - \delta_{ij} \mid \leqq K(a_{ij} - \sum_{k=1}^{\infty}{}' \mid a_{jk} \mid), \tag{113}$$

where $a_{ij}$ is defined by (93).† It is quite possible that, once the condition (*i*) is assumed, the conditions (111), (112) and (113) may be superfluous. That is to say, in some special cases, if the integral equation (100) admits a square-integrable solution $h(s,t)$ it may be possible to prove directly that

$$h(s,t) = \lim_{n\to\infty} \sum_{i=1}^{\infty}\sum_{j=1}^{\infty} h_{ij}{}^{(n)}\psi_i(s)\psi_j(t), \tag{114}$$

in the mean, which immediately implies (97) and (105), thus establish-

* The prime on the summation sign symbolizes omission of the term $j = i$.

† See Appendix D.

ing (107) and leading to (110). However, in the general case, establishment of (114) does not seem possible, nor does finding a sufficient condition for (114), without making the resultant condition excessively implicit and complex.

### 2.5 *Summary*

If

$$\int_T\int_T r_0(s,u)h(u,v)r_1(v,t)\;du\;dv = r_1(s,t) - r_0(s,t)$$

has a solution $h(s,t)$,

$$\int_T\int_T h^2(s,t)\;ds\;dt < \infty,$$

then the set $S_\alpha \in \mathcal{B}_T$, given $\alpha$ $(0 < \alpha < 1)$, can be specified as

$$S_\alpha = \left\{\int_T\int_T x_s(\omega)h(s,t)x_t(\omega)\;ds\;dt \geqq \log\frac{1}{\hat{\beta}}\left(\frac{\alpha}{1-\alpha}\right)^2\right\}, \quad (115)$$

where

$$\hat{\beta} = \lim_{n\to\infty} |\, Q_0^{(n)}\,(Q_1^{(n)})^{-1}\,|,$$

and

$$(Q_0^{(n)})_{ij} = \lambda_i\delta_{ij}, \qquad (Q_1^{(n)})_{ij} = a_{ij}\,; \qquad i,j = 1, \cdots, n,$$

and

$$a_{ij} = \sum_{k=1}^{\infty} \mu_k u_{ki} u_{kj}\,; \qquad i,j = 1, 2, \cdots,$$

$$u_{ij} = \int_T \varphi_i(t)\psi_j(t)\;dt;$$

where $\lambda_1 \geqq \lambda_2 \geqq \cdots$; $\psi_1(t), \psi_2(t), \cdots$, and $\mu_1 \geqq \mu_2 \geqq \cdots$; $\varphi_1(t), \varphi_2(t), \cdots$, are the eigenvalues and the corresponding orthonormal eigenfunctions associated with the given covariance functions $r_0(s,t)$ and $r_1(s,t)$, which are positive-definite and continuous on $T \times T$; provided that

(1) $a_{ii} < 1,\ i = 1, 2, \cdots,$

(2) $a_{ii} > \sum_{j=1}^{\infty}{}' \,|\, a_{ij} \,|, \qquad i = 1, 2, \cdots,$

(3) the following is bounded uniformly in $i, j = 1, 2, \cdots$ :

$$\frac{\left| \dfrac{a_{ij}}{\lambda_i} - \delta_{ij} \right|}{1 - \sum_{k=1}^{\infty} | \delta_{jk} - a_{jk} |} \leqq K.$$

## III. ACKNOWLEDGMENTS

The author gratefully acknowledges his deep indebtedness to D. Slepian and L. A. Shepp, without whose encouragement and assistance this work could not have been completed. Acknowledgment should be made also to I. W. Sandberg, I. Jacobs and W. L. Nelson for their valuable suggestions.

## APPENDIX A

### *Theorem on Optimality*

Let $P_0$ and $P_1$ be probability measures defined on a Borel field $\mathcal{B}$ of subsets of an abstract space $\Omega$. Through the use of Lebesgue decomposition theorem and Radon-Nikodym theorem:* for a nonempty set $H \in \mathcal{B}$ with $P_0(H) = 0$, there exists a nonnegative function $f(\omega)$ integrable over $\Omega$ with respect to $P_0$ such that

$$P_1(\Lambda) = \int_\Lambda f(\omega) dP_0 + P_1(\Lambda \cap H) \tag{116}$$

for an arbitrary $\Lambda \in \mathcal{B}$.†

*Theorem: For an arbitrary constant $k > 0$, define a set $S \in \mathcal{B}$ by*

$$S = \{f(\omega) \geqq k\} \cup H. \tag{117}$$

*Then,*

$$kP_0(S) + P_1(S^c) - kP_0(\Lambda) - P_1(\Lambda^c) \leqq 0 \tag{118}$$

*for an arbitrary set $\Lambda \in \mathcal{B}$ where $S^c$ and $\Lambda^c$ are the complements of $S$ and $\Lambda$ with respect to $\Omega$.*

*Proof:*

Put $\rho = kP_0(S) + P_1(S^c) - kP_0(\Lambda) - P_1(\Lambda^c)$. By adding and subtracting $kP_0(S \cap \Lambda)$ and $P_1(S^c \cap \Lambda^c)$,

* See Loéve,[10] pp. 130–132.

† This paragraph closely parallels Grenander,[1] pp. 209–210.

$$\begin{aligned}\rho = {} & k[P_0(S) - P_0(S \cap \Lambda)] + P_1(S^c) - P_1(S^c \cap \Lambda^c) \\ & \quad - k[P_0(\Lambda) - P_0(S \cap \Lambda)] - P_1(\Lambda^c) + P_1(S^c \cap \Lambda^c) \\ = {} & kP_0(S \cap \Lambda^c) - P_1(S \cap \Lambda^c) + P_1(S^c \cap \Lambda) \\ & \quad - kP_0(S^c \cap \Lambda).\end{aligned} \tag{119}$$

From (116) and (117), with $k > 0$ and $P_0(H) = 0$,

$$\begin{aligned}P_1(S \cap \Lambda^c) - P_1(S^c \cap \Lambda) & \\ = {} & \int_{S\cap\Lambda^c} f(\omega)dP_0 - \int_{S^c\cap\Lambda} f(\omega)dP_0 \\ & + P_1(S \cap \Lambda^c \cap H) - P_1(S^c \cap \Lambda \cap H) \\ \geqq {} & kP_0(S \cap \Lambda^c) - kP_0(S^c \cap \Lambda),\end{aligned} \tag{120}$$

since

$$\begin{aligned}P_1(S^c \cap \Lambda \cap H) & \\ = {} & P_1(\{f(\omega) < k\} \cap H^c \cap \Lambda \cap H) \leqq P_1(H^c \cap H) = 0.\end{aligned}$$

By substituting (120) into (119),

$$\rho \leqq 0,$$

which proves (118). (Q.E.D.)

*Corollary 1. Suppose* $P_0 \equiv P_1$, *and let* $k = [\alpha/(1 - \alpha)]$, $0 < \alpha < 1$. *Then, a set* $S_\alpha$ *defined by*

$$S_\alpha = \{f(\omega) \geqq \alpha/(1 - \alpha)\} \tag{121}$$

*has the property expressed by (118), i.e.,*

$$\alpha P_0(S_\alpha) + (1 - \alpha)P_1({S_\alpha}^c) \leqq \alpha P_0(\Lambda) + (1 - \alpha)P_1(\Lambda^c) \tag{122}$$

*for an arbitrary* $\Lambda \in \mathcal{B}$.

*Proof:*

Note that $P_0 \equiv P_1$ implies $P_1(H) = 0$. Hence, in (118),

$$\begin{aligned}kP_0(S) + P_1(S^c) &= [\alpha/(1 - \alpha)]P_0(S_\alpha \cup H) + P_1(S_\alpha \cup H) \\ &= [\alpha/(1 - \alpha)]P_0(S_\alpha) + P_1(S_\alpha).\end{aligned}$$

Thus, substitution of the above into (118) and multiplication by $1 - \alpha$ proves (122).

*Corollary 2. Take* $\Omega$ *to be* $R_n$, *an n-dimensional Euclidean space, and* $\mathcal{B}$

*to be Borel field of right semi-closed, semi-infinite intervals in* $R_n$*, denoted by* $\mathfrak{F}_n$. *Let* $p_m(x_1, \cdots, x_n)$, $m = 0, 1$ *and* $(x_1, \cdots, x_n) \in R_n$*, be Baire density functions corresponding to* $P_m$, $m = 0, 1$; *i.e.*,

$$P_m\{x_i \leqq \rho_i, i = 1, \cdots, d\} \\ = \int_{-\infty}^{\rho_1} \cdots \int_{-\infty}^{\rho_n} dx_1 \cdots dx_n \, p_m(x_1, \cdots, x_n). \tag{123}$$

*Suppose* $p_1(x_1, \cdots, x_n) = 0$ *whenever* $p_0(x_1, \cdots, x_n) = 0$. *Then* $S_{\alpha,n}$ *defined by*

$$S_{\alpha,n} = \left\{ \frac{p_1(x_1, \cdots, x_n)}{p_0(x_1, \cdots, x_n)} \geqq \frac{\alpha}{1 - \alpha} \right\}$$

*has the property expressed by* (122).*

*Proof:*

Note that $P_0 \equiv P_1$, thus $P_1(H) = 0$. Then, from (116),

$$f(x_1, \cdots, x_n) = \frac{p_1(x_1, \cdots, x_n)}{p_0(x_1, \cdots, x_n)}, \qquad \text{a.e. } (P_0, P_1)$$

Hence, apply Corollary 1.

## APPENDIX B

### *Preliminaries on Integral Operators*†

Let $L$ be an integral operator with a real, symmetric, continuous and positive-definite kernel $r(s,t)$ defined on the rectangle $T \times T$ where $T$ is the closed interval [0,1]. That is,

$$Lf(t) \equiv \int_T r(s, t) f(s) \, ds, \tag{124}$$

where $f(t)$ is an arbitrary real-valued function in the space of square-integrable functions on $T$, which is symbolically denoted by $\mathfrak{L}_2(0,1)$, or simply by $\mathfrak{L}_2$.

Then, according to the theory of linear operators, all the eigenvalues of $L$ are positive, of finite multiplicity, and finite or denumerably infinite in number. Thus, counting each eigenvalue as many times as its multiplicity, we can construct an ordered sequence of eigenvalues,

* This replaces the Neyman-Pearson theorem in the classical theory of testing simple hypotheses when the criterion changes from the Neyman-Pearson's to the minimum error probability. See Cramér,[18] pp. 529–530.

† See Riesz-Nagy,[17] pp. 227–246.

$$\lambda_1 \geqq \lambda_2 \geqq \cdots , \tag{125}$$

and the corresponding sequence of orthonormal eigenfunctions (using the Gram-Schmidt orthonormalization process if necessary),

$$\psi_1(t), \psi_2(t), \cdots . \tag{126}$$

Then, according to Mercer's theorem,

$$r(s,t) = \sum_{i=1}^{\infty} \lambda_i \psi_i(s) \psi_i(t), \tag{127}$$

where the series converges uniformly on $T$. Consequently, $\psi_i(t)$ is continuous on $T$ for all $i$, and

$$\sum_{i=1}^{\infty} \lambda_i = \int_T r(t,t)\, dt < \infty , \tag{128}$$

namely, the sum of all eigenvalues is finite.

Furthermore, because of the positive definiteness of the kernel $r(s,t)$, the set of the eigenfunctions $\{\psi_i(t)\}$ forms an orthonormal basis of $\mathfrak{L}_2$. Let $\{\varphi_i(t)\}$ be another orthonormal basis of $\mathfrak{L}_2$. Then,

$$\psi_j(t) = \sum_{i=1}^{\infty} u_{ij} \varphi_i(t), \quad \text{in the mean,} \tag{129}$$

where

$$u_{ij} = \int_T \varphi_i(t) \psi_j(t)\, dt, \tag{130}$$

which satisfies the following orthogonality conditions:

$$\begin{aligned} \sum_{k=1}^{\infty} u_{ik} u_{jk} &= \int_T \varphi_i(t) \varphi_j(t)\, dt = \delta_{ij} , \\ \sum_{k=1}^{\infty} u_{ki} u_{kj} &= \int_T \psi_i(t) \psi_j(t)\, dt = \delta_{ij} . \end{aligned} \tag{131}$$

## APPENDIX C

### *Density Functions of $\xi_i(\omega)$, $i = 1, \cdots , n$*

It has been established in Section 2.4.1 that the random variables defined by (75), i.e.,

$$\xi_i(\omega) = \int_T x_t(\omega)\, \psi_i(t)\, dt, \qquad i = 1, 2, \cdots, \tag{132}$$

are Gaussian variables with respect to $P_0$ and $P_1$, where

$$E_m\{x_t(\omega)\} = 0, t \in T, m = 0, 1, \quad \text{and} \quad \int_T x_t^2(\omega)\, dt < \infty,$$

$$\text{a.e. } (P_0, P_1) .$$

C.1 *With respect to* $P_0$

Through repeated use of Fubini's theorem,

$$E_0\{\xi_i(\omega)\} = \int_T E_0\{x_t(\omega)\}\, \psi_i(t)\, dt = 0, \qquad i = 1, 2, \cdots, \tag{133}$$

and

$$\begin{aligned} E_0\{\xi_i(\omega)\ \xi_j(\omega)\} &= \int_T\int_T E_0\{x_s(\omega)x_t(\omega)\}\psi_i(s)\psi_j(t)\, ds\, dt \\ &= \int_T\int_T r_0(s,t)\psi_i(s)\ \psi_j(t)\ ds\, dt \\ &= \lambda_i\delta_{ij}\, ; \quad i,j = 1, 2, \cdots, \end{aligned} \tag{134}$$

where Mercer's theorem is used for the third equality. Then, since $\xi_i(\omega)$, $i = 1, \cdots, n$, are Gaussian variables, (133) and (134) immediately give (76) and (77) with $m = 0$.

C.2 *With respect to* $P_1$

By substituting (129) into (132),

$$\xi_i(\omega) = \sum_{k=1}^{\infty} u_{ki}\eta_k(\omega), \qquad \text{a.e. } (P_0, P_1) \tag{135}$$

where

$$\eta_i(\omega) = \int_T x_t(\omega)\varphi_i(t)\, dt, \qquad i = 1, 2, \cdots,\text{*} \tag{136}$$

which exist a.e. $(P_0, P_1)$, and Gaussian variables just as $\xi_i(\omega)$, $i = 1, 2, \cdots$, are. Then, the results in C.1 imply that

$$E_1\{\eta_i(\omega)\eta_j(\omega)\} = \mu_i\delta_{ij}\, ; \qquad i,j = 1, 2, \cdots . \tag{137}$$

Define

$$\xi_j^{(m)}(\omega) = \sum_{k=1}^{m} u_{kj}\eta_k(\omega), \qquad j = 1, \cdots, n, \tag{138}$$

* Note $\eta_i(\omega)$ here must not be confused with the one in Section 2.3.2.

and let $F_1^{(m)}$ be the distribution function of $\xi_1^{(m)}(\omega), \cdots, \xi_n^{(m)}(\omega)$, and let $f_1^{(m)}(\tau_1, \cdots, \tau_n)$, $-\infty < \tau_j < \infty$, $j = 1, \cdots, n$, be its characteristic function with respect to $P_1$, i.e.,

$$f_1^{(m)}(\tau_1, \cdots, \tau_n) = E_1\left\{\exp\left[i \sum_{j=1}^{n} \tau_j \xi_j^{(m)}(\omega)\right]\right\}. \tag{139}$$

Then, according to Levy's continuity theorem,* $\lim_{m\to\infty} F_1^{(m)}$ exists if and only if $\lim_{m\to\infty} f_1^{(m)}(\tau_1, \cdots, \tau_n)$ exists for every $\tau_j$, $-\infty < \tau_j < \infty$, and continuous at $\tau_j = 0$, $j = 1, \cdots, n$; and, furthermore, when $\lim_{m\to\infty} F_1^{(m)} = F_1$ exists, its characteristic function $f_1(\tau_1, \cdots, \tau_n)$ is equal to $\lim_{m\to\infty} f_1^{(m)}(\tau_1, \cdots, \tau_n)$ for all $\tau_j$, $-\infty < \tau_j < \infty$, $j = 1, \cdots, n$. Hence, it suffices to obtain $\lim_{m\to\infty} f_1^{(m)}(\tau_1, \cdots, \tau_n)$, namely, the limit of (139) as $m \to \infty$, and to assure its continuity at the origin.

By substituting (138) into (139),

$$\begin{aligned} f_1^{(m)}(\tau_1, \cdots, \tau_n) &= E_1\left\{\exp\left[i \sum_{j=1}^{n} \tau_j \sum_{k=1}^{m} u_{kj}\eta_k(\omega)\right]\right\} \\ &= E_1\left\{\exp\left[\sum_{k=1}^{m} i\eta_k(\omega) \sum_{j=1}^{n} \tau_j u_{kj}\right]\right. \\ &= \prod_{k=1}^{m} \exp\left[-\frac{1}{2}\mu_k\left(\sum_{j=1}^{n} \tau_j u_{kj}\right)^2\right] \\ &= \exp\left[-\frac{1}{2}\sum_{k=1}^{m}\sum_{i=1}^{n}\sum_{j=1}^{n} \tau_i\tau_j\mu_k u_{ki}u_{kj}\right]. \end{aligned}$$

Note that

$$\sum_{k=1}^{\infty} | \mu_k u_{ki} u_{kj} | = \sum_{k=1}^{\infty} \mu_k | u_{ki} u_{kj} | \leqq \sum_{k=1}^{\infty} \mu_k < \infty, \tag{138}$$

since

$$\begin{aligned} | u_{ki} u_{kj} | &= \left| \int \varphi_k(t)\psi_i(t)\, dt \right| \left| \int \varphi_k(t)\psi_j(t)\, dt \right| \\ &\leqq \int \varphi_k^2(t)\, dt \left[\int \psi_i^2(t)\, dt \int \psi_j^2(t)\, dt\right]^{\frac{1}{2}} \\ &= 1. \end{aligned}$$

[*] See Cramér,[18] p. 102.

Hence,

$$\sum_{k=1}^{\infty}\sum_{i=1}^{n}\sum_{j=1}^{n}\tau_i\tau_j\mu_k u_{ki}u_{kj} = \sum_{i=1}^{n}\sum_{j=1}^{n}\tau_i\tau_j\sum_{k=1}^{\infty}\mu_k u_{ki}u_{kj}.$$

Then, putting

$$(Q_1^{(n)})_{ij} = \sum_{k=1}^{\infty}\mu_k u_{ki}u_{kj}; \qquad i, j = 1, \cdots, n,$$

continuity of exponential functions implies

$$\lim_{m\to\infty} f^{(m)}(\tau_1, \cdots, \tau_n) = \exp\left[-\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}(Q_1^{(n)})_{ij}\tau_i\tau_j\right],$$

which is obviously continuous at $\tau_i = 0$, $i = 1, \cdots, n$. Note that the right-hand side above is the characteristic function of the Gaussian distribution function with the density function (76) and (77) with $m = 1$.

## APPENDIX D

### $P_0$ — *Equivalence between* $\mathcal{B}_T$ *and* $\hat{\mathcal{B}}_\infty$

It is to be proved that, for an arbitrary set $\Lambda \in \mathcal{B}_T$, there exists a nonempty set $\hat{\Lambda} \in \hat{\mathcal{B}}_\infty$ such that $P_0(\Lambda\Delta\hat{\Lambda}) = 0$. Note, however, that the above statement is equivalent to the following:

> Let $\mathfrak{F}_T \subset \mathcal{B}_T$ be a class of all sets $\Lambda \in \mathcal{B}_T$ such that $\Lambda \in \mathfrak{F}_T$ implies existence of a nonempty set $\hat{\Lambda} \in \hat{\mathcal{B}}_\infty$ with $P_0(\Lambda\Delta\hat{\Lambda}) = 0$. Then, $\mathfrak{F}_T = \mathcal{B}_T$.*

The second statement will be proved.

D.1 For every $t \in T$,

$$x_t(\omega) = \sum_{k=1}^{\infty}\xi_k(\omega)\psi_k(t), \qquad \text{a.e.}(P_0). \tag{139}$$

*Proof:*

According to the discussion in Section 2.1, (*ii*), $\xi_k(\omega)$, $k = 1, 2, \cdots$, are equal, a.e. $(P_0)$, to the Riemann integrals in quadratic mean criterion of $x_t(\omega)\psi_k(t)$ on $T$. Hence, from the proper orthogonal decomposi-

* It must be proved first that such an $\mathfrak{F}_T$ is not empty. This will be done in Section D.2.

tion theorem,* the series of (139) converges in quadratic mean with respect to $P_0$ to $x_t(\omega)$ uniformly on $T$. Furthermore, $\xi_k(\omega)$, $k = 1, 2, \cdots$, are mutually independent Gaussian variables with means zero and variances $\lambda_k$, $k = 1, 2, \cdots$, with respect to $P_0$.† Hence, the series converges, a.e. ($P_0$), to a limit for every $t \in T$ since the series of its variances converges for every $t \in T$, i.e.,

$$\sum_{k=1}^{\infty} E_0\{\xi_k^2(\omega)\}\psi_k^2(t) = \sum_{k=1}^{\infty} \lambda_k\psi_k(t)\psi_k(t) = r_0(t,t) < \infty,$$

from Mercer's theorem. Yet, since both the convergence in quadratic mean and the convergence almost everywhere imply the convergence in probability measure, this limit must be equal, a.e. ($P_0$), to $x_t(\omega)$ for every $t \in T$. (Q.E.D.)

D.2 Let $\Lambda_T \in \mathcal{B}_T$ be defined by

$$\Lambda_T = \{x_{t_i}(\omega) \leqq \rho_i\,,\ i = 1, \cdots, n\}. \tag{140}$$

Then there exists a nonempty set $\hat{\Lambda}_T \in \hat{\mathcal{B}}_\infty$ such that

$$P_0(\Lambda_T \Delta \hat{\Lambda}_T) = 0.$$

*Proof:*

Consider a set defined by

$$\hat{\Lambda}_T = \left\{\sum_{k=1}^{\infty} \xi_k(\omega)\psi_k(t_i) \leqq \rho_i\,,\ i = 1, \cdots, n\right\}. \tag{141}$$

Clearly, $\hat{\Lambda}_T \in \hat{\mathcal{B}}_\infty$. Define $\Gamma_t \in \mathcal{B}_T$ by

$$\Gamma_t = \left\{x_t(\omega) = \sum_{k=1}^{\infty} \xi_k(\omega)\psi_k(t)\right\}, \qquad t \in T. \tag{142}$$

Note that (139) implies

$$P_0(\Gamma_t) = 1, \qquad t \in T. \tag{143}$$

Then it is self-evident that, for $t_i \in T$, $i = 1, \cdots, n$,

$$\begin{aligned} \Lambda_T &= \Lambda_T \cap \left(\bigcap_{i=1}^{n} \Gamma_{t_i}\right) + \Lambda_T \cap \left(\bigcup_{i=1}^{n} \Gamma_{t_i}^{\,c}\right), \\ \hat{\Lambda}_T &= \hat{\Lambda}_T \cap \left(\bigcap_{i=1}^{n} \Gamma_{t_i}\right) + \hat{\Lambda}_T \cap \left(\bigcup_{i=1}^{n} \Gamma_{t_i}^{\,c}\right), \end{aligned} \tag{144}$$

where $\Gamma_{t_i}^{\,c}$ is the complement of $\Gamma_{t_i}$. Note that, from (142),

* See Loéve,[10] pp. 478–479.

† See Appendix C.1.

$$\Lambda_T \cap \left( \bigcap_{i=1}^{n} \Gamma_{t_i} \right) = \hat{\Lambda}_T \cap \left( \bigcap_{i=1}^{n} \Gamma_{t_i} \right), \tag{145}$$

and, from (143),

$$P_0\left[ \Lambda_T \cap \left( \bigcup_{i=1}^{n} \Gamma_{t_i}{}^c \right) \right] = 0 = P_0\left[ \hat{\Lambda}_T \cap \left( \bigcup_{i=1}^{n} \Gamma_{t_i}{}^c \right) \right]. \tag{146}$$

Hence, upon combination of 144, 145, and (146),

$$P_0(\Lambda_T \Delta \hat{\Lambda}_T) = 0. \qquad \text{(Q.E.D.)}$$

D.3 $\mathfrak{F}_T = \mathcal{B}_T$.

*Proof:*

First, it is easily seen that the class $\mathfrak{F}_T$ is a field. Moreover, it will now be shown that $\mathfrak{F}_T$ is a Borel field. Let $\Lambda_i \in \mathfrak{F}_T$, $i = 1, 2, \cdots$. Then, from the definition of $\mathfrak{F}_T$, there exists $\hat{\Lambda}_i \in \hat{\mathcal{B}}_\infty$ such that

$$P_0(\Lambda_i \Delta \hat{\Lambda}_i) = 0, \qquad i = 1, 2, \cdots. \tag{147}$$

Define two sequences of null sets $M_i$ and $N_i$, $i = 1, 2, \cdots$, by

$$M_i = \Lambda_i - \hat{\Lambda}_i, \qquad N_i = \hat{\Lambda}_i - \Lambda_i. \tag{148}$$

Then,

$$\hat{\Lambda}_i - N_i \subset \Lambda_i \subset \hat{\Lambda}_i \cup M_i, \qquad i = 1, 2, \cdots.$$

Hence,

$$\bigcup_{i=1}^{\infty} \hat{\Lambda}_i - \bigcup_{i=1}^{\infty} N_i \subset \bigcup_{i=1}^{\infty} \Lambda_i \subset \left( \bigcup_{i=1}^{\infty} \hat{\Lambda}_i \right) \cup \left( \bigcup_{i=1}^{\infty} M_i \right),$$

which implies

$$\bigcup_{i=1}^{\infty} \hat{\Lambda}_i - \bigcup_{i=1}^{\infty} \Lambda_i \subset \bigcup_{i=1}^{\infty} N_i,$$

$$\bigcup_{i=1}^{\infty} \Lambda_i - \bigcup_{i=1}^{\infty} \hat{\Lambda}_i \subset \bigcup_{i=1}^{\infty} M_i.$$

Thus,

$$P_0\left[ \left( \bigcup_{i=1}^{\infty} \Lambda_i \right) \Delta \left( \bigcup_{i=1}^{\infty} \hat{\Lambda}_i \right) \right] = 0,$$

namely,

$$\bigcup_{i=1}^{\infty} \Lambda_i \in \mathfrak{F}_T.$$

Furthermore, since

$$\bigcap_{i=1}^{\infty} \Lambda_i = \left( \bigcup_{i=1}^{\infty} \Lambda_i^c \right)^c$$

and also $\mathfrak{F}_T$ is a field,

$$\bigcap_{i=1}^{\infty} \Lambda_i \in \mathfrak{F}_T .$$

Hence, $\mathfrak{F}_T$ is a Borel field.

Secondly, note that $\mathfrak{F}_T$ contains the generating class of $\mathcal{B}_T$ as shown by (140) and (35). Hence,

$$\mathfrak{F}_T \supset \mathcal{B}_T .$$

Yet, from the definition of $\mathfrak{F}_T$,

$$\mathfrak{F}_T \subset \mathcal{B}_T .$$

Therefore,

$$\mathfrak{F}_T = \mathcal{B}_T . \qquad \text{(Q.E.D.)}$$

## APPENDIX E

### *Equivalence between Two Equations*

### E.1 *Preliminary*

Through Mercer's theorem,

$$r_0(s,t) = \sum_{i=1}^{\infty} \lambda_i \psi_i(s) \psi_i(t) , \tag{149}$$

uniformly.

$$r_1(s,t) = \sum_{k=1}^{\infty} \mu_k \varphi_k(s) \varphi_k(t), \tag{150}$$

Then,

$$\int_T \int_T r_1(s,t) \psi_i(s) \psi_j(t) \, ds \, dt = \sum_{k=1}^{\infty} \mu_k u_{ki} u_{kj} = a_{ij} .$$

Hence,

$$r_1(s,t) = \sum_{i=1}^{\infty} \sum_{j=1}^{\infty} a_{ij} \psi_i(s) \psi_j(t), \qquad \text{in the mean.} \qquad (151)^*$$

* This is a trivial extension of well-known results in the case of functions of one variable. A special case of (151) is found in Courant and Hilbert,[19] pp. 73–74.

### E.2 *Equivalence between Two Equations*

Equation (95) can be rewritten as

$$\sum_{k=1}^{\infty} \lambda_i h_{ik} a_{kj} = a_{ij} - \lambda_i \delta_{ij}, \qquad j = 1, 2, \cdots, \tag{152}$$

where $i = 1, 2, \cdots$. Repeating (100),

$$\int_T \int_T r_0(s,u) h(u,v) r_1(v,t)\, du\, dv = r_1(s,t) - r_0(s,t). \tag{153}$$

(a) If $(h_{i1}, h_{i2}, \cdots)$ is a solution of (152) for each $i = 1, 2, \cdots$, with

$$\sum_{i=1}^{\infty} \sum_{j=1}^{\infty} h_{ij}^2 < \infty, \tag{154}$$

then a square-integrable function $h(s,t)$ with

$$h(s,t) = \sum_{i=1}^{\infty} \sum_{j=1}^{\infty} h_{ij} \psi_i(s) \psi_j(t), \qquad \text{in the mean}, \tag{155}$$

satisfies (153).

*Proof:*

The left-hand side of (153) is clearly square-integrable. Hence, it has the following expansion:

$$\int_T \int_T r_0(s,u) h(u,v) r_1(v,t)\, du\, dv = \sum_{i=1}^{\infty} \sum_{j=1}^{\infty} \sum_{k=1}^{\infty} \lambda_i h_{ik} a_{kj} \psi_i(s) \psi_j(t), \tag{156}$$
$$\text{in the mean},$$

since, through substitution of (149), (150), and (155),

$$\int_T \int_T \left[ \int_T \int_T r_0(s,u) h(u,v) r_1(v,t)\, du\, dv \right] \psi_i(s) \psi_j(t)\, ds\, dt$$
$$= \sum_{k=1}^{\infty} \lambda_i h_{ik} a_{kj}; \qquad i,j = 1, 2, \cdots. \tag{157}$$

Yet, by virtue of $(h_{i1}, h_{i2}, \cdots)$ being a solution of (152) for each $i = 1, 2, \cdots$, the right-hand side of (156) becomes

$$\sum_{i=1}^{\infty} \sum_{j=1}^{\infty} \sum_{k=1}^{\infty} \lambda_i h_{ik} a_{kj} \psi_i(s) \psi_j(t) = \sum_{i=1}^{\infty} \sum_{j=1}^{\infty} (a_{ij} - \lambda_i \delta_{ij}) \psi_i(s) \psi_j(t), \tag{158}$$

the right-hand side of which in turn becomes, from (149) and (150),

$$r_1(s,t) - r_0(s,t)$$

$$= \sum_{i=1}^{\infty} \sum_{j=1}^{\infty} (a_{ij} - \lambda_i \mathcal{E}_{ij})\psi_i(s)\psi_j(t), \qquad \text{in the mean.} \quad (159)$$

Thus, upon combination of (156), (158), and (159),*

$$\int_T \int_T r_0(s,u)h(u,v)r_1(v,t)\, du\, dv = r_1(s,t) - r_0(s,t). \quad (160)$$

(Q.E.D.)

(b) If $h(s,t)$ is a square-integrable solution of (153), then $(h_{i1}, h_{i2}, \cdots)$ satisfies (152) for each $i = 1, 2, \cdots$, where

$$h_{ij} = \int_T \int_T h(s,t)\psi_i(s)\psi_j(t)\, ds\, dt. \quad (161)$$

*Proof:*

Since $h(s,t)$ is square-integrable, it has the expansion of (155) where $h_{ij}$; $i,j = 1, 2, \cdots$, are defined by (161); thus (157) is established. Meanwhile, from (149) and (150),

$$\int_T \int_T [r_1(s,t) - r_0(s,t)]\psi_i(s)\psi_j(t) = a_{ij} - \lambda_i\delta_{ij}; \quad i,j = 1, 2, \cdots. \quad (162)$$

Then, combination of (160), (157) and (162) establishes

$$\sum_{k=1}^{\infty} \lambda_i h_{ik} a_{kj} = a_{ij} - \lambda_i \delta_{ij}. \qquad \text{(Q.E.D.)}$$

## APPENDIX F

### *Alternative Conditions*

Assume

$$a_{ii} < 1, \qquad i = 1, 2, \cdots, \quad (163)$$

and the integral equation (100)

$$\int_T \int_T r_0(s,u)h(u,v)r_1(v,t)\, du\, dv = r_1(s,t) - r_0(s,t) \quad (164)$$

has a square-integrable solution.† Then, the conditions that

* Note that, if a sequence of functions converges in the mean to two limits, the limits are equal almost everywhere. Furthermore, if the limits are continuous, they are equal everywhere. Note also that continuity of the left-hand side of (156) can easily be seen through the use of the Schwartz inequality.

† Recall from Appendix E that this implies $\sum_{i=1} \sum_{j=1} h_{ij}^2 < \infty$.

$$a_{ii} > \sum_{j=1}^{\infty}{}' \mid a_{ij} \mid , \tag{165)*}$$

and there exists a constant $K > 0$, independent of $i,j = 1, 2, \cdots$, such that

$$\mid (a_{ij}/\lambda_i) - \delta_{ij} \mid \leqq K \left( a_{jj} - \sum_{k=1}^{\infty}{}' \mid a_{jk} \mid \right), \tag{166}$$

imply the conditions (98), (99) and (88); namely, for each $i = 1, 2, \cdots$,

$$\sum_{j=1}^{\infty} \mid c_{ij} \mid < 1, \tag{167}$$

and there exists a constant $K_i > 0$ such that

$$b_j(i) \leqq K_i \left( 1 - \sum_{k=1}^{\infty} \mid c_{jk} \mid \right), \qquad j = 1, 2, \cdots , \tag{168}$$

and finally

$$\begin{gathered} \lim_{n \to \infty} \operatorname{tr} [(Q_0^{(n)})^{-1} Q_1^{(n)} - I] < \infty, \\ \lim_{n \to \infty} \operatorname{tr} [Q_0^{(n)} (Q_1^{(n)})^{-1} - I] < \infty. \end{gathered} \tag{169}$$

*Proof:*

First, note that

$$a_{ii} > 0, \qquad i = 1, 2, \cdots , \tag{170}$$

and

$$\sum_{i=1}^{\infty} a_{ii} = \sum_{i=1}^{\infty} \mu_i . \tag{171}$$

For, from (93) and the fact that $\mu_k > 0$, $k = 1, 2, \cdots$,

$$a_{ii} = \sum_{k=1}^{\infty} \mu_k u_{ki}^2 > 0,$$

and, from (131),

$$\sum_{i=1}^{\infty} a_{ii} = \sum_{i=1}^{\infty} \sum_{k=1}^{\infty} \mu_k u_{ki}^2 = \sum_{k=1}^{\infty} \mu_k \sum_{i=1}^{\infty} u_{ki}^2 = \sum_{k=1}^{\infty} \mu_k . \dagger$$

[*] The prime symbolizes omission of the term $j = i$ in the summation.

[†] For justification of interchange of order of summation, see Apostol,[20] pp. 374–375.

Second, through (93) and (94) with (163), (165) and (166),

$$\begin{aligned}1 - \sum_{j=1}^{\infty} | c_{ij} | &= 1 - \sum_{j=1}^{\infty} | \delta_{ij} - a_{ij} | \\ &= 1 - | 1 - a_{ii} | - \sum_{j=1}^{\infty}{}' | a_{ij} | \\ &= a_{ii} - \sum_{j=1}^{\infty}{}' | a_{ij} | > 0,\end{aligned}$$

and

$$\frac{b_j(i)}{1 - \sum_{k=1}^{\infty} | c_{jk} |} = \frac{| (a_{ij}/\lambda_i) - \delta_{ij} |}{a_{jj} - \sum_{k=1}^{\infty}{}' | a_{jk} |} \leqq K; \qquad i,j = 1, 2, \cdots ,$$

which prove (167) and (168).

Last, note from the definition of $h_{ij}^{(n)}$, $i,j = 1, 2, \cdots$,*

$$\operatorname{tr} [(Q_0^{(n)})^{-1}Q_1^{(n)} - I] = \sum_{i=1}^{n} \sum_{j=1}^{n} h_{ij}^{(n)} a_{ij} = \sum_{i=1}^{\infty} \sum_{j=1}^{\infty} h_{ij}^{(n)} a_{ij},$$

$$- \operatorname{tr} [Q_0^{(n)} (Q_1^{(n)})^{-1} - I] = \sum_{i=1}^{\infty} \sum_{j=1}^{\infty} h_{ij}^{(n)} \lambda_i \delta_{ij} = \sum_{i=1}^{\infty} h_{ii}^{(n)} \lambda_i .$$

Yet, according to the theory of infinite systems of equations, for each $i = 1, 2, \cdots$,

$$| h_{ij}^{(n)} | \leqq K_i , \qquad j = 1, 2, \cdots .\dagger$$

By putting $K_i = K$, $i = 1, 2, \cdots$,

$$| h_{ij}^{(n)} | \leqq K; \qquad i,j = 1, 2, \cdots .$$

Then,

$$\sum_{i=1}^{\infty} \sum_{j=1}^{\infty} | h_{ij}^{(n)} a_{ij} | \leqq K \sum_{i=1}^{\infty} \sum_{j=1}^{\infty} | a_{ij} | < \infty,$$

$$\sum_{i=1}^{\infty} | h_{ij}^{(n)} \lambda_i | \leqq K \sum_{i=1}^{\infty} \lambda_i < \infty,$$

since

* Recall:

$$(Q_0^{(n)})^{-1}Q_1^{(n)} - I = [(Q_0^{(n)})^{-1} - (Q_1^{(n)})^{-1}]Q_1^{(n)},$$

$$h_{ij}^{(n)} = \begin{cases}[(Q_0^{(n)})^{-1} - (Q_1^{(n)})^{-1}]_{ij} ; & i,j = 1, \cdots , n, \\ 0; & i,j = n + 1, n + 2, \cdots .\end{cases}$$

† See Kantorovich and Krylov,[15] pp. 26–27.

$$\sum_{i=1}^{\infty}\sum_{j=1}^{\infty} | a_{ij} | = \sum_{i=1}^{\infty}\left(a_{ii} + \sum_{j=1}^{\infty}{}' \, | a_{ij} |\right) < 2\sum_{i=1}^{\infty} a_{ii} = 2\sum_{i=1}^{\infty} \mu_i < \infty .$$

Hence, from (97),

$$\begin{aligned}
\lim_{n\to\infty} | \operatorname{tr} [(Q_0^{(n)})^{-1} Q_1^{(n)} - I] | &= \lim_{n\to\infty} \left| \sum_{i=1}^{\infty}\sum_{j=1}^{\infty} h_{ij}^{(n)} a_{ij} \right| \\
&= \left| \sum_{i=1}^{\infty}\sum_{j=1}^{\infty} h_{ij} a_{ij} \right| \\
&\leqq \left(\sum_{i=1}^{\infty}\sum_{j=1}^{\infty} h_{ij}^{2}\right)^{\frac{1}{2}} \left(\sum_{i=1}^{\infty}\sum_{j=1}^{\infty} a_{ij}^{2}\right)^{\frac{1}{2}} \\
&< \infty , \\
\lim_{n\to\infty} | \operatorname{tr} [Q_0^{(n)} (Q_1^{(n)})^{-1} - I] | &= \lim_{n\to\infty} \left| \sum_{i=1}^{\infty} h_{ii}^{(n)} \lambda_i \right| \\
&= \left| \sum_{i=1}^{\infty} h_{ii} \lambda_i \right| \\
&\leqq \left(\sum_{i=1}^{\infty} h_{ii}^{2}\right)^{\frac{1}{2}} \left(\sum_{i=1}^{\infty} \lambda_i^{2}\right)^{\frac{1}{2}} \\
&< \infty .
\end{aligned}$$

(Q.E.D.)

REFERENCES

1. Grenander, U., Stochastic Processes and Statistical Inference, Arkiv för Matematik, **17,** No. 1, 1950, pp. 195–277.
2. Middleton, D., On the Detection of Stochastic Signals in Additive Normal Noise, I.R.E. Trans., **IT-3,** pp. 86–121.
3. Price, R., Optimum Detection of Stochastic Signals in Noise with Applications to Scatter-Multipath Communications, I.R.E. Trans., **IT-2,** pp. 125–135.
4. Davis, R. C., The Detectability of Random Signals in the Presence of Noise, I.R.E. Trans., **PGIT-3,** pp. 52–62.
5. Bello, P., Some Results on the Problem of Discriminating between Two Gaussian Processes, I.R.E. Trans., **IT-7,** pp. 224–233.
6. Turin, G. L., On Optimal Diversity Reception, I.R.E. Trans., **IT-7,** pp. 154–166.
7. Slepian, D., Some Comments on the Detection of Gaussian Signals in Gaussian Noise, I.R.E. Trans., **IT-4,** pp. 65–68.
8. Parzen, E., Probability Density Functionals and Reproducing Kernel Hilbert Spaces, Proc. of Symposium on Time Series Analysis, John Wiley, New York, 1963, pp. 155–169.
9. Doob, J. L., *Stochastic Processes*, John Wiley, New York, 1953.
10. Loeve, M., *Probability Theory*, 2nd ed., Van Nostrand, Princeton, 1960.
11. Hajek, J., A Property of J-Divergence of Marginal Probability Distributions, Czechoslovak Math. J., **83,** 8, 1958, pp. 460–463.
12. Hajek, J., On a Property of Normal Distribution of Any Stochastic Process (in Russian), Czechoslovak Math., J., **83,** 8, 1958, pp. 610–618. (A translation

appears in Selected Translations in Mathematical Statistics and Probability, Institute of Mathematical Statistics and American Mathematical Society, **1,** pp. 245–252.)

13. Shepp, L. A., The Singularity of Gaussian Measures in Function Space, Proc. Natl. Acad. Sci., **52,** 1964, pp. 430–433.
14. Yaglom, A. M., On the Equivalence and Perpendicularity of Two Gaussian Probability Measures in Function Space, Proc. Symp. on Time Series Analysis, John Wiley, New York, 1963, pp. 327–348.
15. Kantorovich, L. V., and Krylov, V. I., *Approximate Methods of Higher Analysis*, Interscience, New York, 1958.
16. Cooke, R. G., *Infinite Matrices and Sequence Spaces*, MacMillan, London, 1950.
17. Riesz, F., and Sz-Nagy, B., *Functional Analysis*, Frederick Ungar Publishing Co., New York, 1955.
18. Cramer, H., *Mathematical Methods of Statistics*, Princeton University Press, Princeton, 1945.
19. Courant, R., and Hilbert, D., *Methods of Mathematical Physics*, vol. I, Interscience, New York, 1953.
20. Apostol, T. M., *Mathematical Analysis*, Addison-Wesley, Reading, Mass., 1957.

# On the Spectral Properties of Single-Sideband Angle-Modulated Signals

By R. D. BARNARD

(Manuscript received May 11, 1964)

*The representation of single-sideband angle-modulated carriers as originally given by Bedrosian is generalized through the functional and spectral notions of distribution theory. In this treatment the class of related modulating signals is extended to rather general types of distributions, and spectral criteria and iteration algorithms are established by which such modulating signals can be recovered from bandlimited components of the modulated carriers.*

## I. INTRODUCTION

Among the more recent signal transmission techniques for conserving spectral bandwidth is single-sideband angle modulation, first proposed and investigated by Bedrosian.[1] In this scheme a carrier wave is simultaneously angle modulated by an appropriate baseband (bandlimited) signal and amplitude modulated (multiplied) by the negative exponential of the Hilbert transform of the baseband signal, the combined modulation process resulting in an RF spectrum which vanishes identically on the low-frequency side of the carrier frequency and carrier axis crossings which coincide exactly with those of a conventional angle-modulated carrier modulated by the same baseband signal. The single-sideband and axis-crossing properties, although suggesting means with which to obtain ideal bandwidth reduction and compatible detection, are only partially applicable to physical systems.* In general, the RF spectra under the combined and conventional modulation schemes are of infinite extent, and the nonvanishing portion of the spectrum under the former can have, according to any one of several common definitions, a larger effective bandwidth than that under the latter; consequently, single-sideband angle modulation does not necessarily lead to bandwidth reduction, and

* Detection compatibility is suggested by the fact that the output of an ideal limiter depends only on the axis crossings of the input.

the axis-crossing patterns of filtered versions of single-sideband and conventional angle-modulated carriers can differ appreciably. Nevertheless, Bedrosian has shown that the combined form of modulation, as prescribed, offers the possibility of both a reduction in the effective bandwidth over limited ranges of the angle-modulation index and detection compatibility. It is therefore of practical and theoretical interest to establish criteria relating either directly or indirectly to the spectral properties of such carrier waves. In the present paper we specify rather general signal conditions under which the Bedrosian scheme and the associated single-sideband property obtain, and determine spectral conditions under which knowledge of the RF spectrum over a frequency interval slightly wider than half the signal bandwidth provides enough information to recover the baseband signal up to an additive constant.* Signal recovery in this second category is effected by an iterative computation that cannot be carried out exactly in real time; however, the possibility of pure mathematical recovery based on a finite portion of the spectrum constitutes an important spectral property, indicating that the RF spectrum, although infinite in extent, can be viewed theoretically as having an effective bandwidth equal to half the signal bandwidth.† These qualitative results are now restated somewhat more explicitly.

In precise terms, single-sideband angle-modulated carriers are generally assumed to have the form

$$y_c(t) = \exp\left[-\hat{x}(t)\right] \cos\left[2\pi f_c t + x(t)\right]$$

where $x$, $\hat{x}$, and $y_c$ represent respectively a specified angle-modulating signal, its Hilbert transform, and the modulated carrier, the first two functions being periodic or square-integrable, bounded, and bandlimited to some frequency interval $[-f_0, f_0]$. Modulated under these conditions, $y_c$ exhibits the two previously mentioned properties with respect to bandwidth and detection; viz., the corresponding amplitude spectrum (Fourier transform) vanishes over $(-f_c, f_c)$, and the axis crossings as well as the effects that they produce at the output of an ideal limiter coincide exactly with those of the usual angle-modulated carrier

* Contrary to established usage, the term "bandwidth" refers here and throughout to the total frequency spread of the spectrum of the baseband signal over both positive and negative frequencies (cf. Section 2.2).

† Other problems and criteria pertaining to the recoverability of signals subject to nonlinear and bandlimiting operations have received considerable attention recently.[2,3] Beurling's theorem, directly applicable to instantaneous compandors, is perhaps the principal result along these lines.[4] In unpublished work, H. O. Pollak shows by means of Fredholm equation methods that under special conditions the baseband signal of a conventional FM carrier can be recovered mathematically from knowledge of the RF spectrum over an interval of twice the signal bandwidth.

$$y_1 = \cos\,[2\pi f_c t + x(t)].^*$$

To deal with more general modulating signals, i.e., signals which are neither periodic nor square-integrable yet to which the spectral concepts of Fourier transforms and the results above still apply, requires the theory of temperate distributions (generalized functions).[5,6,7] In this paper we treat both $x$ and $y_c$ as special types of distributions and investigate the feasibility of recovering the former from bandlimited components of the latter. Specifically, we: (*i*) generalize the definitions, concepts, and methods of classical Hilbert transform theory to incorporate arbitrary distributions (cf. Section II, Definition 2 and Theorem 3);† (*ii*) extend the class of modulating signals to include all bounded, bandlimited distributions with bounded generalized Hilbert transforms (cf. Section III, Theorem 4); and (*iii*) establish through a standard fixed-point theorem related subclasses for which the spectrum of $y_c$ over any open interval containing $[f_c\,, f_c + f_0]$ furnishes sufficient information for reconstructing derivative $x'(t)$ by iteration (cf. Section IV, Theorems 7–9).‡ It is intended also that this development illustrate the distribution-theoretic approach to be generally employed in connection with other modulation schemes.

## II. PRELIMINARIES

As noted above, characterizing the amplitude spectra and spectral properties of the signals considered in this paper requires the theory of temperate distributions.[6] We discuss here four aspects of this theory: notation and terminology, bandlimited distributions, convolution, and generalized Hilbert transforms.

### 2.1 *Temperate Distributions — Notation and Terminology*

Let $I$ denote a specified, open interval on the real line with $I_\infty$, $I_{+\infty}$, and $I_{-\infty}$ signifying respectively the intervals $(-\infty, \infty)$, $(0, \infty)$, and $(-\infty, 0)$; $\bar{I}$, the closure of $I$; $C^k(I)$, the space of scalar functions of which the derivatives up to and including order $k$ are continuous on $I$; and $C_d$, the space of "rapidly decaying" functions, viz., the linear vector space

* In the first case the nonvanishing portion of the spectrum of $y_c$ is generally so smeared out as to have an effective bandwidth greater than that of $y_1$.

† For detailed examples of Hilbert transform applications in modulation theory, the reader is referred to the expositions of Rowe,[8] Bennett,[9] and Dugundji.[10]

‡ Landau,[2] Miranker,[2] Sandberg,[3] and Beneš[11] have recently made extensive use of fixed-point theorems in a variety of system-theoretic problems relating to recovery and stability.

$$C_d = \{\varphi \mid \varphi \in C^\infty(I_\infty),\ t^j\varphi^{(k)}(t) \to 0(\ |t| \to \infty\ )\forall j,k \geqq 0\}. \tag{1}$$

Also, a topology in $C_d$ is introduced by means of the metric

$$\rho_d(\varphi_1, \varphi_2) = \sum_{k=0}^{\infty} \sum_{j=0}^{k} \frac{2^{-k}}{(k+1)} \frac{\mu(\varphi_1 - \varphi_2, j, k-j)}{[1 + \mu(\varphi_1 - \varphi_2, j, k-j)]} \qquad (2)^* \qquad \varphi_1, \varphi_2 \in C_d$$

where

$$\mu(\varphi, j, k) = \sup_{t \in I_\infty} |\ t^j\varphi^{(k)}(t)\ |.$$

For convenience the convergence of a sequence $\{\varphi_n\}$ relative to this metric is expressed as $\varphi_n \to \varphi(\varphi_n, \varphi \in C_d)$. Since the series in (2) converges uniformly over such sequences, it follows that

$$\varphi_n \to 0 \Leftrightarrow \sup_{t \in I_\infty} |\ t^j{\varphi_n}^{(k)}(t)\ | \xrightarrow[n \to \infty]{} 0 \qquad \forall j,k.$$

As generally defined, temperate distributions are merely the elements of the conjugate space of $C_d$, i.e., the space of linear, continuous functionals on $C_d$.[5,12] In the treatment below we represent this space by $D$ and the corresponding elements by $x\langle\cdot\rangle$. Although mathematically distinct, a distribution $x\langle\cdot\rangle$ and an ordinary function $x(\cdot)$ for which

$$\int_{-\infty}^{\infty} x(t)\varphi(t)dt = x\langle\varphi\rangle \qquad \forall \varphi \in C_d \tag{3}$$

are regarded as characterizing one another, either form being essentially determined by the other.† To extend this notion, we associate every element $x \in D$ with a "generalized function" $x(\cdot)$ (cf. Ref. 6), viz., the totality of sequences $\{x_n(\cdot)\}$ in $C_d$ such that

$$\lim_{n \to \infty} \int_{-\infty}^{\infty} x_n(t)\varphi(t)dt = x\langle\varphi\rangle \qquad \forall \varphi \in C_d. \qquad (4)\ddagger$$

As distributions and generalized functions are in one to one correspondence, it is common to employ all related terms and symbols interchangeably. Also, the ordinary and generalized functions relating to (3) are considered to be equivalent in that both define the same element of $D$.

In connection with the equality of distributions, let $N[\varphi]$ signify the null set of $\varphi \in C_d$, viz.,

* Space $C_d$ constitutes a complete linear topological space in $\rho_d(\cdot,\cdot)$ (cf. Ref. 12, p. 49).

† "Essentially" is used here to indicate that $x\langle\cdot\rangle$ determines $x(\cdot)$ almost everywhere on $I_\infty$ (cf. Ref. 5, pp. 1645–1646).

‡ Sequences satisfying this condition can be shown to exist for an arbitrary distribution (cf. Ref. 6, p. 183).

$$N[\varphi] = \{t \mid t \in \bigcup_\alpha I_\alpha \,;\, \varphi(t) = 0 \;\forall t \in I_\alpha\},$$

and $S[\varphi]$, the support of $\varphi$, viz.,

$$S[\varphi] = I_\infty - N[\varphi].$$

Similarly, let $N_D[x]$ signify the null set of $x \in D$, viz.,

$$N_D[x] = \{t \mid t \in \bigcup_\alpha I_\alpha \,;\, x\langle\varphi\rangle = 0 \;\forall\varphi \ni S[\varphi] \subseteq \bar{I}_\alpha \,,\, \varphi \in C_d\},$$

and $S_D[x]$, the support of $x$, viz.,

$$S_D[x] = I_\infty - N_D[x].$$

Accordingly, if for an interval $I \subseteq I_\infty$ and two elements $x$ and $y$ of $D$, $I \subseteq N_B[x - y]$, we say $x\langle\cdot\rangle = y\langle\cdot\rangle$ and $x(\cdot) = y(\cdot)$ on $I$. This definition also allows one to equate generalized and ordinary functions on arbitrary intervals; that is, if

$$x\langle\varphi\rangle = \int_I v(t)\varphi(t)\,dt \tag{5}$$

for some $v$ and all $\varphi \in C_d$ such that $S[\varphi] \subseteq \bar{I}$, then $I \subseteq N_D[x - v]$ and $x = v$ on $I$.

Among the standard operations associated with distributions, five require special notation:

(*i*) *Products.* With respect to any two distributions $x$ and $y$ of which at least one, say $y$, characterizes an ordinary function $y(\cdot)$ such that $y(\cdot)\varphi \in C_d \;\forall\varphi \in C_d$, let $xy\langle\cdot\rangle$ (and $yx\langle\cdot\rangle$) denote the product of $x$ and $y$ given by

$$xy\langle\varphi\rangle = yx\langle\varphi\rangle \equiv x\langle y\varphi\rangle \qquad \forall\varphi \in C_d\,, \tag{6}$$

and let $x(\cdot)y(\cdot)$ (and $y(\cdot)x(\cdot)$) denote the related generalized function.[6]

(*ii*) *Derivatives.* For any $x \in D$, let $p^n x\langle\cdot\rangle$ denote the $n$th order derivative of $x$ given by

$$p^n x\langle\varphi\rangle \equiv (-1)^n x\langle\varphi^{(n)}\rangle \qquad \forall\varphi \in C_d\,, \tag{7}$$

and $p^n x(\cdot)$, or $(d^n/d(\cdot)^n)x(\cdot)$, the related generalized function.[6]

(*iii*) *Antiderivatives.* For any $x \in D$, let

$$\int_n x\langle\cdot\rangle$$

denote any $n$th order antiderivative of $x$ satisfying

$$p^n \int_n x\langle\varphi\rangle = x\langle\varphi\rangle \qquad \forall\varphi \in C_d,$$

and

$$\int_n x(\cdot),$$

the related generalized function. All $n$th order antiderivatives of a particular element $x \in D$ can differ by only additive polynomials of degree $n - 1$ (cf. Ref. 7, p. 8).

(*iv*) *Limits.* Distribution limits of the form

$$\lim_{\lambda} x_\lambda\langle\varphi\rangle = x\langle\varphi\rangle \in D \qquad \forall\varphi \in C_d$$

are represented in terms of generalized functions by

$$\lim_{\lambda}{}^{(\mathrm{D})}\, x_\lambda(\cdot) = x(\cdot).$$

(*v*) *Fourier Transforms.* For any $x \in D$, let $\tilde{x}\langle\cdot\rangle$ denote the generalized Fourier transform of $x$, viz., the distribution given by

$$\tilde{x}\langle\varphi\rangle \equiv x\langle F\cdot\varphi\rangle \qquad \forall\varphi \in C_d \tag{8}$$

where

$$F\cdot\varphi = \int_{-\infty}^{\infty} \varphi(\xi)e^{-2\pi i f\xi}d\xi \qquad i = \sqrt{-1},$$

and let $\tilde{x}(\cdot)$, or $F\cdot x(\cdot)$, denote the related generalized function (cf. Ref. 6, p. 188). For the right-hand functional in definition (8) to exist, it is required that $F\cdot\varphi \in C_d$, a condition which holds for all $\varphi \in C_d$. Rewriting this relation yields the more suggestive form

$$F\cdot x\langle\varphi\rangle = x\langle F\cdot\varphi\rangle. \tag{9}$$

Similarly,

$$F^{-1}\cdot x\langle\varphi\rangle = x\langle F^{-1}\cdot\varphi\rangle$$

where

$$F^{-1}\cdot\varphi = \int_{-\infty}^{\infty} \varphi(\xi)e^{2\pi i t\xi}d\xi.$$

One property pertaining to operators $\lim^{(\mathrm{D})}$, $p^n$, and $F$ is of paramount importance in applications of distribution theory: the last two commute with the first.[6] The reader is referred to the previously mentioned literature for a detailed discussion of these and other operations as well as the various terms outlined above.

### 2.2 *Bandlimited Distributions*

Let $f$ and $\bar{J}$ denote respectively a point and a compact set on the real frequency line $I_\infty$. A distribution $x$ for which $S_D[F \cdot x] \subseteq \bar{J}$ (i.e., $F \cdot x = 0$ on any $I$ disjoint from $\bar{J}$) is defined to be bandlimited to $\bar{J}$, the space of such elements being designated as

$$B(\bar{J}) = \{x \mid x \in D,\, S_D[F \cdot x] \subseteq \bar{J}\}.$$

Defining, in addition, the space

$$C_g = \{v \mid v \in C^\infty(I_\infty); \forall k \exists j \ni (1 + t^2)^{-j} v^{(k)}(t) \to 0( \mid t \mid \to \infty)\}, \quad (10)$$

we establish the following

*Lemma 1: If $x \in B(\bar{J})$, then $x(\cdot) \in C_g$.*

*Proof:* Construct a real, positive function $\zeta(f) \in C_d$ satisfying the conditions

$$\zeta(f) = \begin{cases} 1 & f \in I_1 \supset \bar{J} \\ 0 & f \notin I_2 \supset \bar{I}_1,\, I_2 \subset I_\infty, \end{cases}$$

and set

$$v(t) = F \cdot x \langle \zeta(f) e^{2\pi i f t} \rangle. \quad (11)$$

We consider first representing $F \cdot x\langle \cdot \rangle$ on $I_2$ by an integral. For this it is necessary to employ the well known result that on any finite interval an arbitrary distribution can be characterized by a multiple derivative of some ordinary, continuous function; more specifically, there exist both a function $\psi \in C(I_2)$ and an integer $N \geqq 0$ such that

$$F \cdot x\langle \varphi \rangle = (-1)^N \int_{I_2} \psi(f) \varphi^{(N)}(f) df \quad (12)$$

for all $\varphi \in C_d$ for which $S[\varphi] \subseteq \bar{I}_2$ (cf. Ref. 7, pp. 11–12). Inasmuch as $S[\zeta] \subseteq \bar{I}_2$, expression (11) becomes

$$v(t) = \int_{I_2} (-1)^N \psi(f) \frac{\partial^N}{\partial f^N} [\zeta(f) e^{2\pi i f t}] df,$$

which in turn gives

$$v^{(k)}(t) = \int_{I_2} (-1)^N \psi(f) \frac{\partial^N}{\partial f^N} [(2\pi i f)^k \zeta(f) e^{2\pi i f t}] df$$
$$= F \cdot x\langle (2\pi i f)^k \zeta(f) e^{2\pi i f t} \rangle = 0(t^N) \qquad (|t| \to \infty)$$

for all $k$; therefore, $v \in C^\infty(I_\infty)$. In order to identify $v$ further, observe that for any $k$

$$(1 + t^2)^{-j} v^{(k)}(t) \to 0 \qquad (|t| \to \infty)$$

for some integer $j$; consequently, by (10), $v \in C_g$.

Regarding the relationship between $v$ and $x$ we note that since

$$\int_{-\infty}^{\infty} v(t)\varphi(t)dt = \int_{-\infty}^{\infty} \int_{I_2} (-1)^N \psi(f) \frac{\partial^N}{\partial f^N} [\zeta(f)\varphi(t)e^{2\pi i f t}] df\, dt$$
$$= (-1)^N \int_{I_2} \int_{-\infty}^{\infty} \psi(f) \frac{\partial^N}{\partial f^N} [\zeta(f)\varphi(t)e^{2\pi i f t}] dt\, df$$
$$= \int_{I_2} (-1)^N \psi(f) \frac{\partial^N}{\partial f^N} \left[ \zeta(f) \int_{-\infty}^{\infty} \varphi(t) e^{2\pi i f t} dt \right] df$$
$$= F \cdot x\langle \zeta F^{-1} \cdot \varphi \rangle \qquad \forall \varphi \in C_d \qquad *$$

and since $S[(\zeta - 1)\theta] \cap S_D[F \cdot x]$ is empty for all $\theta \in C_d$,

$$x\langle\varphi\rangle = F \cdot x\langle F^{-1} \cdot \varphi \rangle + F \cdot x\langle (\zeta - 1) F^{-1} \cdot \varphi \rangle$$
$$= F \cdot x\langle \zeta F^{-1} \cdot \varphi \rangle = \int_{-\infty}^{\infty} v(t)\varphi(t) dt.$$

Hence, in accordance with (3) et seq., $v(\cdot) = x(\cdot)$ with $v \in C_g$.

### 2.3 *Convolution*

A convolution operation sufficiently general for most applications in signal theory is given by

*Definition 1:* For any two distributions $x$ and $y$ of which at least one, say $y$, is such that $F \cdot y(\cdot) \in C_g$ we define a distribution $x*y$, termed the convolution of $x$ and $y$, by the relation

$$x*y = y*x = F^{-1} \cdot [(F \cdot x)(F \cdot y)]. \tag{13}$$

As to the consistency of this definition, observe that with $F \cdot y(\cdot) \in C_g$, $\varphi F \cdot y(\cdot) \in C_d$ for all $\varphi \in C_d$; therefore, according to (6) et seq., both the product $(F \cdot x)(F \cdot y)$ and corresponding convolution exist as dis-

* Interchanging the order of integration in this relation is justified by means of the Tonelli-Hobson theorem (cf. Ref. 13, p. 3).

tributions, and their factors commute. The associative and distributive properties of this operation depend in general on the factors involved, the results in any given case being determined directly from (13). One important consequence of Definition 1 is stated as

*Theorem 1: For any two distributions $x$ and $y$ of which at least one, say $y$, is such that $y(\cdot) \in C_g$*

$$F\cdot(xy) = (F\cdot x)*(F\cdot y).$$

*Proof:* From

$$x\langle\varphi(t)\rangle = x\langle F\cdot F\cdot\varphi(-t)\rangle = F\cdot F\cdot x\langle\varphi(-t)\rangle \qquad \forall x \in D, \quad \forall\varphi \in C_d\,,$$

(6), and (13) it follows that

$$\begin{aligned} F\cdot(xy)\,\langle\varphi(f)\rangle &= xy\langle F\cdot\varphi\rangle = x\langle y(t)F\cdot\varphi\rangle \\ &= F\cdot F\cdot x\langle y(-t)F^{-1}\cdot\varphi\rangle = F\cdot F\cdot x\langle\,[F\cdot F\cdot y(t)]F^{-1}\cdot\varphi\rangle \\ &= [\,(F\cdot F\cdot x)\,(F\cdot F\cdot y)\,]\,\langle F^{-1}\cdot\varphi\rangle \;=\; F^{-1}[(F\cdot F\cdot x)\,(F\cdot F\cdot y)\,]\,\langle\varphi\rangle \\ &= [\,(F\cdot x)*(F\cdot y)\,]\,\langle\varphi\rangle. \end{aligned}$$

We show at this point that Definition 1 relates to a more common but less general form of convolution (cf. Ref. 7, p. 31).

*Theorem 2: If at least one of two distributions $x$ and $y$, say $y$, has a finite support (i.e., $S_D[y] \subseteq \bar{I} \subset I_\infty$), then $x*y$ exists, and*

$$x*y\langle\varphi\rangle = x\langle y\langle\varphi(t+\bar{t})\rangle\rangle \qquad \forall\varphi \in C_d\,.$$

*Proof:* Reversing the roles of $t$ and $f$ in Lemma 1 demonstrates that with $S_D[y]$ finite, i.e., with $y$ time-limited to $\bar{I}$, $F\cdot y(\cdot) \in C_g$ ; hence, by Definition 1, $x*y$ exists. In addition, from (13), (9), and (6) there obtains

$$\begin{aligned} x*y\langle\varphi(t)\rangle &= F\cdot(x*y)\,\langle F^{-1}\cdot\varphi\rangle \;=\; [\,(F\cdot x)\,(F\cdot y)\,]\,\langle F^{-1}\cdot\varphi\rangle \\ &= F\cdot x\langle\,(F\cdot y)\,(F^{-1}\cdot\varphi)\,\rangle = x\langle F\cdot[\,(F\cdot y)\,(F^{-1}\cdot\varphi)\,]\,\rangle \\ &= x\left\langle\int_{-\infty}^{\infty} e^{-2\pi it\xi}(F\cdot y)_\xi(F^{-1}\cdot\varphi)_\xi d\xi\right\rangle \end{aligned}$$

where the subscripts indicate a function of $\xi$. As the integral of this last functional proves to be linear and continuous on $C_d$, i.e., as

$$\int_{-\infty}^{\infty}(F\cdot y)_\xi[e^{-2\pi it\xi}(F^{-1}\cdot\varphi)_\xi]d\xi = F\cdot y\langle(F^{-1}\cdot\varphi)_\xi e^{-2\pi it\xi}\rangle \in D,$$

then

$$\begin{aligned} x*y\langle\varphi(t)\rangle &= x\langle F\cdot y\langle(F^{-1}\cdot\varphi)_\xi e^{-2\pi i\bar{t}\xi}\rangle\rangle \\ &= x\langle y\langle F\cdot[(F^{-1}\cdot\varphi)_\xi e^{-2\pi i\bar{t}\xi}]\rangle\rangle \\ &= x\langle y\langle\varphi(t+\bar{t})\rangle\rangle. \end{aligned}$$

If defined by this expression instead of (13), convolution would not necessarily have commuting factors; e.g., with $y(\cdot)$ and $x(\cdot)$ equal to a Dirac delta function $\delta(\cdot)$ and a constant, respectively, $y\langle x\langle\varphi(t+\bar{t})\rangle\rangle$ is not defined because for this choice of $x$, $x\langle\varphi(t+\bar{t})\rangle$ is constant and not an element of $C_d$.*,†

### 2.4 *Generalized Hilbert Transforms*

In this subsection we extend the applicability of classical Hilbert transform properties and techniques to arbitrary distributions. Required initially are two lemmas relating to antiderivatives.

*Lemma 2: Corresponding to all antiderivatives of an element* $\mathrm{F}\cdot x \in D$ *the distribution limits*

$$\lim_{\lambda\to\infty}{}^{(\mathrm{D})}\left[\tan^{-1}\lambda f\int_N F\cdot x\right] \tag{14}$$

*exist for some* $N \geqq 0$.

*Proof:* Set $I_{\epsilon_1} = (-\epsilon_1, \epsilon_1)$ and $I_{\epsilon_2} = (-\epsilon_2, \epsilon_2)$ with $0 < \epsilon_1 < \epsilon_2 < \infty$, and construct a real, positive function $\eta(f) \in C_d$ satisfying the conditions

$$\eta(f) = \begin{cases} 1 & f \in I_{\epsilon_1} \\ 0 & f \notin I_{\epsilon_2} \supset \bar{I}_{\epsilon_1}. \end{cases}$$

It is convenient to consider first the same type of integral representation as was used in Lemma 1 [cf. (12)]; namely, there exist both a function $\psi \in C(I_{\epsilon_2})$ and an integer $N \geqq 0$ such that

$$F\cdot x\langle\varphi\rangle = (-1)^N\int_{I_{\epsilon_2}}\psi(f)\varphi^{(N)}(f)df \tag{15}$$

* The Dirac distribution is given formally by the equation $\delta\langle\varphi\rangle = \varphi(0)$.

† Commutativity can be forced in such cases by defining the convolution according to the form

$$x*y\langle\varphi\rangle = x\langle y\langle\zeta_0(\bar{t})\varphi(t+\bar{t})\rangle\rangle$$

where $\bar{t}$ corresponds to the distribution of finite support and where $\zeta_0 \in C_d$ equals unity over an open interval containing this support and vanishes outside some finite interval.

for all $\varphi \in C_d$ for which $S[\varphi] \subseteq \bar{I}_{\epsilon_2}$. In agreement with (7) and (5) et seq., relation (15) merely asserts that

$$F \cdot x = \frac{d^N}{df^N} \psi(f) \qquad \text{on } I_{\epsilon_2}$$

and that for all antiderivatives,

$$\int_N F \cdot x = \psi(f) + \sum_{n=0}^{N-1} \alpha_n f^n \qquad \text{on } I_{\epsilon_2} \tag{16}$$

where constants $\alpha_n$ are arbitrary. Since $S[\eta\varphi] \subseteq \bar{I}_{\epsilon_2}$ for all $\varphi \in C_d$, Eq. (16) can be written as

$$\int_N F \cdot x\langle \eta\varphi \rangle = \int_{I_{\epsilon_2}} \left[ \psi(f) + \sum_{n=0}^{N-1} \alpha_n f^n \right] \eta(f)\varphi(f)df.$$

Therefore, by the Lebesgue convergence theorem[13,14]

$$\lim_\lambda \int_N F \cdot x\langle (\tan^{-1} \lambda f)\eta\varphi \rangle$$
$$= \frac{\pi}{2} \int_{I_{\epsilon_2}} (\operatorname{sgn} f) \, [\psi(f) + \sum_n \alpha_n f^n] \eta(f)\varphi(f)df \tag{17}$$

with

$$\operatorname{sgn} f = \begin{cases} 1, & f > 0 \\ -1, & f < 0. \end{cases}$$

On the other hand, since

$$I_{\epsilon_1} \subseteq N\left[ f^j \frac{d^k}{df^k} (1 - \eta)\varphi \right]$$

for all $j$, $k$, and $\varphi \in C_d$,

$$(\tan^{-1} \lambda f) \, (1 - \eta)\varphi \rightarrow (\pi/2) \, (\operatorname{sgn} f) \, (1 - \eta)\varphi \qquad \lambda \rightarrow \infty,$$

and

$$\lim_\lambda \int_N F \cdot x\langle (\tan^{-1} \lambda f) (1 - \eta)\varphi \rangle = \frac{\pi}{2} \int_N F \cdot x\langle (\operatorname{sgn} f) (1 - \eta)\varphi \rangle. \tag{18}$$

Finally, adding limits (17) and (18) yields

$$\frac{\pi}{2} \int_N F \cdot x\langle (\operatorname{sgn} f) (1 - \eta)\varphi \rangle + \frac{\pi}{2} \int_{I_{\epsilon_2}} (\operatorname{sgn} f) \, [\psi + \sum_n \alpha_n f^n] \eta\varphi \, df$$
$$= \lim_\lambda \int_N F \cdot x\langle (\tan^{-1} \lambda f)\varphi \rangle. \tag{19}$$

As both terms on the left are elements of $D$, the distribution limits given by (14) exist.

*Lemma 3: Corresponding to element $x$, integer $N$, and all antiderivatives of Lemma 2 the generalized functions*

$$\frac{d^N}{df^N} \cdot \underset{\lambda\to\infty}{lim}^{(D)} \left[ (tan^{-1} \lambda f) \int_N F \cdot x \right]$$

*differ by only the additive combinations*

$$\pi \sum_{n=0}^{N-1} \alpha_n n!\, \delta(f)^{(N-n-1)}$$

*where $\delta(f)^{(n)}$ represents the nth order derivative of the Dirac function. Furthermore, $\delta(f)^{(N-1)}$ is the highest-order Dirac component which can exist at $= 0$.*

*Proof:* From (19) and (7) there results

$$\begin{aligned} p^N \cdot \underset{\lambda}{\lim}^{(\mathrm{D})} \cdot \int_N F \cdot x\langle(\tan^{-1} \lambda f)\varphi\rangle = {} & \frac{\pi}{2} F \cdot x\langle(\operatorname{sgn} f)(1 - \eta)\varphi\rangle \\ & + (-1)^N \frac{\pi}{2} \int_{I_{\epsilon_2}} (\operatorname{sgn} f)\psi(f)[\eta(f)\varphi(f)]^{(N)} df \qquad (20) \\ & + (-1)^N \frac{\pi}{2} \int_{I_{\epsilon_2}} [(\operatorname{sgn} f) \sum_n \alpha_n f^n][\eta\varphi]^{(N)} df, \end{aligned}$$

the last, only nonunique term reducing to

$$\pi \sum_n \alpha_n n!(-1)^{N-n-1} \varphi(0)^{(N-n-1)} = \pi \sum_n \alpha_n n! \delta\langle\varphi\rangle^{(N-n-1)}.$$

Inspection of the two remaining distributions on the right of (20) shows, in addition, that $\delta(f)^{(N-1)}$ is the highest-order Dirac function possible at $f = 0$; for the support of the first does not include the origin, and the second represents the $N$th derivative of an ordinary, sectionally continuous function.

The preceding two lemmas lead immediately to

*Definition 2:* For any distribution $x$ we define a distribution $\hat{x}$, termed the generalized Hilbert transform of $x$, by the relation

$$\begin{aligned} \hat{x}(\cdot) = -i \frac{2}{\pi} F^{-1} \cdot \Bigg\{ & \frac{d^{N_0}}{df^{N_0}} \cdot \underset{\lambda\to\infty}{\lim}^{(\mathrm{D})} \left[ (\tan^{-1} \lambda f) \int_{N_0} F \cdot x \right] \\ & + \sum_{n=0}^{N_0-1} \beta_n \delta(f)^{(N_0-n-1)} \Bigg\} \end{aligned} \qquad (21)$$

where $N_0$ designates the smallest integer for which Lemma 2 holds and where constants $\beta_n$ are constrained so as to eliminate from $F\cdot\hat{x}$ (or to prescribe) all Dirac distributions at $f = 0$.

As regards ordinary Hilbert transforms it is noted that if

$$F\cdot x(\cdot) \in L_2(I_\infty) \quad \text{(square-integrable)},$$

then

$$\lim_{\lambda}{}^{(\mathrm{D})}\,(\tan^{-1}\lambda f)F\cdot x = (\pi/2)\,(\operatorname{sgn} f)F\cdot x.$$

Consequently, $N_0 = 0$ and

$$\hat{x}(\cdot) = -iF^{-1}\cdot\{(\operatorname{sgn} f)F\cdot x\},$$

a formula which is in agreement with classical theory (cf. Ref. 15, pp. 119–120).

Denoting the linear mapping of (21) by $H$ (i.e., $H : D \rightarrow D$), we list a few of the more significant properties of generalized Hilbert transforms:

(*i*) $H\cdot H\cdot x = -x$ provided there exist in $F\cdot x$ no Dirac components at $f = 0$.

(*ii*) $H\cdot x$ is real provided $x$ is real.

(*iii*) $S_D[F\cdot H\cdot x] \subseteq S_D[F\cdot x]$.

These results follow directly from (20) and Definition 2. Of importance in single-sideband theory is the property given by

*Theorem 3: For any distribution $x$,*

$$S_D[F\cdot(x + i\hat{x})\,] \subseteq \bar{I}_{+\infty}$$

*and*

$$S_D[F\cdot(x - i\hat{x})\,] \subseteq \bar{I}_{-\infty}\,.$$

*That is, $F\cdot(x + i\hat{x})$ and $F\cdot(x - i\hat{x})$ vanish on $I_{-\infty}$ and $I_{+\infty}$, respectively.*

*Proof:* Consider all $\varphi$ such that $S[\varphi] \subseteq \bar{I}_{-\infty}$ ; then,

$$(-1)^N \int_{I_{\epsilon_2}} (\operatorname{sgn} f)[\psi(f) + \sum_n \alpha_n f^n][\eta(f)\varphi(f)]^{(N)} df$$

$$= -p^N \int_N F\cdot x\langle\eta\varphi\rangle = -F\cdot x\langle\eta\varphi\rangle,$$

and from (20) and (21) there obtains

$$F\cdot x\langle\varphi\rangle + iF\cdot\hat{x}\langle\varphi\rangle = F\cdot x\langle\varphi\rangle - F\cdot x\langle(1 - \eta)\varphi\rangle - F\cdot x\langle\eta\varphi\rangle = 0.$$

Similarly, with $S[\varphi] \subseteq \bar{I}_{+\infty}$, $F\cdot x\langle\varphi\rangle - iF\cdot\hat{x}\langle\varphi\rangle = 0$.

## III. SINGLE-SIDEBAND ANGLE MODULATION (SSBΘM)

The notions and results of the previous section apply directly to signals classified as single-sideband angle-modulated, namely, time functions of the form

$$y_c(t) = \exp\,[-\hat{x}(t)\,]\cos\,[2\pi f_c t + x(t)\,].$$

It is the intent here to show that if the modulating signals $x$ correspond to elements of the space

$$S_0 = \{x \mid x \in B(\bar{I}_0),\, I_0 = (-f_0\,, f_0);\, |\,x\,|,|\,\hat{x}\,| < \infty,\, x \text{ real}\},$$

functions $y_c$ characterize distributions, and have, as the term SSBΘM suggests, amplitude spectra (Fourier transforms) which vanish on the interval $(-f_c\,, f_c)$. We begin with three lemmas pertaining to exponentials and convolution.

*Lemma 4: Elements of the spaces*

$$S_1 = \{y \mid y = e^{iz},\, z = x + i\hat{x},\, x \in S_0\},$$

$$S_2 = \{v \mid v = z^N,\, z = x + i\hat{x},\, x \in S_0\,,\, N \geqq 0\}$$

*are equivalent to generalized functions* [*cf.* (4) *et seq.*].

*Proof:* Clearly, since $S_D[F\cdot\hat{x}] \subseteq S_D[F\cdot x]$, both $x$ and $\hat{x}$ are bandlimited as well as bounded, and are, by Lemma 1, elements of $Cg$; hence, $y$ is bounded on $I_\infty$ and integrable over finite intervals, and

$$\left|\int_{-\infty}^{\infty} y(t)\varphi(t)dt\right| \leqq \sup_{t\in I_\infty} |\,(1+t^2)\varphi(t)\,| \int_{-\infty}^{\infty}\left|\frac{y(t)}{1+t^2}\right| dt \qquad (22)$$

$$\forall\varphi \in C_d\,.$$

This latter condition, however, implies that

$$\int_{-\infty}^{\infty} y(t)\varphi_n(t)dt \to 0 \qquad (23)$$

for $\varphi_n \to 0$. Therefore, the left-hand integral of (22) constitutes a continuous, linear functional on $C_d$, i.e., a distribution, and $y$ is equivalent to a generalized function. Precisely the same argument applies to $z^N(N \geqq 0)$, showing that this function is also bounded, integrable over finite intervals, and equivalent to a generalized function.

*Lemma 5: For $x \in S_0$ and $z = x + i\hat{x}$*

$$e^{iz} = \lim_{N\to\infty}{}^{(D)} \sum_{n=0}^{N} \frac{(iz)^n}{n!}\,. \qquad (24)$$

*Proof:* Set

$$y = e^{iz},$$
$$y_N = \sum_{n=0}^{N} \frac{(iz)^n}{n!}.$$

Then, by Darboux's formula,[16]

$$y - y_N = \frac{(iz)^{N+1}}{N!} \int_0^1 e^{i\lambda z}(1 - \lambda)^N d\lambda,$$

and inasmuch as $y$ and $y_N$ represent generalized functions (cf. Lemma 4),

$$\begin{aligned} | y\langle\varphi\rangle - y_N\langle\varphi\rangle | &= \left| \int_{-\infty}^{\infty} [y(t) - y_N(t)]\varphi(t)dt \right| \\ &= \left| \frac{1}{N!} \int_{-\infty}^{\infty} z^{N+1}\varphi \int_0^1 e^{i\lambda z}(1 - \lambda)^N d\lambda \, dt \right| \\ &\leqq \frac{1}{N!} \sup_{t,\lambda} | z^{N+1}e^{i\lambda z} | \int_{-\infty}^{\infty} | \varphi(t) | dt \\ &\leqq \frac{1}{N!} (\sup_t | z |)^{N+1} \exp (\sup_t | \hat{x} |) \\ &\qquad \cdot \int_{-\infty}^{\infty} | \varphi(t) | dt \xrightarrow[N\to\infty]{} 0 \qquad \forall\varphi \in C_d , \end{aligned}$$

a result corresponding to (24).

*Lemma 6: If two distributions $g$ and $h$ are such that*

$$S_D[F\cdot g] \subseteq [0,f_1],$$
$$S_D[F\cdot h] \subseteq [0,f_2],$$

*then*

$$S_D[F\cdot(gh)] \subseteq [0,f_1 + f_2].$$

*Proof:* With respect to any $\varphi \in C_d$ for which $S[\varphi] \subseteq \bar{I}_{-\infty}$, set

$$\varphi_0(t) = F\cdot h\langle\varphi(t + \bar{t})\rangle.$$

As defined, $\varphi_0(t) = 0$ for all $t > 0$; i.e., $I_{+\infty} \subseteq N[\varphi_0]$ and $S[\varphi_0] \subseteq \bar{I}_{-\infty}$. Hence, by Theorems 1 and 2

$$\begin{aligned} [F\cdot(gh)]\langle\varphi\rangle &= [(F\cdot g)*(F\cdot h)]\langle\varphi\rangle \\ &= F\cdot g\langle F\cdot h\langle\varphi(t + \bar{t})\rangle\rangle = F\cdot g\langle\varphi_0(t)\rangle = 0, \end{aligned}$$

which yields

$$S_D[F\cdot(gh)] \subseteq \bar{I}_{+\infty}\,. \tag{25}$$

In a similar manner, consider any $\varphi \in C_d$ for which $S[\varphi] \subseteq [f_1 + f_2\,,\,\infty)$, and set

$$\varphi_1(t) = F\cdot h\langle\varphi(t+\bar{t})\rangle.$$

It follows that $\varphi_1(t) = 0$ for all $t \in (-\infty, f_1)$, i.e., that $S[\varphi_1] \subseteq [f_1\,,\,\infty)$. Therefore,

$$[F\cdot(gh)]\langle\varphi\rangle = F\cdot g\langle\varphi_1\rangle = 0,$$

which yields

$$S_D[F\cdot(gh)] \subseteq (-\infty, f_1 + f_2]. \tag{26}$$

Conditions (25) and (26) prove that

$$S_D[F\cdot(gh)] \subseteq [0, f_1 + f_2].$$

The main result of this section is stated as

*Theorem 4: The amplitude spectra of generalized functions*

$$y_c(t)\ \mathit{exp}\,[-\hat{x}(t)]\ \mathit{cos}\,[2\pi f_c t + x(t)] \qquad x \in S_0$$

*vanish on the interval* $(-f_c\,,f_c)$.

*Proof:* Again, for $x \in S_0$ and $z = x + i\hat{x}$, $S_D[F\cdot z] \subseteq \bar{I}_0$ and, by Theorem 3, $S_D[F\cdot z] \subseteq \bar{I}_{+\infty}$ ; consequently, $S_D[F\cdot z] \subseteq [0, f_0]$. This condition combined with Lemmas 5 and 6 leads to

$$\begin{aligned} &S_D[F\cdot e^{iz}] \\ &\quad = S_D\left[F\cdot \lim_N{}^{(\mathrm{D})} \sum_{n=0}^{N} \frac{(iz)^n}{n!}\right] = S_D\left[\lim_N{}^{(\mathrm{D})} \sum_{n=0}^{N} \frac{1}{n!}\,F\cdot[(iz)^n]\right] \\ &\quad \subseteq \bigcup_n S_D[F\cdot[(iz)^n]] \subseteq \bar{I}_{+\infty}\,. \end{aligned} \tag{27}$$

On the other hand, for $\bar{z} = x - i\hat{x}$

$$S_D[F\cdot e^{-i\bar{z}}] \subseteq \bar{I}_{-\infty}\,. \tag{28}$$

Finally, since $F\cdot[e^{\pm 2\pi i f_c t}y(t)] = \tilde{y}(f \mp f_c)$ for $F\cdot y \equiv \tilde{y}$, then (27) and (28) give

$$S_D[F\cdot y_c] = S_D[F\cdot(e^{iz}e^{2\pi i f_c t} + e^{-i\bar{z}}e^{-2\pi i f_c t})] \subseteq [f_c\,,\,\infty) \cup (-\infty, -f_c],$$

or, equivalently, $F\cdot y_c = 0$ on $(-f_c\,,f_c)$.

## IV. SIGNAL RECOVERY FOR SSBΘM REPRESENTATIONS

In this section we treat the problem of reconstructing signals $x \in S_0$ from bandlimited versions of the associated $SSB\Theta M$ functions $y_c$. Specifically, it is demonstrated that for a large subclass of $S_0$, knowledge of the amplitude spectrum of $y_c$ over any open interval containing $[f_c, f_c + f_0]$ proves sufficient to recover $x$ up to an additive constant. As in the previous section, several lemmas involving the exponential $e^{iz}$ are developed first. To collect notation, we set

$$z = x + i\hat{x} \qquad x \in S_0$$

$$g = iz', \qquad y_a = 1 - e^{iz}, \qquad y_b = (1 - 2\pi it)^{-1} y_a$$

$$y_d = F^{-1}\cdot[(\lambda\tilde{y}_a)*k], \qquad y_n = F^{-1}\cdot[(\lambda F\cdot y_a')*\sigma_n]$$

$$g_n = F^{-1}\cdot[\tilde{g}*\sigma_n], \qquad \sigma_n = n\sigma(nf)(n = 1, 2, \cdots)$$

$$\tilde{y} \equiv F\cdot y \qquad \forall y \in D$$

$$k(f) = F\cdot(1 - 2\pi it)^{-1} = \begin{cases} e^{-f} & f > 0 \\ 0 & f \leqq 0 \end{cases} \tag{29}$$

where $\lambda$ and $\sigma$ are any frequency functions of $C_d$ such that

$$\lambda(f) = \begin{cases} 1 & f \in \bar{I}_c, \quad I_c = (0, f_0 + \epsilon), \quad 0 < \epsilon < \infty \\ 0 & f \notin [-\epsilon, f_0 + 2\epsilon] \end{cases}$$

$$S[\sigma(f)] \subseteq [0,\epsilon], \qquad \int_0^\epsilon \sigma(f)\,df = 1.$$

In addition, let $BV(\bar{I})$ and $UL$ denote respectively the space of scalar functions of bounded variation on a closed interval $\bar{I}$ and the space of scalar functions satisfying a first-order uniform Lipschitz condition on some closed neighborhood of the origin. Finally, define the following subclass of signal space $S_0$:

$$S_{00} = \{x \mid x \in S_0 ; [\tilde{y}_b - \tilde{y}_b(0^+)] \in UL \cap BV(\bar{I}_c); \tilde{y}_b(0^+) \neq 1\}.$$

*Lemma 7: Elements $y_a$, $y_b$, $y_d$, $y_n$, $g_n$, $k$, $\lambda$, and $\sigma_n$ are in $D$.*

*Proof:* This result follows immediately from the corresponding definitions and the test employed in Lemma 4 [cf. (22) and (23)].

*Lemma 8: $S_D[\tilde{y}_b] \subseteq \bar{I}_{+\infty}$.*

*Proof:* Clearly, by (27)

$$S_D[\tilde{y}_a] = S_D[\delta] \cup S_D[F\cdot e^{iz}] \subseteq \bar{I}_{+\infty}. \tag{30}$$

Also, taking any $\varphi \in C_d$ for which $S[\varphi] \subseteq \bar{I}_{-\infty}$, one obtains

$$F^{-1}\cdot[\,(F^{-1}\cdot k)\ (F\cdot\varphi)\,] = 0 \qquad \forall f > 0,$$

or

$$S[F^{-1}\cdot[\,(F^{-1}\cdot k)\ (F\cdot\varphi)\,]] \subseteq \bar{I}_{-\infty}.$$

Hence, for all such $\varphi$

$$\begin{aligned} F\cdot y_b\langle\varphi\rangle &= y_b\langle F\cdot\varphi\rangle = y_a\langle(F^{-1}\cdot k)\ (F\cdot\varphi)\rangle \\ &= F\cdot y_a\langle F^{-1}\cdot[(F^{-1}\cdot k)\ (F\cdot\varphi)]\rangle = 0, \end{aligned}$$

a condition implying that $\tilde{y}_b = 0$ on $I_{-\infty}$.

*Lemma 9: On* $I_c$, $\tilde{y}_b = (\lambda\tilde{y}_a)*k = \tilde{y}_d$ *and* $(F\cdot y_a')*\sigma_n = (\lambda F\cdot y_a')*\sigma_n = \tilde{y}_n$.

*Proof:* Take any $\varphi \in C_d$ for which $S[\varphi] \subseteq \bar{I}_c$. With regard to the first relation

$$S[F^{-1}\cdot[(F^{-1}\cdot k)\ (F\cdot\varphi)]] \subseteq (-\infty, f_0 + \epsilon],$$

and by (30)

$$\begin{aligned} F\cdot y_b\langle\varphi\rangle &= y_b\langle F\cdot\varphi\rangle = y_a\langle(F^{-1}\cdot k)\ (F\cdot\varphi)\rangle \\ &= F\cdot y_a\langle F^{-1}\cdot[(F^{-1}\cdot k)\ (F\cdot\varphi)]\rangle = F\cdot y_a\langle\lambda F^{-1}\cdot[(F^{-1}\cdot k)\ (F\cdot\varphi)]\rangle \\ &= F\cdot[(F^{-1}\cdot k)F^{-1}\cdot(\lambda F\cdot y_a)]\langle\varphi\rangle = [(\lambda\tilde{y}_a)*k]\langle\varphi\rangle. \end{aligned}$$

As to the second relation

$$\int_{-\infty}^{\infty} \sigma_n(\bar{f})\varphi(f+\bar{f})d\bar{f} = \sigma_n\langle\varphi(f+\bar{f})\rangle,$$

$$S[(2\pi if)\sigma_n\langle\varphi(f+\bar{f})\rangle] \subseteq [-\epsilon, f_0 + \epsilon],$$

and by (30) and Theorem 1

$$\begin{aligned} [(F\cdot y_a')*\sigma_n]\langle\varphi\rangle &= F\cdot y_a\langle 2\pi if\sigma_n\langle\varphi(f+\bar{f})\rangle\rangle \\ &= F\cdot y_a\langle 2\pi if\lambda(f)\sigma_n\langle\varphi(f+\bar{f})\rangle\rangle = [(\lambda F\cdot y_a')*\sigma_n]\langle\varphi\rangle. \end{aligned}$$

*Lemma 10:* $\tilde{y}_d \in L_2(I_\infty)$ *and* $\tilde{y}_n \in C_d$.

*Proof:* On the basis of the Tonelli-Hobson theorem

$$F^{-1}\cdot(\lambda F\cdot y_a)\langle\varphi\rangle = y_a\langle F\cdot(\lambda F^{-1}\cdot\varphi)\rangle$$

$$= \int_{-\infty}^{\infty} y_a(\tau) \left\{ \int_{-\infty}^{\infty} e^{-2\pi i \tau f} \lambda(f) \left[ \int_{-\infty}^{\infty} e^{2\pi i f t} \varphi(t)\,dt \right] df \right\} d\tau$$

$$= \int_{-\infty}^{\infty} y_a(\tau) \left\{ \int_{-\infty}^{\infty} \left[ \int_{-\infty}^{\infty} e^{2\pi i f(t-\tau)} \lambda(f)\,df \right] \varphi(t)\,dt \right\} d\tau$$

$$= \int_{-\infty}^{\infty} \left[ \int_{-\infty}^{\infty} y_a(\tau)\,(F^{-1}\cdot\lambda)_{t-\tau}\,d\tau \right] \varphi(t)\,dt \qquad \forall \varphi \in C_d.$$

Therefore, as indicated operationally,

$$F^{-1}\cdot(\lambda F\cdot y_a) = (F^{-1}\cdot\lambda) * y_a = \int_{-\infty}^{\infty} y_a(\tau)\,(F^{-1}\cdot\lambda)_{t-\tau}\,d\tau,$$

and

$$| F^{-1}\cdot(\lambda F\cdot y_a) | \leqq (\sup_{t \in I_\infty} | y_a(t) |) \int_{-\infty}^{\infty} | (F^{-1}\cdot\lambda)_\tau |\,d\tau < \infty,$$

which indicates that $y_d$, $\tilde{y}_d \in L_2(I_\infty)$. Lastly, since

$$(\lambda F\cdot y_a') \in B([-\epsilon, f_0 + 2\epsilon]),$$
$$F^{-1}\cdot\sigma_n \in C_d,$$

by Lemma 1

$$F^{-1}\cdot(\lambda F\cdot y_a') \in C_g,$$
$$y_n = (F^{-1}\cdot\sigma_n)\,[F^{-1}\cdot(\lambda F\cdot y_a')] \in C_d,$$
$$\tilde{y}_n \in C_d.$$

*Lemma 11:* $\mathit{lim}^{(D)}_{n\to\infty} g_n = g$, $g_n \in C_d$, *and* $S_D[\tilde{g}_n] \subseteq \bar{I}_c$.

*Proof:* As

$$\lim_n \int_{-\infty}^{\infty} n\sigma(nf)\varphi(f)\,df = \lim_n \int_{-\infty}^{\infty} \sigma(f)\varphi\left(\frac{f}{n}\right) df = \varphi(0) \qquad \forall \varphi \in C_d,$$

then

$$\lim_n{}^{(\mathbf{D})} \sigma_n = \delta,$$

and

$$\lim_n g_n\langle\varphi\rangle = \lim_n (gF^{-1}\cdot\sigma_n)\langle\varphi\rangle = \lim_n \sigma_n\langle F^{-1}\cdot(g\varphi)\rangle$$

$$= \delta\langle F^{-1}\cdot(g\varphi)\rangle = gF^{-1}\cdot\delta\langle\varphi\rangle = g\langle\varphi\rangle \qquad \forall \varphi \in C_d.$$

Furthermore, with $\sigma_n \in C_d$, all three elements $F^{-1}\cdot\sigma_n$, $g_n$, and $\tilde{g}_n$ are also of $C_d$. Finally, from Lemmas 6 and 7 it follows that

$$S_D[\tilde{g}_n] = S_D[\tilde{g}] \cup S[\sigma_n] \subseteq \bar{I}_c .$$

*Theorem 5: For $x \in S_{00}$, elements $\tilde{g}_n$ satisfy the functional equations*

$$\tilde{g}_n(f) = \gamma\left[\int_0^f \tilde{g}_n(f-\bar{f})\, d\tilde{y}_i(\bar{f}) - \tilde{y}_n(f)\right] \qquad f \in \bar{I}_c; \qquad n = 1, 2, \cdots, \tag{31}$$

*where*

$$\tilde{y}_i(f) = \tilde{y}_d(f) - \tilde{y}_b(0^+) + \int_0^f \tilde{y}_d(\bar{f})\, d\bar{f} \qquad f \in I_c$$

$$\tilde{y}_i(0) = \tilde{y}_i(f_0 + \epsilon) = 0$$

$$\gamma = [1 - \tilde{y}_b(0^+)]^{-1}.$$

*Proof:* According to definitions (29)

$$y_a' F^{-1}\cdot\sigma_n = (1 - 2\pi i t) g_n y_b - g_n \tag{32}$$

with $y_b \in L_2(I_\infty)$, $y_a' F^{-1}\cdot\sigma_n \in C_d$, and $(1 - 2\pi i t) g_n \in C_d$. Expression (32) and Parseval's relation combine to give

$$(F\cdot y_a') * \sigma_n = \tilde{y}_b * (\tilde{g}_n + \tilde{g}_n') - \tilde{g}_n \qquad f \in I_\infty ,$$

which by Lemmas 8 and 11 reduces to

$$(F\cdot y_a') * \sigma_n = \int_0^f \tilde{y}_b(\bar{f}) \left[\tilde{g}_n(f-\bar{f}) - \frac{\partial}{\partial \bar{f}} \tilde{g}_n(f-\bar{f})\right] d\bar{f} - \tilde{g}_n(f).$$

However, if considering $f$ on $I_c$, one need specify $(F\cdot y_a') * \sigma_n$ and $\tilde{y}_b$ on this interval only; consequently, Lemmas 9, 10, and 11 apply, yielding

$$\tilde{y}_n(f) = \int_0^f \tilde{y}_d(\bar{f}) \left[\tilde{g}_n(f-\bar{f}) - \frac{\partial}{\partial \bar{f}} \tilde{g}_n(f-\bar{f})\right] d\bar{f} - \tilde{g}_n(f) \qquad f \in I_c .$$

Clearly, in this equation, $\tilde{y}_d(0)$ and (for the development below) $\tilde{y}_d(f_0 + \epsilon)$ can be set equal to zero without affecting the associated integrals; hence, on integrating by parts, we get (31). To be noted in this theorem is that with respect to signal information, $\tilde{y}_n$, $\tilde{y}_d$, and $\gamma$ derive solely from the bandlimited signal spectrum $\lambda\tilde{y} = \lambda F\cdot e^{ix}$; i.e.,

$$\tilde{y}_d = (\lambda - \lambda\tilde{y}) * k,$$

$$\tilde{y}_n = -(2\pi i f \lambda\tilde{y}) * \sigma_n ,$$

$$\tilde{y}_b(0^+) = \tilde{y}_d(0^+).$$

It is necessary to consider next some general properties of the integral operators in (31), viz., the mappings

$$w = T_n \cdot v = \gamma \left[ \int_0^f v(f - \bar{f}) \, d\tilde{y}_i(\bar{f}) - \tilde{y}_n(f) \right] \qquad f \in \bar{I}_c \tag{33}$$
$$v \in C(\bar{I}_c); \qquad n = 1, 2, \cdots.$$

Relative to the domain of $T_n$, define the norm

$$\| v \|_c = \sup_{f \in \bar{I}_c} | v(f) | \qquad v \in C(\bar{I}_c)$$

and metric

$$\rho_c(v,w) = \| v - w \|_c \qquad v,w \in C(\bar{I}c). \tag{34}$$

Under this scheme the pair $[C(\bar{I}_c), \rho_c] \equiv R_c$ (more precisely, the pair consisting of $C(\bar{I}_c)$ and the metric topology in $C(\bar{I}_c)$) forms a metric space which is complete.[12] We then have

*Theorem 6: Corresponding to any modulating signal $x \in S_{00}$, operators $T_n$ constitute continuous mappings of the complete metric space $R_c$ into itself.*

*Proof:* To show that the range as well as the domain of $T_n$ is in $C(\bar{I}_c)$, take any $v \in C(\bar{I}_c)$ and set $w = T_n \cdot v$ for an arbitrary $n$. With $x \in S_{00}$, $\tilde{y}_b \in BV(\bar{I}_c)$, and by Lemma 10, $\tilde{y}_n \in C_d \subset C(\bar{I}_c)$; moreover, since $\tilde{y}_d = \tilde{y}_b$ on $I_c$ (cf. Lemma 9), and since

$$\tilde{y}_i = \tilde{y}_d - \tilde{y}_b(0^+) + \int_0^f \tilde{y}_d \, d\bar{f} \qquad f \in I_c$$

$$\tilde{y}_i(0) = \tilde{y}_i(f_0 + \epsilon) = 0,$$

function $\tilde{y}_i \in BV(\bar{I}_o)$. Therefore, by the Lebesgue convergence theorem

$$\lim_{f_2 \to f_1} [w(f_2) - w(f_1)] = \lim \left\{ \gamma \int_0^{f_1} [v(f_2 - \bar{f}) - v(f_1 - \bar{f})] \, d\tilde{y}_i(\bar{f}) \right.$$
$$\left. + \gamma \int_{f_1}^{f_2} v(f_2 - \bar{f}) \, d\tilde{y}_i(\bar{f}) - \gamma[\tilde{y}_n(f_2) - \tilde{y}_n(f_1)] \right\} = 0 \qquad \forall f_1, f_2 \in \bar{I}_c.$$

That is, $w \in C(\bar{I}_c)(T_n : C(\bar{I}_c) \to C(\bar{I}_c))$. For establishing the continuity of $T_n$, consider any two functions $v_1, v_2 \in C(\bar{I}_c)$, and set $w_1 = T_n \cdot v_1$, $w_2 = T_n \cdot v_2$, and $v_0 = v_2 - v_1$ for an arbitrary $n$. It follows from (33) and (34) that

$$\rho_c(w_2, w_1) = \sup_{f \in \bar{I}_c} \left| \gamma \int_0^f v_0(f - \bar{f})\, d\tilde{y}_i(\bar{f}) \right|$$

$$\leqq |\gamma|\, V_i(f_0 + \epsilon) \sup_{f \in \bar{I}_c} |v_0(f)|$$

$$= |\gamma|\, V_i(f_0 + \epsilon) \rho_c(v_2, v_1)$$

where $V_i(f)$ signifies the total variation of $\tilde{y}_i$ on the interval $[0,f]$. Consequently, $\rho_c(w_2, w_1) \rightarrow 0$ if $\rho_c(v_2, v_1) \rightarrow 0$; i.e., $T_n$ is continuous.

A basic result relating to the reconstruction of $SSB\Theta M$ signals can now be stated as

*Theorem 7: Corresponding to any modulating signal* $x \in S_{00}$, *each of the equations*

$$\tilde{g}_n = T_n \cdot \tilde{g}_n \qquad \tilde{g}_n \in C(\bar{I}_c); \qquad n = 1, 2, \cdots,$$

*has a unique solution given by*

$$\tilde{g}_n = \lim_{m \to \infty} \tilde{g}_{n,m} \qquad n,m = 1, 2, \cdots$$

$$\tilde{g}_{n,m+1} \equiv T_n^{\ m} \cdot \tilde{g}_{n,1} \qquad \forall n,m$$

$$\tilde{g}_{n,1} \equiv 0 \qquad f \in \bar{I}_c \qquad \forall n$$

*where convergence is uniform on* $\bar{I}_c$. *Furthermore,*

$$\frac{dx}{dt} = Im\,[F^{-1} \cdot \lim_n{}^{(D)} \cdot \lim_m \tilde{g}_{n,m}] = Im\,[\lim_n \cdot \lim_m \cdot g_{n,m}]$$

$$\tilde{g}_n, \tilde{g}_{n,m} \equiv 0 \qquad f \notin \bar{I}_c \qquad \forall n, m$$

*where* $Im[\cdot]$ *indicates the imaginary part of the quantity in brackets.*

*Proof:* We employ here a standard fixed-point contraction-mapping theorem (cf. Ref. 17, p. 50): If $\rho(\cdot, \cdot)$ and $T$ represent respectively a metric in a complete metric space $R = [C,\rho]$ and a continuous mapping of $R$ into itself, and if for some $k$ and any two elements $v,w \in C$

$$\rho(T^k \cdot v, T^k \cdot w) \leqq \alpha\rho(v,w) \qquad \alpha < 1,$$

then there exists a unique solution to the equation $T \cdot v_0 = v_0$. Also, for an arbitrary element $v \in C$ this solution is given by

$$v_0 = \lim_{m \to \infty} T^m \cdot v,$$

where convergence is taken relative to $\rho$. In view of Theorem 6 we need only demonstrate in the present proof that for some $k$ and any two elements $v,w \in C(\bar{I}_c)$ each mapping $T_n$ satisfies the contraction condition

$$\| T_n^k \cdot v - T_n^k \cdot w \|_c \leqq \alpha \| v - w \|_c \qquad \alpha < 1.$$

First, set

$$v - w = u$$

$$M = \sup_{f \in \bar{I}_c} | u(f) | = \| u \|_c .$$

With $\tilde{y}_i \in BV(\bar{I}_c)$ (cf. Theorem 6)

$$| T_n \cdot v - T_n \cdot w | = \left| \gamma \int_0^f u(f - \tilde{f}) \, d\tilde{y}_i(\tilde{f}) \right|$$

$$\leqq M | \gamma | \int_0^f dV_i(\tilde{f}) = M | \gamma | V_i(f) \qquad f \in \bar{I}_c ,$$

where $V_i(f)$ again denotes the total variation of $\tilde{y}_i$ on $[0,f]$. However, inasmuch as

$$[\tilde{y}_b - \tilde{y}_b(0^+)] \in UL \cap BV(\bar{I}_c),$$

$$\tilde{y}_i = \tilde{y}_b - \tilde{y}_b(0^+) + \int_0^f \tilde{y}_b \, d\tilde{f} \qquad f \in I_c ,$$

$$\tilde{y}_i(0) = \tilde{y}_i(f_0 + \epsilon) = 0,$$

then $\tilde{y}_i \in UL \cap BV(\bar{I}_c)$; hence, there exists a positive constant $a_0$ such that

$$V_i(f) \leqq a_0 f \qquad f \in \bar{I}_c .$$

From this last condition it follows that

$$| T_n \cdot v - T_n \cdot w | \leqq M | \gamma | a_0 f$$

$$| T_n^2 \cdot v - T_n^2 \cdot w | \leqq M | \gamma |^2 a_0 \int_0^f (f - \tilde{f}) \, dV_i(\tilde{f}) = M | \gamma |^2 a_0 \int_0^f V_i(\tilde{f}) \, d\tilde{f}$$

$$\leqq M | \gamma |^2 a_0^2 \int_0^f \tilde{f} \, d\tilde{f} = \frac{M | \gamma |^2 a_0^2 f^2}{2} \qquad f \in \bar{I}_c ,$$

and, in general,

$$| T_n^k \cdot v - T_n^k \cdot w | \leqq \frac{M | \gamma |^k a_0^{k-1}}{(k - 1)!} \int_0^f (f - \tilde{f})^{(k-1)} \, dV_i(\tilde{f})$$

$$= - \frac{M | \gamma |^k a_0^{k-1}}{(k - 1)!} \int_0^f V_i(\tilde{f}) \, d(f - \tilde{f})^{(k-1)}$$

$$\leqq -\frac{M \mid \gamma \mid^k {a_0}^k}{(k-1)\,!} \int_0^f \bar{f} d(f - \bar{f})^{(k-1)}$$

$$= \frac{M \mid \gamma \mid^k {a_0}^k}{(k-1)\,!} \int_0^f \bar{f}^{(k-1)}\, d\bar{f} = \frac{M \mid \gamma \mid^k {a_0}^k f^k}{k\,!} \qquad f \in \bar{I}_c .$$

Therefore, for $k$ sufficiently large

$$\frac{M \mid \gamma \mid^k {a_0}^k (f_0 + \epsilon)^k}{k\,!} = \alpha < 1,$$

and

$$\| {T_n}^k \cdot v - {T_n}^k \cdot w \|_c \leqq \alpha \| v - w \|_c \qquad \alpha < 1.$$

The contraction principle as outlined then yields the main statement of the theorem, the last result being an immediate consequence of Lemma 11.

Treated next are two important classes of modulating signals which prove to be contained in $S_{00}$ : periodic functions of $S_0$ and integrable $(L_1(I_\infty))$ functions of $S_0$ having integrable Hilbert transforms. In the following development we represent the space of periodic functions by $P$ and the intersection $H \cdot (L_1(I_\infty)) \cap L_1(I_\infty)$ by $\hat{L}_1(I_\infty)$.

*Theorem 8:* $S_0 \cap P \subset S_{00}$ .

*Proof:* Elements $x \in S_0 \cap P$ must have the form

$$x = \sum_{n=-N}^{N} b_n e^{2\pi i n f_p t}$$

$$N f_p \leqq f_0 , \qquad b_n = \bar{b}_{-n}$$

where $\bar{b}$ signifies the conjugate of $b$. Consequently, in accordance with Definition 2

$$z = x + i\hat{x} = \sum_{n=0}^{N} b_n e^{2\pi i n f_p t}.$$

Putting $z_p = z - b_0$ , we obtain

$$y_b = (1 - 2\pi i t)^{-1} \left[ (1 - e^{ib_0}) - e^{ib_0} \sum_{n=1}^{\infty} (i z_p)^n \right],$$

or

$$\tilde{y}_b = (1 - e^{ib_0}) k - e^{ib_0} \left[ \sum_{n=1}^{\infty} k * F \cdot (i z_p)^n \right],$$

which, since

$$S_D[F\cdot(iz_p)^n] \subseteq [nf_p\,,\ \infty)$$

$$S_D\left[F\cdot\sum_{n=1}^{\infty}(iz_p)^n\right] \subseteq [f_p\,,\ \infty),$$

gives

$$S_D[\tilde{y}_b - (1 - e^{ib_0})k] \subseteq [f_p\,,\ \infty),$$

$$\tilde{y}_b(0^+) = (1 - e^{ib_0}) \neq 1,$$

and

$$[\tilde{y}_b - \tilde{y}_b(0^+)] \in UL \cap BV(\bar{I}_c).$$

*Theorem 9:* $S_0 \cap \hat{L}_1(I_\infty) \subset S_{00}$, *and for all* $x$ *in this intersection*

$$\frac{dx}{dt} = Im\,[\lim_{m\to\infty} h_m] \qquad m = 1, 2, \cdots,$$

*where*

$$h_{m+1} = (F^{-1}\cdot\lambda_0)*[(y_a*F^{-1}\cdot\lambda_0)h_m] - \frac{d}{dt}(y_a*F^{-1}\cdot\lambda_0) \qquad t \in I_\infty$$

$$h_1 \equiv 0 \qquad t \in I_\infty$$

$$\lambda_0(f) = \begin{cases} 1 & f \in \bar{I}c \\ 0 & f \notin \bar{I}_c\,. \end{cases}$$

*Proof:* Considering that $x \in S_0 \cap \hat{L}_1(I_\infty)$, $z$ is a bounded element of $L_1(I_\infty)$; hence, by Darboux's formula

$$|\,y_a(t)\,| = |\,1 - e^{iz}\,| \leqq |\,z\,|\int_0^1 e^{\lambda|\hat{z}|}\,d\lambda \leqq |\,z(t)\,|\exp\,(\sup_{t\in I_\infty}|\,\hat{x}\,|), \quad (35)$$

and

$$|\,\tilde{y}_b(f_2) - \tilde{y}_b(f_1)\,| \leqq |\,f_2 - f_1\,|\exp\,(\sup_t|\,\hat{x}\,|)\int_{-\infty}^{\infty}\left|\,z(t)\left(\frac{2\pi it}{1 - 2\pi it}\right)\cdot\frac{\sin\pi(f_2 - f_1)t}{\pi(f_2 - f_1)t}\right|dt$$

$$\leqq |\,f_2 - f_1\,|\exp\,(\sup_t|\,\hat{x}\,|)\int_{-\infty}^{\infty}|\,z(t)\,|\,dt$$

$$\forall f_1, f_2 \in I_\infty,$$

the latter condition indicating that $\tilde{y}_b(0^+) = \tilde{y}_b(0^-) = 0$ (cf. Lemma 8) and that $\tilde{y}_b \in UL \cap BV(\bar{I}_c)$.* In order to prove the second part of the theorem, we note first that with $x \in S_0 \cap \hat{L}_1(I_\infty)$, $z \in L_2(I_\infty)$ and $\tilde{z} \in L_2 \cap C(I_\infty)$; moreover, as $S_D[\tilde{z}] \subseteq \bar{I}_c$, $\tilde{z} \in L_1(I_\infty)$ and $\tilde{g} = 2\pi i f(i\tilde{z}) \in L_1 \cap L_2 \cap C(I_\infty)$. Similarly, by (35), $y_a \in L_1 \cap L_2(I_\infty)$ and $\tilde{y}_a \in L_2 \cap C(I_\infty)$. As a result,

$$\begin{aligned} |\tilde{g}_n| &= |\tilde{g}*\sigma_n| \\ &= \left| \int_0^f \tilde{g}(f - \bar{f})\sigma(n\bar{f})n\, d\bar{f} \right| \\ &\leqq \sup_{f \in \bar{I}_c} |\tilde{g}(f)| \int_0^\epsilon |\sigma(\bar{f})|\, d\bar{f} \leqq \text{constant} \qquad \forall n, \end{aligned}$$

$$\lim_n{}^{(\mathrm{D})} \tilde{g}_n = \lim_n \tilde{g}_n = \tilde{g},$$

$$\lim_n{}^{(\mathrm{D})} \tilde{y}_n = \lim_n [(2\pi i f \tilde{y}_a \lambda)*\sigma_n] = (2\pi i f \tilde{y}_a \lambda),$$

and by the Lebesgue convergence theorem

$$\tilde{g} = \lim_n{}^{(\mathrm{D})} T_n \cdot \tilde{g}_n = \int_0^f \tilde{g}(f - \bar{f})\, d\tilde{y}_i(\bar{f}) - 2\pi i f \tilde{y}_a \lambda \equiv A \cdot \tilde{g} \qquad f \in \bar{I}_c .$$

This expression asserts that $\tilde{g}$ is a fixed point of the mapping $A: C(\bar{I}_c) \rightarrow C(\bar{I}_c)$. Precisely the same arguments as were used in Theorems 6 and 7 apply here to show that $A$ is continuous with respect to norm $\| \cdot \|_c$, and that

$$\begin{aligned} \frac{dx}{dt} &= \mathrm{Im}\, [\lim_{m\to\infty} \cdot h_m] && m = 1, 2, \cdots \\ \tilde{h}_{m+1} &= A^m \cdot \tilde{h}_1 && \forall m \\ \tilde{h}_1 &\equiv 0 && f \in \bar{I}_c \\ \tilde{h}_m &\equiv 0 \qquad f \notin \bar{I}_c && \forall m \end{aligned}$$

where convergence is uniform on $\bar{I}_c$. On writing $\tilde{y}_i$ as

$$\tilde{y}_i = (\tilde{y}_a\lambda)*k + \int_0^f [(\tilde{y}_a\lambda)*k]\, d\bar{f}$$

* A similar calculation employing the Schwarz inequality shows that $\tilde{y}_b(0^+) = 0$ for $x \in L_2(I_\infty)$ also. Most square-integrable signals of practical interest satisfy the appropriate Lipschitz and bounded variation conditions, and are therefore contained in $S_{00}$.

$$= e^{-f}\int_0^f \tilde{y}_a\lambda e^{\bar{f}}\,d\bar{f} + \int_0^f e^{-\bar{f}}\left[\int_0^{\bar{f}} \tilde{y}_a\lambda e^{f'}\,df'\right]d\bar{f}$$

$$= \int_0^f \tilde{y}_a(\bar{f})\lambda(\bar{f})\,d\bar{f} \qquad f \in \bar{I}_c,$$

we get

$$\tilde{h}_{m+1} = \int_0^f \tilde{h}_m(f-\bar{f})\tilde{y}_a(\bar{f})\lambda(\bar{f})\,d\bar{f} - 2\pi i f\tilde{y}_a\lambda \qquad f \in \bar{I}_c$$

or, more compactly,

$$\tilde{h}_{m+1} = \lambda_0(f)\int_0^f \tilde{h}_m(f-\bar{f})\tilde{y}_a(\bar{f})\lambda_0(\bar{f})\,d\bar{f} - 2\pi i f\tilde{y}_a\lambda_0 \qquad f \in I_\infty. \quad (36)$$

(Since $S_D[\tilde{g}] \subseteq [0, f_0]$, $\lambda_0$ could be defined to have the same support.) Taking the inverse Fourier transform of both sides of (36) yields the second part of the theorem.

## V. SUMMARY

Definition 2 and Theorems 3 through 9, which constitute the principal results of the preceding sections, provide both a distribution-theoretic basis for the spectral representation of single-sideband angle-modulated carriers and a recurrence formulation for reconstructing most of the associated modulating signals of practical interest. It is important to emphasize again that the approach employed in this development applies also to other modulation schemes.

## VI. ACKNOWLEDGMENTS

The author wishes to express his appreciation to P. L. Zador for several fruitful discussions and comments during the course of this work.

## APPENDIX

*Index of Symbols*

| | | | |
|---|---|---|---|
| $k(f)$ | — p. 2827 | $\tilde{y}_i$ | — p. 2830 |
| $\hat{L}_1$ | — p. 2834 | $z$ | — p. 2824 |
| $N, N_D$ | — p. 2814-2815 | $\gamma$ | — p. 2830 |
| $P$ | — p. 2834 | $\delta$ | — p. 2820 |
| $S, S_D$ | — p. 2815 | $\lambda$ | — p. 2828 |
| $S_0, S_1, S_2$ | — p. 2824 | $\rho_c(\cdot, \cdot)$ | — p. 2832 |
| $S_{00}$ | — p. 2827 | $\sigma$ | — p. 2827 |
| $T_n$ | — p. 2831 | $\sigma_n$ | — p. 2827 |
| $UL$ | — p. 2827 | $p^n x$ | — p. 2815 |
| $x\langle \cdot \rangle, x(\cdot)$ | — p. 2814 | $\int_n x$ | — p. 2815 |
| $\hat{x}$ | — p. 2823 | $x * y$ | — p. 2818 |
| $\tilde{x}$ | — p. 2816 | $\lim^{(D)}$ | — p. 2816 |
| $y_a, y_b, y_d, y_n$ | — p. 2827 | $\rightarrow$ | — p. 2814 |
| $y_c$ | — p. 2824 | $\Vert v \Vert_c$ | — p. 2831 |

## REFERENCES

1. Bedrosian, E., The Analytic Signal Representation of Modulated Waveforms, Proc. I.R.E., **50,** Oct., 1962, pp. 2071–2076.
2. Landau, H. J., and Miranker, W. L., The Recovery of Distorted Bandlimited Signals, J. Math. Anal. and Appl., **2,** Feb., 1961, pp. 97–104.
3. Sandberg, I. W., On the Properties of Some Systems that Distort Signals, Parts I and II, B.S.T.J., **42,** Sept., 1963, p. 2033, and **43,** Jan., 1964, p. 91.
4. Landau, H. J., On the Recovery of a Band-Limited Signal, After Instantaneous Companding and Subsequent Band-Limiting, B.S.T.J., **39,** March, 1960, p. 363.
5. Dunford, N., and Schwartz, J. T., *Linear Operators*, Part II, Interscience, 1963.
6. Temple, G., Generalized Functions, Proc. Roy. Soc. (London), **A228,** 1955, pp. 175–190.
7. Halpern, I., *Introduction to the Theory of Distributions*, University of Toronto Press, Toronto, Canada, 1952.
8. Rowe, H. E., *Signals and Noise in Communication Systems*, to be published by Van Nostrand and Company.
9. Bennett, W. R., *Electrical Noise*, McGraw-Hill, New York, 1960.
10. Dugundji, J., Envelopes and Pre-Envelopes of Real Waveforms, Trans. I.R.E., **IT-4,** No. 1, pp. 53–57.
11. Beneš, V. E., Ultimately Periodic Solutions to a Non-Linear Integrodifferential Equation, B.S.T.J., **41,** Jan., 1962, p. 257.
12. Dunford, N., and Schwartz, J. T., *Linear Operators*, Part I, Interscience, 1958.
13. Goldberg, R. R., *Fourier Transforms*, Cambridge University Press, London, 1961.
14. Burkhill, J. C., *The Lebesgue Integral*, Cambridge University Press, London, 1961.
15. Titchmarsh, E. G., *Theory of Fourier Integrals*, Oxford University Press, London, 1950.
16. Whittaker, E. T., and Watson, G. N., *Modern Analysis*, Cambridge University Press, London, 1958.
17. Kolmagorov, A. N., and Fomin, S. V., *Functional Analysis*, Vol. 1, Graylock Press, Rochester, New York, 1957.

# On the Properties of Nonlinear Integral Equations That Arise in the Theory of Dynamical Systems

By I. W. SANDBERG and V. E. BENEŠ

(Manuscript received May 4, 1964)

*This paper reports on some results concerning the properties of integral equations that govern the behavior of a large class of control systems or electrical networks containing linear time-invariant elements and an arbitrary finite number of nonlinear time-varying elements.*

*In particular, for networks containing linear time-invariant elements and an arbitrary finite number of positive-slope nonlinear resistors, it is proved, under reasonable conditions, that the response to a periodic excitation applied at $t = 0$ is ultimately periodic with the same period as the excitation, regardless of the initial state of the network.*

## I. NOTATION AND DEFINITIONS

Let $M$ denote an arbitrary matrix. We shall denote by $M'$, $M^*$, and $M^{-1}$, respectively, the transpose, the complex-conjugate transpose, and the inverse of $M$. The positive square-root of the largest eigenvalue of $M^*M$ is denoted by $\Lambda\{M\}$, and $1_N$ denotes the identity matrix of order $N$.

The set of real, measurable $N$-vector-valued functions of the real variable $t$ defined on $(-\infty, \infty)$ $[[0, \infty)]$ is denoted by $\mathcal{H}_N$ $[\mathcal{H}_{N+}]$, and

$$\mathfrak{L}_{2N} = \left\{ f \mid f \,\varepsilon\, \mathcal{H}_N\,, \int_{-\infty}^{\infty} f'f\,dt < \infty \right\}$$

$$\mathfrak{L}_{2N+} = \left\{ f \mid f \,\varepsilon\, \mathcal{H}_{N+}\,, \int_{0}^{\infty} f'f\,dt < \infty \right\}.$$

The norm of $f = (f_1, f_2, \cdots, f_N)' \,\varepsilon\, \mathfrak{L}_{2N}$ $[\mathfrak{L}_{2N+}]$ is denoted by

$$\| f \| \; [\, \| f \|_+ ];$$

it is defined by

$$\| f \|^2 = \int_{-\infty}^{\infty} f'f \, dt \qquad \left[ \| f \|_+^2 = \int_0^{\infty} f'f \, dt \right]$$

and the norm of a linear transformation $\mathbf{T}$ defined on $\mathcal{L}_{2N}$ [$\mathcal{L}_{2N+}$] is denoted by $\| \mathbf{T} \|$ [ $\| \mathbf{T} \|_+$].

Let $y \; \varepsilon \; (0, \infty)$, and, if $f \; \varepsilon \; \mathcal{H}_N$, let

$$\begin{aligned} f_y &= f \quad \text{for} \quad |t| \leqq y \\ &= 0 \quad \text{for} \quad |t| > y; \end{aligned}$$

if $f \; \varepsilon \; \mathcal{H}_{N+}$, let

$$\begin{aligned} f_y &= f \quad \text{for} \quad t \; \varepsilon \; [0,y] \\ &= 0 \quad \text{for} \quad t > y. \end{aligned}$$

The sets $\mathcal{E}_N$ and $\mathcal{E}_{N+}$ are defined as follows

$$\begin{aligned} \mathcal{E}_N &= \{ f \mid f \; \varepsilon \; \mathcal{H}_N , \quad f_y \; \varepsilon \; \mathcal{L}_{2N} \text{ for } 0 < y < \infty \} \\ \mathcal{E}_{N+} &= \{ f \mid f \; \varepsilon \; \mathcal{H}_{N+} , \quad f_y \; \varepsilon \; \mathcal{L}_{2N+} \text{ for } 0 < y < \infty \}. \end{aligned}$$

With $\mathcal{B}$ the set of $N$-vector-valued functions of $t$ which have the property that each component is uniformly bounded on its domain of definition, let

$$\mathcal{L}_{\infty N} = \mathcal{B} \cap \mathcal{H}_N, \quad \text{and} \quad \mathcal{L}_{\infty N+} = \mathcal{B} \cap \mathcal{H}_{N+}.$$

Let $T$ be a real positive constant and let

$$\mathcal{K}_N = \left\{ f \mid f \; \varepsilon \; \mathcal{H}_N , \quad f(t) = f(t + T) \quad \text{for all } t, \int_0^T f'f \, dt < \infty \right\}.$$

Throughout the paper, $k$ denotes a measurable, real $N \times N$ matrix-valued function of $t$ defined on $(-\infty, \infty)$, with elements $\{k_{mn}\}$ such that

$$\int_{-\infty}^{\infty} | k_{mn}(t) | \, dt < \infty \qquad (m,n = 1, 2, \cdots, N),$$

and $\psi[f(t),t]$, with $f \; \varepsilon \; \mathcal{H}_N$ or $f \; \varepsilon \; \mathcal{H}_{N+}$, denotes the $N$ vector

$$(\psi_1[f_1(t),t], \psi_2[f_2(t),t], \cdots, \psi_N[f_N(t),t])'$$

where $\psi_1(w,t), \psi_2(w,t), \cdots, \psi_N(w,t)$ are real-valued functions of the real variables $w$ and $t$ for $-\infty < w < \infty$ and $-\infty < t < \infty$ such that

(*i*) there exist real numbers $\alpha$ and $\beta$ with the property that

$$\alpha \leqq \frac{\psi_n(w_1, t) - \psi_n(w_2, t)}{w_1 - w_2} \leqq \beta \qquad (n = 1, 2, \cdots, N)$$

for all $t \varepsilon (-\infty, \infty)$ and all real $w_1$ and $w_2$ such that $w_1 \neq w_2$, and

(*ii*) $\psi_n[w(t),t]$ is a measurable function of $t$ whenever $w(t)$ is measurable ($n = 1, 2, \cdots, N$).

The symbol $s$ denotes a scalar complex variable with $\sigma = \text{Re}[s]$ and $\omega = \text{Im}[s]$.

## II. INTRODUCTION

Equations of the form

$$g(t) = f(t) + \int_0^t k(t - \tau)\psi[f(\tau),\tau]d\tau, \qquad 0 \leqq t < \infty \qquad (1)$$

in which $f \varepsilon \mathcal{E}_{N+}$ and $g \varepsilon \mathcal{L}_{\infty N+}$, are frequently encountered in the study of physical systems containing linear time-invariant elements and an arbitrary finite number of time-varying nonlinear elements. Typically, $f$ represents the system response and $g$ takes into account both the independent energy sources and the initial conditions at $t = 0$. For example, (1) governs the behavior of (a) an important type of control system containing linear time-invariant elements and an arbitrary finite number of memoryless time-varying nonlinear amplifiers, or (b) an important type of electrical network containing linear time-invariant elements and an arbitrary finite number of time-varying nonlinear resistors.

The related equation

$$g(t) = f(t) + \int_{-\infty}^{\infty} k(t - \tau)\psi[f(\tau),\tau]d\tau, \qquad -\infty < t < \infty \qquad (2)$$

is also often encountered. It arises when it is convenient for mathematical reasons to formulate a model of the system such that the response and excitation are defined for all $t \varepsilon (-\infty, \infty)$. In (2), usually $g \varepsilon \mathcal{L}_{\infty N}$ and only solutions belonging to $\mathcal{L}_{\infty N}$ are of interest.

One of the classic problems in the analysis of nonlinear physical systems is the determination of the properties of the response of a system, governed by an equation of the form (1), to a periodic input applied at $t = 0$. Usually, the functions $\psi_n(w,t)$, which enter into the definition of $\psi[\cdot, \cdot]$, are independent of $t$; $g$ can be written as $g = g_1 + g_2$ in which $g_1 \varepsilon \mathcal{K}_N \cap \mathcal{L}_{\infty N+}$, $g_2 \varepsilon \mathcal{L}_{2N+}$, and $g_2(t) \to 0$ as $t \to \infty$; and (in accordance with the usual Volterra integral equation theory) it is known that there exists a solution $f \varepsilon \mathcal{E}_{N+}$. In a great many cases of engineering interest it is simply *assumed* that there exists a unique response and that it is ultimately periodic with the period of the input. This is a central assumption associated, for example, with the well-known describing-function technique for the approximate determination of the steady-state response of nonlinear systems.

In connection with the actual determination of the steady-state response, two common engineering assumptions are (in effect) that there exists a unique element of $\mathcal{L}_{\infty N} \cap \mathcal{K}_N$, $\hat{f}$, that satisfies

$$g_1(t) = \hat{f}(t) + \int_{-\infty}^{t} k(t-\tau)\psi[\hat{f}(\tau)]d\tau, \qquad -\infty < t < \infty$$

and that the solution of (1), with $g = g_1 + g_2$, approaches $\hat{f}(t)$ as $t \to \infty$, the principal ideas evidently being that if the physical system is stable in some suitable sense, then the effect of the initial conditions at $t = 0$ should eventually "die out," and, moreover, that the steady-state response of the system should be obtained "at once" if the periodic excitation is applied at "$t = -\infty$."

The purpose of this paper is to report on some mathematical results concerning the properties of (1) and (2) that are pertinent, to a considerable extent, to engineering questions of the type discussed. In particular, as an application of our first theorem, we establish the mathematical validity of the engineering assumptions described above under what amount to reasonable conditions for the case in which $k(\cdot)$ is the matrix-valued weighting function of a passive network and $\psi[\cdot,\cdot]$ represents $N$ positive-slope nonlinear resistors (see Theorem 3 and associated remarks).

Under similar conditions, it is proved that an equation of the type (2) possesses at most one $\mathcal{L}_{\infty N}$ solution. This type of result is of direct interest with regard to the qualitative nature of the solutions of (2), for if our conditions are met, and, as is often the case, (a) $g$ in (2) is periodic with period $T$, (b) the $\psi_n(w,t)$ are periodic in $t$ with period $T$, and (c) $f$ is an $\mathcal{L}_{\infty N}$ solution of (2), then [since $f(t + T)$ is also a solution of (2)] it is clear that $f$ must be periodic with period $T$.

## III. RESULTS

Theorem 1, below, focuses attention on a relation between the solutions of (1) and (2). This theorem is later used in order to obtain conditions under which the solution of (1) approaches a periodic steady state as $t \to \infty$, when $g$ approaches a periodic steady state as $t \to \infty$.

*Theorem 1: Let*

$$h_1(t) = f_1(t) + \int_{-\infty}^{t} k(t-\tau)\psi[f_1(\tau),\tau]d\tau, \qquad -\infty < t < \infty$$

$$h_2(t) = f_2(t) + \int_0^t k(t - \tau)\psi[f_2(\tau),\tau]d\tau, \qquad 0 \leqq t < \infty$$

*in which* $h_1 \varepsilon \mathcal{H}_N$, $f_1 \varepsilon \mathcal{H}_N \cap \mathcal{E}_{N+}$, $h_2 \varepsilon \mathcal{H}_{N+}$, *and* $f_2 \varepsilon \mathcal{E}_{N+}$. *Suppose that*

(*i*) $(h_1 - h_2) \varepsilon \mathfrak{L}_{2N+}$

(*ii*) $\int_{-\infty}^0 k(t - \tau)\psi[f_1(\tau),\tau]d\tau \ \varepsilon\ \mathfrak{L}_{2N+}$

*and that, with*

$$K(s) = \int_0^\infty k(t)e^{-st}dt \quad \textit{for} \quad \sigma \geqq 0,$$

(*iii*) $\det\,[1_N + \tfrac{1}{2}(\alpha + \beta)K(s)] \neq 0 \quad \textit{for} \quad \sigma \geqq 0$

(*iv*) $\tfrac{1}{2}(\beta - \alpha) \sup_{-\infty<\omega<\infty} \Lambda\{[1_N + \tfrac{1}{2}(\alpha + \beta)K(i\omega)]^{-1}K(i\omega)\} < 1.$

*Then* $(f_1 - f_2) \varepsilon \mathfrak{L}_{2N+}$, *and, with*

$$\rho_1 = \sup_{-\infty<\omega<\infty} \Lambda\{[1_N + \tfrac{1}{2}(\alpha + \beta)K(i\omega)]^{-1}\}$$

$$\rho_2 = \sup_{-\infty<\omega<\infty} \Lambda\{[1_N + \tfrac{1}{2}(\alpha + \beta)K(i\omega)]^{-1}K(i\omega)\},$$

$$\| f_1 - f_2 \|_+ \leqq \rho_1[1 - \tfrac{1}{2}(\beta - \alpha)\rho_2]^{-1} \cdot \left\| h_1 - h_2 - \int_{-\infty}^0 k(t - \tau)\psi[f_1(\tau),\tau]d\tau \right\|_+ .$$

*If, in addition to the hypotheses stated above,*

$$h_1(t) - h_2(t) - \int_{-\infty}^0 k(t - \tau)\psi[f_1(\tau),\tau]d\tau \to 0$$

*as* $t \to \infty$, *and*

$$\int_0^\infty | k_{mn}(t) |^2\, dt < \infty \qquad (m,n = 1, 2, \cdots, N),$$

*then* $[f_1(t) - f_2(t)] \to 0$ *as* $t \to \infty$.

Our next result is concerned with the character of the change in the solution of (2) when $g$ is altered by the addition of an element of $\mathfrak{L}_{2N}$.

*Theorem 2: Let*

$$h_1(t) = f_1(t) + \int_{-\infty}^\infty k(t - \tau)\psi[f_1(\tau),\tau]d\tau, \qquad -\infty < t < \infty$$

$$h_2(t) = f_2(t) + \int_{-\infty}^\infty k(t - \tau)\psi[f_2(\tau),\tau]d\tau, \qquad -\infty < t < \infty$$

*in which:* $h_1$, $h_2$ $\varepsilon$ $\mathcal{K}_N$; $f_1$, $f_2$ $\varepsilon$ $\mathcal{L}_{\infty N}$; *and* $(h_1 - h_2)$ $\varepsilon$ $\mathcal{L}_{2N}$. *Suppose that*

(*i*)

$$\int_0^\infty \left| \int_t^\infty | k_{mn}(x) | \, dx \right|^2 dt + \int_{-\infty}^0 \left| \int_{-\infty}^t | k_{mn}(x) | \, dx \right|^2 dt < \infty,$$

$$(m,n = 1, 2, \cdots, N)$$

*and that, with*

$$K(i\omega) = \int_{-\infty}^\infty k(t) e^{-i\omega t} dt,$$

(*ii*) *det* $[1_N + \frac{1}{2}(\alpha + \beta) K(i\omega)] \neq 0$ *for all* $\omega$

(*iii*) $\frac{1}{2}(\beta - \alpha) \underset{-\infty<\omega<\infty}{sup} \Lambda\{[1_N + \frac{1}{2}(\alpha + \beta)K(i\omega)]^{-1} K(i\omega)\} < 1.$

*Then* $(f_1 - f_2)$ $\varepsilon$ $\mathcal{L}_{2N}$, *and, with*

$$\rho_1 = \underset{-\infty<\omega<\infty}{sup} \Lambda\{[1_N + \tfrac{1}{2}(\alpha + \beta)K(i\omega)]^{-1}\}$$

$$\rho_2 = \underset{-\infty<\omega<\infty}{sup} \Lambda\{[1_N + \tfrac{1}{2}(\alpha + \beta)K(i\omega)]^{-1} K(i\omega)\},$$

$$\| f_1 - f_2 \| \leqq \rho_1 [1 - \tfrac{1}{2}(\beta - \alpha)\rho_2]^{-1} \| h_1 - h_2 \|.$$

Observe that Theorem 2 implies that if (*i*), (*ii*) and (*iii*) are satisfied, then (2) possesses at most one $\mathcal{L}_{\infty N}$ solution.

As indicated earlier, in many cases of engineering interest $g$, in (1), can be written as $g = g_1 + g_2$, in which $g_1$ $\varepsilon$ $\mathcal{K}_N \cap \mathcal{L}_{\infty N+}$, $g_2$ $\varepsilon$ $\mathcal{L}_{2N+}$, and $g_2(t) \to 0$ as $t \to \infty$. In such cases it is often of considerable importance to determine whether $f(t)$ approaches a steady-state response that is periodic with period $T$ as $t \to \infty$. As a specific application of Theorem 1, the following result is proved.

*Theorem 3: Let* $g_1$ $\varepsilon$ $\mathcal{K}_N \cap \mathcal{L}_{\infty N+}$, $g_2$ $\varepsilon$ $\mathcal{L}_{2N+}$, $g_2(t) \to 0$ *as* $t \to \infty$, $\psi_n(w,t) = \psi_n(w,t + T)$ *for all* $w$ *and* $t$ *and* $n = 1, 2, \cdots, N$, *and* $\psi[0,t]$ $\varepsilon$ $\mathcal{K}_N$. *Let* $f$ $\varepsilon$ $\mathcal{E}_{N+}$ *satisfy*

$$g_1(t) + g_2(t) = f(t) + \int_0^t k(t - \tau)\psi[f(\tau),\tau] \, d\tau, \qquad 0 \leqq t < \infty.$$

*Suppose that*

(*i*) $\displaystyle\int_0^\infty \left| \int_t^\infty | k_{mn}(x) | \, dx \right|^2 dt < \infty \qquad (m, n = 1, 2, \cdots, N)$

(*ii*) $\displaystyle\int_0^\infty | (1 + t) k_{mn}(t) |^2 \, dt < \infty \qquad (m, n = 1, 2, \cdots, N)$

*and that, with*

$$K(s) = \int_0^\infty k(t)e^{-st}dt \quad for \quad \sigma \geqq 0,$$

(*iii*) $det\,[1_N + \frac{1}{2}(\alpha + \beta)K(s)] \neq 0 \quad for \quad \sigma \geqq 0$

(*iv*) $\frac{1}{2}(\beta - \alpha) \sup_{-\infty<\omega<\infty} \Lambda\{[1_N + \frac{1}{2}(\alpha + \beta)K(i\omega)]^{-1}K(i\omega)\} < 1.$

*Then there exists a unique* $\hat{f} \,\varepsilon\, \mathcal{K}_N$ *such that*

$$g_1(t) = \hat{f}(t) + \int_{-\infty}^t k(t - \tau)\psi[\hat{f}(\tau),\tau]d\tau, \qquad -\infty < t < \infty.$$

*Moreover,* $\hat{f} \,\varepsilon\, \mathcal{L}_{\infty N}$, $(f - \hat{f}) \,\varepsilon\, \mathcal{L}_{2N+}$, *and*

$$[f(t) - \hat{f}(t)] \rightarrow 0 \quad as \quad t \rightarrow \infty.$$

With regard to the hypotheses of Theorems 1 and 3, it can be shown that*

$$\det\,[1_N + \tfrac{1}{2}(\alpha + \beta)K(s)] \neq 0 \quad \text{for} \quad \sigma \geqq 0$$

and

$$\tfrac{1}{2}(\beta - \alpha) \sup_{-\infty<\omega<\infty} \Lambda\{[1_N + \tfrac{1}{2}(\alpha + \beta)K(i\omega)]^{-1}K(i\omega)\} < 1$$

provided that $\alpha \geqq 0$ and $[K(i\omega) + K(i\omega)^*]$ is nonnegative definite for all $\omega$. For this reason our results are particularly relevant to the theory of passive nonlinear electrical networks.

## IV. PROOFS

### 4.1 *Proof of Theorem 1*

Let $\mathbf{K}$ denote the bounded linear mapping of $\mathcal{L}_{2N+}$ into itself defined by

$$\mathbf{K}f = \int_0^t k(t - \tau)f(\tau)d\tau, \qquad f \,\varepsilon\, \mathcal{L}_{2N+}.$$

With $y$ an arbitrary positive number, and $f$ an arbitrary element of $\mathcal{K}_{N+}$, let $\mathbf{P}$ denote the mapping of $\mathcal{K}_{N+}$ into itself defined by $\mathbf{P}f = f_y$, and let $\psi f$ denote the $N$-vector-valued function of $t$ with values

$$\psi[f(t),t] \quad \text{for} \quad 0 \leqq t < \infty.$$

* The validity of the first assertion can be established with a standard argument involving the analyticity of $K(s)$ for $\sigma > 0$. The second statement is a direct extension of a result proved in Ref. 1. In particular, the greatest lower bound (over $n$) of the smallest eigenvalue of the term $[1_N + R_n]^{-1*}\,[1_N + R_n + R_n^*][1_N + R_n]^{-1}$, which appears in (7) of Ref. 1, can easily be shown to be positive. Thus, the conclusion of Theorem 2 of Ref. 1 remains valid if the condition $\alpha > 0$ is replaced by $\alpha \geqq 0$.

Then from

$$\begin{aligned} h_1(t) - h_2(t) - \int_{-\infty}^{0} k(t-\tau)\psi[f_1(\tau),\tau]d\tau \\ = f_1(t) - f_2(t) + \int_0^t k(t-\tau)(\psi[f_1(\tau),\tau] - \psi[f_2(\tau),\tau])d\tau, \\ 0 \leqq t < \infty \end{aligned} \tag{3}$$

and the fact that

$$\begin{aligned} \mathbf{P}\int_0^t k(t-\tau)(\psi[f_1(\tau),\tau] - \psi[f_2(\tau),\tau])d\tau \\ = \mathbf{P}\int_0^t k(t-\tau)(\psi[f_{1y}(\tau),\tau] - \psi[f_{2y}(\tau),\tau])d\tau, \end{aligned}$$

we obtain

$$\begin{aligned} h_y = \mathbf{P}[\mathbf{I} + \tfrac{1}{2}(\alpha+\beta)\mathbf{K}](f_{1y} - f_{2y}) \\ + \mathbf{PK}\{\psi f_{1y} - \psi f_{2y} - \tfrac{1}{2}(\alpha+\beta)(f_{1y} - f_{2y})\} \end{aligned} \tag{4}$$

in which $\mathbf{I}$ denotes the identity operator on $\mathfrak{L}_{2N+}$ and

$$h_y = h_{1y} - h_{2y} - \mathbf{P}\int_{-\infty}^{0} k(t-\tau)\psi[f_1(\tau),\tau]d\tau .$$

In order to proceed we need the following result.[2]

*Lemma 1: Let* $\det[1_N + \frac{1}{2}(\alpha+\beta)K(s)] \neq 0$ *for* $\sigma \geqq 0$. *Then* $[\mathbf{I} + \frac{1}{2}(\alpha+\beta)\mathbf{K}]$ *possesses a bounded inverse on* $\mathfrak{L}_{2N+}$, *and*

$$\| [\mathbf{I} + \tfrac{1}{2}(\alpha+\beta)\mathbf{K}]^{-1} \|_+ \leqq \sup_{-\infty<\omega<\infty} \Lambda\{[1_N + \tfrac{1}{2}(\alpha+\beta)K(i\omega)]^{-1}\}$$

$$\| [\mathbf{I} + \tfrac{1}{2}(\alpha+\beta)\mathbf{K}]^{-1}\mathbf{K} \|_+ \leqq \sup_{-\infty<\omega<\infty} \Lambda\{[1_N + \tfrac{1}{2}(\alpha+\beta)K(i\omega)]^{-1}K(i\omega)\}.$$

*Furthermore,*

$$\mathbf{P}[\mathbf{I} + \tfrac{1}{2}(\alpha+\beta)\mathbf{K}]^{-1} = \mathbf{P}[\mathbf{I} + \tfrac{1}{2}(\alpha+\beta)\mathbf{K}]^{-1}\mathbf{P} \quad \textit{for all} \quad y > 0.$$

Thus, since

$$\mathbf{P}[\mathbf{I} + \tfrac{1}{2}(\alpha+\beta)\mathbf{K}]^{-1}\mathbf{P}[\mathbf{I} + \tfrac{1}{2}(\alpha+\beta)\mathbf{K}](f_{1y} - f_{2y}) = f_{1y} - f_{2y},$$

we obtain from (4)

$$\begin{aligned} f_{1y} - f_{2y} = \mathbf{P}[\mathbf{I} + \tfrac{1}{2}(\alpha+\beta)\mathbf{K}]^{-1}h_y \\ - \mathbf{P}[\mathbf{I} + \tfrac{1}{2}(\alpha+\beta)\mathbf{K}]^{-1}\mathbf{PK}\{\psi f_{1y} - \psi f_{2y} - \tfrac{1}{2}(\alpha+\beta)(f_{1y} - f_{2y})\}. \end{aligned}$$

Using the fact that

$$\| \psi f_{1y} - \psi f_{2y} - \tfrac{1}{2}(\alpha + \beta)(f_{1y} - f_{2y}) \|_+ \leqq \tfrac{1}{2}(\beta - \alpha) \| f_{1y} - f_{2y} \|_+ ,$$

it follows that

$$\begin{aligned} \| f_{1y} - f_{2y} \|_+ \leqq\ & \| \mathbf{P}[\mathbf{I} + \tfrac{1}{2}(\alpha + \beta)\mathbf{K}]^{-1} h_y \|_+ \\ & + \tfrac{1}{2}(\beta - \alpha) \| \mathbf{P}[\mathbf{I} + \tfrac{1}{2}(\alpha + \beta)\mathbf{K}]^{-1}\mathbf{PK} \|_+ \\ & \cdot \| f_{1y} - f_{2y} \|_+ \\ \leqq\ & \| [\mathbf{I} + \tfrac{1}{2}(\alpha + \beta)\mathbf{K}]^{-1} \|_+ \| h_y \|_+ \\ & + \tfrac{1}{2}(\beta - \alpha) \| [\mathbf{I} + \tfrac{1}{2}(\alpha + \beta)\mathbf{K}]^{-1}\mathbf{K} \|_+ \\ & \cdot \| f_{1y} - f_{2y} \|_+ . \end{aligned}$$

Using the inequalities of the lemma,

$$\begin{aligned} \| f_{1y} - f_{2y} \|_+ &\leqq \rho_1[1 - \tfrac{1}{2}(\beta - \alpha)\rho_2]^{-1} \| h_y \|_+ \\ &\leqq \rho_1[1 - \tfrac{1}{2}(\beta - \alpha)\rho_2]^{-1} \\ &\quad \cdot \left\| h_1 - h_2 - \int_{-\infty}^{0} k(t - \tau)\psi[f_1(\tau),\tau] d\tau \right\|_+ \end{aligned} \tag{5}$$

for *all* $y > 0$. Therefore, $(f_1 - f_2) \,\varepsilon\, \mathfrak{L}_{2N+}$ and $\| f_1 - f_2 \|_+$ possesses the upper bound stated in the theorem.

We now show that $(f_1 - f_2) \,\varepsilon\, \mathfrak{L}_{2N+}$ ,

$$h_1(t) - h_2(t) - \int_{-\infty}^{0} k(t - \tau)\psi[f_1(\tau),\tau] d\tau \to 0 \quad \text{as} \quad t \to \infty , \tag{6}$$

and

$$\int_0^\infty | k_{mn}(t) |^2 \, dt < \infty \qquad (m, n = 1, 2, \cdots, N) \tag{7}$$

imply that $[f_1(t) - f_2(t)] \to 0$ as $t \to \infty$.

Assume that $(f_1 - f_2) \,\varepsilon\, \mathfrak{L}_{2N+}$ and that (6) and (7) hold. Then, from (3) it is evident that $[f_1(t) - f_2(t)] \to 0$ as $t \to \infty$ if

$$\int_0^t k(t - \tau)(\psi[f_1(\tau),\tau] - \psi[f_2(\tau),\tau]) d\tau \to 0 \quad \text{as} \quad t \to \infty . \tag{8}$$

To prove that (8) is satisfied, observe first that $(f_1 - f_2) \,\varepsilon\, \mathfrak{L}_{2N+}$ implies that $(\psi f_1 - \psi f_2) \,\varepsilon\, \mathfrak{L}_{2N+}$ . Thus it suffices to show that if $g \,\varepsilon\, \mathfrak{L}_{2N+}$ , then

$$\int_0^t k(t - \tau) g(\tau) d\tau \to 0 \quad \text{as} \quad t \to \infty .$$

Let

$$G(i\omega) = \text{l.i.m.} \int_0^\infty g(t)e^{-i\omega t}dt, \qquad g \, \varepsilon \, \mathfrak{L}_{2N+}.$$

Then, in view of assumption (7), the modulus of any element of the $N$-vector $K(i\omega)G(i\omega)$ is integrable on the $\omega$-set $(-\infty, \infty)$, and hence by the Riemann-Lebesgue lemma

$$\frac{1}{2\pi} \int_{-\infty}^\infty K(i\omega)G(i\omega)e^{i\omega t}d\omega,$$

which is equal to

$$\int_0^t k(t - \tau)g(\tau)d\tau,$$

approaches zero as $t \to \infty$. This completes the proof of Theorem 1.

### 4.2 *Proof of Theorem 2*

In this section, $\mathbf{K}$ denotes the bounded linear mapping of $\mathfrak{L}_{2N}$ into itself defined by

$$\mathbf{K}f = \int_{-\infty}^\infty k(t - \tau)f(\tau)d\tau, \qquad f \, \varepsilon \, \mathfrak{L}_{2N}.$$

With $y$ an arbitrary positive number and $f$ an arbitrary element of $\mathcal{H}_N$, $\mathbf{P}$ denotes the mapping of $\mathcal{H}_N$ into itself defined by $\mathbf{P}f = f_y$, and $\psi f$ denotes the $N$-vector-valued function of $t$ with values

$$\psi[f(t),t] \quad \text{for} \quad -\infty < t < \infty.$$

From the fact that

$$\begin{aligned} h_1(t) - h_2(t) &= f_1(t) - f_2(t) \\ &\quad + \int_{-\infty}^\infty k(t - \tau)(\psi[f_1(\tau),\tau] - \psi[f_2(\tau),\tau])d\tau, \end{aligned} \tag{9}$$

we obtain

$$\begin{aligned} h_y &= f_{1y} - f_{2y} + \mathbf{K}(\psi f_{1y} - \psi f_{2y}) \\ &= [\mathbf{I} + \tfrac{1}{2}(\alpha + \beta)\mathbf{K}](f_{1y} - f_{2y}) \end{aligned} \tag{10}$$

$$+ \mathbf{K}\{\psi f_{1y} - \psi f_{2y} - \tfrac{1}{2}(\alpha + \beta)(f_{1y} - f_{2y})\}, \tag{11}$$

in which $\mathbf{I}$ denotes the identity operator on $\mathfrak{L}_{2N}$, and

$$h_y(t) = h_{1y}(t) - h_{2y}(t) + \int_{-\infty}^{\infty} k(t - \tau)(\psi[f_{1y}(\tau),\tau] - \psi[f_{2y}(\tau),\tau])d\tau$$
$$- \mathbf{P}\int_{-\infty}^{\infty} k(t - \tau)(\psi[f_1(\tau),\tau] - \psi[f_2(\tau),\tau])d\tau.$$

At this point we need[2,†]

*Lemma 2: If det* $[1_N + \frac{1}{2}(\alpha + \beta)K(i\omega)] \neq 0$ *for all* $\omega$, *then*

$$[\mathbf{I} + \tfrac{1}{2}(\alpha + \beta)\mathbf{K}]$$

*possesses a bounded inverse on* $\mathfrak{L}_{2N}$, *and*

$$\| [\mathbf{I} + \tfrac{1}{2}(\alpha + \beta)\mathbf{K}]^{-1} \| \leqq \sup_{-\infty<\omega<\infty} \Lambda\{[1_N + \tfrac{1}{2}(\alpha + \beta)K(i\omega)]^{-1}\}$$
$$\| [\mathbf{I} + \tfrac{1}{2}(\alpha + \beta)\mathbf{K}]^{-1}\mathbf{K} \| \leqq \sup_{-\infty<\omega<\infty} \Lambda\{[1_N + \tfrac{1}{2}(\alpha + \beta)K(i\omega)]^{-1}K(i\omega)\}.$$

Thus from (11),

$$f_{1y} - f_{2y}$$
$$= -[\mathbf{I} + \tfrac{1}{2}(\alpha + \beta)\mathbf{K}]^{-1}\mathbf{K}\{\psi f_{1y} - \psi f_{2y} - \tfrac{1}{2}(\alpha + \beta)(f_{1y} - f_{2y})\}$$
$$+ [\mathbf{I} + \tfrac{1}{2}(\alpha + \beta)\mathbf{K}]^{-1}h_y.$$

Using the fact that

$$\| \psi f_{1y} - \psi f_{2y} - \tfrac{1}{2}(\alpha + \beta)(f_{1y} - f_{2y}) \| \leqq \tfrac{1}{2}(\beta - \alpha) \| f_{1y} - f_{2y} \|,$$

we have

$$\| f_{1y} - f_{2y} \| \leqq \tfrac{1}{2}(\beta - \alpha) \| [\mathbf{I} + \tfrac{1}{2}(\alpha + \beta)\mathbf{K}]^{-1}\mathbf{K} \| \cdot \| f_{1y} - f_{2y} \|$$
$$+ \| [\mathbf{I} + \tfrac{1}{2}(\alpha + \beta)\mathbf{K}]^{-1} \| \cdot \| h_y \|.$$

In view of the inequalities of the lemma,

$$\| f_{1y} - f_{2y} \| \leqq \rho_1[1 - \tfrac{1}{2}(\beta - \alpha)\rho_2]^{-1} \| h_y \|. \tag{12}$$

Assume now that there exists a constant $c$ such that $\| h_y \| \leqq c$ for all $y > 0$. Then, from (12), it is clear that $(f_1 - f_2) \, \varepsilon \, \mathfrak{L}_{2N}$. This implies that $(\psi f_1 - \psi f_2) \, \varepsilon \, \mathfrak{L}_{2N}$. Hence, (9) can be written as

$$h_1 - h_2 = f_1 - f_2 + \mathbf{K}(\psi f_1 - \psi f_2),$$

from which it follows, by essentially the same argument as that used to obtain (12) from (10), that

[†] With no more than a reinterpretation of the functions involved, the proofs of the inequalities of Lemma 1 suffice to establish the inequalities of Lemma 2.

$$\| f_1 - f_2 \| \leqq \rho_1[1 - \tfrac{1}{2}(\beta - \alpha)\rho_2]^{-1} \| h_1 - h_2 \|.$$

Therefore to complete the proof of Theorem 2, it suffices to prove

*Lemma 3: If* $(h_1 - h_2) \varepsilon \mathfrak{L}_{2N}$, $f_1, f_2 \varepsilon \mathfrak{L}_{\infty N}$, *and assumption* (*i*) *of Theorem 2 is satisfied, then there exists a constant c such that* $\| h_y \| \leqq c$ *for all* $y > 0$.

### 4.2.1 *Proof of Lemma 3*

Let $q = (q_1, q_2, \cdots, q_N)' = (\psi f_1 - \psi f_2)$,

$$\begin{aligned} \theta(t) &= 1 \quad \text{for} \quad |t| \leqq y \\ &= 0 \quad \text{for} \quad |t| > y, \end{aligned}$$

and

$$u = (u_1, u_2, \cdots, u_N)' = \int_{-\infty}^{\infty} k(t - \tau)[\theta(\tau) - \theta(t)]q(\tau)d\tau.$$

Then, since $(h_1 - h_2) \varepsilon \mathfrak{L}_{2N}$, it is sufficient to prove that there exists a constant $c_1$ such that $\| u \| \leqq c_1$ for all $y > 0$. Further, since

$$\begin{aligned} \| u \|^2 &= \sum_{m=1}^{N} \int_{-\infty}^{\infty} | u_m(t) |^2 dt \\ &= \sum_{m=1}^{N} \int_{-\infty}^{\infty} \left| \sum_{n=1}^{N} \int_{-\infty}^{\infty} k_{mn}(t - \tau)[\theta(\tau) - \theta(t)]q_n(\tau)d\tau \right|^2 dt \\ &\leqq N \sum_{m=1}^{N} \sum_{n=1}^{N} \int_{-\infty}^{\infty} \left| \int_{-\infty}^{\infty} k_{mn}(t - \tau)[\theta(\tau) - \theta(t)]q_n(\tau)d\tau \right|^2 dt \\ &\leqq \eta N \sum_{m=1}^{N} \sum_{n=1}^{N} \int_{-\infty}^{\infty} \left| \int_{-\infty}^{\infty} | k_{mn}(t - \tau) | \cdot | \theta(\tau) - \theta(t) | d\tau \right|^2 dt, \end{aligned}$$

in which

$$\eta = \max_n \sup_t | q_n(t) |^2,$$

it suffices to show that there exists a constant $c_2$ such that for all $y > 0$

$$\int_{-\infty}^{\infty} \left| \int_{-\infty}^{\infty} | k_{nn}(t - \tau) | \cdot | \theta(\tau) - \theta(t) | d\tau \right|^2 dt \leqq c_2$$

$$(m, n = 1, 2, \cdots, N).$$

Using the fact that

$$\begin{aligned}\int_{-\infty}^{\infty} &| k_{mn}(\tau) | \, | \theta(t-\tau) - \theta(t) | \, d\tau \\ &= \int_{t+y}^{\infty} | k_{mn}(\tau) | \, d\tau + \int_{-\infty}^{t-y} | k_{mn}(\tau) | \, d\tau \quad \text{for} \quad | t | \leqq y \\ &= \int_{t-y}^{t+y} | k_{mn}(\tau) | \, d\tau \qquad\qquad \text{for} \quad | t | > y,\end{aligned}$$

it is a simple matter to verify that

$$\begin{aligned}\int_{-\infty}^{\infty} &\left| \int_{-\infty}^{\infty} | k_{mn}(t-\tau) | \, | \theta(\tau) - \theta(t) | \, d\tau \right|^2 dt \\ &\leqq 2 \int_0^{2y} \left| \int_x^{\infty} | k_{mn}(\tau) | \, d\tau \right|^2 dx + 2 \int_{-2y}^{0} \left| \int_{-\infty}^{x} | k_{mn}(\tau) | \, d\tau \right|^2 dx \\ &\quad + \int_{-\infty}^{0} \left| \int_{x-2y}^{x} | k_{mn}(\tau) | \, d\tau \right|^2 dx \\ &\quad + \int_0^{\infty} \left| \int_x^{x+2y} | k_{mn}(\tau) | \, d\tau \right|^2 dx,\end{aligned}$$

from which it is evident that our assumptions imply that there exists a $c_2$ with the required property. This proves the lemma and completes the proof of Theorem 2.

*Remark:*

Assumption ($i$) of Theorem 2 is satisfied if

$$\int_{-\infty}^{\infty} | t k_{mn}(t) | \, dt < \infty, \qquad (n, m = 1, 2, \cdots, N),$$

for then the (bounded) functions

$$\int_t^{\infty} | k_{mn}(x) | \, dx \quad \text{and} \quad \int_{-\infty}^{t} | k_{mn}(x) | \, dx$$

are integrable on $(0, \infty)$ and $(-\infty, 0)$, respectively.

### 4.3 *Proof of Theorem 3*

We need two lemmas.

*Lemma 4: Let* $\psi[\cdot, \cdot]$ *satisfy the conditions of Theorem 3,* $g_1 \varepsilon \mathcal{K}_N$, *and*

$$K(i\omega) = \int_0^{\infty} k(t) e^{-i\omega t} dt.$$

*Suppose that, with $\mathfrak{N}$ the set of integers,*

$$(i)\ det\left[1_N + \tfrac{1}{2}(\alpha + \beta)K\left(\frac{i2\pi n}{T}\right)\right] \neq 0 \quad for \quad n \,\varepsilon\, \mathfrak{N}$$

$$(ii)\ \tfrac{1}{2}(\beta - \alpha) \sup_{n \varepsilon \mathfrak{N}} \Lambda\left\{\left[1_N + \tfrac{1}{2}(\alpha + \beta)K\left(\frac{i2\pi n}{T}\right)\right]^{-1} K\left(\frac{i2\pi n}{T}\right)\right\} < 1.$$

*Then there exists a unique $\hat{f} \,\varepsilon\, \mathcal{K}_N$ such that*

$$g_1(t) = \hat{f}(t) + \int_{-\infty}^{t} k(t - \tau)\psi[\hat{f}(\tau),\tau]d\tau, \qquad -\infty < t < \infty.$$

*Proof of Lemma 4:*

Theorem 4 of Ref. 1 and the remarks relating to its proof imply that the conclusion of Lemma 4 is valid if the hypotheses of the lemma are satisfied and the condition

$$\sup_{n \varepsilon \mathfrak{N}} \Lambda\left\{\left[1_N + \tfrac{1}{2}(\alpha + \beta)K\left(\frac{i2\pi n}{T}\right)\right]^{-1}\right\} < \infty \tag{13}$$

is met. However, since every element of $K(i2\pi n/T)$ approaches zero as $|\, n \,| \to \infty$, assumption $(i)$ of Lemma 1 implies that

$$\inf_{n \varepsilon \mathfrak{N}} \left| \det\left[1_N + \tfrac{1}{2}(\alpha + \beta)K\left(\frac{i2\pi n}{T}\right)\right]\right| > 0.$$

Therefore, in view of the fact that the elements of $K(i2\pi n/T)$ are uniformly bounded for $n \,\varepsilon\, \mathfrak{N}$, it follows that (13) is satisfied. This proves the lemma.

*Lemma 5: Let $\psi[\cdot,\cdot]$ satisfy the conditions of Theorem 3, let $\hat{f} \,\varepsilon\, \mathcal{K}_N$, and suppose that assumption (ii) of Theorem 3 is satisfied. Then*

$$\int_{-\infty}^{t} k(t - \tau)\psi[\hat{f}(\tau),\tau]d\tau \;\varepsilon\; \mathfrak{L}_{\infty N}\,.$$

*Proof of Lemma 5:*

Let $q(t) = \psi[\hat{f}(t),t]$, and

$$u = (u_1, u_2, \cdots, u_N)' = \int_{-\infty}^{t} k(t - \tau)q(\tau)d\tau.$$

Then $q \,\varepsilon\, \mathcal{K}_N$, and

$$| u_m(t) | \leqq \sum_{m=1}^{N} \int_{-\infty}^{t} | k_{mn}(t - \tau) | \cdot | q_n(\tau) | \, d\tau$$

$$\leqq \sum_{m=1}^{N} \int_{0}^{\infty} | k_{mn}(\tau) | \cdot | q_n(t - \tau) | \, d\tau.$$

Furthermore,

$$\left| \int_0^\infty | k_{mn}(\tau) | \; | q_n(t - \tau) | \, d\tau \right|^2$$

$$\leqq \int_0^\infty [(1 + \tau) k_{mn}(\tau)]^2 d\tau \int_0^\infty \left| \frac{q_n(t - \tau)}{1 + \tau} \right|^2 d\tau,$$

and the last integral can be bounded as follows

$$\int_0^\infty \left| \frac{q_n(t - \tau)}{1 + \tau} \right|^2 d\tau = \sum_{m=0}^{\infty} \int_{mT}^{(m+1)T} \left| \frac{q_n(t - \tau)}{1 + \tau} \right|^2 d\tau$$

$$\leqq \left( 1 + \sum_{n=1}^{\infty} (mT)^{-2} \right) \int_0^T | q_n(t) |^2 \, dt.$$

Thus, the $u_m(t)$ are uniformly bounded on $(-\infty, \infty)$, which proves the lemma.

Theorem 3 follows at once from Lemmas 4 and 5, Theorem 1, and the fact that assumption $(i)$ of Theorem 3 and $\hat{f} \, \varepsilon \, \mathfrak{L}_{\infty N}$ imply that

$$\int_{-\infty}^{0} k(t - \tau) \psi[\hat{f}(\tau), \tau] d\tau \; \varepsilon \; \mathfrak{L}_{2N+} .$$

REFERENCES

1. Sandberg, I. W., On Truncation Techniques in the Approximate Analysis of Periodically Time-Varying Nonlinear Networks, IEEE Trans. Circuit Theory, **CT-11**, No. 2, June, 1964, p. 195.
2. Sandberg, I. W., On the $\mathfrak{L}_2$-Boundedness of Solutions of Nonlinear Functional Equations, B.S.T.J., **43**, July, 1964, p. 1581.

# Applications of a Theorem of Dubrovskii to the Periodic Responses of Nonlinear Systems

By V. E. BENEŠ and I. W. SANDBERG

(Manuscript received May 14, 1964)

*Dubrovskii's theorem on completely continuous operators that are asymptotic to zero is applied to the study of the existence and uniqueness of periodic responses of nonlinear systems to periodic driving signals. Examples of nonexistence and nonuniqueness are given, a relationship between nonuniqueness and subharmonics is noted, and some general existence theorems are proven, giving estimates on the magnitudes of the harmonics.*

## I. INTRODUCTION

In 1939 V. M. Dubrovskii[1] proved the following result:

*Theorem 1: If $A$ is a completely continuous operator which maps a Banach space $X$ into itself, with the property that*

$$\lim_{\|x\|\to\infty} \frac{\| Ax \|}{\| x \|} = 0, \qquad x \,\varepsilon\, X,$$

*then for each scalar $\lambda$ and $y \,\varepsilon\, X$, the equation $x = y + \lambda Ax$ has at least one solution $x \,\varepsilon\, X$.*

Dubrovskii's theorem was stated in the long review article of M. A. Krasnoselskii[2] on problems of nonlinear analysis, but except for a recent application,[3] it seems to have gone largely unnoticed. It is the purpose of this paper to indicate some applications of the basic idea in the theorem to integral equations (and systems thereof) that arise in the study of nonlinear electrical networks and automatic control systems.

The applications to be made all center around the existence and uniqueness of periodic responses of nonlinear systems to periodic driving signals. These properties of the equations governing nonlinear systems are frequently taken for granted. The fact is, though, that these are by no means universal properties of such equations, as simple examples (to be given) will show. Often, the nonexistence of periodic responses is related to instability of the nonlinear system, while their lack of uniqueness is closely connected with the possibility of responses with subharmonic components. Thus it is important, in control and circuit theory, to be able to distinguish nonlinear equations that have unique periodic solutions for periodic inputs from those that possess several such solutions. With the aid of the idea underlying Dubrovskii's theorem, we examine this problem in the present paper for systems described by the nonlinear integral equation

$$x(t) = y(t) + \int_{-\infty}^{\infty} k(t-u)\psi(x(u),u)\,du, \tag{1}$$

(and by a vector analog thereof,) where $y(\cdot)$ is an input, $k(\cdot)$ is an integrable ($L_1$) impulse response of a linear system, and $\psi(\cdot,\cdot)$ represents a periodically time-varying nonlinear element. Periodic solutions of (1) have already been considered in previous work of one of the authors;[4] almost periodic solutions of (1) have been studied in previous joint work[5] of the authors. In both these papers a basically different assumption about the growth of the element $\psi(\cdot,\cdot)$ (from that to be made here) was used.

## II. SUMMARY

A discussion of the abstract Banach space setting for Dubrovskii's theorem appears in Section III. It includes a quick proof of the theorem from Schauder's fixed-point principle. There follows in Section IV an account of mathematical preliminaries, assumptions, definitions, etc., requisite for our remarks about (1). These remarks begin, in Section V, with a simple example showing that (1) may have no periodic solution and continue in Section VI with an existence theorem, for periodic solutions of (1), based on the principle of Dubrovskii's theorem. In Section VII we apply this result in discussing an example of nonuniqueness due to existence of subharmonic solutions. In Section VIII it is shown how the bound on the norm of the solutions obtained in Section VII can be improved. In Section IX, finally, a vector analog of the existence theorem of Section VII is stated and its proof sketched.

## III. BACKGROUND DISCUSSION

We recall[6] that an operator $A$ taking one Banach space into another is termed *completely continuous* if and only if it is continuous and carries every bounded set into a compact one. Dubrovskii's theorem for such operators is a straightforward consequence of Schauder's fixed-point principle:[7] Let $S$ be a bounded, closed, convex set of a Banach space $X$. Let $A$ be a continuous transformation of $S$ into a compact subset of itself. Then there exists at least one point $x \varepsilon S$ such that $x = Ax$.

An operator $A$ satisfying Dubrovskii's condition

$$\lim_{\| x \| \to \infty} \frac{\| Ax \|}{\| x \|} = 0$$

is said to be *asymptotically close to zerc;* explicitly, the condition is that for every $\epsilon > 0$ there is an $r$ such that $\| x \| \geqq r$ implies $\| Ax \| < \epsilon \| x \|$. To prove Dubrovskii's theorem we seek a closed ball, of radius $R$ to be determined, that is mapped into itself by the (completely continuous) operator $G$ defined by

$$Gx = y + \lambda Ax$$

with $\lambda$ and $y \varepsilon X$ fixed. Let $\epsilon$ be a number such that $0 < | \lambda | \epsilon < 1$, and pick (by Dubrovskii's condition) an $r > 0$ such that $\| x \| \geqq r$ implies $\| Ax \| < \epsilon \| x \|$. If now $s$ is a positive number such that

$$s \geqq \frac{\| y \|}{1 - \epsilon | \lambda |}$$

then for $r \leqq \| x \| \leqq s$

$$\begin{aligned} \| Gx \| &\leqq \| y \| + | \lambda | \cdot \| Ax \| \\ &\leqq \| y \| + | \lambda | \epsilon \| x \| \\ &\leqq s. \end{aligned}$$

Since $A$ is completely continuous, the set

$$\{ Ax \colon \| x \| \leqq r \}$$

is compact. Thus the continuous function $\| Ax \|$ defined on $\{ \| x \| \leqq r \}$ is bounded. If $R$ is chosen as

$$R = \max \left\{ \frac{\| y \|}{1 - | \lambda | \epsilon}, \| y \| + | \lambda | \sup_{\| x \| \leqq r} \| Ax \| \right\}$$

then $\| x \| \leqq R$ implies $\| Gx \| \leqq R$. The closed ball of radius $R$ is convex, and the existence of a fixed point of $G$ in the ball follows from Schauder's fixed-point principle. To establish the result for a particular value of $\lambda$ it is not necessary that $A$ be asymptotically close to zero; clearly, it suffices that there be $\epsilon$ such that $0 < \epsilon < | \lambda |^{-1}$ and $r$ such that $\| x \| > r$ implies $\| Ax \| < \epsilon \| x \|$.

## IV. PRELIMINARIES

We shall be concerned throughout with the case in which the functions $x(\cdot)$ and $y(\cdot)$ of interest are periodic and square-integrable over a period. By $L_2(-T,T)$ we denote the Banach space of all functions $x(\cdot)$ of period $2T$ that are real-valued, measurable on $[-T,T]$, and for which the norm

$$\| x \| = \left(\frac{1}{2T} \int_{-T}^{T} | x(t) |^2 \, dt\right)^{\frac{1}{2}}$$

is finite. According to standard results in the theory of Fourier series, such a function is represented in the mean by its Fourier series

$$x(t) = \underset{N\to\infty}{\text{l.i.m.}} \sum_{m=-N}^{N} x_m e^{\pi imt/T}$$

with Fourier coefficients

$$x_m = \frac{1}{2T} \int_{-T}^{T} x(t) e^{-\pi imt/T} \, dt, \qquad -\infty < m < \infty .$$

The norm of $x(\cdot)$ and its Fourier coefficients are related by the Parseval identity

$$\| x \|^2 = \sum_{n=-\infty}^{\infty} | x_n |^2.$$

We shall need the following two facts from the theory of Fourier series: (1) If $z(\cdot)$, $w(\cdot) \varepsilon L_2(-T,T)$, with respective Fourier coefficients $\{z_n\}$, $\{w_n\}$, then

$$\frac{1}{2T} \int_{-T}^{T} z(t - u) w(u) \, du = \sum_n z_n w_n e^{i\pi nt/T},$$

the series on the right converging absolutely and uniformly; (2) the Fourier coefficients of $z(\cdot + \epsilon)$ are $\{e^{\pi in\epsilon/T} z_n\}$.

The notation

$$\| z \|_1 = \frac{1}{2T} \int_{-T}^{T} | z(t) | \, dt$$

is used occasionally.

For a periodic function $z(\cdot) \varepsilon L_2(-T,T)$ we define the functional

$$\mu(z,\epsilon) = \frac{1}{4T} \int_{-T}^{T} | z(t + \epsilon) - z(t) | \, dt,$$

proportional to an "integral modulus of continuity," and we remark that one of the usual arguments for the Riemann-Lebesgue lemma gives, for $n \neq 0$, the inequality

$$\begin{aligned} | z_n | &= \left| \frac{1}{2T} \int_{-T}^{T} z(t) e^{-\pi i n t/T} \, dt \right| \\ &= \left| \frac{1}{4T} \int_{-T}^{T} z(t)[1 - e^{\pi i}] e^{-\pi i n t/T} \, dt \right| \\ &\leqq \frac{1}{4T} \int_{-T}^{T} | z(t) - z(t + T/n) | \, dt = \mu(z,T/n) . \end{aligned}$$

We shall make two assumptions about the nonlinear element $\psi(\cdot\,,\cdot)$, one about its growth and one about its continuity:

(a) there is a function $\lambda(\cdot)$ nondecreasing on $[0,\infty)$ such that for all $v$, $t$

$$| \psi(v,t) | \leqq \lambda( | v | ), \tag{2}$$

(b) the function $\psi(\cdot\,,\cdot)$ is continuous in the first variable uniformly in both variables. Then its *modulus* of continuity $\omega(\cdot)$, defined by

$$\omega(\delta) = \sup_{u,v,t} | \psi(u,t) - \psi(v,t) | \quad \text{for} \quad | u - v | \leqq \delta, \tag{3}$$

is a continuous monotone function, and approaches zero with $\delta \to 0$. When $\psi(v,\cdot)$, considered as a function of $t$, has a modulus of continuity $\omega_0(\cdot)$, so that

$$| \psi(v,t + \epsilon) - \psi(v,t) | \leqq \omega_0(\epsilon)$$

for all $v$ and $t$, we set

$$q_n = \begin{cases} 0 & n = 0 \\ \omega_0(T/n) & n \neq 0. \end{cases}$$

Jensen's inequality for a concave function $\varphi(\cdot)$ reads

$$\varphi\left( \frac{\displaystyle\int_a^b f(x) p(x) \, dx}{\displaystyle\int_a^b p(x) \, dx} \right) \geqq \frac{\displaystyle\int_a^b \varphi(f(x)) p(x) \, dx}{\displaystyle\int_a^b p(x) \, dx} \tag{4}$$

where $\varphi(\cdot)$ is concave in an interval containing the range of $f(\cdot)$ over $[a,b]$, $p(x) \geqq 0$, $p \not\equiv 0$, and all the integrals in question exist.

We now return to $k(\cdot)$ in (1). Since $k(\cdot)$ belongs to $L_1$, it has a bounded Fourier transform

$$K(\omega) = (2\pi)^{-\frac{1}{2}} \int_{-\infty}^{\infty} e^{-i\omega t} k(t)\, dt.$$

The convolution operator $K$ on $L_2(-T,T)$ defined by

$$Kx(t) = \int_{-\infty}^{\infty} x(t - u) k(u)\, du$$

is described in terms of its effect on Fourier coefficients by the identity

$$(Kx)_m = (2\pi)^{\frac{1}{2}} K\left(\frac{\pi m}{T}\right) x_m ,$$

and takes $L_2(-T,T)$ into itself continuously.

## V. NONEXISTENCE OF SOLUTIONS

It is easy to see that in some very simple cases (1) has no periodic solution. An example is furnished by

*Remark 1:* If $\psi(v,t) = v$ for all $v$ and $t$, and if, for some integer $n$, both the $n$th Fourier coefficient $y_n$ of $y(\cdot)$ does not vanish and

$$h_n = (2\pi)^{\frac{1}{2}} K\left(\frac{n\pi}{T}\right) = 1,$$

then (1) has no periodic solution $x(\cdot)$ belonging to $L_2\,(-T,T)$. For if there were such a solution, the left side of (1) would have $n$th Fourier coefficient $x_n$ , while the right-side would have $y_n + x_n \neq x_n$ .

## VI. EXISTENCE OF SOLUTIONS

*Theorem 2: If $\lambda(\cdot)$ and $\omega(\cdot)$ are concave, $y(\cdot)\ \varepsilon\ L_2(-T,T)$,*

$$\kappa^2 = 2\pi \sum_{m=-\infty}^{\infty} \left| K\left(\frac{\pi m}{T}\right)\right|^2 < \infty$$

*and if the scalar equation*

$$r = \| y \| + \kappa\lambda(r) \tag{5}$$

*has a positive solution $r$, then there exists a solution $x(\cdot)$ of (1), with period $2T$, and such that*

$$\| x \| \leqq r,$$

$$| x_m | \leqq | y_m | + (2\pi)^{\frac{1}{2}} K\left(\frac{\pi m}{T}\right) \lambda(r)$$

*where $x_m$ , $y_m$ are the respective* mth *Fourier coefficients of $x$, $y$.*

*Proof:* In the complex sequence space $l_2$ of Fourier coefficients, isometric to $L_2(-T,T)$, consider the set $\mathcal{S}$ of sequences $x = \{x_n, -\infty < n < \infty\}$ such that, with the overbar denoting the complex conjugate,

$$x_{-m} = \bar{x}_m$$

$$| x_m | \leqq | y_m | + \left| K\left(\frac{\pi m}{T}\right) \right| \lambda(r).$$

By Minkowski's inequality, $x \ \varepsilon \ \mathcal{S}$ implies

$$\| x \| = (\sum_m | x_m |^2)^{\frac{1}{2}} \leqq \| y \| + \kappa\lambda(r) = r.$$

The set $\mathcal{S}$ is compact, being an analog of the Hilbert cube or parallelotope. It is easily verified that $\mathcal{S}$ is convex.

Now let $x(\cdot) \ \varepsilon \ L_2(-T,T)$, and consider the magnitudes of the Fourier coefficients of the function $w(\cdot)$ defined by

$$w(t) = \psi(x(t),t) = \psi(x(t+2T),\, t+2T).$$

We find

$$\begin{aligned} | w_n | = \left| \frac{1}{2T} \int_{-T}^{T} \psi(x(t),t) e^{-\pi i n t/T} \, dt \right| &\leqq \frac{1}{2T} \int_{-T}^{T} | \psi(x(t),t) | \, dt \\ &\leqq \frac{1}{2T} \int_{-T}^{T} \lambda(| x(u) |) \, dt \qquad (6) \\ &\leqq \lambda\left( \frac{1}{2T} \int_{-T}^{T} | x(u) | \, du \right) \\ &\leqq \lambda(\| x \|), \end{aligned}$$

where the second inequality follows from the fact that $\lambda(\cdot)$ bounds the growth of $\psi(\cdot,t)$, the third inequality follows from the concavity of $\lambda(\cdot)$ by the Jensen inequality (4), and the fourth inequality follows from Schwarz's and the monotone nature of $\lambda(\cdot)$. Hence if $\| x \| \leqq r$, then $| y_m + (Kw)_m | \leqq | y_m | + (2\pi)^{\frac{1}{2}} K(m\pi/T)\, \lambda(r)$, and it follows that the operator $A$ defined on $L_2(-T,T)$ by

$$Ax(t) = y(t) + \int_{-\infty}^{\infty} k(t-u)\psi(x(u),u) du, \qquad | t | \leqq T$$

maps the ball $\| x \| \leqq r$ into the compact, convex, isometric image of $\mathcal{S}$, that is, into a compact, convex subset of itself. Continuity of $A$ *on the image* follows from that of $K$ and from the inequality

$$\frac{1}{2T}\int_{-T}^{T} |\psi(x_1(u),u) - \psi(x_2(u),u)|\, du \leqq \omega(\| x_1 - x_2 \|),$$

provable by the same method as (6). Existence of a fixed point of $A$ in the isometric image of $\mathcal{S}$ follows from Schauder's theorem.

We remark that if $\lambda(u) = o(u)$ as $u \to \infty$, then a solution $r$ to the scalar equation (5) always exists. This occurs, for example, if

$$\lambda(u) = \beta u^{\alpha} \qquad \beta > 0, \qquad 0 \leqq \alpha < 1.$$

## VII. NONUNIQUENESS AND SUBHARMONICS

It is known that a solution of (1) may have Fourier components, called "subharmonics," of period greater than $2T$. Our purpose is to remark that if this occurs, then (1) does not have a unique solution, and in fact, the greater the period of the subharmonic, the more distinct solutions exist. We start with a simple example: Let $T = \pi/2$, set

$$\left.\begin{aligned} \psi(u,t) &= \operatorname{sgn} u \cdot |u|^{\frac{1}{2}} \\ y(t) &= \tfrac{9}{2} - \tfrac{1}{2}\cos 2t \end{aligned}\right\}, \tag{7}$$

let

$$x(t) = \tfrac{9}{2} + 4\sin t - \tfrac{1}{2}\cos 2t,$$

and for $K(\cdot)$ take any Fourier transform of an integrable function with $K(0) = 0$ and $K(1) = 4(2\pi)^{-\frac{1}{2}}$. For example, the fourth-order filter

$$K(\omega) = \frac{(2\pi)^{-\frac{1}{2}}16(i\omega)^2}{(1 + i\omega)^4}$$

will do. Actually, since we need to prescribe only the two parameters $K(0)$ and $K(1)$, the second-order filter

$$K(\omega) = \frac{(2\pi)^{-\frac{1}{2}} i\omega}{(i\omega)^2 + \frac{1}{4}i\omega + 1} \tag{8}$$

would do as well.

That $x(\cdot)$ as defined is a periodic solution of (1) of period $2\pi$ can be verified from the identity

$$2 + \sin t = (\tfrac{9}{2} + 4\sin t - \tfrac{1}{2}\cos 2t)^{\frac{1}{2}}.$$

This example, in which the solution $x(\cdot)$ contains the subharmonic component $4 \sin t$, is adapted from Hughes,[8] and has been used earlier[4] by the authors merely to illustrate the real possibility of subharmonics in relatively simple systems.

Now since the input $y(\cdot)$ has period $\pi$, while the response $x(\cdot)$ has period $2\pi$, it can be seen that by shifting $x(\cdot)$ by $\pm\pi$, that is, by changing the sign of (all) the odd components of $x(\cdot)$, another solution of (1) for this $\psi(\cdot)$ and $y(\cdot)$ is obtained, because

$$x(t \pm \pi) = y(t) + \int_{-\infty}^{\infty} \operatorname{sgn} x(t \pm \pi - u) \mid x(t \pm \pi - u) \mid^{\frac{1}{2}} k(u) du.$$

Thus, there are at least two solutions of (1) for this example; the two we have identified so far differ only in phase. As an application of Theorem 2 we show that there is at least one more solution, one that has period $\pi$. The following lemma establishes a Hölder condition for the nonlinearity of the example:

*Lemma 1: If*

$$\psi(v) = sgn\ v \mid v \mid^{\frac{1}{2}}$$

*then for all $v$ and $\epsilon$*

$$\mid \psi(v + \epsilon) - \psi(v) \mid \leqq 2^{\frac{1}{2}} \mid \epsilon \mid^{\frac{1}{2}}.$$

*Proof:* First suppose that $\operatorname{sgn}(v + \epsilon) \neq \operatorname{sgn} v$. Then $\mid \epsilon \mid = \mid v + \epsilon \mid + \mid v \mid$, and concavity gives, by Jensen's theorem,

$$\begin{aligned} \mid \psi(v + \epsilon) - \psi(v) \mid &= \mid v + \epsilon \mid^{\frac{1}{2}} + \mid v \mid^{\frac{1}{2}} \\ &\leqq 2 \left( \frac{\mid v + \epsilon \mid + \mid v \mid}{2} \right)^{\frac{1}{2}} = 2^{\frac{1}{2}} \mid \epsilon \mid^{\frac{1}{2}}. \end{aligned}$$

If $\operatorname{sgn}(v + \epsilon) = \operatorname{sgn} v$, there is no loss of generality in supposing that $v + \epsilon > v \geqq 0$, because $\psi(\cdot)$ is odd. Then using concavity again

$$\psi(v) \geqq \frac{v}{v + \epsilon} \psi(v + \epsilon)$$

$$\psi(\epsilon) \geqq \frac{\epsilon}{v + \epsilon} \psi(v + \epsilon).$$

Hence in this case

$$0 \leqq \psi(v + \epsilon) - \psi(v) \leqq \psi(\epsilon) = \mid \epsilon \mid^{\frac{1}{2}}.$$

A direct application of Theorem 2 shows that (1) for the example (7), (8) has a solution of period $\pi$. The scalar equation

$$\| y \| + \kappa r^{\frac{1}{2}} = r$$

is appropriate, and has a positive root $r$.

The example just discussed illustrates the following general principle regarding subharmonics:

*Theorem 3: If* (1) *has a solution* $x(\cdot)$ *with (minimal) period* $2nT$, $n > 1$, *then each of the functions*

$$x(t + 2kT) \qquad k = 1, \cdots, n$$

*is a (distinct) solution of* (1).

*Proof:* Since $y(\cdot)$, and the time-dependence of $\psi(\cdot,\cdot)$ have period $2T$, we have

$$\begin{aligned} x(t + 2kT) &= y(t) + \int_{-\infty}^{\infty} \psi(x(t + 2kT - u), t + 2kT - u)k(u)du \\ &= y(t) + \int_{-\infty}^{\infty} \psi(x(t + 2kT - u), t - u)k(u)du. \end{aligned}$$

## VIII. CLOSER BOUNDS ON FOURIER COEFFICIENTS

By a more penetrating analysis it is possible to strengthen the bounds on the norm and on the Fourier coefficients given by Theorem 2. For example, the inequality (6) merely establishes a uniform bound $\lambda(r)$ for all Fourier coefficients of functions

$$w(t) = \psi(x(t),t)$$

for $\| x \| \leqq r$. However, since the argument for (6) shows that $w(\cdot)$ is absolutely integrable over a period, its Fourier coefficients actually go to zero at infinity, and it should be possible substantially to improve the estimate (6). This can be done with the help of the quantities $\{q_m, -\infty < m < \infty\}$, and the functional $\mu$, defined in Section IV.

Throughout this section, it is assumed that $\psi(v,\cdot)$ has the modulus of continuity $\omega_0(\cdot)$ as a function of $t$, and that

$$\kappa^2 = \sum_{m=-\infty}^{\infty} \left| K\left(\frac{m\pi}{T}\right) \right|^2 < \infty. \tag{9}$$

It follows from (9) that there is a function $h(\cdot) \varepsilon L_2(-T,T)$ such that for any $x(\cdot) \varepsilon L_2(-T,T)$

$$\frac{1}{2T}\int_{-T}^{T} h(t - u)x(u)du = \int_{-\infty}^{\infty} k(t - u)x(u)du;$$

the Fourier coefficients of $h(\cdot)$ are

$$h_m = (2\pi)^{\frac{1}{2}}K\left(\frac{m\pi}{T}\right), \qquad -\infty < m < \infty.$$

For each positive number $s$, and $m \neq 0$, we define

$$a_m(s) = \mu(y,T/m) + \lambda(s)\mu(h,T/m).$$

Since for $x(\cdot) \varepsilon L_2(-T,T)$, $\mu(x,\epsilon) \to 0$ as $\epsilon \to 0$, the numbers $a_m(s)$ are bounded in $m \neq 0$ for each fixed $s$. By the Riesz-Fischer theorem there is for each $s > 0$ a function $u_s(\cdot) \varepsilon L_2(-T,T)$ whose Fourier coefficients are

$$u_m(s) = \begin{cases} \frac{1}{2} \mid h_m \mid \{\omega(a_m(s)) + q_m\}, & m \neq 0 \\ \mid h_0 \mid \lambda(s) & m = 0. \end{cases}$$

with (cf. Section IV)

$$q_n = \begin{cases} 0 & n = 0 \\ \omega_0(T/n) & n \neq 0. \end{cases}$$

*Theorem 4: Let $\lambda(\cdot)$, $\omega(\cdot)$ be concave, and let $r$ be a positive number satisfying the inequality*

$$\| y \| + \| u_r \| \leqq r.$$

*Then there exists a solution $x(\cdot) \varepsilon L_2(-T,T)$ of (1), such that*

$$\begin{aligned} \| x \| &\leqq r, & \\ \mu(x,T/m) &\leqq a_m(r), & m \neq 0 \\ \mid x_m \mid &\leqq \mid y_m \mid + \mid u_m(r) \mid, & \text{all } m. \end{aligned}$$

*Proof:* Let the operator $A$ be defined on the ball $\{ \| x \| \leqq r\}$ in $L_2(-T,T)$ by

$$\begin{aligned} Ax(t) &= y(t) + \int_{-\infty}^{\infty} k(t-u)\psi(x(u),u)du \\ &= y(t) + \frac{1}{2T}\int_{-T}^{T} h(t-u)\psi(x(u),u)du. \end{aligned}$$

The argument of Theorem 2 shows that $A$ maps $\{ \| x \| \leqq r\}$ continuously into $L_2(-T,T)$. Further, by Fubini's theorem and the concavity of $\lambda(\cdot)$,

$$\begin{aligned} \frac{1}{8T^2}\int_{-T}^{T} dt \int_{-T}^{T} \mid h(t+\epsilon-u) - h(t-u)\mid \cdot \mid\psi(x(u),u)\mid du \\ = \mu(h,\epsilon)\cdot\frac{1}{2T}\int_{-T}^{T} \mid \psi(x(u),u)\mid du \\ \leqq \mu(h,\epsilon)\lambda(\| x \|), \end{aligned}$$

and we find that $\| x \| \leqq r$ implies $\mu(Ax,T/m) \leqq a_m(r)$ for $m \neq 0$. Moreover

$$| \psi(x(u+\epsilon),u+\epsilon) - \psi(x(u),u) | \leqq \omega( | x(u+\epsilon) - x(u) | ) + \omega_0(\epsilon),$$

so that the concavity of $\omega(\cdot)$ implies

$$2\mu(\psi(x),\epsilon) \leqq \omega(\mu(x,\epsilon)) + \omega_0(\epsilon).$$

It follows that $\| x \| \leqq r$ implies

$$| (Ax)_m | \leqq | y_m | + \tfrac{1}{2} | h_m | \{\omega(a_m(r)) + q_m\}, \qquad m \neq 0$$

and also

$$| (Ax)_0 | \leqq | y_0 | + | h_0 | \lambda(r).$$

Let $\mathcal{S}$ be the compact set of $l_2$ sequences

$$x = \{x_n , -\infty < n < \infty\}$$

such that

$$x_{-m} = \bar{x}_m ,$$

$$| x_m | \leqq | y_m | + | u_m(r) |.$$

It can be seen that $A$ maps the ball $\{ \| x \| \leqq r\}$ into the isometric image in $L_2(-T,T)$ of $\mathcal{S}$. This image is compact and convex, and Theorem 4 follows from Schauder's fixed-point principle, as did Theorem 2.

*Theorem 5: Let $\lambda(\cdot),\omega(\cdot)$ be concave, let $y(\cdot)$ ε $L_2(-T,T)$, and let there exist a positive number $r$ and a real bounded sequence $b = \{b_m , m \neq 0\}$ satisfying the inequalities*

$$\| y \|_1 + \| h \|_1\lambda(r) \leqq r$$

$$\mu(y,T/m) + \sum_{n\neq 0} \left| \sin \frac{n\pi}{2m} \right| | h_n | \{\omega(b_n) + q_n\} \leqq b_m , \qquad m \neq 0.$$

Then there exists a solution $x(\cdot)$ ε $L_2(-T,T)$ of (1) such that

$$| x_m | \leqq | y_m | + \tfrac{1}{2} | h_m | \{\omega(b_n) + q_n\}, \qquad m \neq 0.$$

$$\| x \|_1 \leqq r$$

*Proof:* Let $R$ be the compact, convex subset of $L_2(-T,T)$ consisting of functions $z(\cdot)$ such that

$$\| z \|_1 \leqq r$$

$$2\mu(z,T/m) \leqq b_m , \qquad m \neq 0.$$

Let $x(\cdot) \varepsilon R$, and let $\{\psi_n, -\infty < n < \infty\}$ be the Fourier coefficients of the function $w(\cdot)$ defined by

$$w(t) = \psi(x(t),t), \qquad \text{all } t.$$

Then the concavity of $\lambda(\cdot)$ implies that

$$|\psi_0| \leqq \lambda\left(\frac{1}{2T}\int_{-T}^{T} |x(t)|\,dt\right) \leqq \lambda(r)$$

and that of $\omega(\cdot)$ implies that for $m \neq 0$

$$\begin{aligned}|\psi_m| &\leqq \frac{1}{2T}\int_{-T}^{T} \omega(|x(t+T/m) - x(t)|)dt + q_m \\ &\leqq \omega(2\mu(x,T/m)) + q_m .\end{aligned}$$

Now

$$\begin{aligned}&Ax(t+\epsilon) - Ax(t) \\ &\quad = y(t+\epsilon) - y(t) + \int_{-T}^{T} \{h(t+\epsilon-u) - h(t-u)\}\psi(x(u),u)du\end{aligned}$$

and the second term on the right is

$$\sum_{n\neq 0} h_n(e^{\pi i n\epsilon/T} - 1)\psi_n e^{\pi i n t},$$

the series converging absolutely and uniformly to a quantity of modulus at most

$$2\sum_{n\neq 0} |h_n| \left|\sin\frac{\pi n\epsilon}{2T}\right| |\psi_n| .$$

Hence, with $\epsilon = T/m$, $m \neq 0$,

$$2\mu(Ax,T/m) \leqq b_m .$$

At the same time

$$\begin{aligned}\frac{1}{2T}\int_{-T}^{T} |Ax(t)|\,dt &\leqq \frac{1}{2T}\int_{-T}^{T} |y(t)|\,dt + \frac{1}{4T^2}\int_{-T}^{T} |h(t-u)|\,|\psi(x(u),u)|du\,dt \\ &\leqq \|y\|_1 + \|h\|_1\lambda(r) \\ &\leqq r.\end{aligned}$$

Thus $Ax(\cdot)$ belongs to $R$. The result follows by Schauder's theorem, as before.

*Remark:* Let $\omega(\cdot)$ be concave, with $\omega(u) = o(u)$ as $u \to \infty$. Let $\{z_n, n \neq 0\}$ belong to $l_2$, and let $\{h_n, n \neq 0\}$ belong to $l_1 \cap l_2$. Then there exists a *minimal* bounded sequence $\{b_n, n \neq 0\}$ satisfying

$$|z_n| + \sum_{m \neq 0} \left| \sin \frac{\pi m}{2n} \right| |h_m| \, \omega(b_m) \leqq b_n, \qquad n \neq 0. \tag{10}$$

The sequence $\{b_n\}$ is minimal in the sense that its components are less than or equal to the corresponding components of any other sequence satisfying (10).

Let $B$ be the set of sequences $v$ satisfying (10). To prove $B$ is nonempty, let

$$u_n = \begin{cases} 0 & n = 0 \\ \displaystyle\sum_{m \neq 0} \left| \sin \frac{\pi m}{2r} \right| |h_m|, & n \neq 0, \end{cases}$$

and let $r$ satisfy $\| z \| + \| u \| \omega(r) \leqq r$. Define $w = \{w_n, n \neq 0\}$ by

$$w_n = |z_n| + u_n \omega(r) \leqq r.$$

Then

$$|z_n| + \sum_{m \neq 0} \left| \sin \frac{\pi m}{2n} \right| |h_m| \, \omega(w_m) \leqq w_n,$$

so that $w \, \varepsilon \, B$ and is bounded. Now set $b_n = \inf\limits_{v \varepsilon B} v_n$. For any $v \, \varepsilon \, B$

$$|z_n| + \sum_{m \neq 0} \left| \sin \frac{\pi m}{2n} \right| |h_m| \, \omega(b_m) \leqq |z_n| + \sum_{m \neq 0} \left| \sin \frac{\pi m}{2n} \right| |h_m| \, \omega(v_m)$$
$$\leqq v_n .$$

Thus $b \, \varepsilon \, B$ and is minimal.

## IX. THE VECTOR EQUATION

In this final section, we consider a vector form of the integral equation (1). Let $k(\cdot)$ be an $N \times N$ *matrix* of real functions of $L_1$, and for each $t$, let $\psi(\cdot, t)$ be a real $N$-*vector* valued function of a real $N$-vector. Let $y(\cdot)$ be a real $N$-vector valued function of time $t$. With these reinterpretations of the notations in mind, we can leave (1) unchanged.

With $M$ a complex matrix, we let $M', \bar{M}$, and $M^*$ denote the transpose, the complex-conjugate, and the complex-conjugate-transpose, respectively, of $M$. The positive square-root of the largest eigenvalue of $M^*M$ is denoted by $\Lambda\{M\}$.

If $v$ is a real or complex $N$-vector, its norm is defined as the "Euclidean" norm

$$\| v \| = \left( \sum_{i=1}^{N} | v_i |^2 \right)^{\frac{1}{2}} = (v^* v)^{\frac{1}{2}}.$$

It is well-known that

$$\Lambda^2\{M\} = \sup_{\|v\|=1} v^* M^* M v \tag{11}$$

and hence that $\| Mv \| \leqq \Lambda\{M\} \| v \|$ for complex $N$-vectors $v$.

As previously, $L_2(-T,T)$ is the space of real-valued, measurable, functions $x(\cdot)$ of the real variable $t$ which satisfy

$(i)$ $x(t + 2T) = x(t)$,

$(ii)$ $\dfrac{1}{2T} \displaystyle\int_{-T}^{T} | x(t) |^2 \, dt < \infty$.

We take as our basic space the $N$th power of $L_2(-T,T)$, i.e.,

$$L_2^N(-T,T),$$

and think of it as composed of column $N$-vector valued functions of time. A norm for $L_2^N(-T,T)$ can be defined by the formula

$$\| x \|^2 = \frac{1}{2T} \int_{-T}^{T} x' x \, dt$$

$$= \frac{1}{2T} \int_{-T}^{T} \sum_{i=1}^{N} | x_i(t) |^2 dt$$

where $x = (x_1, \cdots, x_N)' \, \varepsilon \, L_2^N(-T,T)$. This norm makes $L_2^N(-T,T)$ a Banach space. Further, an element $x(\cdot)$ of $L_2^N(-T,T)$ has the Fourier representation

$$x(t) = \operatorname*{l.i.m.}_{n \to \infty} \sum_{m=-n}^{n} x_m e^{\pi i m t / T}$$

where the $N$-vector $x_m$ of $m$th Fourier coefficients is given by

$$x_m = \frac{1}{2T} \int_{-T}^{T} x(t) e^{-\pi i m t / T} \, dt,$$

and the Parseval identity

$$\sum_{m=-\infty}^{\infty} x_m^* x_m = \sum_{m=-\infty}^{\infty} \| x_m \|^2 = \| x \|^2$$

for $x \, \varepsilon \, L_2^N(-T,T)$ holds.

The matrix convolution operator $K$ is defined on $L_2^N(-T,T)$ by

$$Kx(t) = \int_{-\infty}^{\infty} k(t - u)x(u)du$$

and the operator $\psi$ by

$$\psi x(t) = \psi(x(t),t), \qquad \text{all } t.$$

Equation (1) assumes the concise form

$$x = y + K\psi x.$$

The matrix $K_m$, $m = 0, \pm 1, \cdots$ is defined by the condition

$$K_m = (k_{ij}^m) = \int_{-\infty}^{\infty} e^{-\pi imt/T} k(t)dt.$$

It is assumed that $\sum_m \Lambda^2\{K_m\} < \infty$. This condition is met, e.g., if

$$\sum_m | k_{ij}^m |^2 < \infty.$$

for $1 \leqq i,j \leqq N$. The matrix convolution operator $K$ takes a function $x(\cdot) \, \varepsilon \, L_2^N(-T,T)$ with (vector) Fourier coefficients $x_m$ into the function $z(\cdot)$ whose coefficients are

$$z_m = K_m x_m , \qquad m = 0, \pm 1, \cdots ,$$

and the Riesz-Fischer theorem guarantees that $z(\cdot) \, \varepsilon \, L_2^N(-T,T)$. Further, by formula (11) we have

$$\| z_m \| \leqq \Lambda\{K_m\} \| x_m \|.$$

An analog of Hilbert's cube in $L_2^N(-T,T)$ is described by

*Lemma 2: Let* $\{c_n , -\infty < n < \infty\}$ *be nonnegative real numbers with*

$$\sum_n c_n^2 < \infty.$$

*Then the set*

$$\{x \, \varepsilon \, L_2^N(-T,T): \| x_n \| \leqq c_n , \quad \text{all } n\}$$

*is compact.*

This result is a consequence of a known theorem. (See p. 136 of Ref. 5.)

Analogs of the growth condition (2) and of the uniform continuity condition (3) will be used. These are (*i*) $\| \psi(u,t \| \leqq \lambda( \| u \| )$, for all $t$ and all real $N$-vectors $u$ where $\lambda(\cdot)$ is a monotone function.

(*ii*) $\psi(\cdot,\cdot)$ is continuous in the first variable uniformly in both variables; its *modulus* of continuity $\omega(\cdot)$, defined by

$$\omega(\delta) = \sup_{u,v,t} \| \psi(u,t) - \psi(v,t) \| \quad \text{for} \quad \| u - v \| \leqq \delta,$$

is a continuous monotone function that approaches zero with $\delta$.

*Theorem 6: If $\lambda(\cdot)$ and $\omega(\cdot)$ are concave, $y$ belongs to $L_2^N(-T,T)$,*

$$\kappa^2 = \sum_m \Lambda^2\{K_m\} < \infty,$$

*and if the scalar equation*

$$r = \| y \| + \kappa N \lambda(r)$$

*has a positive solution $r$, then there exists an element $x \varepsilon L_2^N(-T,T)$ satisfying*

$$x = y + K\psi x$$

$$\| x \| \leqq r$$

$$\| x_m \| \leqq \| y_m \| + \Lambda\{K_m\} N \lambda(r),$$

*with $x_m$, $y_m$ the respective* m*th* (*vector*) *Fourier coefficients of $x,y$.*

The proof of Theorem 6 is an exact analog of that of Theorem 2, using the compact set

$$\{x \,\varepsilon\, L_2^N(-T,T): \| x_m \| \leqq \| y_m \| + \Lambda\{K_m\}\lambda(r), \quad \text{all } m\}$$

and with $w(t) = \psi(x(t),t)$, the inequality, (analogous to (6),)

$$\| w_m \| \leqq N\lambda(\| x \|),$$

provable by observing first that for all $t$

$$\sum_{j=1}^{N} | w_j(t) | \leqq N^{\frac{1}{2}} \| w(t) \|$$

$$\leqq N^{\frac{1}{2}} \lambda(\| x(t) \|),$$

so that trivially

$$| w_j(t) | \leqq N^{\frac{1}{2}} \lambda(\| x(t) \|)$$

and by concavity of $\lambda(\cdot)$,

$$\frac{1}{2T} \int_{-T}^{T} | w_j(t) | \, dt \leqq N^{\frac{1}{2}} \lambda(\| x \|).$$

Squaring both sides and summing over $j = 1, \cdots, N$ we obtain

$$\| w_m \|^2 \leqq N^2 \lambda^2(\| x \|).$$

REFERENCES

1. Dubrovskii, V. M., Sur Certaines Equations Integrales Non-Lineaires, Uc. Zap. Moskov. Gos. Univ., **30,** 1939, pp. 49–60.
2. Krasnoselskii, M. A., Some Problems of Nonlinear Analysis, Amer. Math. Soc. Translations, Ser. 2, **10,** 1958, pp. 345–409.
3. George M. D., Completely Well-Posed Problems for Nonlinear Differential Equations, Proc. Amer. Math. Soc., **15,** 1964, pp. 96–100.
4. Sandberg, I. W., On Truncation Techniques in the Approximate Analysis of Periodically Time-Varying Nonlinear Networks, IEEE Trans. PGCT, **CT-11,** 2, June, 1964.
5. Beneš, V. E., and Sandberg, I. W., On the Response of Time-Variable Nonlinear Systems to Almost Periodic Signals, to appear in J. Math. Anal. and Appl.
6. Lyusternik, L. A., and Sobolev, V. J., *Elements of Functional Analysis,* Ungar, New York, 1961, p. 129.
7. Simmons, G. S., *Introduction to Topology and Modern Analysis,* McGraw-Hill, New York, 1963, Appendix.
8. Hughes, W. L., *Nonlinear Electrical Networks,* Ronald Press, New York, 1960.

# Equivalence Relations among Spherical Mirror Optical Resonators

By J. P. GORDON and H. KOGELNIK

(Manuscript received May 11, 1964)

*The frequencies, field patterns, and losses of the resonant modes of spherical mirror optical resonators can be obtained to good accuracy as the solutions of the integral equations of Fresnel diffraction theory. By a simple transformation of the variables and parameters of the integral equations, we have found certain families of resonators which have the same diffraction loss at each mirror, and whose field patterns are scaled versions of each other. In the case of the infinite strip resonator, this reduces from five to three the number of parameters necessary to specify the losses and mode patterns.*

## I. INTRODUCTION

The resonant frequencies, field patterns, and losses of the modes of spherical mirror optical resonators can be obtained to good accuracy as the solutions of the integral equations of Fresnel diffraction theory.[1] The equations are particularly applicable when the separation between the two mirrors forming the resonator is large compared with the dimensions of the mirrors. Unfortunately, the equations are usually not soluble analytically, and require numerical (machine) computation. There are many parameters involved: the dimensions and curvatures of the mirrors and their separation. By a simple transformation of the variables and parameters of the integral equations, we have found certain families of resonators which have the same diffraction loss at each mirror, and whose field patterns are scaled versions of each other. In the case of the infinite strip resonator, this reduces from five to three the number of parameters necessary to specify the losses and mode patterns.

## II. THE TRANSFORMATION

The equations which determine the field patterns, resonant frequencies, and losses of an infinite strip resonator (see Fig. 1) are[1]

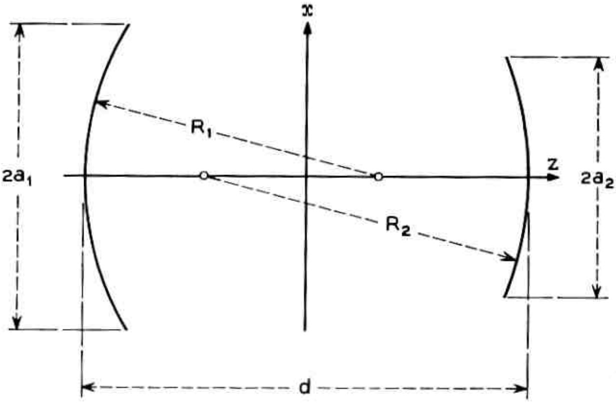


Fig. 1 — Spherical mirror resonator with mirrors of curvature radii $R_1$ and $R_2$, and of widths $2a_1$ and $2a_2$. Mirror spacing is $d$.

$$\gamma_1 u_1(x_1) = \frac{\sqrt{j}}{\sqrt{\lambda d}} \int_{-a_2}^{a_2} K(x_1, x_2)\; u_2(x_2) dx_2 \tag{1a}$$

$$\gamma_2 u_2(x_2) = \frac{\sqrt{j}}{\sqrt{\lambda d}} \int_{-a_1}^{a_1} K(x_2, x_1)\; u_1(x_1) dx_1 \tag{1b}$$

with the complex symmetric kernel

$$K(x_1, x_2) = K(x_2, x_1) = \exp\,[-j(\pi/\lambda d)(g_1 x_1^2 + g_2 x_2^2 - 2x_1 x_2)]. \tag{1c}$$

Here

$$g_i = 1 - (d/R_i), \qquad i = 1, 2,$$

the mirror separation is $d$, $R_1$ and $R_2$ are the radii of mirror curvature, $2a_1$, $2a_2$ are the corresponding mirrors widths, and $\lambda$ is the wavelength in the resonator medium. Also, $u_1(x_1)$ is the (generally complex) normalized field distribution on the left-hand mirror of Fig. 1, while $u_2(x_2)$ is the normalized field distribution on the right-hand mirror. If the two functions are normalized so that

$$\int_{-a_1}^{a_1} |\; u_1(x_1) \;|^2 \, dx_1 = \int_{-a_2}^{a_2} |\; u_2(x_2) \;|^2 \, dx_2 \tag{2}$$

then one notes* that the power reflection coefficient of the left mirror

* According to (1b) a light beam with a field distribution $u_1(x_1)$ across the left mirror causes a field $\gamma_2 u_2(x_2)$ across the right mirror. Therefore, the power reflected from this latter (perfectly reflecting) mirror is proportional to

is $|\gamma_1|^2$ and the reflection coefficient of the right mirror is $|\gamma_2|^2$. Therefore the loss at the left mirror is $1 - |\gamma_1|^2$, the loss at the right mirror is $1 - |\gamma_2|^2$, and the round-trip loss is $1 - |\gamma_1\gamma_2|^2$. The condition for resonance is that $\gamma_1\gamma_2 \exp[-j2\pi(d/\lambda)]$ be real and positive.

We presume on the weight of much experimental and theoretical[2,3] evidence that a sequence of solutions to (1) does exist. Suppose now that we have found a mode of some resonator; i.e., we have found a solution for $u_1(x_1)$ and $u_2(x_2)$ which satisfies (1) for one set of values of the five resonator parameters $a_1$, $a_2$, $g_1$, $g_2$ and $d$, and have found the corresponding eigenvalues $\gamma_1$ and $\gamma_2$. Our present concern is to find a family of resonators, each of which will have a similar mode; that is, a mode with the *same* values of $\gamma_1$ and $\gamma_2$ and with similar (scaled) eigenfunctions. For this purpose we rewrite (1) in terms of dimensionless variables and eigenfunctions by substituting

$$x_i = a_i\xi_i, \qquad i = 1, 2 \tag{3}$$

and

$$v_i(\xi_i) = u_i(x_i)\cdot\sqrt{a_i}, \qquad i = 1, 2. \tag{4}$$

By this transformation we obtain a generalized set of integral equations for the modes of the resonator

$$\gamma_1 v_1(\xi_1) = \sqrt{jN}\int_{-1}^{+1} d\xi_2\, v_2(\xi_2)\, K(\xi_1, \xi_2) \tag{5a}$$

$$\gamma_2 v_2(\xi_2) = \sqrt{jN}\int_{-1}^{+1} d\xi_1\, v_1(\xi_1)\, K(\xi_1, \xi_2) \tag{5b}$$

with the kernel

$$K(\xi_1, \xi_2) = \exp[-j\pi N(-2\xi_1\xi_2 + G_1\xi_1^2 + G_2\xi_2^2)]. \tag{5c}$$

In (5) only three independent resonator parameters occur

$$N \equiv a_1a_2/\lambda d, \tag{6a}$$

$$G_1 \equiv g_1(a_1/a_2), \tag{6b}$$

$$|\gamma_2|^2 \int_{-a_2}^{a_2} |u_2(x_2)|^2\, dx_2 .$$

The power of the beam as it left the left mirror was, of course, proportional to

$$\int_{-a_1}^{a_1} |u_1(x_1)|^2\, dx_1 .$$

$$G_2 \equiv g_2(a_2/a_1). \tag{6c}$$

$N$ is the Fresnel number of the resonator, while $G_1$ and $G_2$ are generalized $g$ factors which describe the mirror curvatures.* A geometrical interpretation of the $G$'s is shown in Fig. 2.

Note that the above transformation maintains the normalization of the eigenfunctions

$$\int_{-1}^{+1} d\xi_1 \mid v_1(\xi_1) \mid^2 = \int_{-1}^{+1} d\xi_2 \mid v_2(\xi_2) \mid^2 \tag{7}$$

and therefore the physical meaning of the eigenvalues $\gamma_1$ and $\gamma_2$ .

## III. DISCUSSION

The integral equations of any spherical mirror resonator can be transformed into the form of (5a, b, c), which describe completely the mode patterns, diffraction losses and resonant frequencies. It is clear that two resonators have the same scaled eigenfunctions, the same diffraction losses at each mirror and corresponding resonant frequencies whenever they are described by the same characteristic parameters $N$, $G_1$ , and $G_2$ . Two resonators are therefore equivalent if

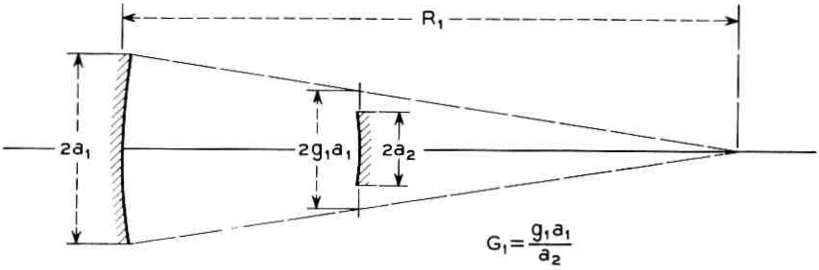


Fig. 2 — Geometrical interpretation of $G_1$ .

$$a_1a_2/\lambda d = \bar{a}_1\bar{a}_2/\lambda\bar{d} = N \tag{8a}$$

$$g_1(a_1/a_2) = \bar{g}_1(\bar{a}_1/\bar{a}_2) = G_1 \tag{8b}$$

$$g_2(a_2/a_1) = \bar{g}_2(\bar{a}_2/\bar{a}_1) = G_2 \tag{8c}$$

where the overbar indicates the dimensions of a resonator equivalent to the original resonator.

* Note added in proof: in recently published perturbation analyses of optical resonators, Gloge[4] and Streifer and Gamo[5] have arrived at the same three resonator parameters.

Quantities which can be expressed in terms of $N$, $G_1$, and $G_2$ also remain invariant when we change from one resonator to an equivalent one. Some of these are listed in Table I. They are used in later computations.

The set of characteristic parameters $N$, $G_1$, and $G_2$ provides some insight into the behavior of resonators. The quantity

$$G^2 \equiv G_1 G_2 = g_1 g_2 \tag{9}$$

may be called the "stability number." One finds from its value whether the resonator is intrinsically of the "stable" or "unstable" type.[1,6] To be stable the resonator must satisfy

$$0 < G^2 < 1. \tag{10}$$

TABLE I—SOME INVARIANTS OF EQUIVALENT RESONATORS

| Resonator Parameters<br>(See Section III and Fig. 1) | | Field Parameters<br>(See Section X and Fig. 3) | |
|---|---|---|---|
| (1.1) | $N \equiv a_1 a_2/\lambda d$ | (2.1) | $a_1 x/\lambda z$ |
| (1.2) | $G_1 \equiv g_1(a_1/a_2)$ | (2.2) | $g_{1z}(a_1/x)$ |
| (1.3) | $G_2 \equiv g_2(a_2/a_1)$ | (2.3) | $g_z(x/a_1)$ |
| (1.4) | $G^2 \equiv g_1 g_2 = G_1 G_2$ | (2.4) | $g_z g_{1z}$ |
| (1.5) | $a_1^2 g_1/a_2^2 g_2 = G_1/G_2$ | (2.5) | $a_1^2 g_{1z}/x^2 g_z$ |
| (1.6)* | $g_1(a_1^2/\lambda d) = G_1 N$ | (2.6) | $g_{1z}(a_1^2/\lambda z)$ |
| (1.7)* | $g_2(\lambda d/a_1^2) = G_2 N$ | (2.7) | $g_z(\lambda z/a_1^2)$ |
| | | (2.8) | $g_z(x^2/\lambda z)$ |

* Quantities like 1.6–1.7 but with subscripts 1 and 2 interchanged are also invariants.

The quantity $N$ is the well-known "Fresnel number." For $N \gg 1$ the diffraction loss of stable resonators is typically very small indeed, and the increase of loss in crossing the boundary from a stable to an unstable type is abrupt. As $N$ decreases toward unity, the loss of the stable resonators increases, and the boundary becomes less sharp until, as $N \ll 1$, all resonators have high loss.

Finally, at least for stable resonators with not too small Fresnel numbers, we can see that the mirror with the larger $G$ has the smaller diffraction loss. From Ref. 6, or from Section VIII of this paper, we know that the radii* $w_i$ of the fundamental mode "spots" on the mirrors are related by

$$w_1^2/w_2^2 = g_2/g_1 . \tag{11}$$

* We use the word "radius" here and later to mean half the width of the mode pattern, as defined in Section VIII.

We therefore have

$$(w_1^2/a_1^2)/(w_2^2/a_2^2) = g_2 a_2^2/g_1 a_1^2 = G_2/G_1 . \tag{12}$$

The ratio $w_i/a_i$ between spot radius and mirror half-width can be taken as a measure for the diffraction loss at a mirror. According to (12) the mirror with the smaller $G_i$ has the larger ratio $w_i/a_i$ , and thus the larger loss. In the special case $G_1 = G_2$ , one sees that (5a) and (5b) become identical, and so the diffraction losses must be equal.

## IV. SPECIAL ADDITIONAL EQUIVALENCES

There are two previously known[1,6] special equivalences which exist in addition to the new ones we have been discussing. These are:

(a) reversal of sign of both $g_1$ and $g_2$

(b) interchange of both $g_1$ and $g_2$ , and $a_1$ and $a_2$ ; i.e., interchange of the mirrors.

The first of these special equivalences changes the sign of both $G_1$ and $G_2$ and does not alter $N$. This equivalence results because the allowed field patterns split up into those of odd and even symmetry in the $x$'s.[1] The equivalent field patterns are complex conjugates of the old ones, but the losses are unchanged.

The second special equivalence corresponds to an interchange of the two mirrors. It leaves $G^2$ and $N$ unchanged, but interchanges $G_1$ and $G_2$ . It also obviously interchanges the mode patterns and the losses of the two mirrors. Combined with the equivalence relations which we have discussed before, this interchanging of the two mirrors means that two resonators are also equivalent if

$$N = \bar{N} \tag{13a}$$

$$G_1 = \bar{G}_2 \tag{13b}$$

and

$$G_2 = \bar{G}_1 . \tag{13c}$$

From these relations one deduces some rather curious equivalent resonator pairs if one postulates that the mirror curvature should be left unchanged ($g_1 = \bar{g}_1$ and $g_2 = \bar{g}_2$) and only the apertures $a_1$ and $a_2$ varied to form an equivalent resonator. With (13) one finds that

$$\bar{a}_1 = a_2(g_1/g_2)^{\frac{1}{2}} \tag{14a}$$

$$\bar{a}_2 = a_1(g_2/g_1)^{\frac{1}{2}} \tag{14b}$$

is necessary for equivalence. Note that the equivalent resonator was found simply by changing the mirror apertures. The mode pattern that appears on the left mirror of this new resonator is a scaled version of the pattern that appeared on the right mirror of the original resonator, and the pattern that was on the left is switched to the right mirror of the resonator.

Similarly, one obtains a pair of resonators equivalent in the above sense when the mirror apertures are kept constant and the curvatures are changed in accordance with (13).

## V. THE CONFOCAL RESONATOR

The resonator commonly known as "the" confocal resonator is actually a very special confocal* resonator for which $R_1 = R_2 = d$, and hence $g_1 = g_2 = 0$, and $G_1 = G_2 = 0$. All of the equivalence transformations we have mentioned transform one confocal resonator into another. Our relations bring out the known fact that the losses and field patterns (apart from scale factors) of the confocal resonator depend only on the Fresnel number $N$ and not at all on the ratio of the mirror apertures.[6]

## VI. RESONATORS WITH EITHER $G_1$ OR $G_2$ EQUAL TO ZERO

When, in a system with mirrors of unequal curvature, the mirror spacing is equal to the radius of curvature of one of the mirrors, then one of the $g$'s is zero and we have $G_1 = 0$, or $G_2 = 0$. Let $g_2 = G_2 = 0$. As a transformation to an equivalent resonator leaves $G_2$ invariant, we have for the equivalent resonator $\bar{g}_2 = 0$. In the stability diagram,[1,6] which shows the stable and unstable resonator regions versus $g_1$ and $g_2$, our transformation yields equivalent resonators that are represented by points on a straight line (in the general case, one has a branch of a hyperbola $g_1 g_2 = \text{const}$).

The parameters of equivalent resonators with $G_2 = 0$ are related by

$$G_1 = \bar{G}_1 = g_1(a_1/a_2) = \bar{g}_1(\bar{a}_1/\bar{a}_2). \tag{15}$$

This relation allows one to find for each resonator with $g_2 = 0$ and *unequal* apertures an equivalent resonator with $\bar{g}_2 = 0$ and *equal* apertures, which is discussed in Ref. 1. Resonators of the former type have

* Any resonator whose mirrors have coincident foci may be termed confocal, whether or not the mirrors have equal curvature. As has been noted,[6] only "the" confocal resonator is a low-loss resonator.

been of interest for the selection of transverse modes in optical maser oscillators.[7] The mode selection properties of equivalent resonators are, of course, the same. For a resonator formed by a spherical mirror and a small plane mirror at its center of curvature[7] ($g_2 = 0$) one finds equivalent resonators of equal mirror apertures which are the closer to the confocal resonator the smaller the flat mirror [compare (15)]. As it appears that the confocal resonator has the best mode selection properties of all spherical mirror resonators, the above behavior would imply that reducing the size of the flat mirror will improve the mode selectivity of the above system. T. Li[7] has indeed found this to be so on the basis of computer calculations.

## VII. RESONATORS WITH RECTANGULAR OR CIRCULAR MIRRORS

The integral equations which determine the modes of resonators with rectangular mirrors decompose into two sets of equations identical to (1), each set involving a single one of the two transverse Cartesian coordinates.[1,6] Hence all of the above applies immediately to such resonators, including resonators with astigmatic mirrors, provided the principal directions of the astigmatism are parallel to the edges of the mirrors.

Equivalent families of resonators with circular mirrors can also easily be found by a similar method, starting from the appropriate integral equations which are indicated in the Appendix. The resulting parameters are of the same form as (6), but with the $a_i$ now redefined as the radii of the mirrors.

## VIII. DETERMINATION OF SPOT RADII

If the apertures of the mirrors are sufficiently large, i.e., if $N \gg 1$, and if $G^2$ is not too close to 0 or 1, then the field patterns of the modes approach closely to Hermite Gaussian functions and lose their dependence on the apertures. Then one can define a "spot size", or spot radius,[6,8] where the Gaussian part of the function has dropped to $e^{-1}$ of its maximum. In the transformations among equivalent resonators, the mode patterns scale in proportion to the apertures; hence two other invariants of equivalent resonators are obtained by replacing $a_1$ and $a_2$ in (6b) and (6c) with the spot radii $w_1$ and $w_2$. Now any quantity which is an invariant of equivalent resonators must be expressible as a function of the basic parameters $N$, $G_1$ and $G_2$. But since the values of $N$ and $G_1/G_2$, which depend on the apertures, do not influence the spot radii, these two invariants of the equivalence transformations can be functionally dependent only on $G^2 = G_1G_2$. Hence we obtain the relations

$$w_1w_2/\lambda d = f(G^2) \tag{16a}$$

$$(w_1/w_2)^2(g_1/g_2) = 1. \tag{16b}$$

Equation (16b) follows, since we know that for mirrors of equal curvature (and hence with $g_1 = g_2$), the spot radii are also equal. The function $f(G^2)$ on the right side of (16a) may be evaluated by comparison with a known result[8] for mirrors of equal curvature

$$(g_1 = g_2 = g; \qquad w_1 = w_2 = w).$$

Equation (27) of Ref. 8 can be conveniently expressed in our present notation as

$$w^2/\lambda d = (1/\pi)(1 - g^2)^{-\frac{1}{2}} \tag{17}$$

which, on comparison with (16a), identifies $f(G^2)$ as

$$f(G^2) = (1/\pi)(1 - G^2)^{-\frac{1}{2}}. \tag{18}$$

Equations (16a) and (16b) can be rewritten with the help of (18) as

$$w_1/w_2 = (g_2/g_1)^{\frac{1}{2}} \tag{19a}$$

$$w_1w_2 = (\lambda d/\pi)(1 - g_1g_2)^{-\frac{1}{2}}. \tag{19b}$$

These last equations are identical with (39) and (40) of Ref. 6 and together determine the two spot radii. Their derivation here is included because of its relative simplicity, and as an example of the use of the invariants.

## IX. FACTORS OF THE GENERAL TRANSFORMATION

Given the parameters (dimensions and curvatures) of one resonator, specification of $\bar{a}_1$ and $\bar{g}_1$ for an equivalent resonator completely determines all parameters of the equivalent resonator, apart from the special equivalences discussed in Section IV. The general transformation from the original to the equivalent resonator can be factored into a succession (product) of two simpler transformations, in the first of which $a_1$ is changed but $g_1$ is not, followed by a second for which $g_1$ is changed but $a_1$ is not.

The first of these simpler transformations effects a rather simple squeezing of all resonator dimensions, all transverse dimensions (apertures) being multiplied by the same factor $\epsilon$, say, while all longitudinal dimensions (radii of curvature, mirror separation) are multiplied by $\epsilon^2$. To see this we note that $R_1$ and $d$ must change proportionally to leave $g_1$ unchanged. $R_2$ must change in proportion with these because of the invariance of $g_1g_2$, i.e., of $G^2$. Finally, $a_1$ and $a_2$ must change proportionally to leave $G_1$ invariant, and they must change as $d^{\frac{1}{2}}$ to leave $N$ invariant.

The second simpler transformation leaves the aperture $a_1$ unchanged. Suppose it changes the radius of curvature $R_1$ in accordance with the relation

$$1/\bar{R}_1 = (1/R_1) + (1/f). \tag{20}$$

In practice a thin lens of focal length $f$ inserted directly in front of the mirror can produce such a transformation. By using the invariants $NG_1$, $N/G_2$, and $N$ [listed as (1.6), (1.7), and (1.1) of Table I] in succession, one can derive the following relations between the parameters of the transformed and original resonators

$$1/\bar{d} = (1/d) + (1/f) \tag{21a}$$

$$1/(\bar{d} - \bar{R}_2) = [1/(d - R_2)] + (1/f) \tag{21b}$$

$$\bar{a}_2/\bar{d} = a_2/d. \tag{21c}$$

Equations (21a), (21b) and (21c) show respectively that the position, *center* of curvature, and aperture of the original second mirror are changed to those of the new one by imaging them through the lens. In this imaging process, objects on the side of the lens toward the second mirror are taken as virtual objects, while objects on the other side of the lens are taken as real objects.

## X. TRANSFORMATION OF THE FIELD INSIDE AND OUTSIDE THE RESONATOR

The mode patterns on the mirrors of two equivalent resonators are scaled versions of each other, and one expects also a correspondence of the fields of a mode inside and outside the equivalent systems. This correspondence is studied in this section.

With the assumptions of the diffraction theory of optical resonators the fields inside or outside the resonator structure can be expressed in terms of the field pattern on one of the mirrors via Fresnel's formula. For fields independent of $y$ (this restriction can be removed easily; compare Appendix) we have for the field traveling to the right, say,

$$u(x,z) = \frac{\sqrt{j}}{\sqrt{\lambda z}} \int_{-a_1}^{a_1} dx_1\, u_1(x_1, 0) \cdot \exp\left[-j\frac{\pi}{\lambda z}(g_{1z}x_1^2 + g_z x^2 - 2xx_1)\right] \tag{22}$$

where $u_1(x_1, 0)$ is the given field pattern on the left mirror and $u(x, z)$ is the field on a spherical references surface that intersects the optics axis a distance $z$ away from the mirror (see Fig. 3). The quantities

$$g_{1z} = 1 - (z/R_1) \tag{23a}$$

$$g_z = 1 - (z/R) \tag{23b}$$

are again used to describe the curvatures of the mirror (curvature radius $R_1$), and that of the reference surface (curvature radius $R$). The mirror width is $2a_1$.

The transformation to an equivalent resonator changes the aperture and curvature of the mirror under consideration, and scales the field pattern on it accordingly, i.e., if

$$a_1 \rightarrow \bar{a}_1 \tag{24a}$$

$$g_1 \rightarrow \bar{g}_1 \tag{24b}$$

then

$$\sqrt{a_1}\, u_1(x_1, 0) = \sqrt{\bar{a}_1}\, \bar{u}_1(\bar{x}_1, 0) \tag{24c}$$

where

$$x_1/a_1 = \bar{x}_1/\bar{a}_1 . \tag{24d}$$

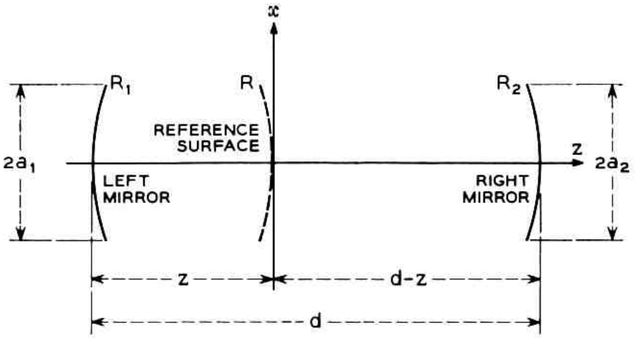


Fig. 3 — Reference surface for description of the fields inside the resonator.

We seek a new surface, described by $\bar{z}$, $\bar{g}_z$ (or $\bar{R}$), on which a scaled "image" $\bar{u}(\bar{x}, \bar{z})$ of $u(x, z)$ will be found. To do this we find, just as in Section II, a set of invariants necessary so that (22) retains its form in the transformed parameters. Essentially these invariants are the three terms in the exponential of (22), with the added fact that since $x_1$ transforms like $a_1$ we replace $x_1$ in those terms by $a_1$. By manipulation we obtain the set of invariants listed as (2.1) through (2.8) in Table I. The terms (2.6), (2.8), and (2.1) come directly from the three terms in the exponential of (22); the others may be derived from them. Finally, the transformed function is given by

$$(z/a_1)^{\frac{1}{2}}u(x, z) = (\bar{z}/\bar{a}_1)^{\frac{1}{2}}\bar{u}(\bar{x}, \bar{z}). \tag{25}$$

From this set of invariants one can find the new position ($\bar{z}$), curvature ($\bar{g}_z$) and transverse scale factor ($\bar{x}/x$) of the scaled function. First, consider the position. The invariant (2.6) may be expanded, using (23a), as

$$\frac{a_1^2}{\lambda}\left(\frac{1}{z} - \frac{1}{R_1}\right) = \frac{a_1^2}{\lambda d}\left(\frac{d}{z} - 1 + g_1\right). \tag{26}$$

But now the term $(a_1^2 g_1/\lambda d)$ is itself an invariant (1.6, Table I) of the resonator transformation, and hence the remaining part of (26), i.e.,

$$\frac{a_1^2}{d\lambda}\left(\frac{d - z}{z}\right)$$

also forms an invariant. Finally we can simplify this a bit by dividing by $N$(1.1, Table I) to yield the invariant

$$\frac{a_1}{a_2}\left(\frac{d - z}{z}\right). \tag{27}$$

We see that the ratio $z/(d - z)$ transforms like $a_1/a_2$. From (27), we obtain the equation from which the new position $\bar{z}$ may be derived

$$\frac{a_1}{a_2}\left(\frac{d - z}{z}\right) = \frac{\bar{a}_1}{\bar{a}_2}\left(\frac{\bar{d} - \bar{z}}{\bar{z}}\right). \tag{28}$$

Once we have found the new position, the new transverse scale factor and curvature may be found most easily using the invariants (2.1) and (2.4), respectively, of Table I; i.e.,

$$(\bar{x}/x) = (\bar{z}/z)(a_1/\bar{a}_1) \tag{29}$$

and

$$\bar{g}_z = g_z g_{1z}/\bar{g}_{1z} \tag{30}$$

and the scaled function is as given in (25).

To provide a more physical picture of the field transformation, it is interesting to note that the simple squeezing and imaging transformations discussed in Section IX apply to the arbitrary reference surface and its field as well as to the second mirror and its field.

Finally we note that the transformations of Fresnel's formula we have been discussing do not depend on the fact that $u_1(x_1, 0)$ is an eigenfunction of a resonator. The preceding discussion, with the exception of the derivation of (26), all applies equally well to the fields generated by *any* prescribed field distribution over an aperture.

## XI. ACKNOWLEDGMENT

The authors would like to thank T. Li for valuable comments.

## APPENDIX

### *Resonators with Spherical Mirrors of General Shape*

Within the assumptions of the theory of optical resonators[1–9] the modes of a resonator formed by two spherical mirrors of quite general shape are governed by the integral equations

$$\sigma_1 E_1(x_1, y_1) = \frac{j}{\lambda d} \int_{A_2} dA_2 \cdot K(x_1, x_2; y_1, y_2) \cdot E_2(x_2, y_2) \tag{31}$$

and

$$\sigma_2 E_2(x_2, y_2) = \frac{j}{\lambda d} \int_{A_1} dA_1 \cdot K(x_1, x_2; y_1, y_2) \cdot E_1(x_1, y_1) \tag{32}$$

with the kernel

$$K(x_1, x_2; y_1, y_2) = \exp\left\{-j\frac{\pi}{\lambda d}[g_1(x_1^2 + y_1^2) + g_2(x_2^2 + y_2^2) - 2(x_1x_2 + y_1y_2)]\right\}. \tag{33}$$

Here $(x_1, y_1)$ and $(x_2, y_2)$ are coordinates in planes perpendicular to the optic axis, $d$ is the mirror separation, and $g_1$ and $g_2$ describe the mirror curvatures as in Section II. Subscript "1" indicates quantities associated with the mirror on the left-hand side, and "2" refers to the mirror on the right. $\sigma_1$ and $\sigma_2$ are the eigenvalues corresponding to $\gamma_1$ and $\gamma_2$ discussed in Section II. The integration has to be performed over the reflecting areas $A_1$ and $A_2$ of the mirrors where $dA_1$ and $dA_2$ are the area elements. No assumptions on the curves bounding the reflecting areas have been made, and the formulation (31) and (32) includes mirrors of quite general shape. Special cases are, of course, strip mirrors, square mirrors, rectangular mirrors, and mirrors of circular shape. These are of main practical interest.

Let us compare (31), (32) and (33) with (1a), (1b), and (1c) of Section II. It is clear that the discussion of two-dimensional resonators systems given in Section II can be extended to the three-dimensional case in which we are interested now. The only difference is that we now have two transverse coordinates $(x, y)$. If they are subjected to the transformation

$$x_i = \epsilon_1 \bar{x}_i \, ; \qquad y_i = \epsilon_i \bar{y}_i \, ; \qquad i = 1, 2 \tag{34}$$

and the mirror areas and area elements are scaled like

$$A_i = \epsilon_i^2 \bar{A}_i \, ; \qquad dA_i = \epsilon_i^2 dA_i \tag{35}$$

then the mirror curvatures and the mirror separation of two equivalent resonators are related by the same invariants as before. All we have to do is to replace $a_i^2$ by $A_i$ in the table of invariants. For the special case of circular mirrors, $a_i$ can be redefined as the mirror radius and retained in the invariance relations.

Note that we have used the same scaling factors $\epsilon_i$ for the $x$ and $y$ coordinates. If different scaling factors are used one obtains, of couse, equivalent resonators with mirrors that are not spherical but astigmatic.

REFERENCES

1. Fox, A. G., and Li, Tingye, Modes in a Maser Interferometer with Curved and Tilted Mirrors, Proc. IEEE, **51,** 1963, pp. 80–89.
2. 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,** January, 1964, pp. 113–126.
3. Cochran, J. A., The Existence of Eigenvalues for the Integral Equations of Laser Theory, to be published in B.S.T.J.
4. Gloge, O., Analysis of Fabry-Perot Resonators by Means of Scattering Matrices, Arch. El. Ü., **18,** March, 1964, pp. 197–203.
5. Streifer, W., and Gamo, H., On the Schmidt Expansion for Optical Resonator Modes, Proc. Symp. on Quasi-Optics, Polytechnic Inst. of Brooklyn, June, 1964.
6. Boyd, G. D., and Kogelnik, H., Generalized Confocal Resonator Theory, B.S.T.J., **41,** July, 1962, pp. 1347–1369.
7. Li, Tingye, Mode Selection in an Aperture-Limited Concentric Maser Interferometer, B.S.T.J., **42,** November, 1963, pp. 2609–2620.
8. Boyd, G. D., and Gordon, J. P., Confocal Multimode Resonator for Millimeter Through Optical Wavelength Masers, B.S.T.J., **40,** March, 1961, pp. 489–508.
9. Fox, A. G., and Li, Tingye, Resonant Modes in a Maser Interferometer, B.S.T.J., **40,** March, 1961, pp. 453–488.

# Modes in a Sequence of Thick Astigmatic Lens-Like Focusers

By E. A. J. MARCATILI

(Manuscript received June 1, 1964)

*Maxwell's equations are solved for a periodic sequence of lens-like focusers separated by gaps. Each focuser consists of an arbitrarily thick slab of dielectric in which the dielectric constant tapers off radially with different quadratic laws in two perpendicular directions. Since there are no limitations on the thickness of the slabs, the solutions cover the complete gamut from a sequence of infinitely thin lenses with astigmatism to a continuous dielectric waveguide, and from spherical to cylindrical lenses.*

*The field configurations of the modes and their propagation constants, as well as the transmission and cutoff bands, are calculated. Any arbitrary input field distribution can then be expanded in terms of the normal modes, and the expansion determines the field everywhere.*

*Formulas derived for sequences of weak lenses turn out to give very good results even for lenses whose thickness and separation are equal to the focal length.*

## I. INTRODUCTION

One possible long distance transmission medium for optical waves consists of a periodic sequence of converging lenses. In order to negotiate unwanted but unavoidable bends of the axis of the sequence it is necessary to space the lenses as closely as possible.[1] Nevertheless, ordinary dielectric lenses exhibit substantial surface scattering, and therefore the minimum spacing between lenses depends on the tolerable transmission loss.

D. W. Berreman has shown that an effective lens can be made using gas with thermal gradients,[2,3] thus avoiding the solid-to-gas transition problems. D. W. Berreman and S. E. Miller[4] proposed a gaseous lens consisting of a tube with hot walls through which a mild gas current at lower temperature is forced to flow. At any cross section the temperature increases from the center to the wall. The density and consequently the dielectric constant is then maximum on the axis and decreases radially roughly with a square law. Without the problem of scattering at the interfaces, tubular gas lenses can be closely spaced and the gaps may be comparable to the thickness of the lenses.

The advent of such a new transmission medium makes it opportune and important to generalize the theory of modes in a sequence of thin lenses by determining the normal modes in an idealized structure which consists of a periodic sequence of arbitrarily thick slabs of dielectric whose dielectric constant tapers off radially with quadratic law.

The preferential direction of gravity creates convection currents that may introduce astigmatism in the gaseous lenses. Such an aberration is included in our model by making the radial quadratic law of the dielectric different in two perpendicular directions.

We calculate the modes of propagation of the idealized structure without including the solid walls surrounding the medium. Taking them into account would perturb the modes only slightly, introducing diffraction losses. Just as in the case of a waveguide with perfect metallic walls, the idealized modes considered here are not attenuated, but their discussion is similarly expected to be useful in approximating: (a) the propagation constants; (b) the range of dimensions over which transmission is permitted or forbidden; (c) the extent of mode conversion at discontinuities or imperfections; and (d) the field at any point due to an arbitrary input such as an off-axis or tilted beam. Of these, (a) and (b) are treated in this article.

The calculations are general enough that by changing the lens parameters and the length of the gaps it is possible to cover uninterruptedly all the range from a sequence of thin lenses[5,6,7,8] to a continuous dielectric guide,[1,9,10] and from spherical to cylindrical lenses. Up to now only the extreme cases, that is, thin lenses or dielectric guide and spherical or cylindrical lenses, have been considered in the literature; this article bridges the gaps.

## II. DESCRIPTION OF THE PROBLEM

Consider a periodic sequence of dielectric slabs, shown in Fig. 1. The refractive index $\nu$ of each slab is independent of $z$, but varies with different quadratic laws in the $x$ and $y$ directions as

$$\nu = \sqrt{\frac{\epsilon}{\epsilon_0}} = n\left[1 - \left(\frac{\pi x}{L_1}\right)^2 - \left(\frac{\pi y}{L_2}\right)^2\right]^{\frac{1}{2}}. \tag{1}$$

The refractive index $n$ on the $z$ axis and the characteristic parameters

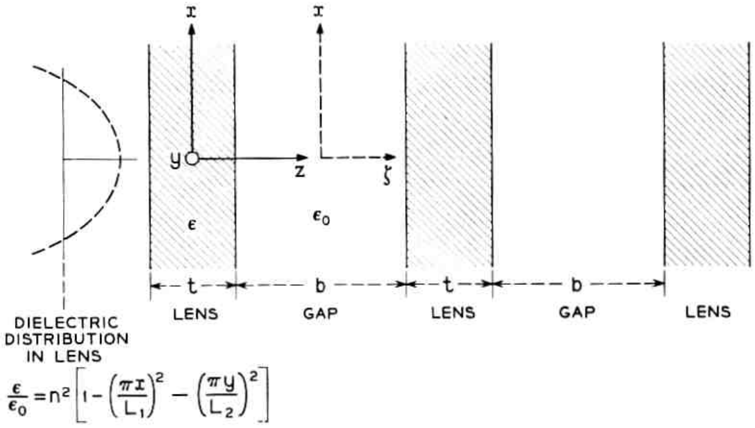

x
x
y
z
ζ
ε
ε₀
t
b
t
b
LENS
GAP
LENS
GAP
LENS
DIELECTRIC DISTRIBUTION IN LENS
ε/ε₀ = n²[1 - (πx/L₁)² - (πy/L₂)²]

Fig. 1 — Periodic sequence of arbitrarily thick and astigmatic lenses.

of the lens-like medium, $L_1$ and $L_2$, permit adjustment of the parabolic distributions. The physical significance of $L_1$ and $L_2$ will be treated below.

In spite of the fact that it reaches negative values for large $x$ or $y$, this dielectric distribution is useful because it matches the dielectric distribution of the gaseous lens, especially for small values of $\pi x/L_1$ and $\pi y/L_2$. Besides, it turns out that the field of most modes is negligible in the region where the dielectric constant is small or negative, and consequently that region does not contribute essentially to the guidance of the modes.

In the Appendix we solve approximately Maxwell's equations. The sequence of lens-like focusers supports hybrid modes $\mathrm{EH}_{pq}$, characterized by the indexes $p$ and $q$. These integers indicate that the intensity of each transverse field component passes through $p$ zeros in the $x$ direction and $q$ zeros in the $y$ direction.

The only approximation in the solution of Maxwell's equations consists in neglecting terms of the order of $p\lambda/L_1$ and $q\lambda/L_2$ compared to unity. $\lambda$ is the free-space wavelength. Typically $\lambda/L_1$ and $\lambda/L_2$ are of the order of $10^{-5}$; therefore, except for very high-order modes ($p$ and/or $q$ very large), the results must be satisfactorily precise.

The modes have no electric field in the $y$ direction nor magnetic field in the $x$ direction. The remaining components — $\mathrm{E}_x$, $\mathrm{E}_z$, $\mathrm{H}_y$ and $\mathrm{H}_z$ in the dielectric slabs and $\mathrm{E}_{xg}$, $\mathrm{E}_{zg}$, $\mathrm{H}_{yg}$ and $\mathrm{H}_{zg}$ in the gaps — are found assuming as normal modes only those field configurations that repeat themselves periodically at each lens. Therefore the equiphase surfaces of each mode are planes at $z = 0$ and $\zeta = 0$, as shown in Fig. 1. We reproduce here only $\mathrm{E}_x$ (55) and $\mathrm{E}_{xg}$ (58).

All the other components can be deduced from them with the help of (45).

$$\begin{aligned}\mathrm{E}_x = \exp\Bigg\{&-i\left[kn\left(z - \frac{x^2}{2R_1} - \frac{y^2}{2R_2}\right) - \left(p + \frac{1}{2}\right)\tan^{-1}\left(\frac{w_1^2}{s_1^2}\tan\frac{\pi z}{L_1}\right)\right.\\ &\left.- \left(q + \frac{1}{2}\right)\tan^{-1}\left(\frac{w_2^2}{s_2^2}\tan\frac{\pi z}{L_2}\right)\right] - \left(\frac{x}{\rho_1}\right)^2 - \left(\frac{y}{\rho_2}\right)^2\Bigg\} \qquad (2)\\ &\cdot \mathrm{H}_p\left(\sqrt{2}\,\frac{x}{\rho_1}\right)\mathrm{H}_q\left(\sqrt{2}\,\frac{y}{\rho_2}\right);\end{aligned}$$

$$\begin{aligned}\mathrm{E}_{xg} = \exp\Bigg\{&-i\left[k\left(\zeta - \frac{x^2}{2R_{g1}} - \frac{y^2}{2R_{g2}}\right) - \left(p + \frac{1}{2}\right)\tan^{-1}\frac{2\zeta}{ks_{g1}^2}\right.\\ &\left.- \left(q + \frac{1}{2}\right)\tan^{-1}\frac{2\zeta}{ks_{g2}^2}\right] - \left(\frac{x}{\rho_{g1}}\right)^2 - \left(\frac{y}{\rho_{g2}}\right)^2\Bigg\} \qquad (3)\\ &\cdot \mathrm{H}_p\left(\sqrt{2}\,\frac{x}{\rho_{g1}}\right)\mathrm{H}_q\left(\sqrt{2}\,\frac{y}{\rho_{g2}}\right)\end{aligned}$$

where [see (55) to (67)], $k = \omega\sqrt{\epsilon_0\mu} = 2\pi/\lambda$ is the free-space propagation constant and $\mathrm{H}_\mu(\alpha)$ is the Hermite polynomial of order $\mu$.

The physical significance of the symbols $s_1$, $s_2$, $s_{g1}$, $s_{g2}$, $R_1$, $R_2$, etc. will be developed below. We give first their mathematical meaning and in order to avoid repetition, from now on the letter $m$ will stand for either the subindex 1 or 2, depending on whether the symbol under consideration refers to a dimension in the plane $y = 0$ or $x = 0$ respectively. Calling the thickness of each dielectric slab $t$, and the gap between them $b$,

$$s_m = w_m\left(\frac{1 + C_m \operatorname{ctn}\varphi_m}{1 - C_m \tan\varphi_m}\right)^{\frac{1}{2}} \qquad (4)$$

$$s_{gm} = w_m(1 + C_m \operatorname{ctn}\varphi_m)^{\frac{1}{2}}(1 - C_m \tan\varphi_m)^{\frac{1}{2}} \qquad (5)$$

$$w_m = \frac{1}{\pi}\sqrt{\frac{\lambda L_m}{n}} \qquad (6)$$

$$C_m = n\,\frac{\pi}{2}\,\frac{b}{L_m} \qquad (7)$$

$$\varphi_m = \frac{\pi}{2}\,\frac{t}{L_m} \qquad (8)$$

$$R_m = \frac{L_m}{\pi}\left\{\frac{1+\left(\frac{w_m}{s_m}\right)^4}{\left[1-\left(\frac{w_m}{s_m}\right)^4\right]\sin\frac{2\pi z}{L_m}} + \operatorname{ctn}\frac{2\pi z}{L_m}\right\} \tag{9}$$

$$\rho_m = s_m\sqrt{\frac{1}{2}\left\{1+\left(\frac{w_m}{s_m}\right)^4+\left[1-\left(\frac{w_m}{s_m}\right)^4\right]\cos\frac{2\pi z}{L_m}\right\}} \tag{10}$$

$$R_{gm} = \frac{k^2 s_{gm}^{\ 4}}{4\zeta}\left[1+\left(\frac{2\zeta}{k s_{gm}^{\ 2}}\right)^2\right] \tag{11}$$

$$\rho_{gm} = s_{gm}\sqrt{1+\left(\frac{2\zeta}{k s_{gm}^{\ 2}}\right)^2} \tag{12}$$

Let us find the physical significance of $R_m$, $R_{gm}$, $\rho_m$, $\rho_{gm}$, $s_m$, $s_{gm}$, $w_m$ and $L_m$. Equating in (2) and (3) the imaginary parts of the exponents to constants we obtain two equations of equiphase surfaces (wavefronts), one applicable within a lens and the other in a gap. At the $z$ axis, each wavefront has a radius of curvature in the plane $y = 0$ which in general is different from that in the plane $x = 0$. Within a lens those main radii of curvature are $R_1$ and $R_2$ [see (9)], while those in a gap are $R_{g1}$ and $R_{g2}$ [see (11)]. If $L_1 = L_2$, then $R_1 = R_2$ and $R_{g1} = R_{g2}$.

For the fundamental mode $p = q = 0$, at a given abscissa $z$ or $\zeta$ the field amplitudes (2) and (3) decrease with different Gaussian laws in the $x$ and $y$ directions. The distances at which the field is $1/e$ of the maximum occurring on the $z$ axis are the beam sizes $\rho_1$ and $\rho_2$ [see (10)] within a lens, and $\rho_{g1}$ and $\rho_{g2}$ [see (12)] in a gap.

For $z = 0$ and $\zeta = 0$ we find from (10) and (12) that $\rho_m = s_m$ and $\rho_{gm} = s_{gm}$. Therefore $s_m$ and $s_{gm}$ are the beam sizes at the planes of symmetry of each lens and each gap respectively.

The physical significance of $w_m$ becomes obvious on reducing the gaps between lenses to zero. Then instead of a sequence of lenses we have an uninterrupted dielectric waveguide and we derive from (7), (4), (5) (10) and (12) that

$$\rho_m = \rho_{gm} = s_m = s_{gm} = w_m . \tag{13}$$

Therefore in the continuous guide the propagating normal modes do not change size along $z$, and for the fundamental mode $w_1$ and $w_2$ measure the beam sizes in the $x$ and $y$ directions.

From (10) we find that within a lens the beam sizes $\rho_1$ and $\rho_2$ in the $y = 0$ and $x = 0$ planes vary periodically along $z$; their periods are $L_1$ and $L_2$ respectively.

For the particular case in which $L_1 = L_2$ the field in the gap (3) coincides with that found by Boyd and Gordon[6] for the resonator made with confocal mirrors of infinite aperture.

## III. TRANSMISSION AND CUTOFF CONDITIONS

Both $s_m$ [see (4)] and $s_{gm}$ [see (5)] must be real quantities, otherwise the fields given in (2) and (3) become infinite as $x$ or $y \to \infty$. This establishes that a mode can propagate in the sequence of lenses either when

$$C_m \leqq \operatorname{ctn} \varphi_m \tag{14}$$

or when

$$C_m \leqq -\tan \varphi_m \,. \tag{15}$$

Their equivalents in explicit form are

$$b \leqq (2L_m/n\pi) \operatorname{ctn} (\pi t/2L_m) \tag{16}$$

and

$$b \leqq -(2L_m/n\pi) \tan (\pi t/2L_m). \tag{17}$$

Which equation must we use? Since $b$ and $L_m$ are positive, (14) or (16) must be used when $\varphi_m = \pi t/2L_m$ falls in an odd quadrant and (15) or (17) when it falls in an even quadrant. Naturally, if these equations are satisfied for only one of the two indexes, that sequence of lenses cannot propagate any nonattenuating mode.

If $b = 0$, the sequence of lenses is reduced to a continuous waveguide and transmission takes place, as it must, no matter what the values of $\varphi_1$ and $\varphi_2$ are. If now we increase the gap $b$, transmission will take place as long as (16) or (17) is satisfied.

## IV. DISCUSSION OF THE FIELD INSIDE AND OUTSIDE THE LENSES

The sequence of lenses admits a complete set of modes. For each mode, the field inside (2) and outside (3) the lenses is a wave traveling in the $z$ direction whose amplitude, period and equiphase surfaces (wavefronts) vary along $z$.

The amplitude depends on $x$ as a product of a Gaussian function and a Hermite polynomial (parabolic cylinder function) whose degree depends on the mode under consideration. A similar type of variation occurs along $y$.

In Fig. 2 we plot qualitatively the beam sizes $\rho_m$ and $\rho_{gm}$ for $\varphi_m = \pi t/2L_m$ in the first, second and third quadrants. For $\varphi_m$ in an odd quadrant, as in Figs. 2(a) and 2(c), the maximum and minimum beam sizes within each lens are

$$\rho_{m\,\max} = s_m = w_m \left( \frac{1 + C_m \operatorname{ctn} \varphi_m}{1 - C_m \tan \varphi_m} \right)^{\frac{1}{4}} \tag{18}$$

and

$$\rho_{m\,\min} = \frac{w_m^{\;2}}{s_m} = w_m \left( \frac{1 - C_m \tan \varphi_m}{1 + C_m \operatorname{ctn} \varphi_m} \right)^{\frac{1}{4}}. \tag{19}$$

The period between two successive maxima is $L_m$. The square root of the product of the maximum and minimum beam sizes in the dielectric is a constant

$$(\rho_{m\,\max}\, \rho_{m\,\min})^{\frac{1}{2}} = w_m$$

and coincides with the beam size $w_m$ of the lens-like medium.

In the gap, the only extremum for the beam size is a single minimum which occurs at $\zeta = 0$ and, from (5) and (12), corresponds to

$$\rho_{gm\,\min} = s_{gm} = w_m (1 + C_m \operatorname{ctn} \varphi_m)^{\frac{1}{4}} \, (1 - C_m \tan \varphi_m)^{\frac{1}{4}}. \tag{20}$$

If $\varphi_m = \pi t / 2L_m$ falls in an even quadrant, as in Fig. 2(b), the minimum and maximum beam sizes interchanged from the odd quadrant are (18) and (19) respectively. Again (20) corresponds to the unique minimum in each gap.

## V. SPECIAL CASES

Let us consider the field in a gap assuming

$$L_1 = L_2 = L$$

and

$$t/L = \eta \quad \text{or} \quad \varphi_1 = \varphi_2 = \eta(\pi/2) \tag{21}$$

where $\eta$ is an integer. Then unless the gap $b = 0$, the minimum beam size in the gap $\rho_{gm\,\min}$ (20) becomes infinitely large and the electric field (3) is reduced to a plane wave travelling in the $z$ direction. If more generally only $\varphi_1 = \eta(\pi/2)$, but $\varphi_2$ is unrestricted, then the wave fronts are cylindrical surfaces parallel to the $x$ axis.

Consider again

$$L_1 = L_2 = L$$

but

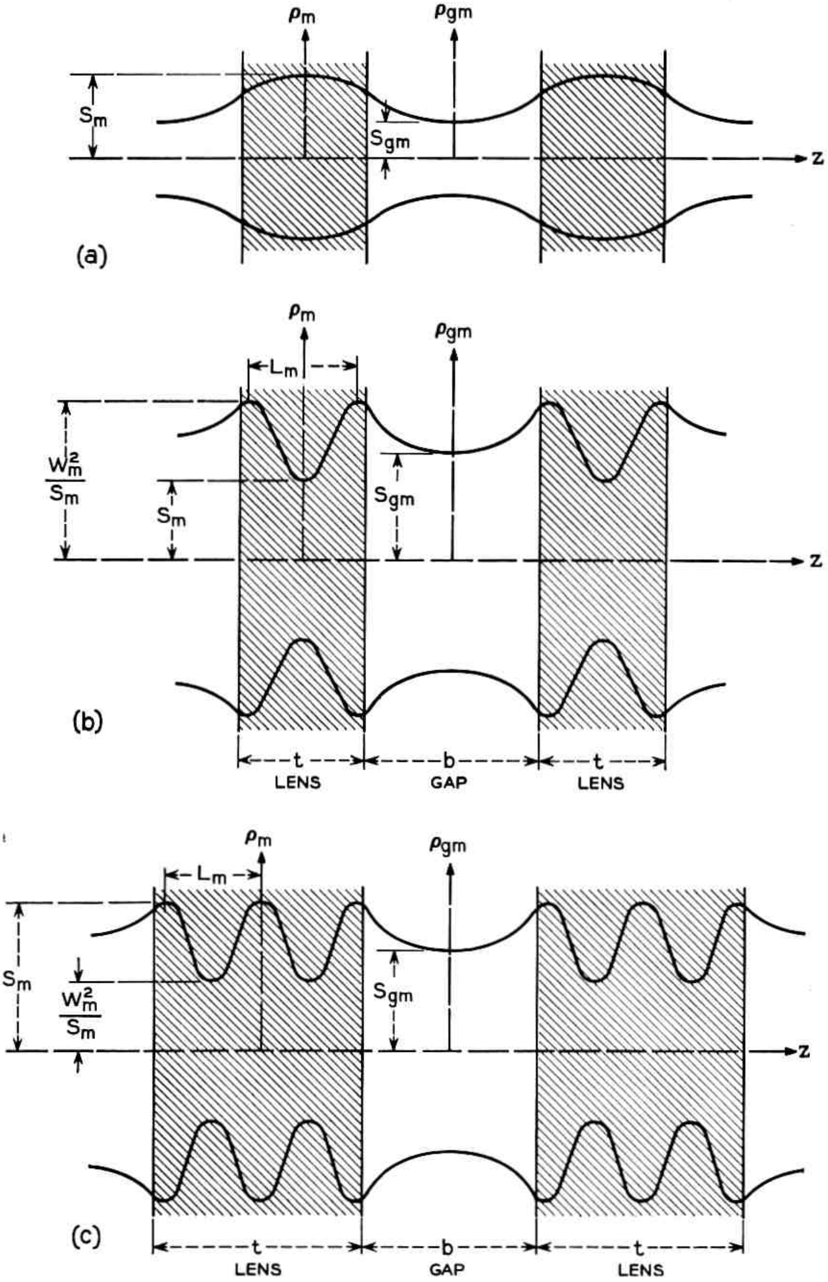


Fig. 2 — (a) Beam size for $\varphi_m = \frac{\pi}{2}\frac{t}{L_m}$ in first quadrant, $t/L_m < 1$; (b) beam size for $\varphi_m = \frac{\pi}{2}\frac{t}{L_m}$ in second quadrant, $1 < t/L_m < 2$; (c) beam size for $\varphi_m = \frac{\pi}{2}\frac{t}{L_m}$ in third quadrant, $2 < t/L_m < 3$.

$$C_1 = C_2 = \operatorname{ctn} \varphi_1 = \operatorname{ctn} \varphi_2 \tag{22}$$

or

$$C_1 = C_2 = -\tan \varphi_1 = -\tan \varphi_2 \,. \tag{23}$$

Then according to (20) the minimum beam size $\rho_{g1\,\min} = \rho_{g2\,\min} = 0$ and the field in the gap (3), for $p = q = 0$, becomes

$$E_{xg} = \exp -ik\zeta \left(1 + \frac{x^2 + y^2}{2\zeta^2}\right). \tag{24}$$

The wavefronts close to the $\zeta$ axis are concentric spheres and their centers coincide with the point $x = y = \zeta = 0$.

Therefore the two conditions indicated above correspond either to plane waves in the gap or to concentric waves (if one observes only the field in the region close to the $\zeta$ axis). They are equivalent to those in Fabry-Perot resonators with plane and concentric mirrors.[5,6,7]

The condition under which the beam is closely concentrated on the $z$ axis is found by minimizing the maximum beam size within a lens, $s_m$ [see (18)] or $w_m^2/s_m$ [see (19)] depending on whether $\varphi_m$ is in an odd or even quadrant.

If the gap $b$ decreases, the value of $s_m$ or $w_m^2/s_m$ also decreases; for $b = 0$ the sequence of lenses becomes a dielectric waveguide, the beam size does not vary with $z$, and its value is $s_m = w_m^2/s_m = w_m$. On the other hand, if the thickness $t$ of each lens is the only variable, the minimum of $s_m$ or $w_m^2/s_m$ is achieved by making

$$\frac{\partial s_m}{\partial t} = \frac{\partial s_m}{\partial \varphi_m} = 0 \qquad \text{if } \varphi_m \text{ is an odd quadrant}$$

or (25)

$$\frac{\partial}{\partial t}\frac{1}{s_m} = \frac{\partial}{\partial \varphi_m}\frac{1}{s_m} = 0 \qquad \text{if } \varphi_m \text{ is in an even quadrant.}$$

These conditions lead to the same requirement, namely:

$$C_m = \operatorname{ctn} 2\varphi_m \tag{26}$$

or its equivalent

$$b = (2L_m/n\pi) \operatorname{ctn} (\pi t/L_m) \tag{27}$$

which, replaced in (18) or (19), determines the minimized value of the maximum beam size within each lens

$$s_{m\,\min} = (w_m^2/s_m)_{\min} = w_m[(1 + C_m^2)^{\frac{1}{2}} + C_m]^{\frac{1}{2}}. \tag{28}$$

For the same condition (26) or (27), the beam size in the gap at any abscissa $\zeta$ is derived from (5), (12) and (26)

$$\rho_{gm} = w_m(1 + C_m^2)^{\frac{1}{4}} \left[ 1 + \frac{C_m^2}{1 + C_m^2} \left( \frac{2\zeta}{b} \right)^2 \right]^{\frac{1}{2}}. \tag{29}$$

## VI. SEQUENCE OF WEAK ASTIGMATIC LENSES

Before considering weak lenses, let us relate the characteristic lengths, $L_1$ and $L_2$ of the lens-like focusers to their focal lengths in the planes $y = 0$ and $x = 0$. To calculate the focal length in the $y = 0$ plane (see Fig. 3) the ray trajectory is determined from the equation

$$d^2x/dz^2 = (1/\nu)\ (d\nu/dx). \tag{30}$$

Taking the refractive index $\nu$ from (1)

$$\frac{d^2x}{dz^2} = \frac{1}{\sqrt{1 - (\pi x/L_1)^2}} \frac{d}{dx} \sqrt{1 - (\pi x/L_1)^2}. \tag{31}$$

For paraxial rays

$$\pi x/L_1 \ll 1 \tag{32}$$

and within a lens the trajectory of a ray entering parallel to the $z$ axis at a distance $x_0$ is

$$x = x_0 \cos\ (\pi z/L_1).$$

The angle of refraction at the output surface is

$$\theta_r = (n\pi/L_1)x_0 \sin\ (\pi t/L_1). \tag{33}$$

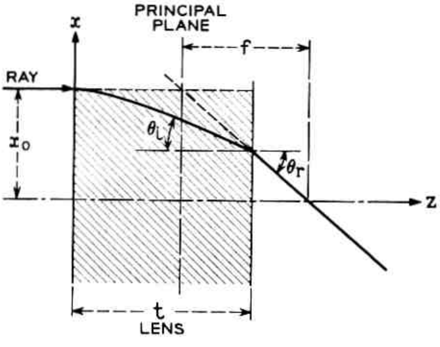


Fig. 3 — Ray trajectory in the plane $y = 0$ of a lens.

Then from simple geometric considerations in Fig. 3, the focal length $f_1$ results

$$f_1 = \frac{L_1}{\pi n \sin(\pi t/L_1)}. \tag{34}$$

We assume now weak lenses. They are characterized by

$$\varphi_m = (\pi/2)(t/L_m) \ll 1, \tag{35}$$

and in all previous results each circular function can be replaced by its leading term.

Because of the inequality (35) the characteristic length of the focusing medium $L_1$ in (34) can be calculated explicitly by

$$L_1 = \pi\sqrt{ntf_1}. \tag{36}$$

Similarly, for the plane $x = 0$

$$L_2 = \pi\sqrt{ntf_2}. \tag{37}$$

The weak lens requirement (35) then becomes

$$\varphi_m = \frac{1}{2}\sqrt{\frac{t}{nf_m}} \ll 1. \tag{38}$$

Using (36) and (37) together with the simplifying assumption (38) we re-evaluate the maximum and minimum beam sizes (18), (20) for weak lenses ($\varphi_m$ in first quadrant),

$$s_m = \left(\frac{tf_m\lambda^2}{n\pi^2}\right)^{\frac{1}{4}}\left(\frac{1 + (nb/t)}{1 - (b/4f_m)}\right)^{\frac{1}{4}} \tag{39}$$

$$s_{gm} = \left(\frac{tf_m\lambda^2}{n\pi^2}\right)^{\frac{1}{4}}\left(1 + \frac{nb}{t}\right)^{\frac{1}{4}}\left(1 - \frac{b}{4f}\right)^{\frac{1}{4}}. \tag{40}$$

The distance $h$ between the principal planes may be of interest. Using (33) and (34), this distance turns out to be

$$h_m = \frac{2L_m}{\pi n}\tan\frac{\pi t}{2L_m} - t. \tag{41}$$

Expanding the circular function in series, keeping only the first two terms and substituting $L_m$ by their equivalents (36) and (37) we obtain,[14]

$$h_m = t\left(\frac{1}{n} - 1\right) + \frac{t^2}{12n^2f_m}. \tag{42}$$

### 6.1 *Example*

Let us assume a sequence of gaseous lenses such that

$$b = t = f_1 = f_2 = 0.25 \text{ m}$$

$$\lambda = 0.6328\ 10^{-6} \text{ m}$$

$$n \approx 1.$$

For these dimensions $\varphi_1 = \varphi_2 = 0.5$, and therefore the weak lens inequality (38) is hardly satisfied. Nevertheless, let us go ahead and calculate extreme beam sizes $s_1 = s_2$ and $s_{g1} = s_{g2}$ as well as the characteristic length $L_1 = L_2$ of the lens using (39), (40) and (36)

$$\begin{aligned} s_1 = s_2 &= 0.286 \text{ mm} \\ s_{g1} = s_{g2} &= 0.248 \text{ mm} \\ L_1 = L_2 &= 0.785 \text{ m}. \end{aligned} \tag{43}$$

Let us calculate again the extreme beam sizes using the exact expressions (18), (20), deriving $L$ from (34)

$$\begin{aligned} s_1 = s_2 &= 0.276 \text{ mm} \\ s_{g1} = s_{g2} &= 0.224 \text{ mm} \\ L_1 = L_2 &= 0.704 \text{ m}. \end{aligned} \tag{44}$$

The two sets of results (43) and (44) are reasonably similar and show the usefulness of weak lens formulas even for lenses with comparable values of $t$ and $f$.

## VII. CONCLUSIONS

The properties of the modes in a sequence of thick, astigmatic and unbounded lens-like focusers are similar to those in a sequence of thin infinitely large lenses.

The modes are hybrid and described by parabolic cylinder functions (product of Gaussians times Hermite polynomials). Transmission takes place as long as the gap between lenses is smaller than a value given in (16) or (17).

The maximum beam size can be reduced by decreasing the distance between dielectric slabs. Nevertheless, if the gap is fixed, the minimization of the maximum beam size can be obtained by selecting the dielectric properties or the thickness of each focuser according to (27).

Simplified formulas derived for sequences of weak lenses yield good approximations even for lenses whose thickness, separation and focal length are comparable.

## VIII. ACKNOWLEDGMENT

I am grateful to J. W. Tukey for his constructive criticism of this article.

## APPENDIX

### *Solution of Maxwell's Equations in a Sequence of Thick Astigmatic Lenses*

We will obtain, first, a general enough solution of Maxwell's equations for one of the lenses; see Fig. 1. Then by making $n = 1$ and $L_1 = L_2 \to \infty$, we will deduce a general solution for the uniform gap between lenses, and finally we will match the tangential fields to satisfy the boundary conditions. For modes with only four field components, $\mathrm{E}_x$, $\mathrm{E}_z$, $\mathrm{H}_y$ and $\mathrm{H}_z$, Maxwell's equations become

$$
\begin{aligned}
\frac{\partial E_x}{\partial z} - \frac{\partial E_z}{\partial x} &= j\omega\mu \mathrm{H}_y \\
\frac{\partial E_x}{\partial y} &= -j\omega\mu \mathrm{H}_z \\
\frac{\partial H_z}{\partial y} - \frac{\partial H_y}{\partial z} &= -j\omega\epsilon \mathrm{E}_x \\
\frac{\partial H_y}{\partial x} &= -j\omega\epsilon \mathrm{E}_z
\end{aligned}
\tag{45}
$$

where $\mu$, the magnetic permeability, is a constant;

$$
\frac{\epsilon}{\epsilon_0} = n^2\left[1 - \left(\frac{\pi x}{L_1}\right)^2 - \left(\frac{\pi y}{L_2}\right)^2\right] \tag{46}
$$

$\epsilon$, $L_1$ and $L_2$ are arbitrary constants; and $\omega\sqrt{\epsilon_0\mu} = 2\pi/\lambda = k$ is the free-space propagation constant.

By eliminating variables and by neglecting terms of the order of $\lambda/L_1$ and $\lambda/L_2$* as compared to unity we obtain identical equations for $\mathrm{E}_x$ and $\mathrm{H}_y$. For $\mathrm{E}_x$,

$$
\frac{\partial^2 E_x}{\partial x^2} + \frac{\partial^2 E_x}{\partial y^2} + \frac{\partial^2 E_x}{\partial z^2} + (kn)^2\left[1 - \left(\frac{\pi x}{L_1}\right)^2 - \left(\frac{\pi x}{L_2}\right)^2\right]\mathrm{E}_x = 0. \tag{47}
$$

* In practice $\lambda/L_1$ and $\lambda/L_2$ are of the order of $10^{-5}$.

This equation is separable and a general solution is

$$\begin{aligned} E_x &= \exp\left[-\left(\frac{x}{w_1}\right)^2 - \left(\frac{y}{w_2}\right)^2\right] \\ &\quad \times \sum_{\nu=0}^{\infty}\sum_{\mu=0}^{\infty} A_{\nu\mu} \exp\left[-iknz\left(1 - \frac{2\nu+1}{2n}\frac{\lambda}{L_1} - \frac{2\mu+1}{2n}\frac{\lambda}{L_2}\right)^{\frac{1}{2}}\right] \qquad (48) \\ &\quad \times H_\nu\left(\frac{\sqrt{2}x}{w_1}\right) H_\mu\left(\frac{\sqrt{2}y}{w_2}\right). \end{aligned}$$

where $\nu$ and $\mu$ are integers, and $A_{\nu\mu}$ is an arbitrary constant. Using $m$ to indicate either subscript 1 or 2,

$$w_m = \frac{1}{\pi}\sqrt{\frac{\lambda L_m}{n}}. \qquad (49)$$

The function

$$H_\nu(\xi) = (-1)^\nu e^{\xi^2}\,(d^\nu/d\xi^\nu)e^{-\xi^2}$$

is the Hermite polynomial[12] of order $\nu$. Hermite polynomials of lowest degree are $H_0(\xi) = 1$; $H_1(\xi) = 2\xi$; $H_2(\xi) = 4\xi^2 - 2$; and $H_3(\xi) = 8\xi^3 - 12\xi$.

Expression (48) can be simplified provided that the important terms of the summation are those for which

$$\nu\lambda/L_1 \ll 1$$

and (50)

$$\mu\lambda/L_2 \ll 1.$$

Then the square root in the exponent can be replaced by the first two terms of a power series expansion and

$$\begin{aligned} E_x \cong \exp\left(-i\theta - \frac{x^2}{w_1^2} - \frac{y^2}{w_2^2}\right)&\left[\sum_{\nu=0}^{\infty} A_\nu \exp\left[i(\pi\nu z/L_1)\right]H_\nu\left(\frac{\sqrt{2}x}{w_1}\right)\right] \\ &\times\left[\sum_{\mu=0}^{\infty} B_\mu \exp\left[i(\pi\nu z/L_1)\right]H_\mu\left(\frac{\sqrt{2}y}{w_2}\right)\right] \end{aligned} \qquad (51)$$

where

$$\theta = knz\left[1 - \frac{\lambda}{4n}\left(\frac{1}{L_1} + \frac{1}{L_2}\right)\right].$$

We will look for a periodic field configuration that reproduces itself at each lens. For reasons of symmetry, then, the planes of symmetry of the lenses ($z = 0$) and gaps ($\zeta = 0$) must be equiphase surfaces.

We choose the field at the plane $z = 0$ to be

$$\exp\left(-\frac{x^2}{s_1^2}-\frac{y^2}{s_2^2}\right) H_p\left(\frac{\sqrt{2}x}{s_1}\right) H_q\left(\frac{\sqrt{2}y}{s_2}\right)$$

where $p$ and $q$ are integers and $s_1$ and $s_2$ are arbitrary parameters for the time being. Therefore for $z = 0$, we obtain from (51)

$$\sum_{\mu=0}^{\infty} A_\nu \exp\left(-\frac{x^2}{w_1^2}\right) H_\nu\left(\frac{\sqrt{2}x}{w_1}\right) = \exp\left(-\frac{x^2}{s_1^2}\right) H_p\left(\frac{\sqrt{2}x}{s_2}\right) \quad (52)$$

and

$$\sum_{\mu=0}^{\infty} B_\mu \exp\left(-\frac{y^2}{w_2^2}\right) H_\mu\left(\frac{\sqrt{2}y}{w_2}\right) = \exp\left(-\frac{y^2}{s_2^2}\right) H_q\left(\frac{\sqrt{2}y}{s_2}\right). \quad (53)$$

Using the orthogonality properties of the Gaussian-Hermitian product,[12] we obtain

$$A_\nu = \frac{\sqrt{2}}{\sqrt{\pi 2^\nu \nu!}\, w_1} \int_{-\infty}^{\infty} \exp\left[-\xi^2\left(\frac{1}{s_1^2}+\frac{1}{w_1^2}\right)\right] \cdot H_p\left(\frac{\sqrt{2}\xi}{s_1}\right) H_\nu\left(\frac{\sqrt{2}\xi}{w_1}\right) d\xi \quad (54)$$

and a similar expression for $B_\mu$. Replacing the result in (51) and performing, as in Ref. 13, first the summation in $\nu$ and $\mu$ and then the integration in $\xi$, the transverse field component inside a lens expressed in closed form results

$$\begin{aligned} \mathrm{E}_x = \exp\Big\{-i\Big[kn\left(z-\frac{x^2}{2R_1^2}-\frac{y^2}{2R_2^2}\right)-\left(p+\frac{1}{2}\right) \\ \cdot \tan^{-1}\left(\frac{w_1^2}{s_1^2}\tan\frac{\pi z}{L_1}\right)-\left(q+\frac{1}{2}\right)\tan^{-1}\left(\frac{w_2^2}{s_2^2}\tan\frac{\pi z}{L_2}\right)\Big] \\ \times \exp\left[-\left(\frac{x}{\rho_1}\right)^2-\left(\frac{y}{\rho_2}\right)^2\right] H_p\left(\sqrt{2}\,\frac{x}{\rho_1}\right) H_q\left(\sqrt{2}\,\frac{y}{\rho_2}\right) \end{aligned} \quad (55)$$

where, for $m = 1$ or $2$

$$R_m = \frac{L_m}{\pi}\left\{\frac{1+\left(\dfrac{w_m}{s_m}\right)^4}{\left[1-\left(\dfrac{w_m}{s_m}\right)^4\right]\sin\dfrac{2\pi z}{L_m}}+\operatorname{ctn}\frac{2\pi z}{L_m}\right\} \quad (56)$$

and

$$\rho_m = s_m \sqrt{\frac{1}{2}\left\{1 + \left(\frac{w_m}{s_m}\right)^4 + \left[1 - \left(\frac{w_m}{s_m}\right)^4\right] \cos \frac{2\pi z}{L_m}\right\}}. \tag{57}$$

The electric field in the uniform dielectric gap between two lenses can be derived from the previous expression by making

$$n = 1$$

and

$$1/L_m = 0$$

and by substituting another symbol, $s_{gm}$, for $s_m$. Again we demand the plane of symmetry of the gap, $z = (b + t)/2$ (see Fig. 1), to be an equiphase surface. This is achieved by substituting $\zeta = z - (b + t)/2$ for $z$.

The electric field in the gap is then

$$\begin{aligned} E_{xg} = \exp \Bigg\{ -i \Bigg[ k\left(\zeta - \frac{x^2}{2R_{g1}} - \frac{y^2}{2R_{g2}}\right) - \left(p + \frac{1}{2}\right) \tan^{-1} \frac{2\zeta}{k{s_{g1}}^2} \\ - \left(q + \frac{1}{2}\right) \tan^{-1} \frac{2\zeta}{k{s_{g2}}^2}\Bigg] - \left(\frac{x}{\rho_{g1}}\right)^2 - \left(\frac{y}{\rho_{g2}}\right)^2\Bigg\} \\ \cdot \mathrm{H}_p\left(\sqrt{2}\,\frac{x}{\rho_{g1}}\right) H_q\left(\sqrt{2}\,\frac{y}{\rho_{g2}}\right) \end{aligned} \tag{58}$$

where

$$R_{gm} = \frac{k^2{s_{gm}}^4}{4\zeta}\left[1 + \left(\frac{2\zeta}{k{s_{gm}}^2}\right)^2\right] \tag{59}$$

and

$$\rho_{gm} = s_{gm} \sqrt{1 + \left(\frac{2\zeta}{k{s_{gm}}^2}\right)^2}. \tag{60}$$

To match the fields (55) and (58) at the interfaces, the $x$ and $y$ dependences of the field at both sides must coincide. The fact that it can be matched guarantees that Maxwell's equations are satisfied simultaneously in lenses and gaps. It can be verified that if the tangential electric field continuity is satisfied, the tangential magnetic field continuity is also guaranteed. By considering waves propagating in both directions, it could be possible to take into account reflections at the interfaces, but we shall instead assume that at each interface there is a matching mechanism that prevents reflections. Notice that in the case of gaseous lenses the small changes of dielectric constants automatically insure negligible reflection at the interfaces.

The exact matching of the fields at the interfaces is achieved by making equal the coefficients of $x$, $y$, $x^2$ and $y^2$ in both expressions (55) and (58) at the boundary $z = t/2$ of the lens and $\zeta = -b/2$ of the gap. Then

$$R_{m(z = t/2)} = R_{gm(\zeta = -b/2)} \tag{61}$$

$$\rho_{m(z = t/2)} = \rho_{gm[\zeta = -(b/2)]} \, . \tag{62}$$

From them, together with (56), (57), (59) and (60), we deduce the values of $s_m$ and $s_{gm}$ that guarantee the matching at the interfaces. They are:

$$s_m = w_m \left[ \frac{1 + C_m \operatorname{ctn} \varphi_m}{1 - C_m \tan \varphi_m} \right]^{\frac{1}{4}} \tag{63}$$

and

$$s_{gm} = w_m(1 + C_m \operatorname{ctn} \varphi_m)^{\frac{1}{4}} \, (1 - C_m \tan \varphi_m)^{\frac{1}{4}} \tag{64}$$

where

$$C_m = n(\pi/2) \, (b/L_m) \tag{65}$$

$$\varphi_m = (\pi/2) \, (t/L_m) \tag{66}$$

$$w_m = (1/\pi) \sqrt{\lambda L_m / n}. \tag{67}$$

## REFERENCES

1. Pierce, J. R., Modes in Sequences of Lenses, Proc. Nat. Acad. Sci., **47,** 1961, pp. 1808–31.
2. Berreman, D. W., A Lens or Light Guide Using Convectively Distorted Thermal Gradients in Gases, B.S.T.J., **43,** July, 1964, pp. 1469–1475.
3. Berreman, D. W., A Gas Lens Using Unlike, Counter-Flowing Gases, B.S.T.J., **43,** July, 1964, pp. 1476–1479.
4. Marcuse, D., and Miller, S. E., Analysis of a Tubular Gas Lens, B.S.T.J., **43,** July, 1964, pp. 1759–1782.
5. Fox, A. G., and Li, Tingye, Resonant Modes in a Maser Interferometer, B.S.T.J., **40,** March, 1961, pp. 453–488.
6. Boyd, G. D., and Gordon, J. P., Confocal Multimode Resonator for Millimeter through Optical Wavelength Masers, B.S.T.J., **40,** March, 1961, pp. 489–508.
7. Boyd, G. D., and Kogelnik, H., Generalized Confocal Resonator Theory, B.S.T.J., **41,** July, 1962, pp. 1347–1369.
8. Goubau, G., and Schwering, F., On the Guided Propagation of Electromagnetic Wave Beams, Trans. I.R.E., **AP-9,** May, 1961, pp. 248–255.
9. Tonks, L., Filamentary Standing-Wave Pattern in a Solid State Maser, J. Appl. Phys., June, 1962, pp. 1980–1986.

10. Schachter, H., and Bergstein, L., Resonant Modes of Optic Cavities with Inhomogeneous Host Media, J. Opt. Soc. Am., **54,** April, 1964, p. 567.
11. Tien, P. K., private communication.
12. Jahnke, Emde and Lösch, *Tables of Higher Functions*, McGraw-Hill, New York, 1960, pp. 101–104.
13. Erdélyi, Magnus, Oberhettinger and Tricomi, *Higher Transcendental Functions*, McGraw-Hill, New York, 1953, vol. 2, pp. 192–196.
14. Kogelnik, H., Modes in Optical Resonators, to appear in *Advances in Lasers*, Dekker Publications.

# Substitution of Laminated Low-Carbon Steel for Silicon Steel in the Cores of Wire Spring Relays

By WILLIAM C. SLAUSON

(Manuscript received May 13, 1964)

*This article describes the analytical and laboratory studies undertaken to determine if low-carbon steel could be substituted for the more expensive 1 per cent silicon steel in the cores of general-purpose wire spring relays. Not only is this silicon steel more costly, but its hardness characteristics are such that tool maintenance for manufacture is an appreciable item. It was found that this substitution can be made without degrading the performance of these relays, provided the new core is made up of two laminations. When two laminations are used, the eddy current time constant of low-carbon steel matches that of the silicon steel. This is necessary to achieve the fast operate and release times now obtained and to permit satisfactory operation of present circuits when the substitution is made.*

*This substitution will result in substantial annual manufacturing savings for the general-purpose relays. These savings could be further increased if use of the new core could be extended to other, more special, relays of the wire spring relay family. These applications are now under study.*

## I. INTRODUCTION

The wire spring family of relays (see Fig. 1) was designed to serve as the basic components of modern switching circuits. It was first introduced in the No. 5 crossbar switching system and later in other systems, including a wide variety of switching applications for the Bell System. The design provides an electromagnetic device of high efficiency and reliability with excellent operating characteristics and suited to a high degree of automation in manufacture. The relays are obtainable in a wide variety of codes with different coil resistances and are capable of controlling from 1 to 24 contact sets per relay in various combinations of makes, breaks, transfers, and operating and releasing time intervals, ranging from a few milliseconds to longer than one-half second in slow-

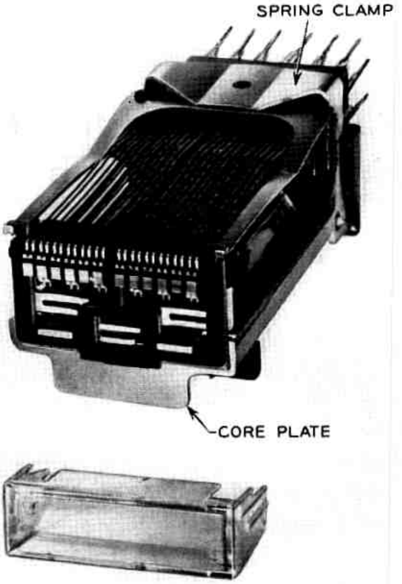


Fig. 1 — Wire spring relay.

release applications. The operating life of these relays approaches approximately one billion operations. In view of these considerations and the outstanding performance record of the relays now in use, the demand has been continually rising over the past years. Since the production of general-purpose wire spring relays began in 1950, more than one hundred million have been manufactured, and approximately twenty million of these were produced in 1963.

One per cent silicon steel was chosen originally as the magnetic core material for these relays because of its high resistivity, good magnetic properties, low aging characteristics, and its ability to achieve fast, efficient and stable operating characteristics. This material has proven satisfactory but over the years has presented some manufacturing and procurement problems. Therefore, consideration has been given to the use of alternate materials, particularly low-carbon steel for the relay cores. This article will discuss the theoretical and the manufacturing aspects involved in substituting laminated low-carbon steel for the 1 per cent silicon steel cores. Only the general run of fast-operate and fast-release relays will be considered, since the major effect of the material substitution is due to the difference in the eddy current time constants (because of the low resistivity of low-carbon steel). The slow-release relays will not be discussed, since they have built-in eddy current inducers, such as short-circuited windings or copper sleeves, which reduce the eddy current conductance of the magnet core to secondary importance.

## II. PRESENT CORE DESIGN

When the general-purpose wire spring relay was developed in the late 1940's, the best material available for magnetic cores was 1 per cent silicon steel. As a result, the magnetic design of the relay was based on the use of this material. One per cent silicon steel has relatively high electrical resistivity, which keeps the eddy current time constant of the structure to a minimum, thus permitting fast operate and release times. The material has good magnetic properties with relatively high flux densities, which permit the development of ample pull forces between the relay armature and core. Also, it is a stable magnetic material that does not exhibit any significant change in properties with time. As a result, general-purpose wire spring relays with silicon steel cores have delivered reliable performance with good operating margins for the past fourteen years.

The disadvantages of silicon steel have been in the manufacturing and procurement areas. This steel is relatively difficult to fabricate by punching, because it has an abrasive action on punches and dies which necessitates frequent tool maintenance. Also, it is not as readily available and is more expensive than the low-carbon steels, such as S.A.E. 1010 steel. As a result of these considerations, there has been a continuing effort to find or adapt a substitute magnetic material for cores of general-purpose wire spring relays. This has led to the development described in the subsequent sections of this article.

## III. LAMINATED CORE PROPOSAL

Recently a new and relatively inexpensive method of annealing low-carbon steel to obtain good and stable magnetic characteristics has been developed using "wet forming" gas. This annealing technique produces in low-carbon S.A.E. 1010 steel magnetic properties comparable to those obtained with 1 per cent silicon steel. Thus, low-carbon steel, which is much less expensive, could be considered as a substitute for silicon steel in the cores of general-purpose wire spring relays. However, S.A.E. 1010 steel has a nominal electrical resistivity of 12 microhm $\mathrm{cm}^{-1}$, whereas the comparable value for 1 per cent silicon steel is 25 microhm $\mathrm{cm}^{-1}$.

The eddy current conductance of a material is inversely proportional to the electrical resistivity; hence, the eddy current conductance of S.A.E. 1010 steel is about twice as great as that for 1 per cent silicon steel. Since the operate and release times of a relay are directly affected by the eddy current conductance, a direct substitution of S.A.E. 1010 steel for 1 per cent silicon steel would materially increase these times in general-purpose wire spring relays. Such a change in performance would be intolerable, since many switching circuits are designed to take full advantage of the fast operate and release times obtainable with the present relays. An increase in these times would result in circuit-race conditions or add to circuit holding times, thereby increasing the number of common control units needed in a central office.

Theoretically, if the volume and shape of a piece of magnetic material are not changed but the material is laminated with equal-size laminations, the effective eddy current conductance is reduced by an amount which is inversely proportional to the number of laminations. Also, if the cross-sectional area is rectangular, the eddy current conductance is further reduced as the ratio of width to thickness is increased. Thus, laminating a rectangular cross-section magnetic part reduces the effective eddy current conductance to between $1/N$ and $1/N^{\frac{3}{2}}$ of the unlaminated value, where $N$ is the number of equal-size laminations.

Taking advantage of the recent development in annealing and the concept of laminations, it was therefore proposed that the 1 per cent silicon steel core of the general-purpose relay be replaced with a core made up of two equal laminations of S.A.E. 1010 low-carbon steel. The effect of this material change on the operation of the relay will be discussed from both the practical and theoretical aspects, with a view to showing that the change results in a relay equal in performance to those produced during the past several years, at a considerable cost saving.

## IV. PRACTICAL ASPECTS OF THE PROPOSAL

In order to introduce a substitute design for a functional part of the relay, the following factors must be considered:

### 4.1 *Operational Factors*

(1) time characteristics
  (a) electrical operate time
  (b) electrical release time
(2) magnetic pull vs ampere turns
(3) heating (watts input vs temperature rise)

(4) life

(a) wear vs number of operations

(b) wear effect on operate and release characteristics

(5) core plate tightness (due to staking efficiency of softer 1010 steel material)

(6) corrosion protection (effectiveness of plating along laminated seam if the laminations are welded before plating).

### 4.2 *Manufacturing Factors*

(1) Shape of laminations

(a) economic considerations (punching properties and tool life)

(b) assembly considerations (dimensional considerations for spoolhead and core plate areas)

(2) cost of material

(3) loose vs attached laminations

(a) handling ease

(b) assembly ease

(4) method of punching

(a) single (one at a time)

(b) double (two at a time, i.e., one on top of the other)

(5) welding or mating of laminations

(a) before punching

(b) after punching

(c) location of welds.

To have the laminated S.A.E. 1010 steel core accepted for use in the relay, the new design must perform as well as the old design with regard to all of the operational factors and should have definite advantages with regard to the manufacturing factors. In order to obtain the maximum improvement in the manufacturing area without affecting the over-all relay, it was necessary to introduce the minimum number of changes to the structure. As a result, the object of the laminated core proposal was to match, as nearly as possible, all of the characteristics of the present general-purpose wire spring relay having a silicon steel core with a new core of the same physical outside dimensions.

Analysis of the magnetic properties of S.A.E. 1010 steel annealed by the wet forming gas method indicated that sufficient magnetic pull would be developed with this material provided the efficiency of the relay's magnetic circuit was maintained. However, as indicated previously and analyzed in detail in the next section, laminating the core was necessary in order to reduce the eddy current conductance to tolerable levels. To be assured that the benefits of laminating the core would always be present, it was at first believed necessary to physically separate the laminations by depressions or an insulating film to prevent the flow of eddy currents between the laminations. From a manufacturing standpoint, it was decided that if this were necessary, it would be more practical to depress a large section of one of the laminations instead of using an insulating finish. As a result, the first sample cores were made this way. For comparison, a standard core is shown in Fig. 2 and laminated cores of the first design in Figs. 3 and 4.

Fig. 4, a side view of the laminations, shows the recessed section in the upper lamination to provide an air gap over the greater portion of the length of the core. The two laminations are in intimate contact at the two ends to provide a low-reluctance path for the magnetic flux to pass from the bottom lamination through the upper to the relay armature. However, timing tests of various combinations of recessed laminations, as well as flat laminations, in relays have shown that it is not necessary to create a positive or visual air gap between the parts. Apparently, the surface resistance of the laminations due to normal oxide films is sufficient to keep the eddy currents of the laminations from combining.

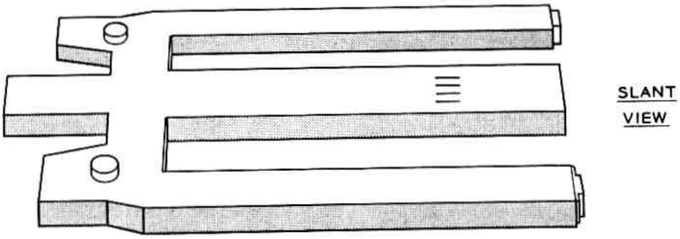


Fig. 2 — Standard one-piece core.

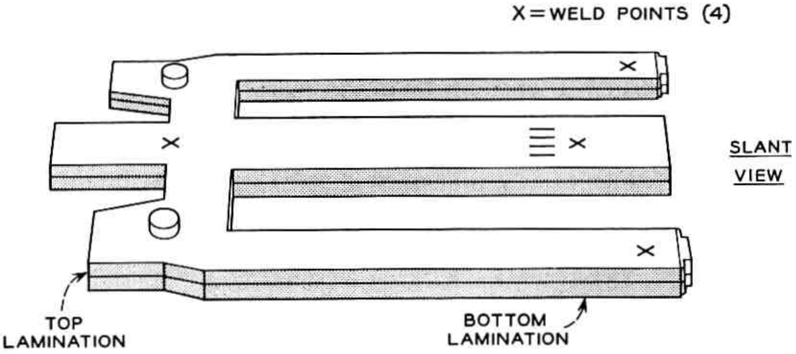


Fig. 3 — Laminated core.

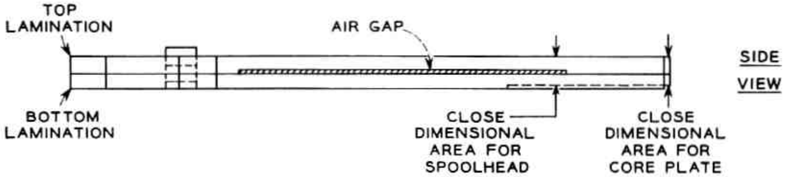

TOP LAMINATION
AIR GAP
SIDE VIEW
BOTTOM LAMINATION
CLOSE DIMENSIONAL AREA FOR SPOOLHEAD
CLOSE DIMENSIONAL AREA FOR CORE PLATE

Fig. 4 — Side view of laminated core.

In Fig. 3, a slant view of the laminations, four welds are shown at points marked (X). The welding was done before magnetic annealing or plating of the parts. The welds are located at the front and rear, where they have the least effect on the eddy currents in each lamination. The welds are proposed only to associate the two laminations punched together and to facilitate the assembly of the relay. Relays were assembled and tested and satisfactory results obtained when the two laminations were not welded together. The laminations in this case were held together by the core plate at the front end and the spring clamp at the rear end after the relay had been assembled (see Fig. 1).

## V. THEORETICAL ASPECTS OF THE PROPOSAL

### 5.1 *Symbols*

The following is a list of the symbols used in the theoretical discussions presented in the balance of this article. Where variant forms of these symbols, as distinguished by subscripts, are used in the text, they are defined in connection with the specific use.

$a_1$ — inner surface area of coil
$a_2$ — outer surface area of coil
$E$ — applied voltage
$F$ — magnetic pull
$G$ — total equivalent single-turn conductance
$G_c$ — equivalent single-turn conductance of coil — $N^2/R$
$G_e$ — equivalent single-turn eddy current conductance
$H_c$ — coercive force (oersteds)
$I$ — steady-state current
$i$ — instantaneous current
$K$ — total thermal conductance
$k_1$ — thermal conductance per unit area of coil inner surface
$k_2$ — thermal conductance per unit area of coil outer surface
$L_1$ — single-turn inductance — $(4\pi/\mathcal{R}_1)$ or $(\phi/NI)$
$N$ — number of turns in coil

| | |
|---|---|
| $NI$ | — steady-state ampere turns |
| $Ni$ | — just operate or just release ampere turns |
| $NI_c$ | — ampere-turn coercive force — $(H_c l/0.4\pi)$ |
| $R$ | — coil resistance |
| $t$ | — time |
| $t_1$ | — waiting time |
| $v$ | — ratio of flux at time $t$ to steady-state flux |
| $W$ | — power — $I^2R$ |
| $\mathfrak{F}$ | — magnetomotive force — $4\pi NI$ or $\mathcal{R}\phi$ |
| $l$ | — length of magnetic path |
| $\mathcal{R}$ | — reluctance of relay |
| $\mathcal{R}_c$ | — reluctance of core |
| $\mathcal{R}_r$ | — initial incremental reluctance |
| $\phi$ | — flux |
| $\phi_1$ | — steady-state flux |
| $\phi''$ | — saturation flux |
| $\phi_0$ | — residual flux |
| $\theta_0$ | — ambient temperature |
| $\bar{\theta}$ | — mean coil temperature. |

### 5.2 *General*

Since the only change proposed in the relay is the core material, the armature, contact springs, balance spring and other operating parts will be unchanged. The mass and spring forces of these parts control the travel time of the relay in both the operate and release directions. The change in core material will, because of a different eddy current conductance, affect only the electrical waiting time of the relay. Therefore, the electrical waiting time on the operate and then on the release of the relay will be considered first.

#### 5.2.1 *Electrical Waiting Time — Operate*

The electrical waiting time on operate of a relay is defined as the time from the application of potential to the relay coil until the magnetic pull on the armature equals the back force on the armature and it starts to move. During this period the flux development in the structure follows the relationship:

$$\mathfrak{F} - \mathfrak{F}_s + 4\pi G \frac{d\phi}{dt} = 0 \tag{1}$$

where $\mathfrak{F}_s$ is the steady-state magnetomotive force (mmf) or $4\pi NI$, $\mathfrak{F}$

is the effective mmf and is equal to $\mathcal{R}\phi$, and $G$ is equal to the coil conductance ($G_c$) plus the eddy current conductance ($G_e$). The reluctance ($\mathcal{R}$)is a function of the armature-core air gap ($X$) and the flux ($\phi$). With the armature at rest against the back stop and using the initial conditions $X_1$ and $\mathcal{R}_1$, (1) can be rewritten in the integral form with $\phi_1\mathcal{R}_1$ substituted for $\mathfrak{F}_s$ as follows:

$$t_1 = \frac{4\pi G}{\mathcal{R}_1} \int_0^\phi \frac{d\phi}{\phi_1 - \phi} \tag{2}$$

which on integration gives:

$$t_1 = \frac{4\pi G}{\mathcal{R}_1} \ln \frac{1}{1 - v}. \tag{3}$$

Since $v$, the ratio of flux attained at time $t$ to the steady-state flux, can be written as the ratio of the just operate ampere turns to the steady-state ampere turns $Ni/NI$, and $4\pi/\mathcal{R}_1$ is equal to $L_1$, the single-turn inductance, (3) may be rewritten as follows:

$$t_1 = L_1(G_c + G_e) \ln \frac{1}{1 - (Ni/NI)} \tag{4}$$

which is the general form of the equation for the electrical waiting time of a relay on operate.

### 5.2.2 *Electrical Waiting Time — Release*

The release waiting time of a relay is defined as the time from the opening of the coil circuit to the beginning of motion of the armature from the operated position. The opening of the circuit results in a decaying magnetic field which is sustained only by eddy currents. The release waiting time is described by the same relationship as the operate waiting time except that since $\mathfrak{F}_s = 0$ with the coil circuit open (1) becomes:

$$\mathfrak{F} + 4\pi G \frac{d\phi}{dt} = 0. \tag{5}$$

The normal-release waiting time of a relay without copper sleeves can only be determined approximately because of the variable distribution of the magnetic field sustained only by eddy currents ($G_e$). However, if the flux decay is retarded by the introduction of a conductance much larger than $G_e$, such as a copper sleeve or short-circuited coil turns, the decaying field is nearly uniform and a relatively close approximation to (5) can be made. Equations thus derived for the slow-release case can be used for the approximate analysis of the normal-release time case. Then (5) can be written in the integral form:

$$t_1 = 4\pi G \int_{\phi}^{\phi_1} \frac{d\phi}{\mathfrak{F}} \tag{6}$$

where $t_1$ is the waiting time for the flux to decay from the steady-state value ($\phi_1$) to the value ($\phi$) at which the magnetic pull is equal to the retractile force.

For reliable and repetitive release times, the steady-state flux ($\phi_1$) of a relay should be in the region of flux saturation ($\phi''$). Therefore, release times will only be considered from this condition. Since $\mathfrak{F} = \mathcal{R}\phi$ and the relationship between $\phi$ and $\mathfrak{F}$ is in the demagnetization curve, the following equation results:

$$\mathfrak{F} = \frac{(\phi'' - \phi_0)(\phi - \phi_0)}{(\phi'' - \phi)} \mathcal{R}_r \tag{7}$$

where $\mathcal{R}_r$ is the incremental reluctance when $\mathfrak{F} = 0$.

If (7) is substituted in (6), the expression for release waiting time becomes:

$$t_1 = \frac{4\pi G}{\mathcal{R}_r} \int_{\phi}^{\phi''} \left( \frac{1}{\phi - \phi_0} - \frac{1}{\phi'' - \phi_0} \right) d\phi. \tag{8}$$

Integration of this equation results in:

$$t_1 = \frac{4\pi G}{\mathcal{R}_r} [\ln Z - 1 + (1/Z)], \tag{9}$$

where

$$Z = \frac{\phi'' - \phi_0}{\phi - \phi_0}.$$

To have a more readily usable relationship for release waiting time, it is necessary to obtain an expression for $Z$ in terms of $\mathfrak{F}$ or $4\pi Ni$. Equation (7) can be rewritten to give the following expression for:

$$Z = 1 + \frac{\phi'' - \phi_0}{4\pi Ni} \mathcal{R}_r. \tag{10}$$

Substituting the expression for $\mathcal{R}_r$ as obtained from (10) in (9), and recognizing that with the coil circuit open and no sleeves or short-circuited turns the only conductance involved is the eddy current conductance ($G_e$), the following expression for release waiting time is obtained:

$$t_1 = \frac{G_e(\phi'' - \phi_0)}{Ni}\left(\frac{\ln Z}{Z - 1} - \frac{1}{Z}\right). \tag{11}$$

## VI. EVALUATION OF OPERATE AND RELEASE WAITING TIME

Equations (4) and (11), for operate and release waiting times respectively, can be evaluated by obtaining values for the variables experimentally and graphically. In this section the procedures for the establishment of the relay parameters will be discussed.

The first data needed are magnetization curves of flux vs ampere turns with the armature in the unoperated position for the evaluation of operate waiting time and in the operated position for release waiting time. Typical curves are shown in Figs. 11 through 16 (see Section VIII). With measured values of just operate, just release and steady-state current, all of the flux values for (4) and (11) can be read from the curves. The inductance per turn ($L_1$) may be found by drawing a line through the origin of the unoperated magnetization curve tangent to the nearly flat or linear portion of the curve (see Fig. 5). The slope of the tangent is a reliable value for $L_1$ if the just operate flux falls on the linear portion of this curve.

Since $G$ in (2) and (6) is equal to the sum of $G_e$ and $G_c$, the effective eddy current conductance $G_e$ can be determined reasonably well graphically and experimentally by holding the values of the integrals of the two equations fixed and making timing measurements as $G_c$, which is equal

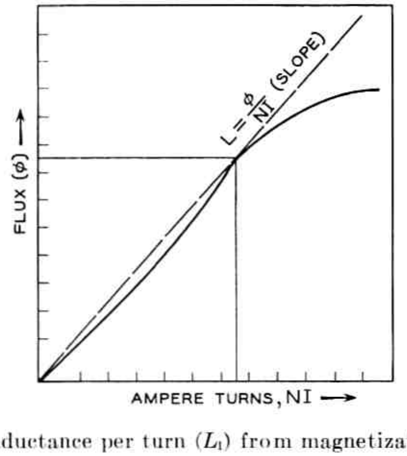


Fig. 5 — Inductance per turn ($L_1$) from magnetization curves.

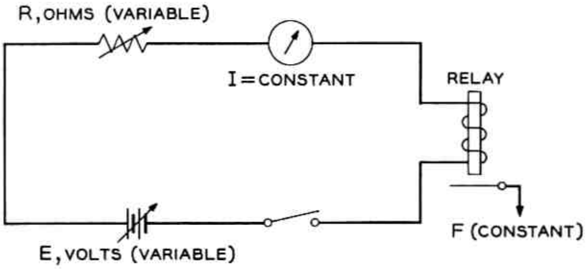


Fig. 6 — Circuit for varying $G_c$ or $N^2/R$ by changing $R$ (coil in operate condition).

to $N^2/R$, is varied. A fixed value for either integral can be assured by having the relay adjusted so that its steady-state current and either just operate or just release current are maintained constant throughout the experiment. $G_c$ or $N^2/R$ can then be varied by changing the resistance in series with the coil as shown in Fig. 6 for the operate condition and Fig. 7 for the release condition. Since all factors except $G_c$ are held constant, a plot of $G_c$ versus waiting time will be linear, and when extrapolated to $t = 0$ will have a negative intercept on the $G_c$ axis equal to $G_e$ as shown in Fig. 8.

Values for all of the variables in (4) and (11) were determined for both one-piece 1 per cent silicon steel cores and laminated S.A.E. 1010 steel cores and the waiting time for operate and release computed. The computed values are compared to measured values in a later section.

## VII. EXPERIMENTAL DETERMINATION OF OPERATE AND RELEASE TIMES

In production, permissible dimensional tolerances of the parts and differences in the magnetic characteristics of the cores due to material

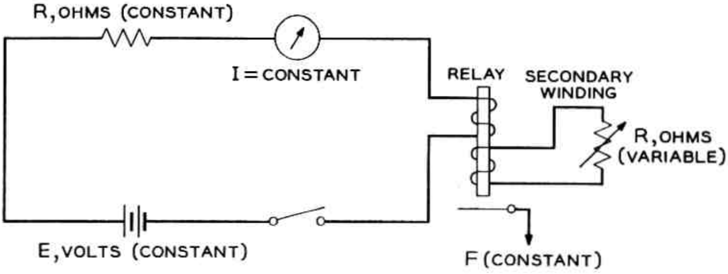


Fig. 7 — Circuit for varying $G_c$ or $N^2/R$ by changing $R$ (coil in release condition).

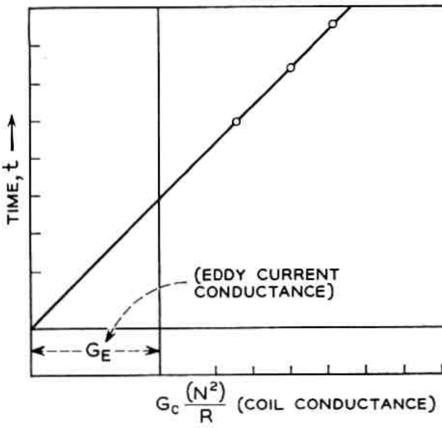


Fig. 8 — Waiting time vs coil conductance.

and annealing variations will result in variations in the operational characteristics of the relays. To evaluate these effects, two sample groups of relays with each core material were constructed. Group SP was made with minimum-dimension parts having a poor anneal and group LG had maximum-dimension parts with a good anneal. Fig. 9 shows the operate time of these relays as a function of coil conductance with an unoperated air gap of 0.032 inch and with input powers of 2 and 10 watts. It will be noted that the laminated S.A.E. 1010 steel core relays are approximately 1.3 to 3.4 per cent faster than the 1 per cent silicon steel core relays.

Fig. 10 shows the release times and release pull values for the same groups of relays. Here, it is noted that the laminated S.A.E. 1010 steel core relays release from 13 to 27 per cent faster than the 1 per cent silicon steel core relays.

## VIII. COMPARISON OF CALCULATED AND MEASURED TIMES

The measured times shown in Figs. 9 and 10 all include some armature movement. However, all of the relays had essentially the same mechanical adjustments, and the same SP and LG armatures were used on both the laminated S.A.E. 1010 steel and 1 per cent silicon steel cores. Therefore, it can be assumed that the mechanical armature travel times of the relays were essentially the same. As a result, a comparison of like sets of relays, i.e., SP laminated versus SP one-piece, etc., should reflect the difference in electrical waiting time of the groups being compared.

A comparison of the calculated times for the like sets of relays was also made using the values of $L$, $G_e$, $\phi$, $\phi_0$, etc., obtained from Figs. 11

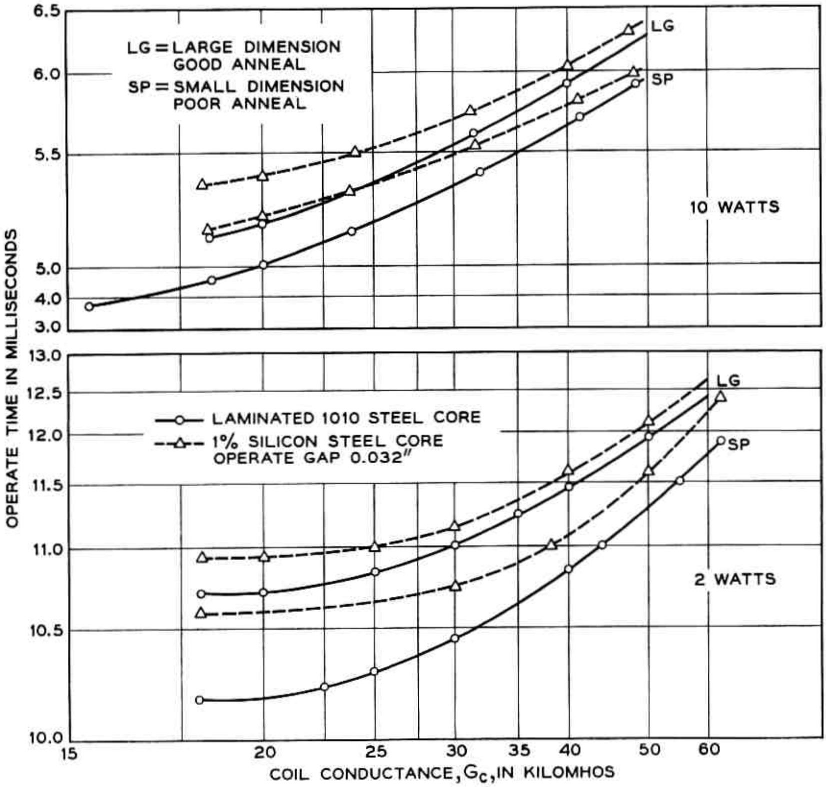


Fig. 9 — Change in relay operate time due to variation in relay dimensions and quality of anneal. Relays had cores of either laminated 1010 steel or 1% silicon steel.

through 16 as described earlier. Tables I and II show the comparison between the laminated S.A.E. 1010 core relays and the 1 per cent silicon steel core relays as determined by measurement and by calculation. In all cases good agreement was found between the measured and calculated values.

## IX. COMPARISON OF EDDY CURRENT CONDUCTANCE

In Tables I and II are listed the values of eddy current conductance obtained for the SP and LG groups of relays. 7.20 and 8.25 kilomhos respectively were found for the 1 per cent silicon steel cores and 5.55 and 6.65 kilomhos respectively were found for the laminated S.A.E. 1010 steel cores. Since the S.A.E. 1010 steel has approximately twice the con-

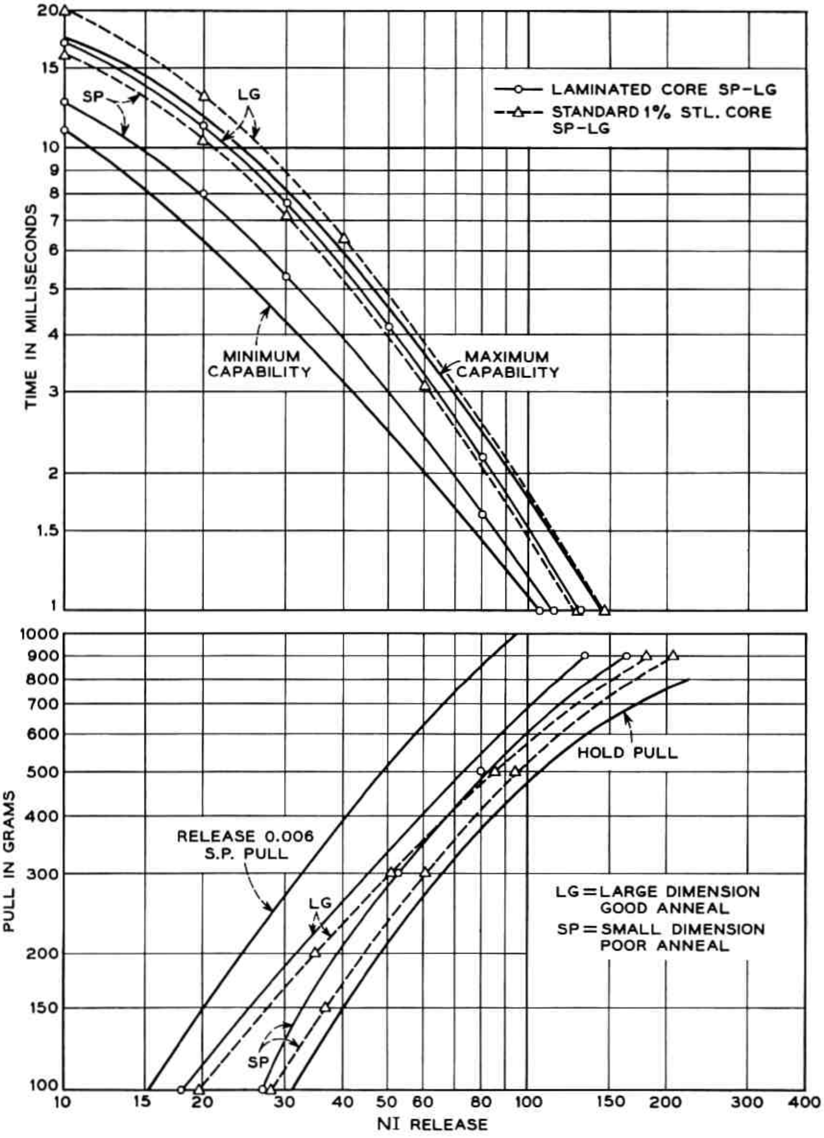


Fig. 10 — Release pull values (grams) and release times (ms) for relays with cores differing in dimension, quality of anneal, and type of steel used.

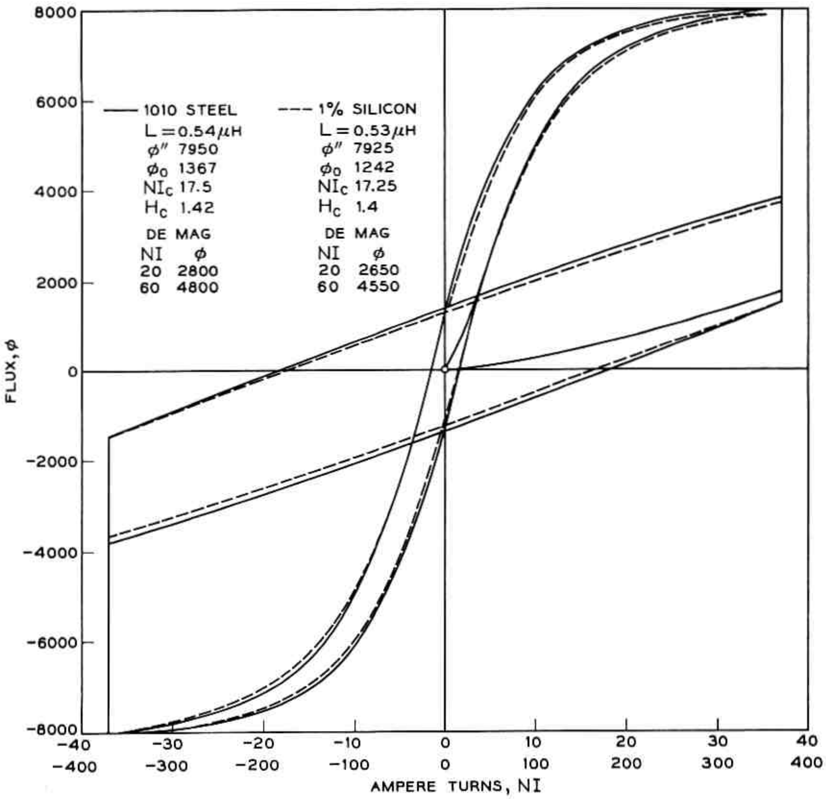


Fig. 11 — Magnetization curves of cores with small dimensions and poor anneal (0.006 air gap).

ductivity of 1 per cent silicon steel, the eddy current conductance of the S.A.E. 1010 cores would have been about 14.4 and 16.5 kilomhos for the SP and LG groups respectively if the cores had not been laminated. Thus laminating reduced the eddy current conductance of the cores by the ratios of 5.55/14.4 or 38.5 per cent for the SP group of relays and 6.65/16.5 or 40.3 per cent for the LG group of relays. Since it was expected that the use of two laminations would reduce the effective eddy current conductance by a ratio of between $1/N$ and $1/N^{\frac{3}{2}}$ or between 50 and 35.4 per cent, good agreement with the theoretical analysis is indicated.

## X. COIL HEATING (POWER INPUT VS TEMPERATURE RISE)

Another major consideration is the effect of using a laminated core and a new material on the dissipation of heat from the relay coil. The

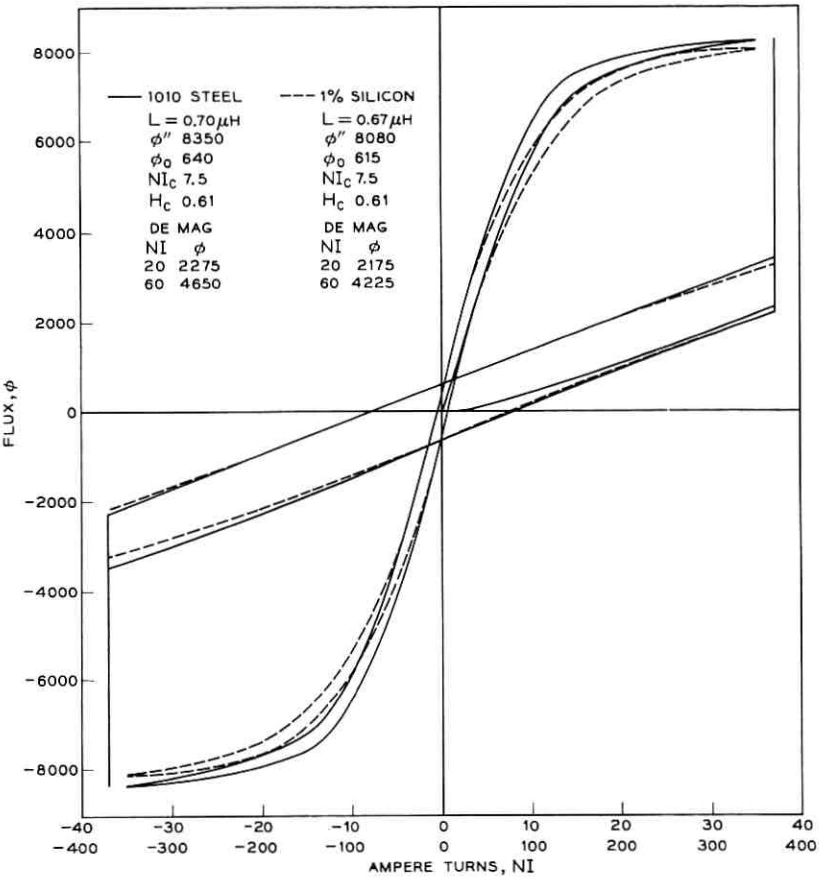


Fig. 12 — Magnetization curves of cores with large dimensions and good anneal (0.006 air gap).

allowable mean temperature rise of a relay coil is limited by two factors. The first of these is that in normal operation the temperatures should not rise to a point that is dangerous to personnel in case of physical contact. This limit has been established for many years in the Bell System at 225°F. The temperature rise in normal operation is a function of the duty cycle of the relay and is influenced, therefore, by its circuit application. The second limitation on temperature rise — that the relay shall not become a fire hazard in case of indefinite energization — is of more direct interest from an apparatus standpoint. With the normal wire insulations and coil insulating materials used in Bell System relays, it has been found that a maximum mean coil temperature of 360°F can

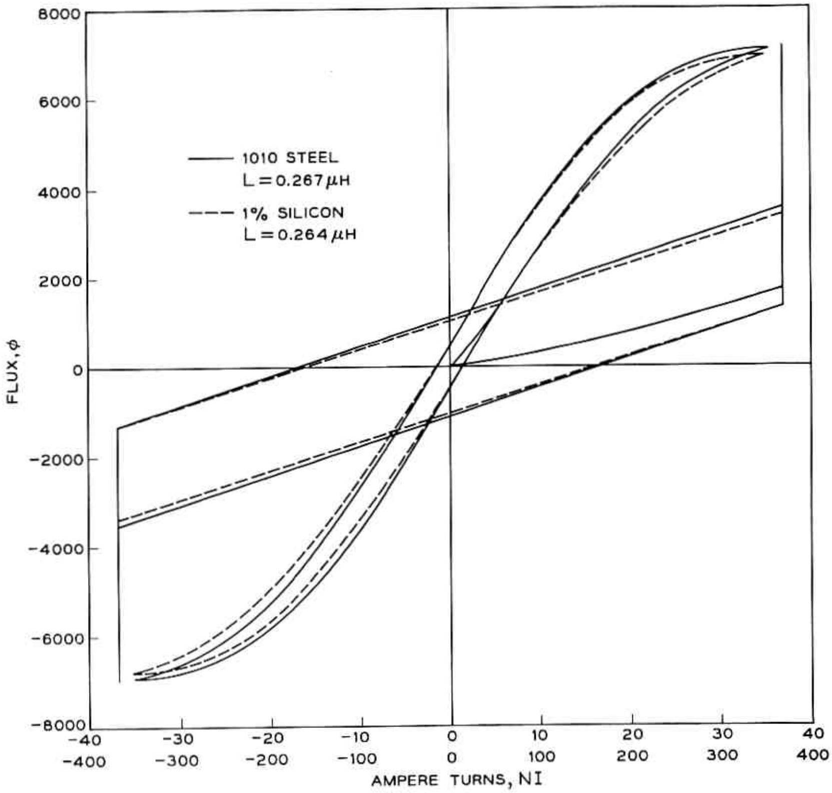


Fig. 13 — Magnetization curves of cores with small dimensions and poor anneal (0.032 air gap).

be allowed with essentially no risk of insulation breakdown which would produce a fire hazard. It is recognized, though, that prolonged exposure of a relay to such a temperature could result in permanent damage.

The dissipation of heat from a relay coil occurs mainly from the inner and outer surfaces by a combination of conduction, convection and radiation, with negligible dissipation at the coil ends. Convection and radiation are principal factors at the outer surface and conduction through the insulation and core is the principal factor at the inner surface.

The dissipation of heat is therefore through parallel paths which can be represented by the electrical circuit analogy shown in Fig. 17. The imposed voltage is equivalent to the temperature difference between the coil and ambient, the electrical current is equivalent to the heat

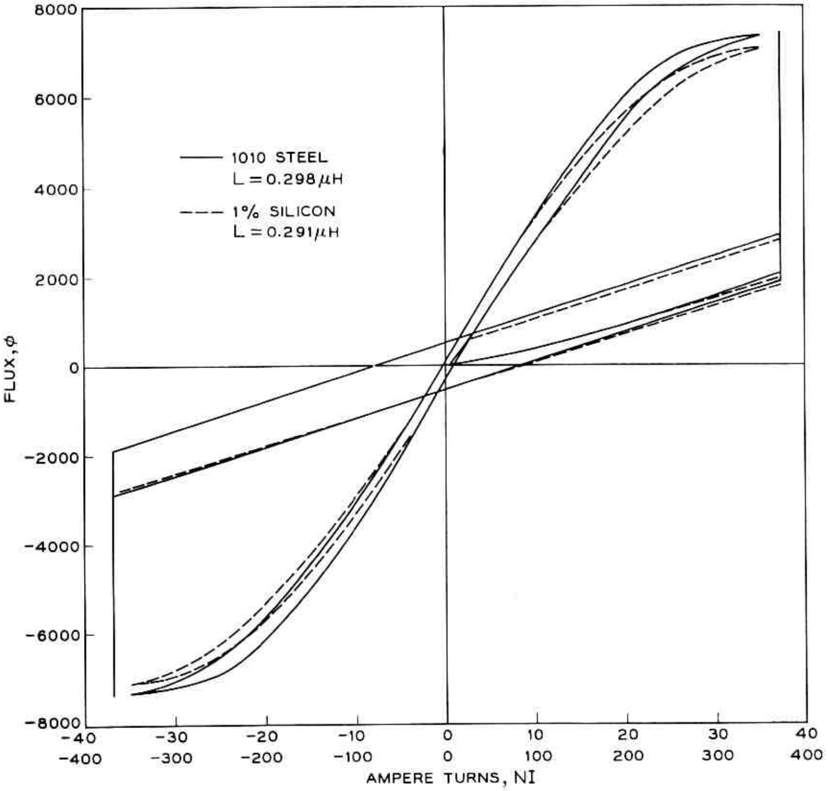


Fig. 14 — Magnetization curves of cores with large dimensions and good anneal (0.032 air gap).

flow in the branches, and the electrical conductance is equivalent to the heat conductance. As shown in Fig. 17, there is a circuit of two branches: one from the coil through the outer core surface to ambient and the other from the coil through the inner coil surface and the core to ambient.

From this analogy it has been found that a good approximation of the mean coil temperature can be obtained from the relationships:

$$K = k_2a_2 + \frac{1}{R_c + \dfrac{1}{k_1a_1}} \tag{12}$$

and

$$\frac{E^2}{R_0} = K(\bar{\theta} - \theta_0)\left(1 + \frac{\bar{\theta} - \theta_0}{390 + \theta_0}\right) \tag{13}$$

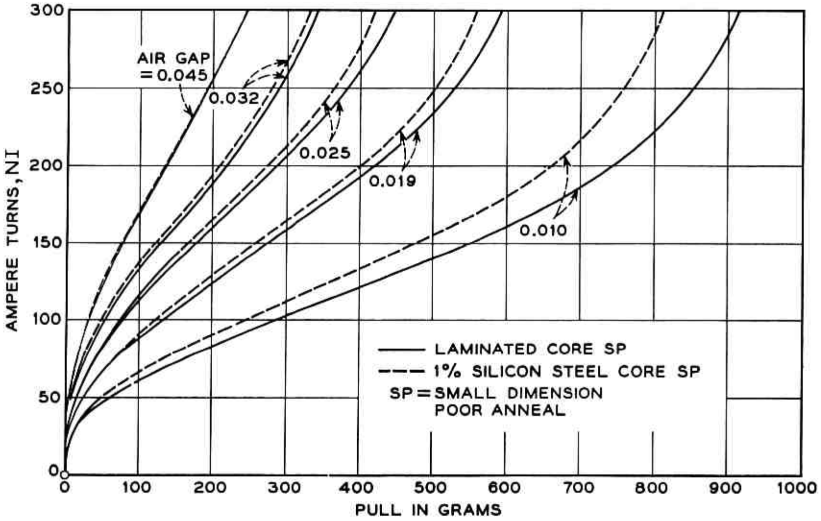


Fig. 15 — Ampere turns vs operate pull values for cores with small dimensions and poor anneal.

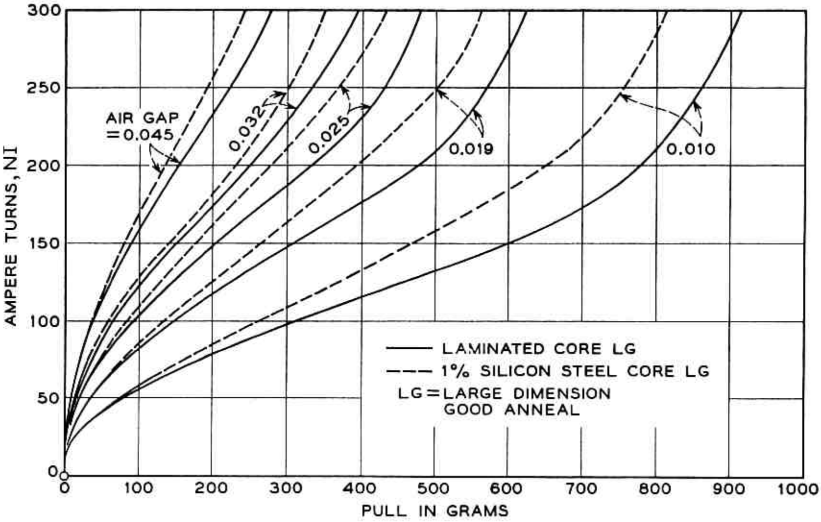


Fig. 16 — Ampere turns vs operate pull values for cores with large dimensions and good anneal.

TABLE I — EFFECT OF USING LAMINATED CORES ON EDDY CURRENT CONDUCTANCE AND OPERATE TIMES

| Relay Dimensions | Core | | $L$ $\mu H$ | $G_e$ Kmho | Per Cent Decrease in Operate Times — $t_L$ (laminated core) vs $t_s$ (1% silicon core) | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | 2 watts $G_c$ = 25 Kmho | | 2 watts $G_c$ = 40 Kmho | | 10 Watts $G_c$ = 25 Kmho | | 10 watts $G_c$ = 40 Kmho | |
| | Material | Type | | | Calc. | Meas. | Calc. | Meas. | Calc. | Meas. | Calc. | Meas. |
| SP | SAE 1010 | lam. | 0.267 | 5.55 | 3.7% | 3.3% | 2.5% | 2.8% | 3.4% | 3.4% | 2.3% | 2.1% |
| SP | 1% silicon | solid | 0.264 | 7.20 | | | | | | | | |
| LG | SAE 1010 | lam. | 0.298 | 6.65 | 2.3% | 2.8% | 1.1% | 1.3% | 2.3% | 2.7% | 3.5% | 3.3% |
| LG | 1% silicon | solid | 0.291 | 8.25 | | | | | | | | |

TABLE II — EFFECT OF USING LAMINATED CORES ON EDDY CURRENT CONDUCTANCE AND RELEASE TIMES

| Relay Dimension | Core | | $G_e$ Kmho | Per Cent Decrease in Release Times — $t_L$ (laminated core) vs $t_s$ (1% silicon core) | | | |
|---|---|---|---|---|---|---|---|
| | | | | 20 Ni Release | | 60 Ni Release | |
| | Material | Type | | Calc. | Meas. | Cal. | Meas. |
| SP | SAE 1010 | lam. | 5.55 | 24% | 25% | 27% | 27% |
| SP | 1% silicon | solid | 7.20 | | | | |
| LG | SAE 1010 | lam. | 6.65 | 16% | 13% | 19% | 14% |
| LG | 1% silicon | solid | 8.25 | | | | |

where $R_c$ = the thermal resistance of the core in "Fahrenheit ohms,"
$R_0$ = the coil resistance at ambient temperature,
$\theta_0$ and $\bar{\theta}$ are expressed in degrees Fahrenheit
and other symbols are as defined earlier.

The exact thermal conductance of complete relay structures will vary

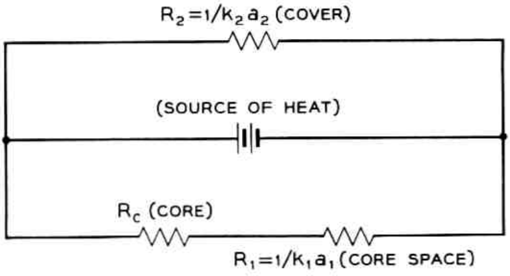


Fig. 17 — Electrical circuit analogy of dissipation of heat.

considerably from unit to unit because of variations in relay assembly, such as the tightness of fit of the coil on the core. However, nominal values for $a_1$, $a_2$, $k_1$, $k_2$, and $R_c$ have been established for general-purpose wire spring relays with 1 per cent silicon steel cores. These values along with the theoretical difference in $R_c$ for laminated S.A.E. 1010 steel cores can be used to calculate mean coil temperatures. The established mean values are as follows:

$$k_1 = 0.01 \text{ watt/F°/in.}^2$$

$$k_2 = 0.0055 \text{ watt/F°/in.}^2$$

$$a_1 = 2.32 \text{ in.}^2$$

$$a_2 = 5.54 \text{ in.}^2$$

$$R_{cs} = 35 \text{ Fahrenheit ohms (solid 1 per cent silicon steel core).}$$

Then with an applied voltage of 100 volts dc, a coil resistance of 1451 ohms, and an ambient temperature of 77°F, the mean coil temperature of a relay with a 1 per cent silicon steel core was calculated as follows:

$$k_1a_1 = 0.0232 \text{ watt/F°}$$

$$k_2a_2 = 0.0304 \text{ watt/F°}$$

$$K_s = k_2a_2 + \frac{1}{R_{cs} + \dfrac{1}{k_1a_1}} = 0.0304 + \frac{1}{35 + \dfrac{1}{0.0232}} = 0.0432 \text{ watts/F°}$$

and

$$\frac{E^2}{R_0} = K_s(\bar{\theta} - \theta_0)\left(1 + \frac{\bar{\theta} - \theta_0}{390 + \theta_0}\right)$$

$$\frac{(100)^2}{1451} = 0.0432(\bar{\theta} - 77)\left(1 + \frac{\bar{\theta} - 77}{390 + 77}\right)$$

$$\bar{\theta}_s = 204°\text{F (1 per cent silicon steel core).}$$

The mean thermal conductivity of S.A.E. 1010 is 53 watts/C°/cm² while the mean thermal conductivity of 1 per cent silicon steel is 28 watts/C°/cm², so that they are in a ratio of 53/28 or 1.89.

$$\therefore R_{cl} = \frac{35}{1.89} = 18.5 \text{ Fahrenheit ohms}$$

for S.A.E. 1010 steel cores. With the value of $R_{c1}$ for S.A.E. 1010 steel cores and the other constants with the same as above, except that with this core the coil resistance was 1443 ohms, the mean coil temperature was calculated as follows:

$$K_1 = 0.0304 + \frac{1}{18.5 + \dfrac{1}{0.0232}} = 0.0467 \text{ watt/F}^\circ$$

and

$$\frac{(100)^2}{1443} = 0.0467(\bar{\theta} - 77)\left(1 + \frac{\bar{\theta} - 77}{390 + 77}\right)$$

$$\bar{\theta}_1 = 195^\circ\text{F (laminated S.A.E. 1010 steel core).}$$

For confirmation, heat tests were conducted on sample relays with both types of cores using applied voltages of 50 and 100 volts dc. Fig. 18 shows the mean coil temperature as a function of time. With an applied potential of 100 volts dc the calculated mean coil temperatures are 204°F and 195°F for the 1 per cent silicon steel and S.A.E. 1010 steel cores respectively, while the measured values are 205.5°F and 199.5°F in the same order. Thus the measured values are found to be in close agreement with the calculated values. Although there is nearly a two-to-one ratio between the thermal conductances of the two materials, there is only a 3 to 4 per cent difference in the mean coil temperatures. However, this small difference is in favor of the relays with laminated S.A.E. 1010 steel cores.

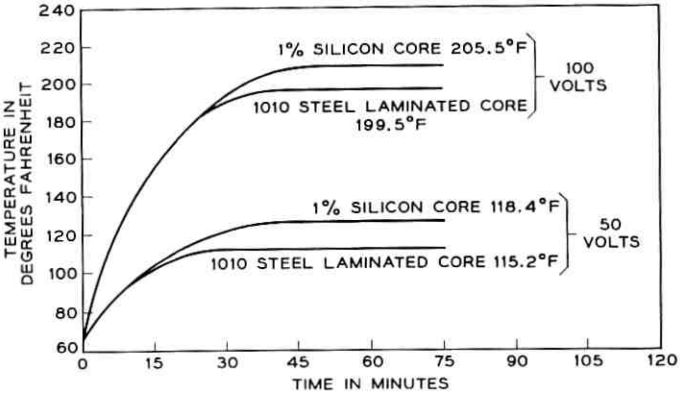


Fig. 18 — Mean coil temperature vs time.

## XI. OPERATING LIFE

Since S.A.E. 1010 steel is softer than 1 per cent silicon steel, life tests were conducted to determine whether relays with laminated low-carbon steel cores have at least as long an operating life as currently manufactured relays. In operation, the armature stop disks and the hinge spring at the heel of the armature wear or pound into the core (see Fig. 19). If significant wear occurs at either of these points, it will cause a corresponding reduction in the release current and an increase in release time.

Fig. 20 shows the measured wear at the armature heel and the change in release time as a function of the number of operations. It is noted that the wear of the laminated S.A.E. 1010 steel cores is about equal to or less than that of the 1 per cent silicon steel cores. This is probably due to the lack of abrasiveness of the low-carbon steel as compared to the silicon steel after the finish has been worn through.

## XII. CORE PLATE ASSEMBLY

The core plate of the relay (see Figs. 1 and 19) serves as a back stop for the armature, a positioning stop for the fixed contact molded block, an aligning fixture for the three core legs and a means of mass adjustment of the contacts. Therefore, it is essential that the core plate be securely fastened in place. The core plate is held in place by staking the ends of the two outer legs of the core as shown in Fig. 19. Tests show that the pull-off force of the core plate on the laminated low-carbon steel cores is approximately twice that of the pull-off force from silicon steel

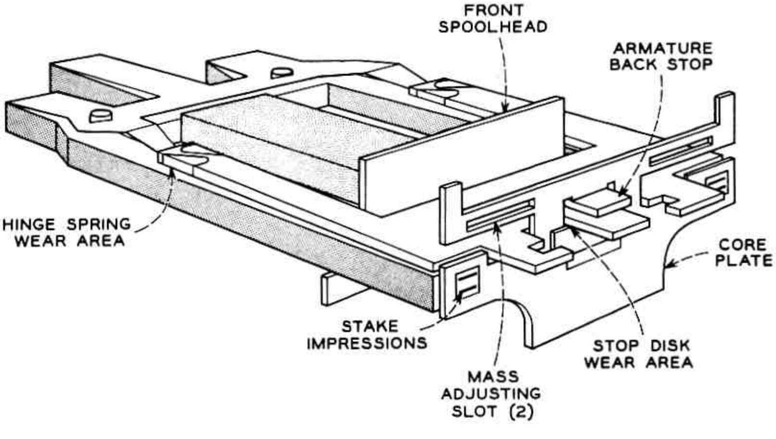


Fig. 19 — Wear areas of stop disks and hinge spring.

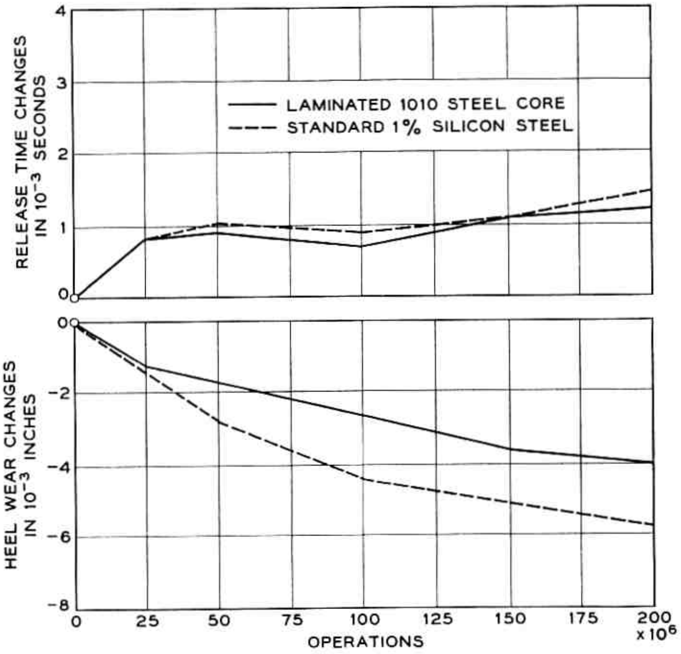


Fig. 20 — Heel wear and release time vs number of operations.

cores. The softer S.A.E. 1010 steel is upset more by the staking operation than the harder 1 per cent silicon steel.

## XIII. CORROSION PROTECTION

If the two laminations were to be welded together immediately after punching, it was deemed necessary to establish the reliability of corrosion protection, along the laminated seam, obtained with the standard zinc-chromate finish on such an assembly. Therefore, a number of laminated core assemblies of S.A.E. 1010 steel were punched and welded together, magnetically annealed and then plated in the normal manner. The laminated assemblies, along with standard zinc-chromate plated 1 per cent silicon steel cores, were subjected to extensive corrosion studies under extremes of temperature and humidity. The extent of corrosion on both materials was within tolerable limits.

## XIV. MANUFACTURING FACTORS

The manufacture of a relay with a two-piece laminated core presents a number of problems of dimensional control and parts handling which must be overcome before its introduction is economically feasible. Close dimensional control of core thickness is necessary in the areas where the front coil spoolhead and core plate are mounted, to assure tight and stable assemblies (see Figs. 4 and 19). With the one-piece core the thickness dimensions in these areas are controlled accurately by a squeezing operation on the part as the last step of the fabrication. With two separate laminations these areas would require closer control of the material thickness and the depressed areas, since with two parts fabricated separately the thickness tolerances would be additive on assembly.

A proposal to overcome this problem was suggested (see Fig. 21) in which one lamination is undercut so that only the thickness of one lamination appears in the critical areas. This would require a redesign of the core plate and spoolhead to fit the new core. Tests of sample relays with cores of this design revealed an appreciable degradation in pull and time characteristics.

An alternate proposal was suggested whereby two strips of S.A.E. 1010 steel are welded together at prearranged locations and then punched simultaneously as a single part. The welds between the two sheets are located so that after punching they appear on the core at the locations shown in Fig. 3. With this method of welding and punching, the two laminations of a core remain together throughout the fabrication proc-

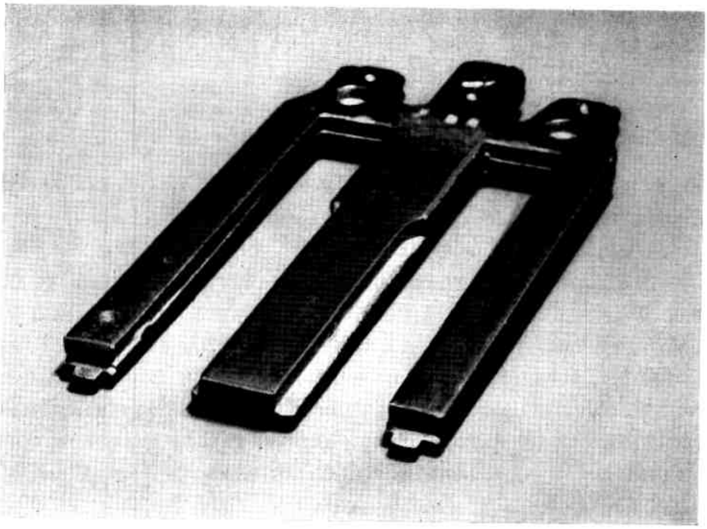


Fig. 21 — SAE 1010 steel laminated cores for general-purpose relays (rejected design).

ess, and the assembly is flattened and squeezed to size in the same operation as is now used with the one-piece core. Fig. 22 shows the two laminations without welding in the upper view, and a welded laminated core compared to a one-piece core in the lower view. Fig. 23 shows a schematic layout of the proposed process of welding the strip and punching the cores.

There is no appreciable difference in the operational quality of these

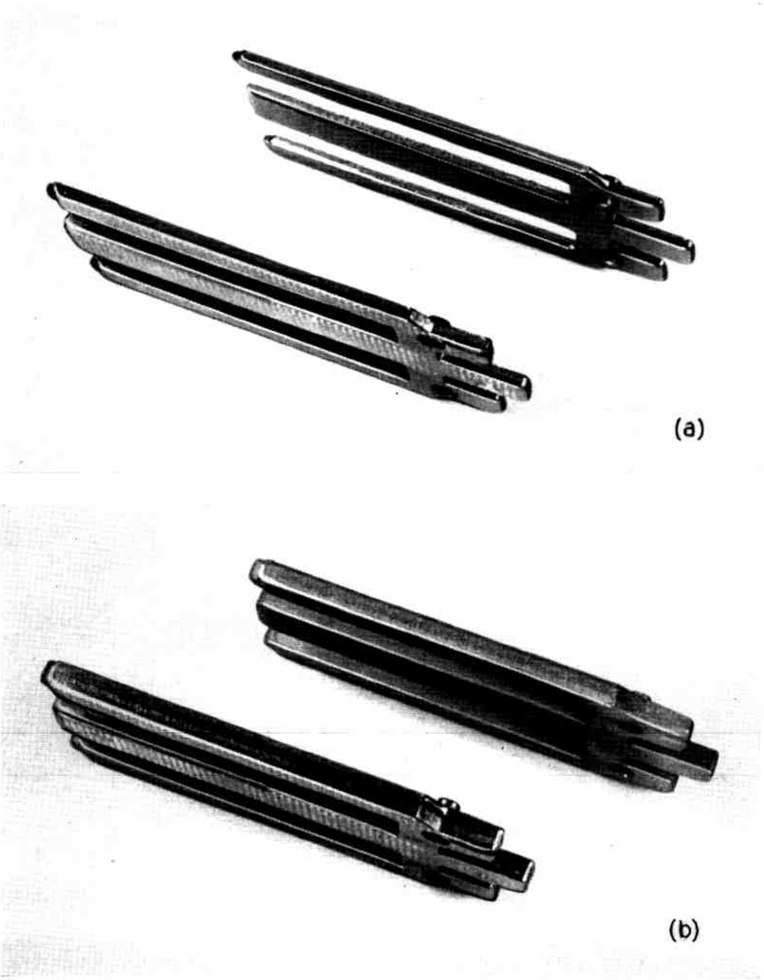


Fig. 22 — (a) Two core laminations before welding; (b) a welded laminated core compared to a one-piece core.

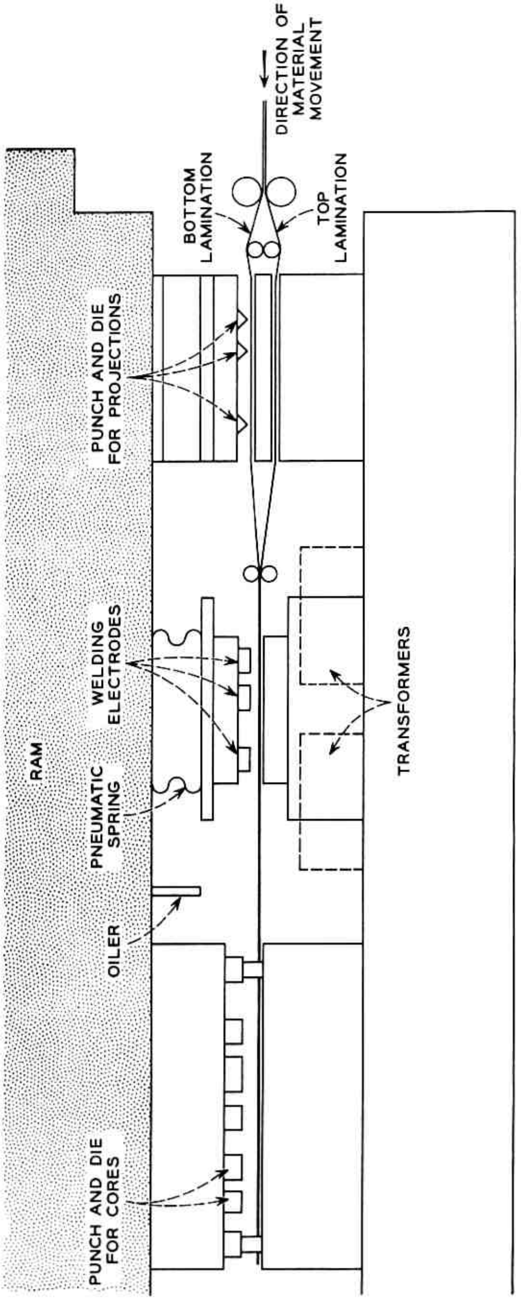


Fig. 23 — Schematic layout of the proposed process of welding the strip and punching the core.

relays whether the core laminations are welded or not. Therefore, the ultimate design in this respect will be determined by manufacturing considerations.

## XV. SUMMARY AND CONCLUSIONS

(1) Wire spring relays made with laminated 1010 steel cores are at least as good as and in some respects better than the present standard 1 per cent silicon steel core relays.

(2) Operate times of the proposed laminated 1010 steel core relays are in general slightly faster than those of the present one-piece silicon steel core relays (1.3–3.4 per cent).

(3) Release times of the proposed laminated low carbon steel core relays are considerably faster than those of the present silicon core relays (13–27 per cent).

(4) Heat studies indicate slightly better heat dissipating qualities in the laminated core relays than in the present silicon steel core relays.

(5) Core plates are tighter on the laminated 1010 steel core relays than on the present silicon steel core relays.

(6) Resistance to corrosion with the present finish (zinc chromate) is as good on the laminated 1010 steel core design as on the one-piece silicon steel core design.

(7) Life tests show that the laminated low carbon steel cores are at least as resistant to wear as the present silicon steel cores.

(8) Substantial manufacturing savings can be realized by changing to the laminated 1010 steel core for the general-purpose wire spring relay.

## XVI. ACKNOWLEDGMENTS

The author wishes to acknowledge the assistance given by J. A. Cooper, J. J. Dunbar, F. J. Gibson and E. R. Priday in obtaining the data used and C. B. Brown and H. M. Knapp for the helpful suggestions in the preparation of this article.

## REFERENCES

1. Bell Telephone System Monograph 2180, *Design of Relays*, American Telephone and Telegraph Company, New York, 1954.
2. Peek, R. L., Jr., and Wagar, H. N., *Switching Relay Design*, D. Van Nostrand, New York, 1955.

# A Model For Mobile Radio Fading Due to Building Reflections: Theoretical and Experimental Fading Waveform Power Spectra

By JOSEPH F. OSSANNA, Jr.

(Manuscript received May 20, 1964)

*Fluctuations in received signal amplitude occur during mobile communications because of the motion of the mobile station through the spatial standing-wave pattern resulting from the interaction of direct and reflected signals. A model is presented which permits a theoretical calculation of the power spectrum of these fluctuations and satisfactorily predicts the features of spectra computed from experimental fading data except for an observed rise at low frequencies.*

*The model is based on the geometry of the reflections from nearby randomly placed vertical plane reflectors. Vertical polarization is assumed. Both the standing-wave pattern and the Doppler shift view of fading are used to obtain nearly identical results. The detailed shape and in particular the sharp cutoff frequency of the spectrum are shown to depend crucially on the angle $\alpha$ between the direction of vehicle motion and the direction to the fixed station. Detailed comparisons are made of theoretical spectra with experimental spectra representing a range of the angle $\alpha$.*

*The collection, digitization, calibration, plotting, and digital processing to obtain power spectra of actual recorded fading waveforms are described.*

## I. INTRODUCTION AND SUMMARY

The mobile radio fading phenomenon discussed herein is the fluctuation in the received signal amplitude during mobile communication due to the motion of the mobile station through the spatial standing-wave pattern resulting from the interaction of direct and reflected signals. Knowledge of the statistical behavior of such fading signals can permit more meaningful mobile communication system studies and design effort. For example, it can aid in the choice and design of automatic gain control systems and systems involving data transmission. The particular statistical description of fading to be discussed is the power spectrum of the signal amplitude.

First, this article presents a theoretical model for mobile radio fading due to building reflections, which permits a theoretical calculation of the power spectrum of the fading waveform. Second, the collection, digital processing, and power spectral analysis of actual fading waveforms are described. Then detailed comparisons are made between the theoretical and experimental spectra. Limitations and extensions of the model are explored.

Historically, the evolution of the model followed a study of the power spectra of fading waveforms recorded on one particular street. These spectra had an unexpected and interesting shape. The model indicated that the shape of the spectrum and particularly the frequency at which the power density fell abruptly would depend on the vehicle's direction of travel with respect to the direction toward the fixed station. Subsequent fading waveforms, recorded on other streets having various relative directions, produced power spectra which collectively exhibited all the features predicted by the model.

### 1.1 *Characteristics of Experimental Power Spectra*

Fading waveforms were recorded on 13 streets in New Providence, N. J. on Sept. 13, 1962, using a carrier frequency of 838 mc and a nominal vehicle speed of 15 mph. Vertical polarization was used. The fade rate corresponding to motion through standing-wave minima a half-wavelength apart is 37.5 cps. After digitization, calibration, and smoothing, power spectra were computed. Almost all of these spectra exhibited: (1) a very sharp cutoff at a frequency somewhere between 20 and 40 cps, followed, after a drop of between 10 and 20 db, by a 12 db/octave fall-off; (2) a narrow peak between 3 and 10 db high just prior to the sharp fall; (3) a broader, shorter intermediate peak; and (4) a rise at the low-frequency end of 10–15 db. Many spectra exhibited a shelf in the frequency range between the sharp fall and the gradual fall-off which extended to about twice the sharp-fall frequency. Other sometimes subtle features were noted which will be mentioned later.

### 1.2 *The Model*

The simple model to be described predicts a theoretical power spectrum for the fading waveform having all the features of a corresponding experimental spectrum except a low-frequency rise. In particular, the frequency, shape, and size of two peaks, the sharp fall, the following shelf, and the subsequent gradual fall-off are satisfactorily predicted.

The model is based on the geometry of the reflections from nearby randomly placed, vertical, plane, good reflectors. Vertical polarization is assumed (the use of horizontal polarization is discussed in Section XIV). The vehicle is assumed to move through the standing-wave pattern caused by the reflections. A virtually identical result obtains if the vehicle is assumed to encounter appropriate Doppler shifts for each reflected signal. The computed spectrum depends on: (1) the radio carrier frequency, (2) the vehicle speed, and (3) the angle $\alpha$ between the vehicle direction of motion and the direction to the fixed station.

If $f_m$ is the fade frequency which would be experienced by the vehicle moving directly across standing-wave minima spaced $\lambda_c/2$ apart ($\lambda_c$ = carrier wavelength), the model indicates that the spectrum will peak and then fall sharply at some frequency $f_p$ between $f_m/2$ and $f_m$ and that an intermediate peak will occur at a frequency $f_m - f_p$. In terms of the angle $\alpha$ defined in the previous paragraph, $f_p$ is equal to the larger of $f_m \sin^2 \alpha/2$ or $f_m \cos^2 \alpha/2$. As $\alpha$ varies between 0 (or 180°) and 90°, $f_p$ varies between $f_m$ and $f_m/2$.

### 1.3 *Comparison between Experimental and Theoretical Spectra*

Because the angle $\alpha$ varies during a typical data run along any street, theoretical spectra were determined by averaging spectra computed for sample $\alpha$'s along the run. The agreement between experimental and theoretical spectra is generally quite good. The sharp-fall frequency agrees perfectly in almost all cases. Other details are in good agreement in many cases. The main discrepancy is the absence of a theoretical forecast of the rise in the observed spectral density at low frequencies. One street which, unlike all the others, had few buildings produced experimental spectra only vaguely similar to theoretical expectations. Mechanisms not considered in the theoretical model which would contribute low-frequency energy include: (1) shadowing by buildings, (2) variations in and the shadowing of ground reflections, (3) the intermodulation of concurrent reflections and (4) nonrandom reflector orientation.

### 1.4 *Usefulness of Model*

The model in the form offered successfully predicts fading waveform power spectra in a suburban residential environment. The possibility certainly exists that an extension of the model can be made to work elsewhere. The usefulness of power spectra is not unlimited, and many statistical properties of fading cannot easily be deduced from spectra. One of the main values of the agreement between these theoretical and experimental spectra is its verification of the physical basis of the model.

## II. THE MODEL

We will begin by considering the fading experienced by a mobile receiver moving through a standing-wave pattern due to a single reflector when the transmission is vertically polarized. Reciprocity will insure application of the results to the case where the receiver is fixed and the transmitter is moving. First, the fade rate will be related to the vehicle trajectory. Then the same result will be obtained using the Doppler point of view. Then the relative contribution from reflectors in different directions will be determined. Next, a theoretical spectrum will be constructed for the case of many reflectors; its shape will depend strongly on the vehicle direction relative to the fixed station.

## III. STANDING-WAVE PATTERN DUE TO A SINGLE REFLECTOR

If the mobile antenna is a fixed height above the ground, only the larger (many wavelengths) vertical plane reflectors in the vicinity of the mobile station are of major importance in determining the local standing-wave pattern fluctuations. Reflectors of ordinary size which are not in the vicinity of either the mobile or fixed stations are of lesser importance because their reflected signals will be of smaller amplitude. Reflectors near the fixed station can contribute large amplitude reflections, but their effect is more that of modifying the directivity pattern of the fixed antenna; their effect is to put slow multiplicative trends into the standing-wave pattern at the mobile station. Furthermore, fixed station antennas are usually mounted above local obstacles that would not only reflect but shield some direction.

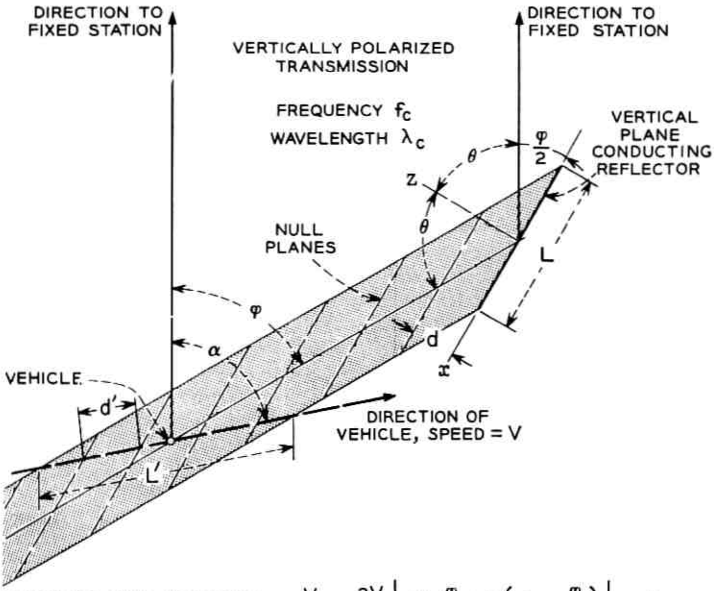
We will assume that the transmitting and receiving antennas are vertically polarized. Then the local standing-wave pattern due to a single vertical plane conducting reflector is as shown in Fig. 1 where:

$\varphi$ = angle between the direction to the fixed station and the direction to the reflector as seen from the mobile antenna,

$\alpha$ = angle between the direction to the fixed station and the direction of vehicle travel,

$\theta$ = angle of incidence at the reflector,

$d$ = perpendicular distance between null planes,


DIRECTION TO FIXED STATION
VERTICALLY POLARIZED TRANSMISSION
DIRECTION TO FIXED STATION
FREQUENCY $f_c$
WAVELENGTH $\lambda_c$
VERTICAL PLANE CONDUCTING REFLECTOR
$\frac{\varphi}{2}$
$\theta$
z
$\theta$
NULL PLANES
L
$\varphi$
d
$\alpha$
x
VEHICLE
d′
DIRECTION OF VEHICLE, SPEED = V
L′

$$\text{OBSERVED FADE FREQUENCY} = \frac{V}{d'} = \frac{2V}{\lambda_c}\left|\sin\frac{\varphi}{2}\sin\left(\alpha-\frac{\varphi}{2}\right)\right| = f$$

$$\text{RELATIVE WEIGHT} = \frac{L'}{L} = \left|\frac{\sin\frac{\varphi}{2}}{\sin(\alpha-\varphi)}\right| = W$$

Fig. 1 — Vehicle moving through standing-wave pattern due to single reflector.

$d'$ = spacing between null planes observed along direction of mobile station travel,

$L$ = length of reflector,

$L'$ = length of vehicle path in reflected beam, and

$\lambda_c$ = carrier frequency wavelength.

The fixed station is assumed to be far enough away to permit taking the incident waves on the mobile antenna and on the reflector as parallel. The reflector is assumed to be large compared to a wavelength ($L \gg \lambda_c$) and close enough to the mobile station to neglect divergence of the reflected beam.

### 3.1 *Null Plane Spacing*

Note that a reflector in a direction $\varphi$ must be oriented such that

$$\theta + (\varphi/2) = 90° \tag{1}$$

for the reflected beam to be directed toward the mobile station (for the angle of reflection to equal the angle of incidence). To determine the spacing $d$ between null planes, refer to Fig. 2 and observe that

$$a - b = \lambda_c , \tag{2}$$

$$b/a = \cos \varphi, \tag{3}$$

and

$$d/a = \sin (\varphi/2). \tag{4}$$

From (2) and (3), $a$ is found to be

$$a = \frac{\lambda_c}{1 - \cos \varphi} = \frac{\lambda_c}{2 \sin^2 \frac{\varphi}{2}} . \tag{5}$$

Then (4) and (5) yield

$$d = a \sin \frac{\varphi}{2} = \frac{\lambda_c}{2 \sin \frac{\varphi}{2}} , \tag{6}$$

or, using (1)

$$d = \frac{\lambda_c}{2 \cos \theta} . \tag{7}$$

This is of course a common result (see Ref. 1, p. 293 ff.).

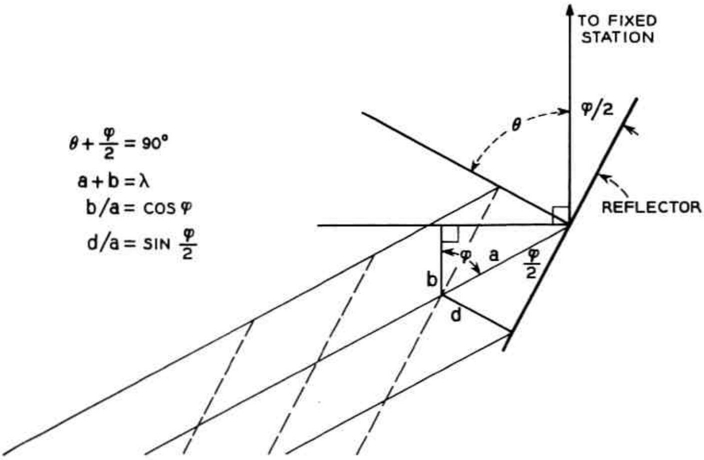


Fig. 2 — Portion of Fig. 1 in vicinity of reflector.

### 3.2 *Fading Rate*

Because of its direction of travel, the mobile antenna observes a null spacing $d'$, which from Fig. 3 is

$$d' = \frac{d}{\sin\left(\alpha - \frac{\varphi}{2}\right)}, \tag{8}$$

which by using (6) becomes

$$d' = \frac{\lambda_c/2}{\sin\frac{\varphi}{2}\sin\left(\alpha - \frac{\varphi}{2}\right)}, \tag{9}$$

which holds for the case where $\alpha > \varphi/2$. Consideration of various values for $\alpha$ and $\varphi/2$ leads to the general relation

$$d' = \frac{\lambda_c/2}{\left|\sin\frac{\varphi}{2}\sin\left(\alpha - \frac{\varphi}{2}\right)\right|}. \tag{10}$$

If the vehicle speed is $V$, the fading frequency is

$$f' = \frac{V}{d'} = \frac{2V}{\lambda_c}\left|\sin\frac{\varphi}{2}\sin\left(\alpha - \frac{\varphi}{2}\right)\right|. \tag{11}$$

Then $f'$ has the maximum value $f_m' = 2V/\lambda_c$, where $\varphi = 180°$ and

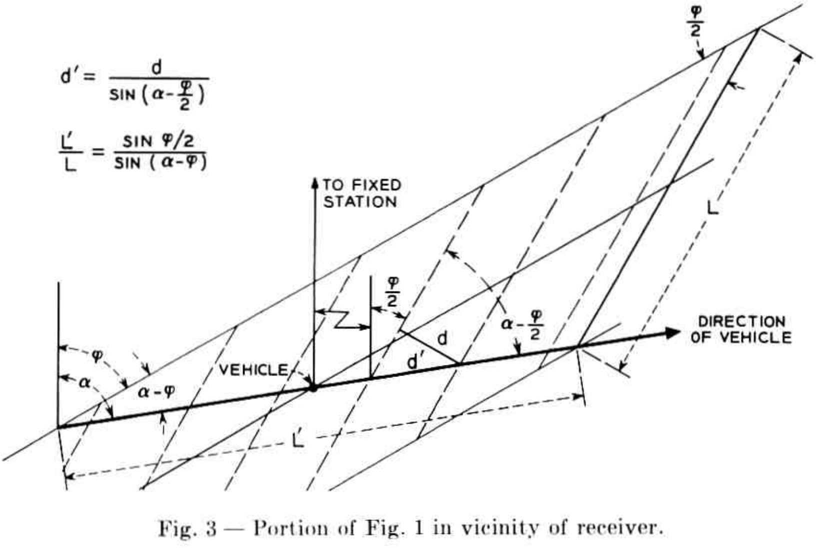

$d' = \frac{d}{\sin(\alpha - \frac{\varphi}{2})}$
$\frac{L'}{L} = \frac{\sin \varphi/2}{\sin(\alpha-\varphi)}$
$\frac{\varphi}{2}$
TO FIXED STATION
L
$\frac{\varphi}{2}$
d
$\alpha - \frac{\varphi}{2}$
DIRECTION OF VEHICLE
$\varphi$
$\alpha$
$\alpha - \varphi$
VEHICLE
d'
L'

Fig. 3 — Portion of Fig. 1 in vicinity of receiver.

$\alpha = 0°$ or $180°$, when the vehicle moves perpendicularly across null planes spaced $\lambda_c/2$ apart. For $f_c = 838$ mc and $V = 15$ mph, $f_m' = 37.5$ cps. For convenience, we will usually use a normalized fading frequency $f = f'/f_m'$. When $\alpha \neq 0°$ or $180°$, $f$ is zero at $\varphi = 0, 2\alpha$, and $360°$, and has the maximum values $f_1 = \sin^2 (\varphi/2)$ at $\varphi = \alpha$ and $f_2 = \cos^2 (\varphi/2)$ at $\varphi = 180 + \alpha$; note that $f_1 + f_2 = 1$. Thus, for a particular $\alpha$, the maximum possible fade rate is $f_{\max} = \max (f_1, f_2)$* and is due to a vehicle motion either toward or away from a reflector. The minimum possible value for $f_{\max}$ occurs when $f_1 = f_2 = \frac{1}{2}$, which corresponds to $\alpha = 90°$. The variation of $f$ with $\varphi$ is shown in Fig. 4 for $\alpha = 0°, 30°$, $60°$, and $90°$.

### 3.3 *Fading Waveform*

If the reflector is perfectly conducting as assumed above, the actual waveform observed at the output of an envelope detector in the mobile vehicle would be the familiar result of beating two frequencies of equal amplitude — a full-wave rectified sine wave. Thus, in addition to the fundamental fade rate discussed above, significant harmonics will also be present. If the reflector is not perfectly conducting or is only a dielectric, minima will occur instead of nulls; the spacing between them will remain the same as for the nulls, and the waveform will tend to be more nearly sinusoidal.

## IV. THE DOPPLER POINT OF VIEW

We can instead consider fading as due to the beating within the receiver of different carrier frequencies arising from the different Doppler shifts occurring for the directly incident and reflected waves. The carrier frequency observed at the vehicle will in general be

$$f_o = f_c + (v/\lambda_c) \tag{12}$$

where $f_c$ and $\lambda_c$ are the transmitted carrier frequency and wavelength and $v$ is the relative velocity of closure between the two stations. From Fig. 1, the observed frequency of the directly incident signal is

$$f_i = f_c + \frac{V \cos \alpha}{\lambda_c}, \tag{13}$$

where $V$ is the vehicle speed.

* Max (a,b,c, ⋯) = the algebraically largest of the sequence a,b,c, ⋯ . Similarly, min (a,b,c, ⋯) = the algebraically smallest of a,b,c, ⋯ .

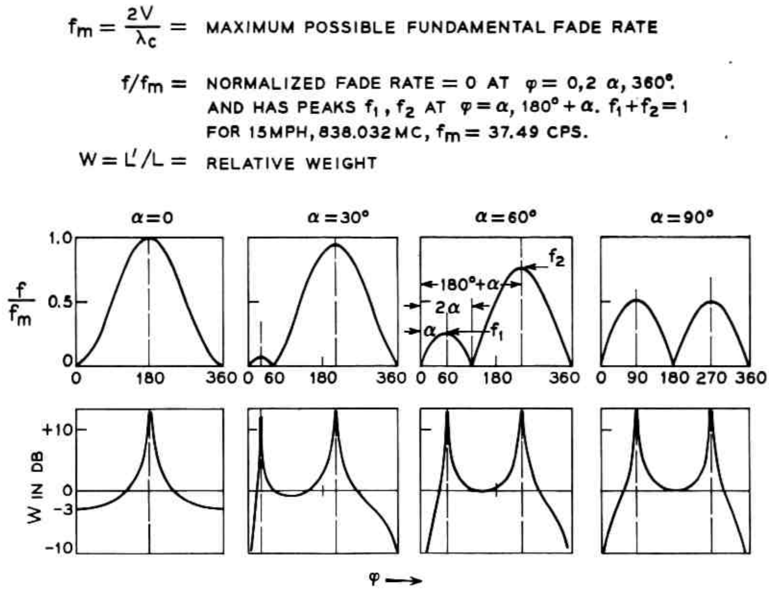


Fig. 4 — Variation of fade frequency $f$ and weighting function $W$ with relative vehicle direction $\alpha$ and relative reflector direction $\varphi$.

The observed frequency of the reflected signal is

$$f_r = f_c + \frac{V \cos(\alpha - \varphi)}{\lambda_c}. \tag{14}$$

The fading rate will then be the beat frequency between $f_i$ and $f_r$:

$$f = f_r - f_i = \frac{V}{\lambda_c}[\cos(\alpha - \varphi) - \cos\alpha] \tag{15}$$

$$= \frac{V}{\lambda_c}[\sin\alpha \sin\varphi - \cos\alpha(1 - \cos\varphi)]$$

$$= \frac{V}{\lambda_c}\left(2\sin\frac{\varphi}{2}\right)\left[\sin\alpha\cos\frac{\varphi}{2} - \cos\alpha\sin\frac{\varphi}{2}\right]$$

$$= \frac{2V}{\lambda_c}\left[\sin\frac{\varphi}{2}\sin\left(\alpha - \frac{\varphi}{2}\right)\right]. \tag{16}$$

After absolute value signs are added to (16) to account for the various relative values of $\alpha$ and $\varphi$, the result is identical to (11).

The Doppler point of view has one important advantage over the standing-wave pattern point of view. It is easier to see what fade rates will occur when more than one reflector is involved; when $n$ reflectors are simultaneously effective, $n(n + 1)/2$ beat frequencies are possible.

## V. THEORETICAL POWER SPECTRUM

Since the observed fading rate is a function of the parameters $\alpha$ and $\varphi$ discussed above, the power spectrum of the fading waveform at the output of an envelope detector in the mobile vehicle will evidently be a function of time, even if the vehicle speed is constant. In this section, we will develop an approximation to the power spectrum of a finite duration of fading waveform. We will assume: (1) the vehicle speed is constant; (2) the transmission is vertically polarized; (3) an unmodulated carrier is being transmitted; (4) the reflectors are large, stationary, vertical, plane conductors; and (5) only reflectors in the vicinity of the mobile vehicle are important. Other assumptions inherent in the development will be stated when appropriate.

### 5.1 *More Than One Reflector*

A major step in simplifying the analysis is to assume that, although many reflected beams will be encountered by the vehicle during the finite run, only one such beam is important at any one time. This eliminates the necessity of considering beats between reflections. The effect of this assumption on the theoretical spectrum will be discussed later. Actually, there is a strong tendency for this assumption to be true in a suburban residential environment because the houses are well spread out.

### 5.2 *The Relative Importance of Different Reflectors*

The energy in a particular small frequency band in the finite sample of fading waveform will be proportional to the time that frequencies in that band are present. The corresponding power spectral density will be proportional to the corresponding fraction of the total run time. Thus the contribution of a particular reflector to the appropriate frequency band will be proportional to the time it takes the vehicle to cross the reflected beam or, if the vehicle speed is constant, proportional to the length of its path through the beam, which is the length $L'$ in Fig. 1.

If we assume that all the reflectors are the same size ($L$ in Fig. 1), then the contribution of a particular reflector to the power spectrum will be proportional to a weighting function $W = L'/L$. The assumption of equal-size reflectors is another assumption that has a strong tendency to be true in suburban residential areas, where all the houses in a given locale tend to be the same size. From Fig. 3 it is evident that

$$W = \frac{L'}{L} = \frac{\sin(\varphi/2)}{\sin(\alpha - \varphi)}. \tag{17}$$

Consideration of the various values of $\alpha$ and $\varphi$ yields the general result

$$W = \frac{L'}{L} = \left| \frac{\sin \varphi/2}{\sin(\alpha - \varphi)} \right|. \tag{18}$$

Because $L'$ cannot exceed the total run length $L_T$, the physical maximum value of $W$ is $W_{\max} = L_T/L$. Fig. 4 shows the variation of $W$ with $\varphi$ for $\alpha = 0°, 30°, 60°$, and $90°$; the plots of $W$, which are in db ($10 \log_{10} W$) are shown directly below corresponding plots of the fade rate $f$. $W$ has the value zero ($-\infty$ db) at $\varphi = 0°$ (and $360°$), except when $\alpha = 0$ where $W = 0.5$ ($-3$ db) at $\varphi = 0°$. $W = 1$ (0 db) at $\varphi = 2\alpha$. And $W$ is truncated to $W_{\max}$ at $\varphi = \alpha$ and $180° + \alpha$; the peaks of $W$ are coincident with the peaks of the fade rate $f$. Not only do the reflectors directly ahead or behind the vehicle cause the most rapid fades, but they are the most important contributors to the power spectrum.

It is interesting to note that the weighting function $W$ can be arrived at in another way. Consider the portion of the $f$ vs $\varphi$ curve between $\varphi = 0°$ and $\varphi = \alpha$. The small range of reflector directions $d\varphi$ contributing to a small frequency band $df$ can be found by differentiating (11) to get

$$df/d\varphi = \tfrac{1}{2} \sin(\alpha - \varphi). \tag{19}$$

The projected length of a reflector in a direction $\varphi$ is $L_p = L \cos \theta = L \sin \varphi/2$ (assuming $L$ constant). Suppose that the contribution to the power in a band $df$ due to the reflectors in a range $d\varphi$ is

$$P(f)df = C \frac{d\varphi}{2\pi} \cdot \frac{L_p}{L}, \tag{20}$$

where $C$ is a constant. Substituting for $d\varphi/df$ and $L_p/L$ then gives

$$P(f) = \frac{C}{\pi} \frac{\sin \varphi/2}{\sin(\alpha - \varphi)}, \tag{21}$$

which except fcr the constant is identical to (17).

### 5.3 *The Theoretical Spectrum Method*

Consider again the plots of fade rate $f$ vs $\varphi$ shown in Fig. 4. For any particular $f < f_{\max} = \max(f_1, f_2)$, there are either two or four corresponding values of reflector directions $\varphi$. For each of these $\varphi$'s the corresponding value of the weighting function can be found from (18) and is seen on the $W$ vs $\varphi$ plot directly under the $f$ vs $\varphi$ plot. We will assume that the mobile station is under the influence of one reflector at a time; this condition has a strong tendency to be true in suburban residential areas. Then, if we further assume that all reflector directions are equally likely, the power density at the frequency $f$ will be proportional to the sum of the two or four values of $W$.

The basic procedure for generating a theoretical spectrum for comparison with a spectrum computed from experimental data is, if $\alpha$ is constant:

(*i*) Select a list of frequencies $f = n\Delta f$, where $\Delta f = (f_{\text{fold}}/M)/(2V/\lambda_c)$ and $n = 0,1,2 \cdots$ ; $f_{\text{fold}}$ is the folding frequency (Ref. 2, p. 117 ff.) of the experimental data, $M$ is the number of lags (Ref. 2, p. 120 ff.) used in computing its spectrum, and $(2V/\lambda_c)$ is the corresponding $f_m'$. In other words, select the same frequencies, normalized by dividing by $f_m'$, at which spectral estimates were computed for the experimental data. The reason for this matching of frequencies will be discussed below.

(*ii*) For each frequency, determine the two or four reflection directions $\varphi$, using (11) or Fig. 4. Then for each frequency determine the corresponding two or four weighting functions $W$ from (18); each $W$ should of course be limited to $W_{\max}$.

(*iii*) At each frequency, sum the two or four corresponding values of the weighting function $W$ to get the spectral power density.

The solution of (11) in step (*ii*) above can be accomplished by Newton's iteration procedure. Also, the symmetry of the $f$ vs $\varphi$ function about $\varphi = \alpha$ and $180 + \alpha$ can be used; if $\varphi_1 < \alpha$ is a solution, $\varphi_2 = 180 - \varphi_1$ ; and, if $2\alpha < \varphi_3 < 180 + \alpha$ is another solution, $\varphi_4 = 360 + 2\alpha - \varphi_3$ . In the case where $\alpha = 0$ (or 180°) an explicit solution for the two $\varphi$'s and the sum of the two $W$'s can be obtained. Setting $\alpha = 0$ in (11) gives

$$\frac{f'}{f_m'} = f = \sin^2 \frac{\varphi}{2}, \tag{22}$$

and the solution $\varphi/2 = \arcsin f^{\frac{1}{2}}$. Setting $\alpha = 0$ in (18) gives

$$W = \left| \frac{1}{2 \cos \varphi/2} \right|. \tag{23}$$

Because (23) is symmetrical about $\varphi = \pi/2$, the two $W$'s are equal, and (22) and (23) combine to give the spectral density as

$$P(f) = \min\left(2W_{\max}, \frac{1}{\sqrt{1-f}}\right), \tag{24}$$

where the physical limit on $W$ is included.

Implicit in the procedure thus far is the assumption that fading waveform contributed by each reflector is sinusoidal. The resulting power spectrum is zero above $f_{\max}$. Actually, the fading waveform due to a single reflector has an harmonic content which depends on the relative amplitudes of the direct and reflected signals; when one is much smaller than the other the fading tends to be sinusoidal, and when they are equal the fading waveform is a full-wave rectified sine wave. This can be seen by superposing the incident and reflected electric field components; in terms of the $z$ coordinate of Fig. 1, the resultant electric field of a vertically polarized wave has the form (see Ref. 1, p. 296)

$$|E| = \left|K \sin\left(\frac{2\pi z}{\lambda_c}\cos\theta\right)\right|, \tag{25}$$

where $K$ is a constant and $\theta$ is the angle of incidence. The spectra of the experimental data all exhibit fall-offs subsequent to the fall corresponding to $f_{\max}$. The harmonic content of the fading can be included in the theoretical spectrum by determining the harmonic power corresponding to each original theoretical spectral estimate and adding this power in at the corresponding set of harmonic frequencies. Arbitrarily, the coefficients for a full-wave rectified sine wave were used to determine the relative power at the harmonic frequencies; this will provide a maximum of harmonic power. It is an interesting fact that inclusion of harmonic power does not significantly alter the shape of the theoretical spectrum at frequencies below $f_{\max}$.

The final step in generating the theoretical spectrum is to smooth it in an appropriate way. Because the spectra computed from experimental data are estimates of smoothed versions of the true power spectra (see Ref. 2), the theoretical spectra should be smoothed in a corresponding way. Therefore the theoretical spectra to be shown will have been smoothed by hanning.[2] This is the reason for matching the theoretical and experimental spectral estimate frequencies.

Finally, if the relative path angle $\alpha$ varies during the run, its variation can be represented by a weighted list of sample $\alpha$'s. The spectrum for each $\alpha$ can be determined and the resulting spectra averaged. The smoothing can be done after averaging.

### 5.4 *Theoretical Spectra for Various Constant $\alpha$'s*

Fig. 5 shows theoretical spectra for $\alpha = 0°, 30°, 60°$, and $90°$, $V = 15$ mph, $f_c = 838$ mc, and $W_{\max} = 15$. The corresponding $f_m' = 37.5$ cps. Consider first the curve for $\alpha = 60°$. The peaks corresponding to the relatively heavily weighted frequencies in the vicinity of $f_1$ and $f_2$ ($f_1' = 9.4$ cps, $f_2' = 28.1$ cps) are clearly evident. Following $f_2'$, the power density falls sharply and levels off abruptly to form a shelf. The shelf, which arises primarily from second-harmonic power, peaks around 56 cps prior to a second sharp fall. Following the second-harmonic shelf is one due primarily to third harmonics which ends at about 84 cps. Because the points in the fundamental frequency portion of the spectrum

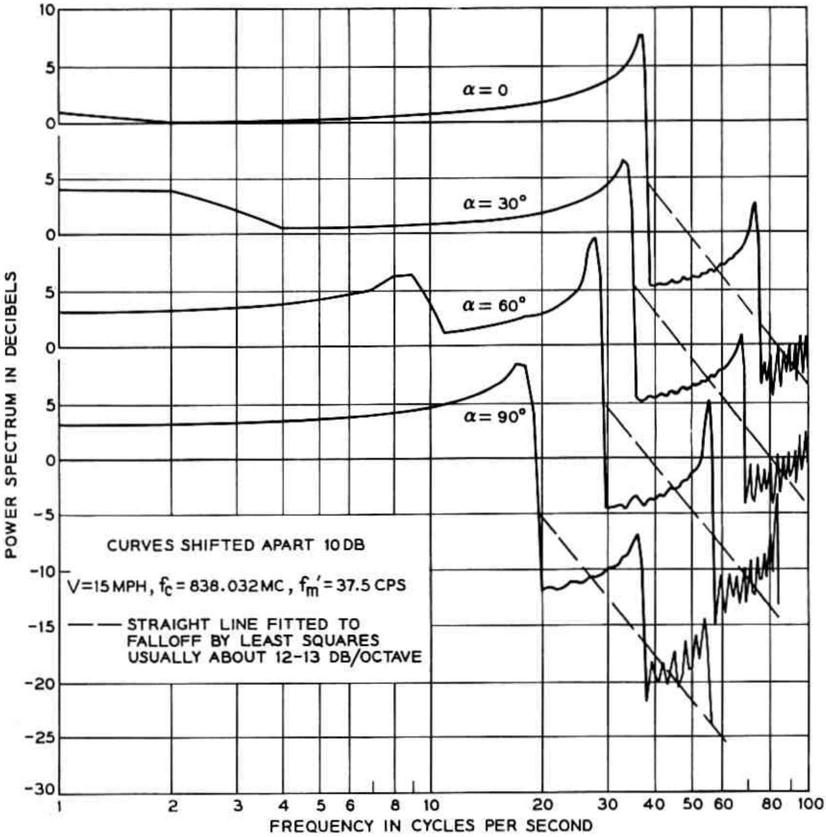


Fig. 5 — Theoretical fading waveform power spectrum vs relative vehicle direction $\alpha$. Theoretical spectral estimates are 1-cps apart.

were computed at 1-cps spacings, the harmonic shelves get increasingly jagged-looking at higher frequencies. The dashed line sloping down through the shelves is a least squares straight-line fit to the portion of the spectrum following the first sharp fall. The falloff line in the $\alpha = C0°$ case has a slope of $-13.0$ db/oct.

The peaks in the $\alpha = 30°$ case correspond to $f_1' = 2.5$ cps and $f_2' = 35.0$ cps. When $\alpha = 0$, $f_1' = 0$ and $f_2' = 37.5$ cps. And when $\alpha = 90°$ the peaks unite at $f_1' = f_2' = 18.75$ cps. The least square fall-off lines have slopes that generally fall between 12–13 db/oct.

### 5.5 *Theoretical Spectra for $\alpha$ Uniformly Distributed*

Fig. 6 shows the result of averaging the spectra for $\alpha$'s uniformly distributed 0–360° (spectra for $\alpha = n2°$, $n = 0, 1, \cdots, 45$, were averaged). The spectral density is quite flat out to 37.5 cps, where it drops abruptly about 16 db to the second-harmonic shelf. The harmonic shelves in this case are also quite flat. The least squares fall-off line is shown and has a slope of $-13.2$ db/oct.

## VI. DATA COLLECTION

The fading waveforms due to vehicle motion were recorded (on FM tape with an Ampex FR100) for 17 runs on 13 different streets (runs on some streets were made in both directions) in New Providence, N. J., on Sept. 13, 1962. Transmission at 838.032 mc was from the mobile

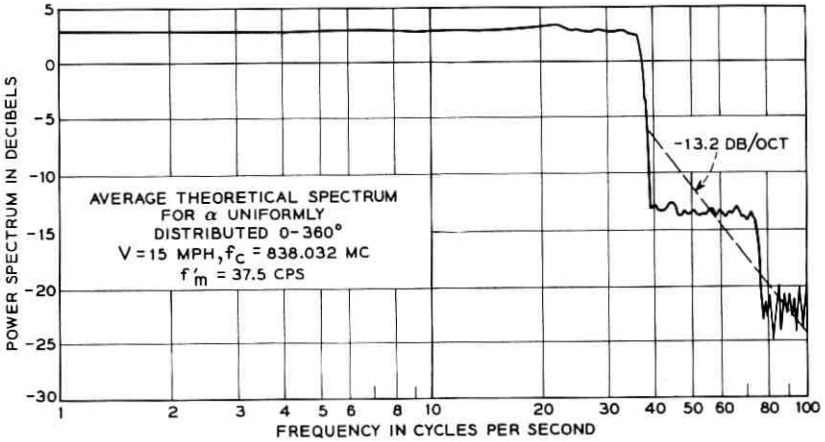


Fig. 6 — Average theoretical power spectrum for a uniform distribution of relative vehicle direction $\alpha$.

vehicle (a Volkswagen Kombi) traveling nominally 15 mph, to a fixed station at the Murray Hill Laboratories. The range was between 1 and 2 miles and varied little during any run. The duration of the recorded waveforms ranged from about 20 to 150 seconds. Values obtained for the parameter angle $\alpha$ ranged from 6° to 90°. All of the streets were in suburban residential areas except Central Avenue, which serves open fields and a few single-story industrial and commercial buildings. The weather was clear and dry.

The vertically polarized transmitting antenna atop the vehicle was a stack of $2\frac{1}{2}$ coaxial dipoles with a net gain of about 4.5 db. The interaction with a second similar antenna several wavelengths away is not known, but is believed to be small. The receiving antenna was a vertically polarized 13-element coaxial array mounted atop a rooftop elevator house. It had about 11 db gain and a 3-db beamwidth of about 6°.

A voice channel was recorded simultaneously with the fading signal on a second FM tape channel. This channel carried a running commentary describing the data and included start- and end-of-data announcements. Also recorded on this same channel were tone bursts triggered every nominal 0.01 mi by a cam attached to the speedometer cable. The exact vehicle speed was ultimately recovered from these bursts.

To permit over-all calibration of the static transfer characteristic of the system, calibration levels 3 db apart over a 60-db range were recorded both prior and subsequent to the recording of the fading signals. The two stations were directly connected by coax for calibration. Each level was recorded for a few seconds along with appropriate voice announcements.

A complete set of Visicorder records were then made from the FM magnetic tape of both the fading signals and the tone bursts for a preliminary examination of the data and for later determination of the vehicle speeds.

The pertinent parameters for data runs whose power spectra are shown in this article are given in Table I. The system bandwidth was limited by the receiver, which was 3 db down at 310 cps. The angular elevation above the horizon of the fixed station as viewed from the mobile station was usually between 1° and 3°.

### 6.1 *Vehicle Speed*

Four timed test runs were made in the vehicle over a fixed, level course of 1443 ft., to determine typical speed variations during a run and

Table I — Parameters of Recorded Data for Which Comparisons Are Shown between Theoretical and Experimental Spectra*

| Case No. | Street | Avg. Speed mph | Alpha (deg) | | | $f'_{\max}$ cps | $f'_m$ cps | Fig. No. |
|---|---|---|---|---|---|---|---|---|
| | | | min | max | avg. | | | |
| 1 | Commonwealth | 15.8 | 82.2 | 84.5 | 83.3 | 22.4 | 39.5 | 10 |
| 2 | Charnwood | 15.8 | 73.0 | 75.8 | 74.4 | 25.5 | 39.5 | 11 |
| 3 | Whitman | 15.2 | 67.8 | 80.4 | 74.2 | 26.2 | 38.0 | 12 |
| 4 | Elkwood | 16.2 | 41.0 | 42.6 | 41.8 | 35.4 | 40.4 | 13 |
| 5 | Ridgeview | 15.8 | 167.8 | 168.4 | 168.1 | 39.2 | 39.6 | 14 |
| 6 | Ridge | 15.9 | 15.0 | 15.8 | 15.4 | 38.9 | 39.7 | 15 |
| 7 | Central | 16.0 | 68.7 | 72.9 | 70.8 | 27.2 | 39.9 | 16 |

* All spectra were computed using 5000 sample points (20 sec at 250 points/sec).

to calibrate the tone burst rate. The nominal speed for each run was 15 mph = 22 ft./sec. Visicorder recordings were made of the bursts from the FM tape to enable counting them and measuring their spacing; the Visicorder paper speed was determined to be 1.019 in./sec using a 10-cps square wave (set by frequency counter). The tone burst rate was found to be 51.98 ± 0.16 ft. between beginnings of bursts. Using this burst rate, the averages and standard deviations of the speed during these four runs are shown in Table II.

The actual average vehicle speed during each data run or part of a run was determined from the Visicorder records which have the tone bursts plotted alongside the fading signal. Let $N_B$ be the number of bursts occurring during a part of the run and $D_R$ (inches) be the corresponding length of Visicorder paper. The average vehicle speed $S_R$ for that part of the run was then computed from $S_R = 53.0\ N_B/D_R$ ft./sec.

### 6.2 *Location of Data Runs*

The precise location of each run was carefully marked on a set of topographic maps (100 ft = 1 inch) which showed actual street and house shapes. Typically, street intersections and poles were used as starting and ending points. Except in the case of Whitman Road, which is slightly S-shaped, the vehicle was driven in a straight line.

## VII. PATH TRAJECTORY DATA

The set of topographic maps referred to previously have a common coordinate grid. By determining the coordinates of the starting and

TABLE II — AVERAGES AND STANDARD DEVIATIONS OF SPEEDS DURING FOUR TEST RUNS

| Test Run | Avg. Speed (ft./sec) | Std. Dev. (ft./sec) |
|---|---|---|
| 1 | 23.15 | 0.63 |
| 2 | 23.08 | 1.09 |
| 3 | 23.16 | 0.54 |
| 4 | 23.43 | 0.41 |
| Average | 23.21 | 0.67 |

ending points of each run, and the coordinates of the fixed station, it is possible to compute the vehicle azimuth (path azimuth), the azimuth of the direction from the fixed station to the mobile station (fixed azimuth), the angle $\alpha$ between the direction of the vehicle and the direction to the fixed station (positive if fixed station is to the left of the vehicle), and the range at various points along the run. Except in the case of Whitman Road, the end points were connected with a straight line which was then divided into 50-ft. intervals (the last interval usually extending past the original end point). The value of $\alpha$ was then tabulated for the distances $n50$ ft. ($n = 0,1,2, \cdots$) along each run. It should be noted that even with a straight path $\alpha$ varies because of the finite distance to the fixed station.

In the case of Whitman Road, where the path trajectory follows the shape of the road and is not straight, a larger map (50 ft. = 1 in.) was used which showed the actual azimuth variation along the street.

## VIII. DIGITAL PROCESSING

Following digitization of the fading data, the calibration, plotting, filtering and spectral analysis were accomplished on an IBM 7094. Many computer programs and subroutines were written for these purposes as well as for such auxiliary purposes as calibration curve fitting, magnetic tape searching (subroutines that can conveniently retrieve requested data pieces), spectra equalizing and plotting, and vehicle path angle determination. An available set of time series processing subroutines[3] was extensively used; this set included subroutines for tapering and detrending data, and for computing auto- and cross-covariances and Fourier transforms. A large arsenal of subroutines was eventually amassed, and writing a program for some particular task became the relatively simple job of writing a program to call appropriate subprograms.

## IX. INITIAL DATA PROCESSING

### 9.1 *Digitization*

Both the fading waveforms and the recorded fixed calibration levels were digitized on an analog-to-digital converter within the Laboratories,[4] using 11 bits/sample and sampling at 500 cps. The procedure for digitizing consisted of playing back the analog tape, listening to the voice-channel announcements, and manually triggering the digitizer on and off at the indicated times. Approximately 2-second intervals of each calibration level were digitized. The signals were filtered prior to sampling by a passive filter which was 3 db down at 250 cps, 10 db down at 300 cps, and subsequently fell 36 db/octave. The folding or Nyquist frequency (see Ref. 2, p. 30 ff.) of $500/2 = 250$ cps was chosen to safely contain the expected power spectra.

### 9.2 *Microfilm Plotting*

The digital data was read into the 7094 and completely plotted on microfilm on a peripheral General Dynamics 4020 microfilm printer. This provided a good visual record of the raw data as well as a check on the digitizing process. A computer subroutine was developed which generates a long continuous plot down the length of the microfilm. Such plots were produced for monitoring after every step in processing or transcribing the data. The comparative ease with which large quantities of digital data can be monitored by viewing microfilm considerably reduces the chance of the accidental processing and use of data containing errors. The 17 runs of recorded fading waveforms, which totaled over 920 seconds, yielded over 460,000 data points when digitized at 500 cps. When plotted at 480 points per 35 mm frame, the complete data comprising 960 frames could be viewed in detail on a roll film viewer in about an hour. Fig. 7, which exhibits a typical data section, was traced from a print from one frame of microfilm.

### 9.3 *Calibration*

The communication system nonlinearities, including that of the linear-to-log converter used during analog recording, had to be removed to obtain true signal amplitude. The before and after (the data) sequences of calibration level records were read on the 7094, and each record was averaged to remove noise and obtain a calibration point. Any system net drift during original data recording or during digitization would make the before and after curves different. Fortunately, they were quite similar and they were averaged to obtain the adopted calibration curve. A suitable function was then fitted, using a least squares criterion, to the list of calibration points. The adopted calibration function is

$$\begin{aligned} Y = {} & -60.255 + 0.8282\,(X - 170.6)^{\frac{1}{2}} - 0.01614\,(X - 170.6) \\ & + 8.474 \cdot 10^{-6}\,(X - 170.6)^2 - 9.658 \cdot 10^{-10}\,(X - 170.6)^3, \end{aligned} \tag{26}$$

where $X$ is the digital sample value and $Y$ is the true signal in relative db. This function, which has a maximum error of 0.47 db near $Y = -3$ db and rms error of 0.21 db, is shown in Fig. 8 plotted along with the original calibration points. Input values outside expected limits of $X = 170.60$ and 4041.06 were clipped to these values. The calibration program kept a statistical history of any clipped regions. The signal amplitude is then exp $(0.11512926Y)$.

## X. INITIAL ANALYSIS AND SECOND-STAGE PROCESSING

### 10.1 *Preliminary Power Spectra*

These were computed for several pieces of the data to determine whether any smoothing and decimating [Ref. 2, pp. 129–135] was

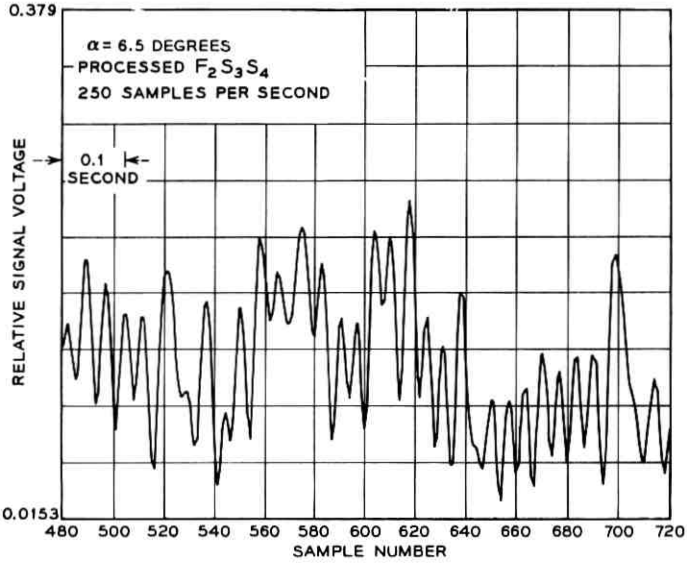


Fig. 7 — Typical section of data.

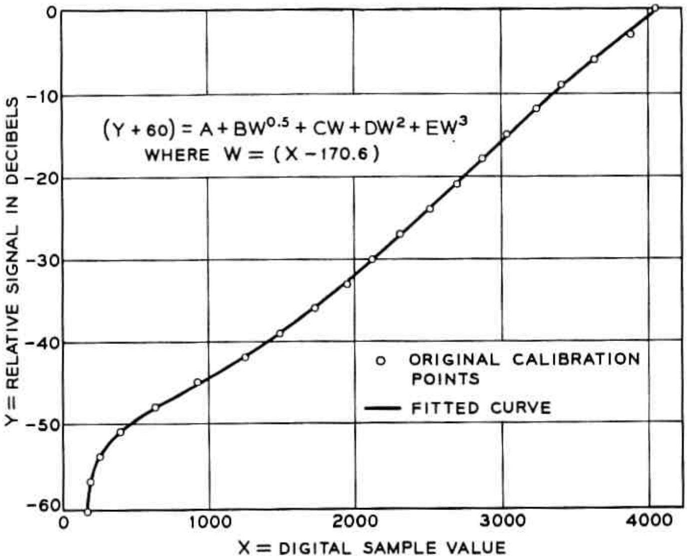


Fig. 8 — Original calibration points and fitted curve from (26).

necessary or desirable. The power spectra computation will be discussed later. Being sure to pick runs expected to have the widest band spectra, it was determined that the significant portions of the spectra were safely below one-half the folding frequency of 250 cps. Decimation by two (retaining every other point), which would reduce the folding frequency to 125 cps, would be safe and would reduce computation time. Suitable smoothing prior to decimating can also remove or reduce the 120-cps and higher hum peaks which were observed. Removing this hum is not essential for the spectral analysis, but doing so makes the data more suitable for level crossing analysis.

### 10.2 *Smoothing and Decimating*

The decimation of data retaining every $J$th point, symbolically indicated by $F_J$, multiplies the folding frequency by $1/J$. To prevent power, including noise power and hum power, at frequencies above the new lower folding frequency from folding over and appearing below this frequency, the data must be smoothed (or filtered) before decimation (see Ref. 2, pp. 129–135).

The most economical type of smoothing in digital analysis is to compute straight running means of $L$ consecutive values. Usually, simple sums which differ from the means by a factor are used to obviate division by $L$. This smoothing, symbolically indicated by $S_L$, is then

$$Y_i = \sum_{j=i-L+1}^{i} X_i, \tag{27}$$

and has the power transfer function

$$S_L(f) = \left[ \frac{\sin \dfrac{L\pi f}{2f_F}}{\sin \dfrac{\pi f}{2f_F}} \right]^2, \tag{28}$$

where $f$ is the frequency and $f_F$ is the folding frequency. $S_L(f)$ has periodic transmission nulls at $(f/f_F) = 2n/L$, where $n = 1, 2, \cdots$. Because the loss between nulls is usually not too great, a common procedure is to smooth twice with $S_L$ followed by $S_{L\pm1}$ (indicated by $S_{L\pm1}S_L$); the second smoothing will have nulls tending to fall between those of the first.

The processing chosen for the present data was $F_2S_3F_4$ — smoothing by threes and fours and then retaining every other point. The new folding frequency is $250/2 = 125$ cps. The smoothing loss is plotted in Fig. 9 as a function of the fraction of the new folding frequency; the folded portion of the loss curve is also shown. Maximum loss occurs at 125 (near 120), 167 (near 180), and 250 cps. The loss peak at 125 will make it impossible to obtain accurate spectral estimates close to the folding frequency, but this is not serious. Spectra can now be computed with the same resolution, stability and duration of data, for one-fourth of the computer time.

## XI. POWER SPECTRA

The method employed in determining power spectral estimates is that described by Blackman and Tukey;[2] another good reference is Ref. 5.

### 11.1 *Spectral Computation Parameters*

The spectra to be shown were computed using 5000 points (20 seconds at 250 samples/second). The mean and a least squares linear trend was removed from the data sections used, and the first and last 5 per cent of each section were raised cosine tapered to zero. The autocovariances (mean lagged products) were determined for 100 lags (i.e. for lags of $n\Delta t$, where $n = 0, 1, \cdots, 100$, and $\Delta t$ is the sample spacing). A finite cosine transform of the autocovariances then provides spectral estimates $125/100 = 1.25$ cps apart from zero to the folding frequency. The spectra

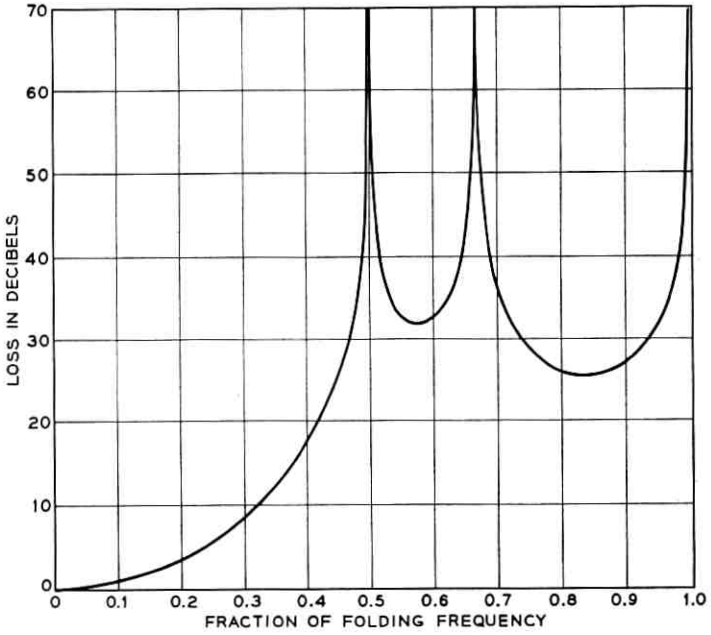


Fig. 9 — Smoothing loss for $S_3S_4$ vs fraction of folding frequency.

were smoothed by hanning (see Ref. 2, p. 98). Under these conditions, each estimate has a 90 per cent chance of being within about a 2-db range of the true spectrum. Or, the difference between the estimate and the true spectra has a variance of about $(0.3 \text{ db})^2$.

### 11.2 *Computed Spectra*

Almost fifty different spectra were computed from the data collected. A representative set of these are plotted (circles) on Figs. 10–16, where they may be compared to corresponding theoretical spectra. The comparison will be discussed later. All of the plotted spectra were equalized for the smoothing loss before plotting. It may be noted that most of the curves are not plotted beyond some frequency between 80–100 cps. The smoothing and decimating ($F_2S_3S_4$) produced an infinite-loss notch in the spectrum at 125 cps. When the spectrum is subsequently computed this hole is filled in by computation noise. The plot was automatically ended at the frequency where equalization of this noise started to produce a meaningless result. Even so, the last few points plotted are inaccurate and tend to be lower than they should be.

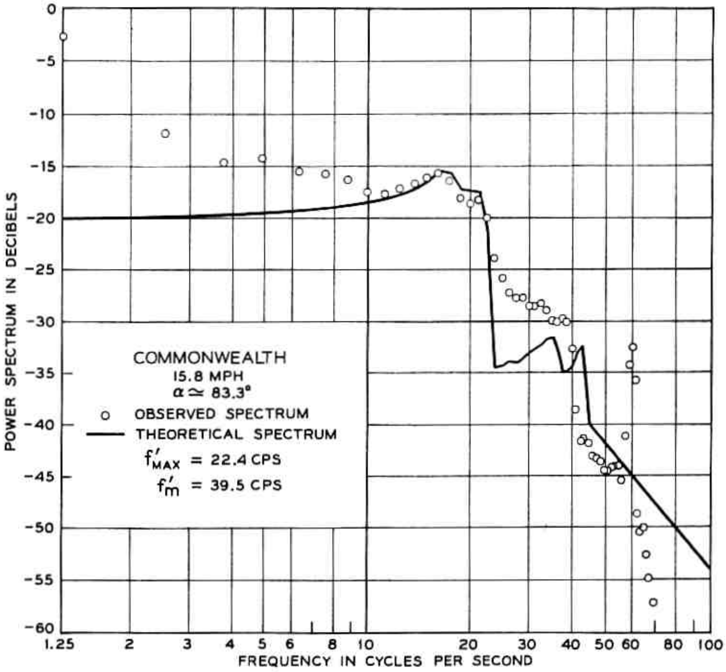


Fig. 10 — Experimental fading waveform spectrum compared with corresponding theoretical spectrum. Relative amplitudes are arbitrary; see discussion in text.

### 11.3 *Spectral Density Curve Shapes*

The spectra all exhibit significant power density out to some frequency between about 20 to 40 cps, where the density falls sharply between 10 and 15 db and then more gradually at about 12 db/oct. Many show a distinct shelf between the sharp-fall frequency and the subsequent slow fall-off. The shelf generally ends with a noticeable sharp drop at a frequency about twice the earlier sharp-fall frequency. The shelf in Fig. 14 has a noticeable peak prior to its fall at about 75 cps. The peak at 60 cps in all the plots is power supply hum. Another significant common feature is the relatively narrow peak immediately preceding the sharp-fall frequency. Many of the spectra exhibit a noticeable broad peak below the narrow one. All of the spectra rise 10 to 15 db at low frequencies.

## XII. COMPARISON OF THEORETICAL AND EXPERIMENTAL SPECTRA

To compute a theoretical spectrum corresponding to a particular experimental spectrum, it is necessary to know the carrier frequency

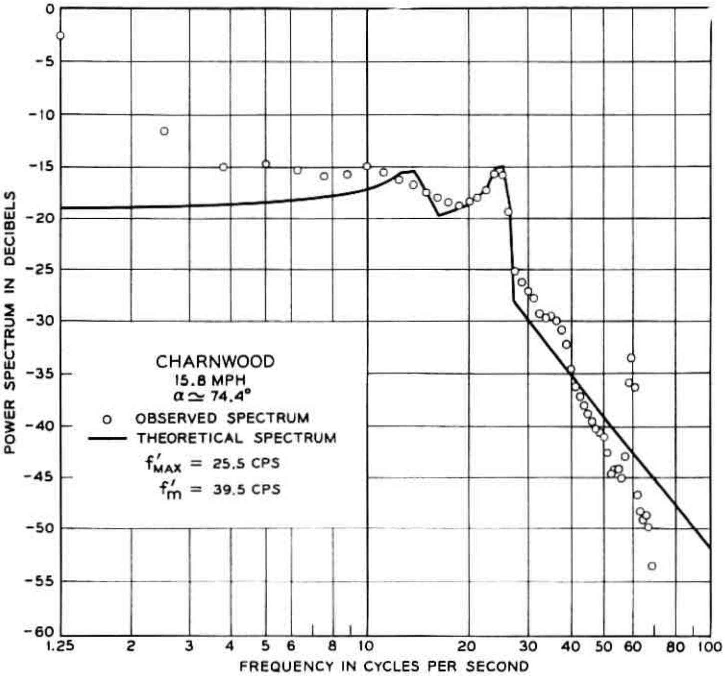


Fig. 11 — Experimental fading waveform spectrum compared with corresponding theoretical spectrum. Relative amplitudes are arbitrary; see discussion in text.

and vehicle speed in order to compute $f_m$, and to know the variation in the parameter angle $\alpha$. For the present comparisons, the carrier frequency is 838.032 mc, and the average vehicle speed and corresponding $f_m'$ are shown in Table I. The range of $\alpha$'s represented in the comparisons is from about 12° to 83°; remember that the spectrum for $\alpha = 90° + x$ is the same as one for $\alpha = 90° - x$. The weighting function $W$ was limited to $W_{\max} = 15$ (compatible with the typical run length of 450 feet and typical house side length of 30 feet).

For each theoretical spectrum, the corresponding list of $\alpha$'s was used. Spectra were computed for each $\alpha$ and the final spectrum was the hanned weighted average. Harmonic power was included in each $\alpha$'s spectra before averaging. A value of $W_{\max} = 15$ was used in all cases. Table I also lists the $f_{\max}'$ corresponding to the value of $\alpha$ occurring during run that is nearest to 0° or 180°. As an example of how the list of $\alpha$'s was used, consider Commonwealth Avenue (case 1). These data are actually points 1–5000 (250 pts/sec) of a longer piece. The vehicle speed according to Table I was 15.8 mph; thus the run was (20 sec)(15.8 mph)(22

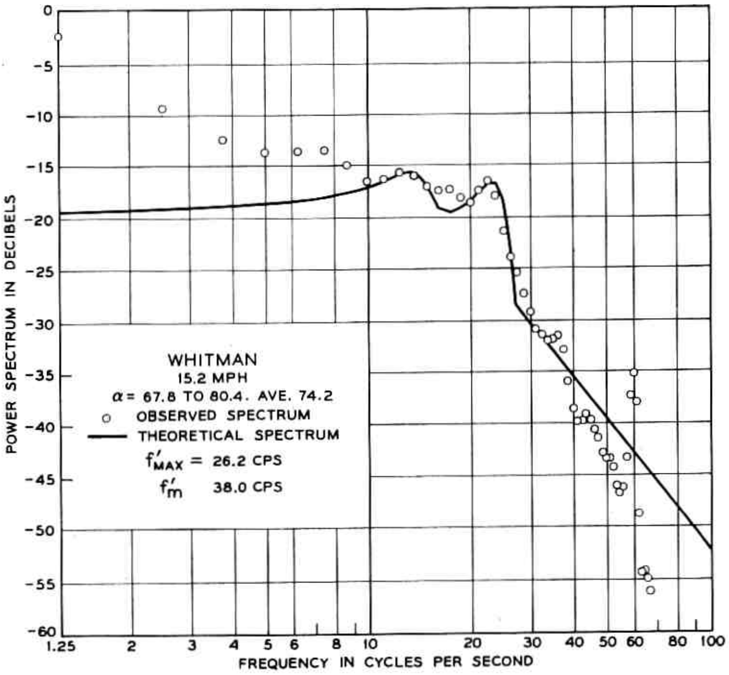


Fig. 12 — Experimental fading waveform spectrum compared with corresponding theoretical spectrum. Relative amplitudes are arbitrary; see discussion in text.

ft./sec)/(15 mph) = 463 ft. in length. The angles $\alpha_n$ are values computed for distances $d_n = 50\ n$ ft., where $n = 0, 1, 2, \cdots$ and may be considered to represent the distance ranges $d_n \pm 25$ ft. Thus the relative weights for the $\alpha_n$ and the corresponding spectra are $w_1 = 0.5$, $w_2$ to $w_9 = 1.0$, and $w_{10} = 0.76$; the latter is $(463 - 25 - 8.50)/50$. When the data section does not begin at sample 1, the distance between sample 1 and the starting sample must be determined using the proper average vehicle speed for that interval.

The change in $\alpha$ during the data section is small enough in many cases to permit using the average $\alpha$ to compute the spectrum. For example, during the run of case 4 the angle $\alpha$ varies only from 41.0° to 42.6°; a spectrum computed from the average value of about 41.8° differs little from one determined by averaging. In other cases — Whitman Road for example — the spectrum determined by averaging has much broader peaks than one corresponding to the average $\alpha$. All the theoretical spectra to be shown were determined by averaging, whether this was necessary or not.

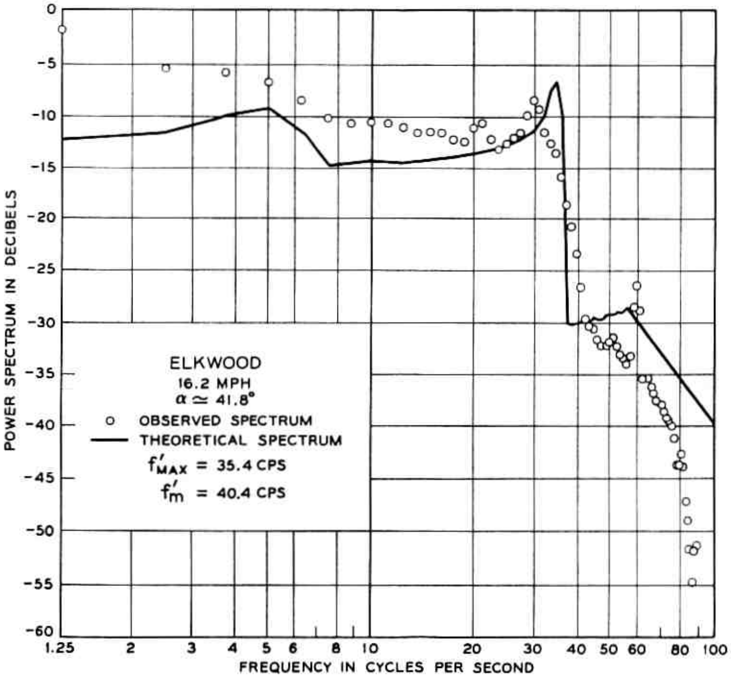


Fig. 13 — Experimental fading waveform spectrum compared with corresponding theoretical spectrum. Relative amplitudes are arbitrary; see discussion in text.

### 12.1 *Comparing Theoretical and Experimental Curves*

Figs. 10–16 show theoretical spectra (solid curves) superimposed on experimental spectra (circles); Table I lists the data sections involved and gives corresponding figure numbers. The only thing arbitrary in comparing the theoretical and experimental spectra is their relative amplitude. Thus the theoretical curve has been shifted vertically to produce some sort of fit. In all cases a transition is shown from the basic theoretical spectrum to a fall-off line fitted by least squares to the portion of the theoretical spectrum above $f_{\max}'$; this fall-off typically has a slope of $-12$ to $-13$ db/oct. The last few points of each experimental curve are not very accurate and tend to be low, as previously discussed.

### 12.2 *General Results of the Comparison*

Before reading further, the reader should make a superficial scan of Figs. 10–16. The agreement between the theoretical and experimental spectra is generally quite good. The main discrepancy is that the ob-

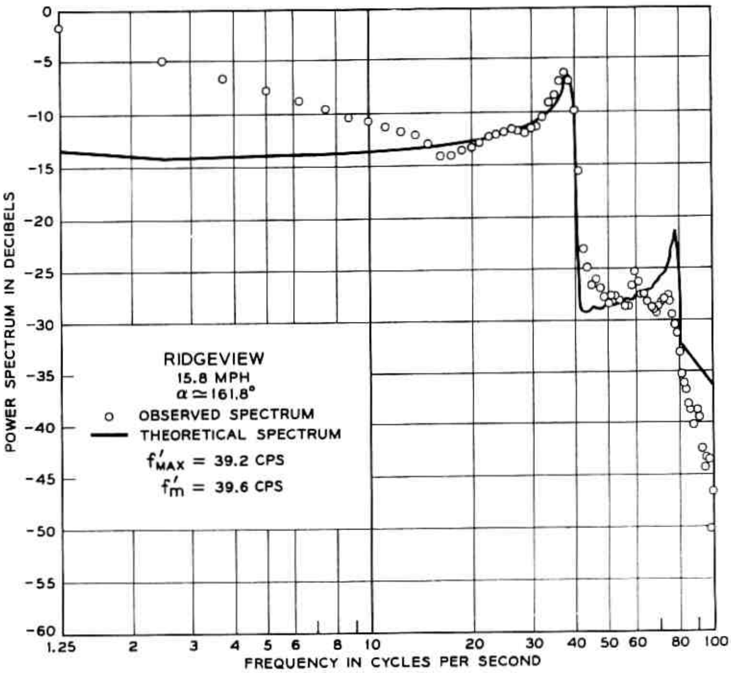


Fig. 14 — Experimental fading waveform spectrum compared with corresponding theoretical spectrum. Relative amplitudes are arbitrary; see discussion in text.

served rise in spectral density at low frequencies is not predicted by the theory. The sharp-fall frequency agrees very well in almost all cases. The peak prior to this sharp fall fits well in many cases. In cases where intermediate peaks are predicted ($\alpha$ not too close to zero), the experimental spectra usually exhibit them. The second harmonic shelf is well formed in many cases. The following are some comments on specific comparisons:

#### 12.2.1 *Case 1; Fig. 10*

This street had an average $\alpha$ of about 83°. The two peaks have nearly merged and have formed a double peak which the experimental spectrum exhibits in agreement. The second-harmonic shelf is higher than predicted and ends somewhat early; more will be said about this later.

#### 12.2.2 *Case 2; Fig. 11*

Here $\alpha$ averages about 74°. The upper peak and the sharp fall agree well. The intermediate peaks are in only fair agreement. Because of

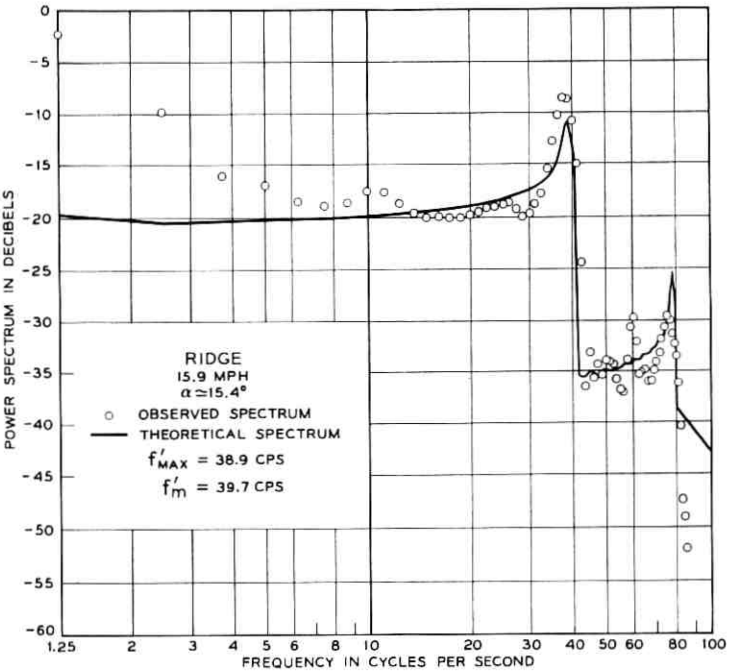


Fig. 15 — Experimental fading waveform spectrum compared with corresponding theoretical spectrum. Relative amplitudes are arbitrary; see discussion in text.

the absence of a second-harmonic shelf in the experimental spectrum, the theoretical fall-off line is plotted beginning with its intersection with the first fall. A slight rise above this line occurs out to nearly 40 cps; a second-harmonic shelf would have to extend to about 50 cps.

### 12.2.3 *Case 3; Fig. 12*

The comparison here is similar to that discussed for case 2. This street, however, has an $\alpha$ which varies between 67.8° and 80.4° and averages 74.7°. The comparatively broader theoretical and experimental peaks may be noted.

### 12.2.4 *Case 4; Fig. 13*

This street has $\alpha$ averaging about 42°. The intermediate peak at about 5 cps is discernible. The sharp fall occurs at about the right frequency but is not as steep as expected. The second-harmonic shelf is not noticeable. The theoretical fall-off line is picked up at its intersection with the theoretical shelf.

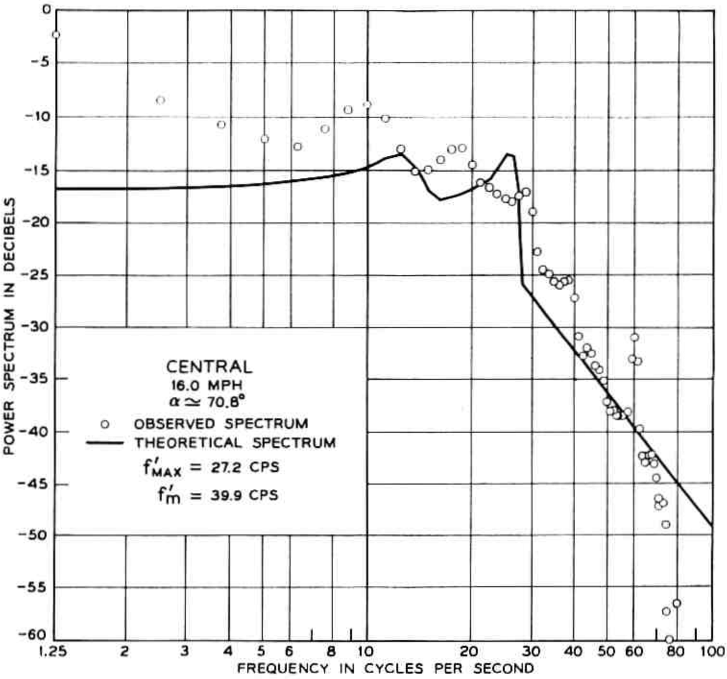


Fig. 16 — Experimental fading waveform spectrum compared with corresponding theoretical spectrum. Relative amplitudes are arbitrary; see discussion in text.

### 12.2.5 *Cases 5 and 6; Figs. 14 and 15*

These streets have a small $\alpha$ (about 12° and 15° respectively) and the intermediate peak is too close to zero frequency (below 1 cps) to show up with the present resolution. The main peak and the sharp fall agree exceptionally well and the second harmonic shelves are well exhibited. Furthermore, the predicted peak on the shelf appears in the experimental spectrum of Fig. 15 (and to a lesser extent in Fig. 14).

### 12.2.6 *Case 7; Fig. 16*

The agreement between the theoretical and experimental curves is in this case relatively poor. This street, however, is not in a suburban residential area, and has only a few low industrial buildings spaced well back from the curb. The section of the street corresponding to Fig. 16 had an average of $\alpha$ of about 71°. It was observed that, if experimental spectra were computed for other portions of the street with different $\alpha$'s, these spectra were quite similar if their ripples were ignored.

### 12.3 *Additional Observations*

It has been observed that the streets having the best agreement between experimental and theoretical spectra in the vicinity of the higher peak, sharp fall, and harmonic shelf are those with $\alpha$ near 0° or 180°. A good reason why this is not too surprising is offered in Section XIII under a discussion of nonrandom reflector orientation.

A cause for the small but discernible drop in Figs. 10, 11, 12 and 16 at the frequency corresponding to $f_m'$ is offered in Section XIII under a discussion of simultaneous reflections.

When the harmonic content was included in the theoretical spectra it was assumed that the reflectors were perfect conductors. The observation of second-harmonic shelves at the predicted amplitude level in many cases indicates that the assumption was reasonable. The use of an aluminum foil vapor barrier integral with outside wall insulation is common in current house construction and may explain their good reflectivity.

## XIII. LIMITATIONS OF THE MODEL

In the preceding section it was seen that a major deficiency of the theoretical model is its failure to forecast the rise in spectral density at the low-frequency end of the spectrum. Some of the mechanisms that can contribute low-frequency energy are discussed in the following paragraphs.

### 13.1 *Shadowing by Buildings*

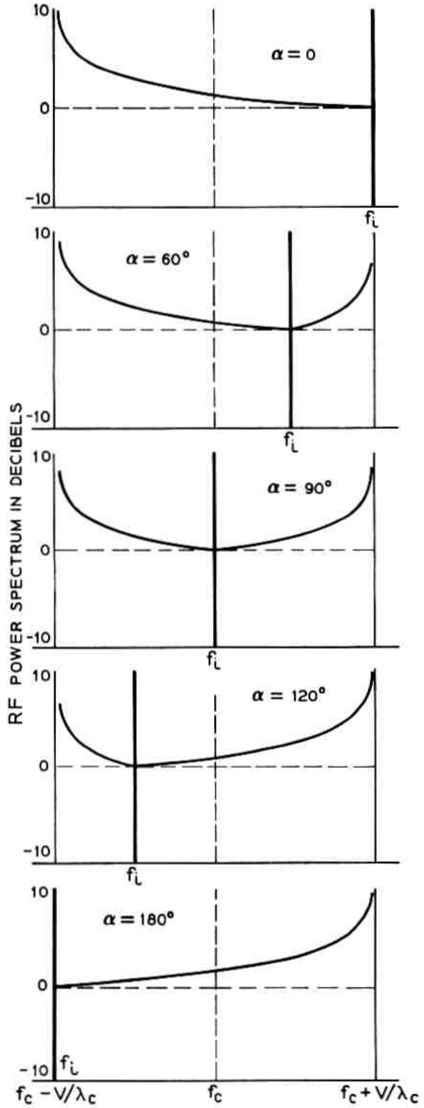
The shadowing of the direct signal by buildings introduces into the fading waveform a low-frequency multiplicative function (likely with some harmonic content) with a fundamental spectrum probably not extending much beyond about ¼ cps (houses spaced 80–100 ft. apart and a vehicle speed of about 15 mph). The resulting fading waveform spectrum would be the convolution of the spectrum without shadowing with that of the low-frequency function. Such a low-frequency multiplicative function was observable in some portions of microfilm plots of recording fading waveforms. It is not likely that this effect explains the entire extent and shape of the low-frequency rise.

### 13.2 *Ground Reflections*

These cause standing-wave patterns which normally vary only in a vertical direction. But as the vehicle moves the point on the ground causing the ground reflection moves and the reflectivity will vary. This and the shadowing of ground reflections by buildings introduce low-frequency variations in the direct signal.

### 13.3 *Simultaneous Reflections*

An assumption inherent in the construction of the theoretical spectra was that the mobile vehicle was under the influence of one reflector at a time. The simultaneous presence of more than one reflected signal will give rise to additional beat frequencies in the fading waveform between the reflected signals (see Section IV). Here the Doppler point of view is useful. From (14) the Doppler-shifted reflected signal is seen to lie in the frequency range $f_r = f_c \pm V/\lambda_c$; thus, if all possible reflections are always present, the radio frequency spectrum would have a bandwidth $2V/\lambda_c = f_m'$. The Doppler shifted direct signal has from (13) a frequency of $f_i = f_c + (V/\lambda_c) \cos \alpha$. The shape of the spectrum is not symmetrical about $f_c$. This shape can be obtained by picking frequencies $f_r$ between $f_c - V/\lambda_c$ and $f_c + V/\lambda_c$, solving (14) for the two values of $\varphi$ corresponding to each frequency, and then summing the two corresponding values of the weighting function $W$ obtained from (18). The result is a spectrum having a broad minimum at $f_r = f_i = f_c + (V/\lambda_c) \cos \alpha$ and peaks at $f_r = f_i \pm V/\lambda_c$. Thus, as $\alpha$ varies from $\alpha = 0$ to 90° to 180°, the direct signal $f_i$ moves from the upper-frequency end of the spectrum to the center and to the lower end. Fig. 17 shows the radio-frequency spectrum as a function of $\alpha$ (the peaks appear sharp because no smoothing has been applied). If the direct signal is large in amplitude compared to all the reflected signals, the spectrum of the envelope would essentially be that obtained previously (Section V), except for the lack of harmonic content. If the direct signal is ignored, the spectrum of the envelope would be the convolution of the radio-frequency spectrum with itself (see Ref. 6, Chap. 12); this spectrum, which would vary from a maximum at zero frequency to zero at $2V/\lambda_c$, would be virtually independent of $\alpha$. The fact that the convolution would carry the spectra only out to $2V/\lambda_c$ suggests that beats between reflections may be responsible for the partial filling in of the second-harmonic shelf when $\alpha$ is not near 0 or 180°. This effect is noticeable to various degrees in Figs. 10, 11, 12 and 16, where a perceptible drop occurs at $f_m'$ independent of any termination of the second-harmonic shelf at $2f_{\max}'$. Similarly, this effect probably decreased the steepness of the observed steep fall in Fig. 13. It remains a fact that in a suburban residential environment, the spacing of houses is such that there is a very strong tendency for a mobile vehicle to experience only one domi-



$f_c$ = TRANSMITTED CARRIER FREQUENCY

$\lambda_c$ = WAVELENGTH CORRESPONDING TO $f_c$

$V$ = MOBILE STATION SPEED

$f_i$ = DOPPLER SHIFTED DIRECT SIGNAL = $f_c + \frac{V}{\lambda_c} \cos \alpha$

Fig. 17 — Radio-frequency power spectra for a range of relative vehicle direction $\alpha$.

nant reflection at a time. There certainly is some continual overlap of reflected beams and this may in part be responsible for the low-frequency rise in the experimental spectra.

### 13.4 *Nonrandom Reflector Orientation*

Houses are generally built with their larger flat sides parallel and perpendicular to the street. It has been shown from (11) that a zero fade rate obtains when $\varphi = 0°$, $2\alpha$, and $180°$; low-frequency fades obtain when $\varphi$ is near these values. The weighting function $W$ is zero for $\varphi = 0°$ and $180°$ (except when $\alpha = 0°$ or $180°$), but is unity at $\varphi = 2\alpha$. Clearly, it is the reflectors whose flat sides are roughly parallel to the direction of vehicle travel that contribute the low-frequency fade rates; each such reflector makes its contribution when the vehicle position is such that its $\varphi \approx 2\alpha$. Conventional house orientation obviously increases the supply of reflectors causing low-frequency fades above that under random conditions. It is felt that this effect may be the most important reason for the low-frequency rise in experimental fading spectra.

The same nonrandomness will affect the experimentally observed spectral peaks. These peaks occur when $\varphi = \alpha$ and $\alpha + 180°$; the corresponding required reflector orientations, relative to the direction $\alpha$, are $\alpha/2$ and $90 - \alpha/2$. When $\alpha$ is small or near 180°, these required relative orientations are near 0° and 90° — i.e., near parallel and perpendicular to the direction of travel. When $\alpha$ is near 90°, the required relative orientations are both near 45°. Thus observed peaks for $\alpha$'s near 90° may be relatively subdued by the relative absence of required reflectors. This effect has been observed in several comparisons of theoretical and experimental spectra.

### 13.5 *Other Low-Frequency Effects*

Nonuniformity of the fixed station antenna pattern and reflectors in the vicinity of the fixed antenna can produce some very low-frequency variations in the standing wave pattern. Motion of the fixed antenna and trees due to wind are additional sources of low frequencies in the standing wave pattern.

## XIV. ADDITIONAL TOPICS

### 14.1 *Moving Reflectors*

All of the preceding discussion was concerned with fixed reflecting objects. The experimental data were taken with the streets devoid of other moving vehicles. What effect will the motion of other vehicles have on the spectrum? Let us limit this discussion to vehicles moving in the same or opposite direction on the same street as the mobile station; let $V_1$ and $V_2$ be the speeds of the mobile station and moving reflector respectively with a positive $V_2$ corresponding to closure between the vehicles. The moving reflector encounters and reflects a frequency

$$f_2 = f_c - \frac{V_2 \cos \alpha}{\lambda_c}, \tag{29}$$

where $f_c$, $\lambda_c$, and $\alpha$ have their previous meanings, and $V_2$ has a component away from the fixed station. The mobile station encounters a reflected beam of frequency

$$f_1 = f_2 + \frac{V_1 + V_2}{\lambda_2}, \tag{30}$$

where $\lambda_2$ corresponds to $f_2$, and encounters a Doppler shifted direct signal of frequency $f_i$ given by (13). The beat frequency $f$ between $f_i$ and $f_1$ is then

$$\begin{aligned} f = f_1 - f_i = f_2 + \frac{V_1 + V_2}{\lambda_2} - f_c - \frac{V_1 \cos \alpha}{\lambda_c} \\ = \frac{V_1 + V_2}{\lambda_2} - \frac{(V_1 + V_2) \cos \alpha}{\lambda_c} \\ \approx \frac{V_1 + V_2}{\lambda_c} (1 - \cos \alpha), \end{aligned} \tag{31}$$

where $c$ is the velocity of light. This result corresponds to (16) with $\varphi = \alpha$. Thus an oncoming vehicle with $V_2 = V_1$ could double the maximum observed fade rate.

### 14.2 *Horizontal Polarization*

In terms of the $x$ and $z$ coordinates of Fig. 1, the electric field components in the reflected beam region will take the form (see Ref. 1, p. 295)

$$E_x = jK \cos \theta \sin \left( \frac{2\pi z}{\lambda_c} \cos \theta \right) \exp \left( -j \frac{2\pi x \sin \theta}{\lambda_c} \right) \tag{32}$$

$$E_z = K \sin \theta \cos \left( \frac{2\pi z}{\lambda_c} \cos \theta \right) \exp \left( -j \frac{2\pi x \sin \theta}{\lambda_c} \right), \tag{33}$$

where $K$ is a constant and $\theta$ is the angle of incidence. If a nondirectional receiving antenna is used, such as is approximated by a "turnstile" consisting of two perpendicular dipoles connected together by a $\lambda_c/4$ stub, the received signal can be shown to be proportional to

$$|E| = \left| K \sin \left( \frac{2\pi z}{\lambda_c} \cos \theta - \theta \right) \right| \tag{34}$$

irrespective of the angular orientation of the turnstile. This result has the same form as (25) except for the spatial phase shift $\theta$. Thus the use of horizontal polarization together with the assumed antenna produces an effective standing-wave pattern that is identical to that for vertical polarization except for the spatial translation. Thus the fading situation would also be identical.

### 14.3 *Field Component Diversity*

Equations (32) and (33) show that $E_z$ is a maximum where $E_x$ is zero and conversely. Suppose the two dipoles of the turnstile antenna are not connected together with a stub but are offered simultaneously to the receiver. Then, if the receiver electronically switched to the dipole offering the greater signal, the receiver would in many cases never experience a null. However, when the direction to the reflector $\varphi$ approaches 180°, $\theta$ approaches 0° and the component $E_z$ becomes smaller and vanishes; likewise, $E_x$ vanishes when $\varphi$ approaches 0. Thus any diversity scheme dependent on choosing between $E_x$ and $E_z$ would be most successful near $\varphi = 90°$ ($\theta = 45°$) and unsuccessful near $\varphi = 0°$ or 180° ($\theta = 90°$ or 0°).

At every point in the reflected beam region it would be possible to rotate a dipole to a position where a maximum signal is picked up. This is not possible only when $\varphi = 0°$ or 180° exactly. It is possible, therefore, for a mobile dipole that is mechanically or electronically rotated continuously to receive a maximum signal, to reduce the amplitude of the fading due to vehicle motion, if horizontal polarization is employed. If the angular position of a mobile single dipole is fixed, the fundamental fading rate experienced is still the same function of $\alpha$ and $\varphi$ as before, except that the amplitude of the fading will vary because of the directivity of the dipole. For example, if $\varphi = 90°$ and the dipole is physically oriented perpendicular to the fixed station, the mobile dipole may be translated anywhere in the reflected beam region without any fading.

## XV. ACKNOWLEDGMENTS

Much credit is due J. L. Wenger and H. H. Hoffman for their part in recording the raw data; the former skillfully manned the fixed station equipment at Murray Hill and the latter operated the mobile equipment and had the difficult job of driving the Volkswagen at something approaching a constant speed. Thanks are due C. A. Sjursen and J. L. Wenger for their part in digitizing the data. Miss A. B. Strimaitis made the Visicorder record measurements. The direction of this effort benefited from B. P. Bogert's experience in time series analysis.

## REFERENCES

1. Ramo, S., and Whinnery, J. R., *Fields and Waves in Modern Radio*, John Wiley & Sons, New York, 1953.
2. Blackman, R. B., and Tukey, J. W., *The Measurement of Power Spectra*, Dover Publications, Inc., New York, 1959.
3. Healy, M. J. R., and Bogert, B. P., FORTRAN Subroutines for Time Series Analysis, Comm. ACM, **6**, Jan., 1963.
4. David, E. E., Mathews, M. V., and McDonald, H. S., Hi-Speed Data Translator for Computer Simulation of Speech and Television Devices, Proc. Western Joint Computer Conf., Mar., 1959.
5. Technometrics, **3**, May, 1961.
6. Davenport, W. B., and Root, W. L., *Random Signals and Noise*, McGraw-Hill Book Co., New York, 1958.

# Digital Data Signal Space Diagrams

By J. R. DAVEY

(Manuscript received April 20, 1964)

*Signal space diagrams are described which show the pattern of amplitude and phase variation for several kinds of modulated carrier signals commonly used in digital data transmission. Such diagrams illustrate important similarities and differences among the various modulation methods. Oscilloscope pictures of actual data signal patterns are presented, and it is shown that these patterns can be used to detect the presence of amplitude and delay distortions in the transmission channel.*

## I. INTRODUCTION

Many different kinds of modulated carrier signals are being used in digital communication systems. All these various data signals can be expressed in the general form $A(t) \cos [\omega_c t + \phi(t)]$ where a carrier $\cos \omega_c t$ is varied in amplitude by $A(t)$ and in phase by $\phi(t)$. The various modulation methods impart different patterns of amplitude and phase variation. The characteristic pattern of a given modulation method can be portrayed by a polar plot of $A$ and $\phi$ in which the angular reference is $\omega_c t$. This type of plot will be referred to as a "signal space diagram."

Signal space diagrams will be described for several kinds of carrier modulation. Only synchronous signals consisting of sequences of evenly spaced symbols will be considered. In each case the received symbols can be thought of as a sequence of carrier pulses or bursts, each with an envelope shape determined by the channel characteristic. In such a view the differences among the types of modulation are due to the number of pulse amplitudes and phases which are used, the particular phase sequences used, and the spacing between pulses. In order to obtain the simplest signal space diagram it is desirable to choose the reference $\omega_c$ as the center of the received pulse spectrum so that the phase variation of a single isolated pulse will be minimized. With a symmetrical pulse spectrum and a linear phase characteristic this reduces the pattern of a single isolated pulse to a radial line. When the phase of the carrier varies from pulse to pulse and the pulses overlap, more complicated patterns are formed. By transmitting a random sequence, a pattern of all the permitted amplitude and phase variations of that particular type of modulation is obtained.

Space diagrams will be presented for a number of commonly used data signals. In these examples the pulse envelope has been taken to have a raised-cosine shape in time in order to make the diagrams consist of circles and straight lines. This is a close approximation to the case of a raised-cosine pulse spectrum which is often typical in actual data systems. The pulse spacing, $T$, for the double-sideband examples is equal to the reciprocal of the half-amplitude width of the pulse spectrum. This corresponds to the maximum rate which avoids intersymbol interference as described by Nyquist. Pulse spacings of $T/2$ are used in FM and VSB methods where special conditions are established to avoid intersymbol interference.

## II. SIGNAL SPACE DIAGRAMS FOR VARIOUS TYPES OF MODULATION

### 2.1 *Amplitude Modulation*

The first example is for on-off AM where mark is represented by a pulse and space by no pulse. The carrier phase remains the same from pulse to pulse, thus resulting in a straight-line pattern as shown in Fig. 1. The signal positions at the mid-symbol sampling instants are indicated by points $M$ and $S$ which are separated by the pulse amplitude $A$. The shape of the pulses is indicated at the right in the figure.

The diagram for suppressed-carrier AM or two-phase signals is shown in Fig. 2. In this case a pulse is sent for both mark and space, but the carrier phase for space is opposite to that for mark. Again the diagram is a straight line, but the mark and space sampling points are separated

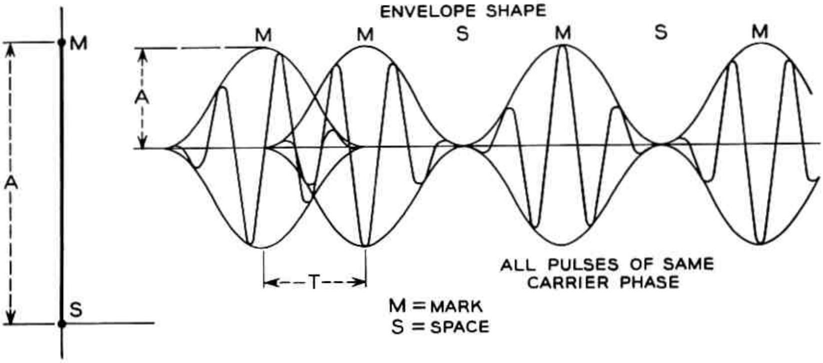


Fig. 1 — On-off AM.

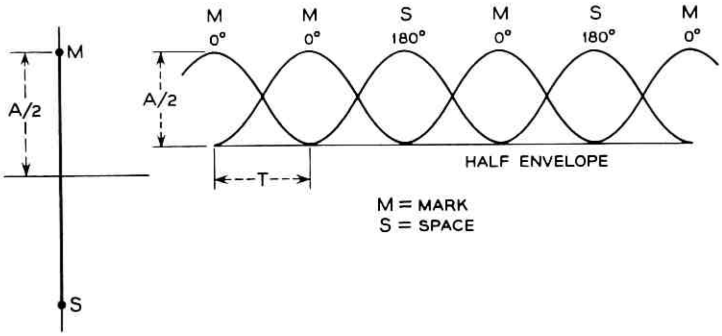


Fig. 2 — Suppressed-carrier AM or binary PM with 0° and 180° changes.

by twice the distance of the pulse amplitude. The positive half of the envelope and the carrier phase for each pulse are indicated at the right. A minimum separation of $A$ is obtained with a pulse amplitude of $A/2$. As compared with the on-off case of Fig. 1, the same margin against noise is obtained with 3 db less average power and 6 db less peak power.

### 2.2 *Phase Modulation*

Binary phase modulation where the choice of phase change is 0° or 180° results in the diagram of Fig. 2, as noted above. Alternatively, the choice of phase change can be ±90°. This has the advantage of symmetry and less amplitude variation. The diagram for this type of signal is a square, as shown in Fig. 3. The signal can move in either direction around the square and at the centers of the symbols is at one of the corners. Since there is always a 90° change between symbols, the signal alternates between corners marked with dots and those marked with circles. For a peak signal of $A/2$ there is a minimum separation between dot positions or circle positions of $A$, as was the case in Fig. 2.

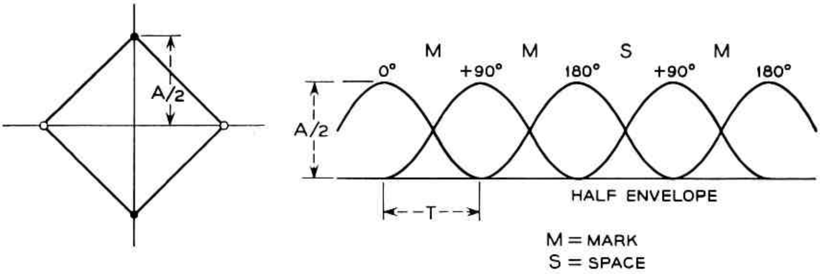


Fig. 3 — Binary PM with ±90° changes.

Note that the separation of interest is that between alternative choices for a given pulse rather than that between successive pulses.

Diagrams for two cases of quaternary phase modulation are presented. When the phase change between symbols is 0°, ±90° or 180° the diagram is a square with diagonals as shown in Fig. 4. The signal can progress around the square in either direction, go across a diagonal or remain at one corner with no restrictions. The four possible positions at the centers of the symbols are indicated by dots. For a minimum separation, $A$, between states the pulse peak becomes $A/\sqrt{2}$. This indicates that for this quaternary system to have the same noise margin per decision as the two-phase signal of Fig. 2 the power must be increased 3 db. This type of signal is equivalent to the sum of two AM suppressed-carrier signals at quadrature phase.

When phase changes of ±45° or ±135° are used between symbols there are eight possible phases for the pulses. The possible positions of the signal vector at the symbol centers are shown as dots and small circles on the diagram of Fig. 5. There is always a phase change between symbols, and the signal must alternate between dot positions and circle positions. With a peak pulse amplitude of $A/\sqrt{2}$ the minimum separation between dots or between circles is again $A$, as in the previous case.

### 2.3 *Vestigial Sideband*

It is assumed that the pulse spectrum for vestigial sideband has the same raised cosine shape used in the previous examples. It is also assumed that the pulse rate is twice the Nyquist rate for double-sideband operation and that the pulses originate from the modulation of a suppressed carrier higher in frequency than midband by an amount equal to one quarter of the pulse rate, as indicated in Fig. 6. This results in a phase change between adjacent pulses of ±90°. As shown in Fig. 6, the pulses overlap to the extent that at the peak of one pulse the adjacent pulses

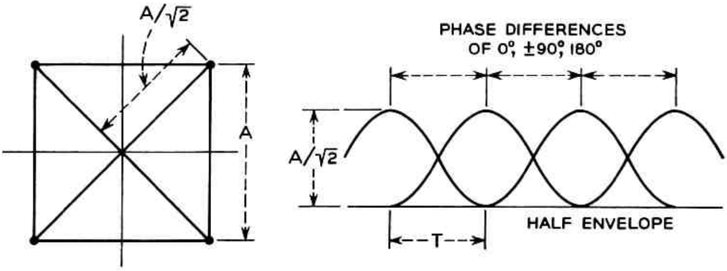


Fig. 4 — Quadrature AM or PM with 0°, ±90°, or 180° changes.

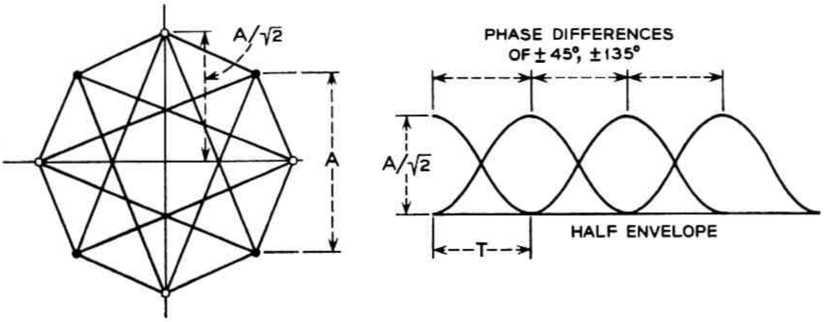


Fig. 5 — PM with ±45° or ±135° changes.

are each at half amplitude. This severe interference is at quadrature phase to the wanted pulse and is eliminated by the use of coherent detection. The signal phase at the center of a symbol is not affected if the two adjacent pulses are of opposing quadrature phases but is perturbed ±45° if the adjacent pulses are of the same quadrature phase. For example, on the diagram of Fig. 6 the center of a marking symbol can occur at any of the three dot positions at the top of the diagram de-

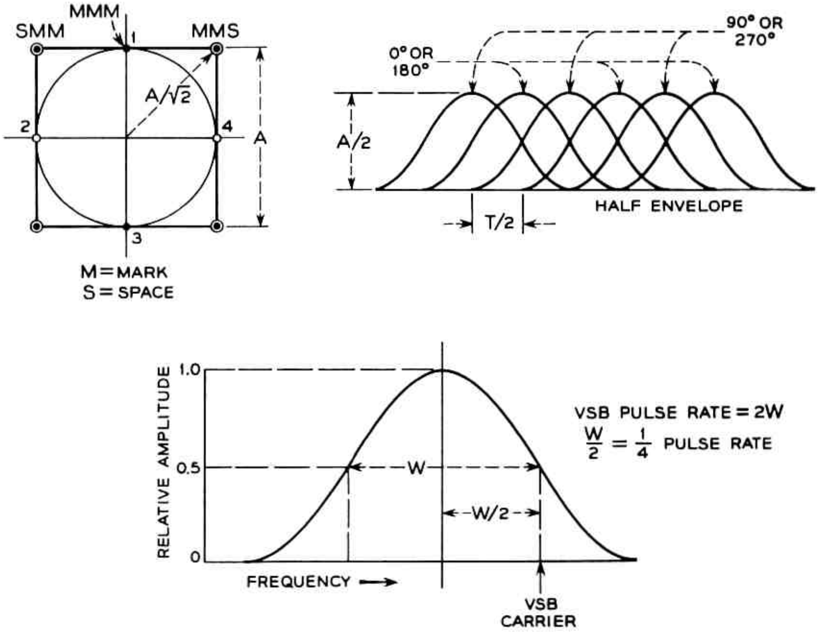


Fig. 6 — VSB: raised cosine pulse spectrum showing location of VSB carrier.

pending on the adjacent symbols as indicated. The phase of the coherent carrier used for detection advances around the diagram by 90° each symbol in the point sequence 1, 2, 3, 4. A continuous marking signal consequently follows this same sequence. A continuous spacing signal likewise advances 90° during each symbol but remains opposite in phase to the coherent reference.

For continuous mark-space alternations each symbol pulse is retarded by 90° from the preceding pulse, and the signal moves around the circle in the opposite direction from steady mark or space. The signal always alternates between dot and small circle points. The corners of the square portion of the diagram are both dot and circle points, and the signal may rest at such a point continuously and represent a MMSSMMSS sequence. All changes in direction of rotation about the diagram occur at the corners of the square; otherwise, the only restriction is for the alternation of the dot and small circle positions. Here again, with a peak signal of $A/\sqrt{2}$, a minimum separation of $A$ between mark and space dots or mark and space circles is obtained. Thus the speed is doubled at a cost of 3 db more power, as in the case of quaternary phase modulation. Note that the individual pulse amplitudes are $A/2$, as for the two-phase case, but that the pulse spacing is halved. For vestigial sideband operation these pulses are sent serially, while for the quaternary phase case of Fig. 4 the pulses can be considered to be of amplitude $A/2$ sent two at a time.

### 2.4 *Frequency Modulation*

The binary rectangular wave frequency modulation case to be presented here is the ideal one where the bit rate is equal to the frequency shift between mark and space. For a continuous mark or space signal this results in the signal changing phase 180° between successive symbols. Again it is assumed that the signal is shaped to give a raised cosine pulse spectrum. Such an FM signal can be resolved into two components, a two-phase signal carrying the binary information and a quadrature component consisting of steady mark and space as indicated by the vector diagram of Fig. 7. This quadrature component can be considered to consist of alternating ±90° carrier pulses located between the 0° and 180° pulses carrying the information. The diagram for such an FM signal is shown in Fig. 7. A continuous mark condition (lower frequency) causes the signal to move around the circle clockwise. A continuous space causes a counterclockwise rotation. At the center of the symbols the signal is at either point A or B. A frequency transition causes the signal to swing out to one of the points "x" and reverse the

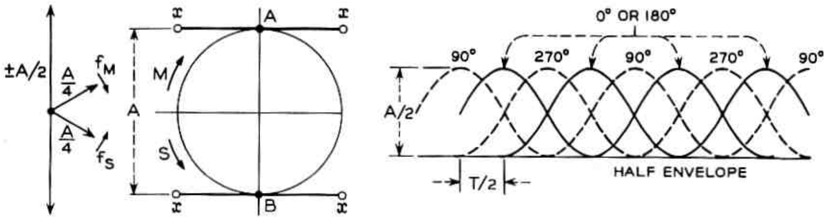


Fig. 7 — FM.

direction of rotation. For continuous reversals the signal swings back and forth through point A or B along a horizontal line. For such a sequence of reversals the phase swing is ±45°.

Although the steady mark and space frequency components which impart the horizontal component of motion in the diagram carry no information, they do permit the detection of the signal on a frequency basis. The mark and space conditions are indicated by the direction of rotation at points A and B. The quadrature component of the signal represents half of the total power. Consequently an FM signal requires twice the power of a two-phase signal to produce the same minimum separation of the points A and B. The two-phase component of the FM wave can be detected by a coherent carrier to determine whether the signal is at point A or B. It will be seen, however, that this leads to a polarity ambiguity because of the nature of the encoding. Reversals of either phase can be represented by the signal being at point A or at point B for successive symbols. A change from point A to point B indicates no transition of the information wave.

### 2.5 *Duobinary Frequency Modulation*[1]

The duobinary technique developed by Lender is a means of doubling the rate of sending binary information. The data are first differentially encoded so that a transition is made for a space symbol and no transition for a mark symbol. The resulting double-speed binary signal is then passed through a frequency shift channel of the type just described for ordinary binary operation with no change in frequency shift or channel shaping. The signal can change phase a maximum of ±90° during these half-length intervals. This results in both the in-phase and quadrature pulses carrying information. The diagram of Fig. 7 applies in part, but because of the double rate we are interested in more points of the pattern. For instance, for steady mark the signal moves around the circle in either direction and the receiver samples the signal not only at points A and B but also at points C and D as shown in Fig. 8. The occurrence

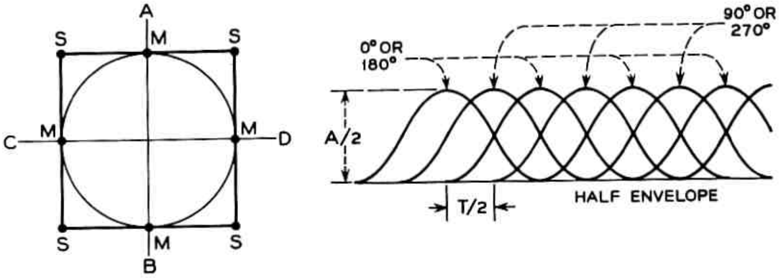


Fig. 8 — Duobinary FM.

of a space symbol causes a frequency transition and the signal leaves the circle and reverses direction at one of the points labeled S. If there are two successive space symbols causing two frequency transitions, the signal pauses at an S point for one symbol interval and then continues on in the same direction of rotation. An odd number of successive space symbols leads to a reversal of rotation while an even number does not. The signal can thus proceed around the circle clockwise or counterclockwise or pause at one of the S points. The rotating conditions represent the high- and low-frequency states while the pausing represents the midband frequency. When the signal is detected on a frequency basis, a three-level baseband output is obtained, with the outer levels representing mark and the center level space.

The complete duobinary diagram of Fig. 8 is seen to be the same as that of Fig. 6 for a vestigial sideband signal. This indicates that the two kinds of line signals are of the same form although the encoding is different. Experimental verification of this identity has been demonstrated by transmitting a vestigial sideband signal to an FM receiver and obtaining a three-level baseband signal such as received in duobinary FM. Fig. 9 shows a photograph of the received eye pattern.

## III. OSCILLOSCOPE PRESENTATION OF SIGNAL SPACE DIAGRAMS

Signal space diagrams of actual data signals can be displayed by coherently detecting both the in-phase and the quadrature components with respect to a midband reference frequency and applying them to the $X$ and $Y$ deflection circuits of an oscilloscope. Such an arrangement was constructed in the laboratory and used to obtain the signal pattern photographs shown in Fig. 10. Three kinds of signals are shown, (a) binary FM, (b) quaternary PM, (c) binary VSB. These were all voiceband data signals within a band centered near 1800 cps. Appropriate filters were used to shape the pulse spectrum closely to a raised cosine.

To appraise the possible value of such signal patterns as a measure of

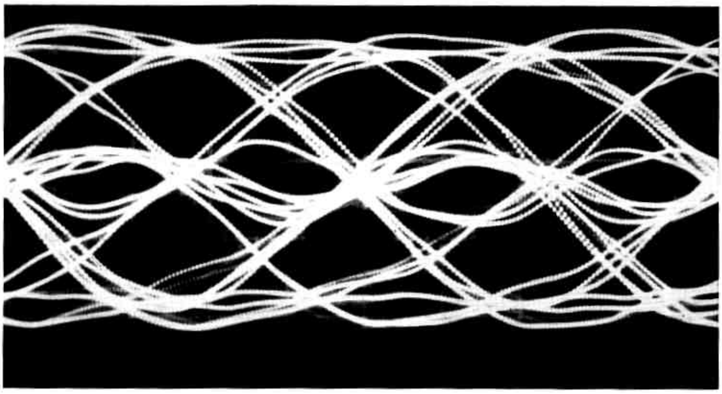


Fig. 9 — Oscilloscope picture of three-level eye obtained by receiving a VSB signal on a 202B FM receiver.

signal quality, the effects of amplitude slope and delay distortion were observed. Examples of the results are shown by the oscilloscope pictures of Fig. 11. The simulated line distortion characteristics which produced these patterns are given in Fig. 12. The effect of amplitude slope is readily apparent for FM and VSB, where portions of the transmitted sequence result in the signal resting at the high-loss end of the band. This accounts for the smaller inner circular portion of the patterns. In the case of PM the pattern is changed but not at the mid-symbol sampling points. The effect of high-end delay distortion shows up as a rotation of one part of the pattern with respect to others. This is readily seen in the PM examples, where the portion of the pattern formed by repeated phase advances is rotated with respect to the portion formed by repeated phase retardations.

## IV. CONCLUDING REMARKS

Signal space diagrams have been described for a number of commonly used data signals. These diagrams are useful in comparing data signals

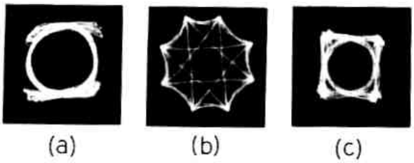


Fig. 10 — Oscilloscope pictures of signal space patterns for a 63-bit pseudo-random sequence: (a) binary FM, 1000 bits/sec, (b) quaternary PM, 2000 bits/sec, (c) binary VSB, 2400 bits/sec.

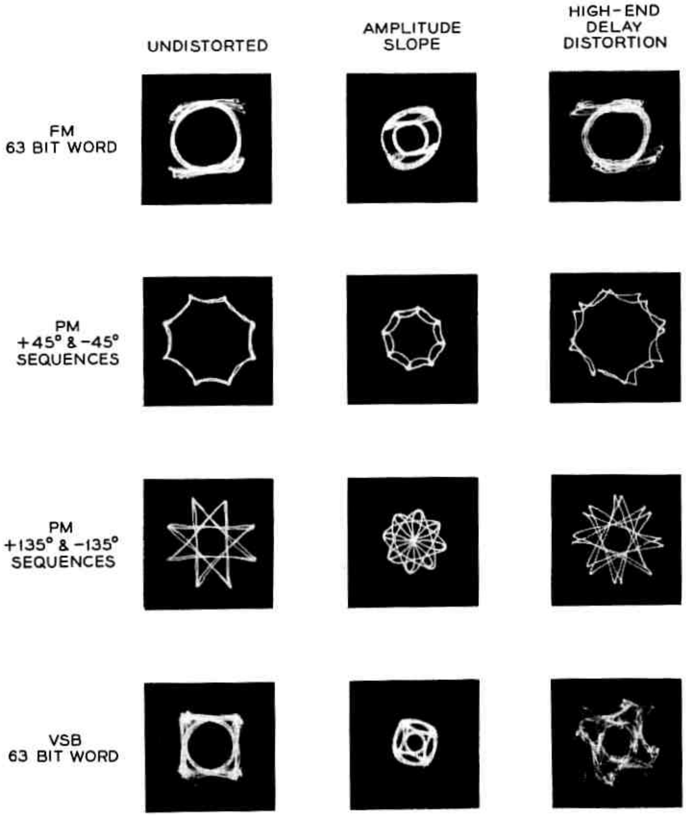


Fig. 11 — Oscilloscope pictures of signal space patterns showing effect of amplitude slope and envelope delay distortion.

on a common basis without regard to specific detection techniques. Similarities and differences are revealed which may not otherwise be apparent and various possible detection methods can be visualized. The margin against noise with ideal detection methods is also indicated by the spatial separation of the sampling points. The signal power is indicated by the pulse amplitude and repetition rate. For example, the foregoing diagrams illustrate that binary FM, quaternary PM and binary VSB signals all give the same margin against noise for a given transmitted power. The binary FM system, however, operates at half the speed of the other two for a given bandwidth. It has also been shown that a duobinary FM signal has the same pattern as a binary VSB signal. The relative simplicity afforded by FM detection of such a signal

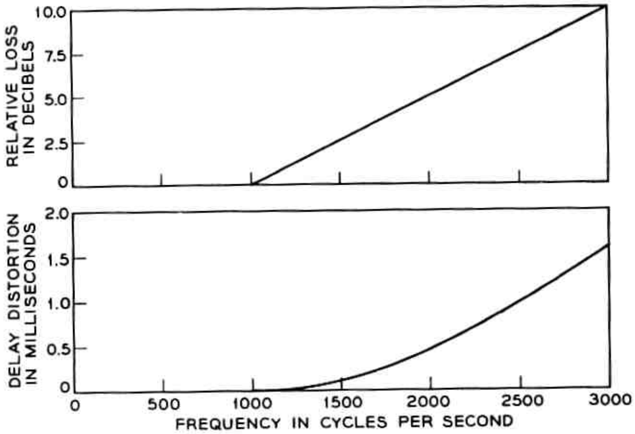


Fig. 12 — Amplitude and delay distortion characteristics used to distort the signal patterns shown in Fig. 11.

as against coherent detection is accomplished at a loss of approximately 6 db in margin against noise.

Considerable information about the nature of the channel characteristic is also indicated by the signal diagrams. The use of signal diagrams as an indication of signal quality is primarily limited, however, to the laboratory. The required synchronization with the midband frequency and symbol rate of the signals to be observed tends to make the method unsuitable for field measurements.

## V. ACKNOWLEDGMENTS

The construction of circuitry for resolving data signals into quadrature components and the photography of the oscilloscope patterns were carried out by John Grason under the direction of F. K. Becker.

## REFERENCE

1. Lender, A., The Duobinary Technique for High-Speed Data Transmission IEEE Trans. Comm. and Elect., **82,** May, 1963, pp. 214–218.

# Using Digit Statistics to Word-Frame PCM Signals

By J. R. GRAY and J. W. PAN

(Manuscript received June 5, 1964)

*Framing of PCM signals can be accomplished by statistical means. For signal samples whose probability distribution tends to be concentrated at the center of the coding range, the second digit of the Gray code generated has a probability of mostly 1's. This information can be used to frame PCM words. Three circuits are proposed that test this probability. Reliability and reframe time for each circuit are obtained either analytically or experimentally. The first circuit uses a pair of racing counters: one counts 0's in the second digit and the other 0's in the third digit of the Gray code. When the system is in-frame, the first counter seldom reaches full count before the second, whereas during out-of-frame either counter can reach full count first with equal probability. The second circuit uses a reversible counter which advances on a 0 and retards on a 1. When connected to the second digit of the Gray code, the preponderance of 1's will keep the counter at or near zero count; when connected to any other digit, where the probability of a 1 is at most 0.5, the counter will reach full count in a finite time. The third circuit uses an RC integrator in place of the reversible counter: each 0 of the second digit generates a pulse to charge the capacitor and each 1 permits the accumulated charge on the capacitor to decay. The action is similar to that of the reversible counter but is difficult to analyze. Experimental framing performance is given for this circuit.*

## I. INTRODUCTION

When a signal is transmitted by PCM, the receiver must be able to group the serial pulse train into code words before it can properly recover the original signal. This process is called "framing." It is also called "word synchronization," as distinguished from bit synchronization where the time base of the individual pulses is sought. When the pulse train contains several PCM signals multiplexed together, there is also the task of multiplex framing or frame synchronization whereby the individual channels must be identified. Word synchronization can be derived from frame synchronization if the words are always arranged in a definite order within a multiplex frame; otherwise, word synchronization is acquired independently. This article will consider only the problem of word synchronization, hereafter simply called "framing."

Framing is ordinarily accomplished by using supplementary framing pulses inserted among the information-bearing pulses at predetermined intervals. The receiver will then find these framing pulses by searching and testing for the unique pattern of these pulses. If the framing pulses are inserted between every word, a substantial loss of channel capacity will result; on the other hand, if framing pulses are inserted only occasionally, the PCM words will not be uniformly spaced, which is inconvenient for a sampled-data system. When the PCM signal contains known redundancies, it is possible to accomplish framing without the use of supplementary pulses. The signal is then said to be framed "statistically." The receiver now searches for the word grouping which will yield the expected statistics for the signal. A simple example of such a statistic is the intelligibility of voice. Voice transmitted by PCM is intelligible only when the PCM words are grouped correctly. Other criteria, easier to instrument than intelligibility, are available. Most signals have amplitude distributions other than the uniform distribution or have frequency spectra other than the flat spectrum. Both of these properties will be altered when framing is incorrect. One of the easiest statistics to measure is the average occurrence of 1's and 0's in the code words. Measurement of this statistic for the case of a linear coder operating on a Gaussian signal source will be the main theme of this article. The next section will elaborate on the digit probabilities, followed by descriptions and analyses of framing circuits which acquire framing by comparing the probabilities of 1's and 0's in the second digit of the Gray code.

## II. PROPERTIES OF THE GRAY CODE

If the amplitude of the signal before PCM encoding is centrally distributed — Gaussian, for example — and the Gray code is used to convert this signal into PCM, then the individual digits of each code word will not have equal probability of being either a 1 or a 0. This fact can be demonstrated by observing the Gray code assignments illustrated in Fig. 1. Because the signal amplitudes are centrally distributed, the center codes will be used more frequently than the codes at the extremes; the second digit, being a 1 for the center codes, will thus be dominated by 1's. It should be noted that this redundancy is the result of a linear coder

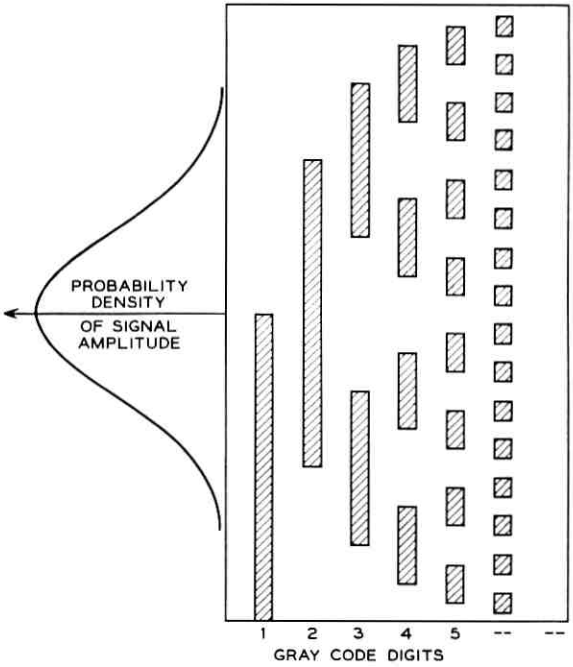


Fig. 1 — Gray code digit assignments.

operating on a Gaussian source. If a more efficient digitizer is used for this source, as for example (1) a nonlinear coder or (2) a linear coder followed by a digital processor to produce variable length codes or block codes, then this redundancy can be removed. The amount of redundancy in question is approximately one bit. Efficient coding would therefore exclude the use of statistical framing.

Fig. 2 illustrates the probabilities of 1's for all the digits; we can see that the probabilities of each digit being a 1 obey the following inequalities:

$$P(D_3 = 1) < P(D_4 = 1) < \cdots < P(D_1 = 1) < P(D_2 = 1) \quad (1)$$

or, equivalently, the probabilities of each digit being a 0 conform to

$$P(D_3 = 0) > P(D_4 = 0) > \cdots > P(D_1 = 0) > P(D_2 = 0). \quad (2)$$

Any out-of-frame condition is represented by a cyclic permutation of the digits so that one of the inequality signs in (1) will be reversed and similarly for (2). Any circuit which examines the validity of (1) or (2) is therefore a framing detector. A few such circuits will be listed here.

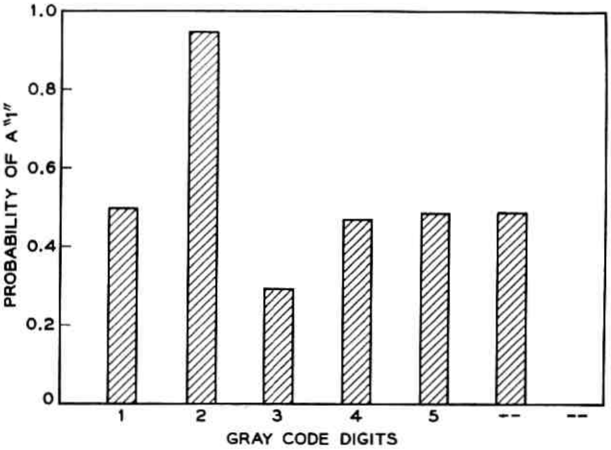


Fig. 2 — Gray code digit probabilities.

(1) Racing counters. In this scheme two counters are connected as shown in Fig. 3. When either counter reaches the full count of $N$, both counters are reset to zero. Now if the upper counter is connected in such a way that its count is advanced for every 0 in digit 2 and the lower counter is similarly connected for digit 3, then according to (2) the lower counter will reach full count and reset both counters most of the time. However, if the signal is out-of-frame, the counters will be actually counting the 0's of the digit pairs 3-4, 4-5, $\cdots$ or 1-2, and according to (2) the upper counter will now be able to reach full count and reset both counters much more frequently. The reset signal from the upper counter can thus be used as an out-of-frame signal. The probability of a false out-of-frame signal can be made small by increasing $N$, the size of the counters.

(2) Reversible counters. A single reversible counter, shown in Fig. 4,

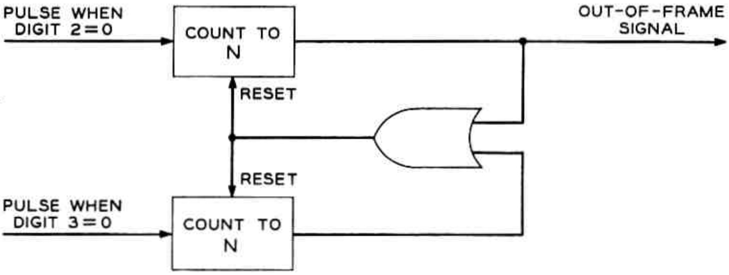


Fig. 3 — Racing counters.

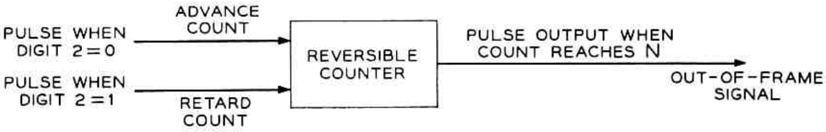


Fig. 4 — Reversible counter.

can also be used to detect the framing status. The count is increased by a 0 and reduced by a 1. When digit 2 is connected to this counter, the preponderance of 1's will keep the counter at or near the zero-count state and prevent it from reaching full count. When the receiver goes out of frame, this counter will be controlled by pulses of some other digit which, as can be seen from Fig. 2, has at least 50 per cent zeros; therefore full count will be reached within a finite time. Framing can be accomplished by searching for a word grouping such that the counter does not reach full count in a certain time interval.

(3) *RC* circuit. If a random pulse train is connected to an *RC* circuit, shown in Fig. 5, then the presence of a pulse will charge the capacitor and the absence of a pulse will permit the accumulated charge on the capacitor to discharge somewhat. The process is similar to that of the reversible counter, except that the charge and discharge rate is now a function of the accumulated charge. A threshold circuit monitoring the voltage on the capacitor can be used to indicate the framing status. A pulse train derived from the received signal such that each pulse indicates a 0 and each space indicates a 1 in the second digit of the Gray code is used as an input to the *RC* circuit. When the receiver is in frame, the pulse pattern at the input to the *RC* circuit will be sufficiently

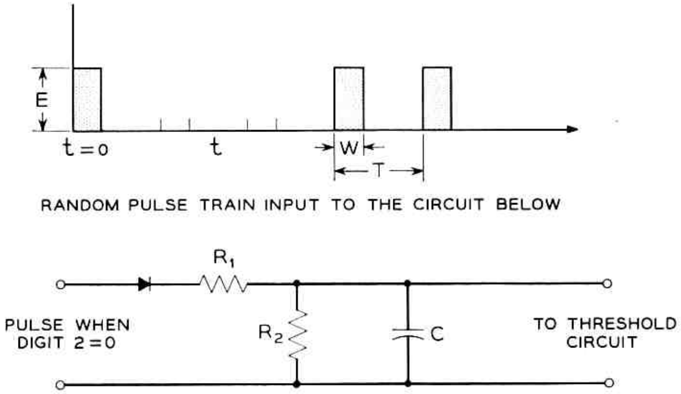


Fig. 5 — Framing with *RC* circuit.

sparse so that the accumulated charge will result in an output voltage that seldom builds up to the threshold. However, an out-of-frame condition will result in at least 50 per cent pulses present at the input, and the output of the *RC* circuit will reach threshold in a finite time.

## III. FRAMING CIRCUIT CHARACTERIZATION

Two figures of merit are commonly used to characterize framing circuit performance, (1) misframe rate and (2) reframe time. Misframe rate is measured in terms of the probability that the circuit will indicate an out-of-frame condition when in fact the receiver is in frame. Reframe time is characterized by the probability distribution of the time required for the receiver to achieve correct framing; this includes the time taken to detect the out-of-frame condition. In a conventional framing circuit, wherein a known framing pulse pattern is monitored, misframe rate and reframe time are sensitive only to the error rate of the transmission medium. Performance is degraded due to masking of the framing pulses by noise. With statistical framing, performance is more dependent on signal statistics. Let the probability of a 0 in digit 2 be 0.05 at the transmitter; with an error rate of 10 per cent, the probability of a 0 will increase to about 0.14, which is still different enough from 0.5 to keep the circuit in frame. The signal itself, of course, will hardly be usable at this error rate. On the other hand, a significant change in signal statistics at the transmitter may cause a collapse of framing. Care must therefore be exercised when the performance of statistical framing circuits is to be compared with that of conventional circuits.

To evaluate the misframe rate and the reframe time of the statistical framing circuits, the response of these circuits to random inputs must be determined. Unfortunately, the statistical properties of the transient response of analog circuits such as the *RC* circuit excited by a random signal have not yet been completely solved. Therefore analytical results for framing schemes using only digital counters will be presented here; even with these circuits the results are approximate.

An experimental approach is used to determine the performance of the framing scheme using *RC* circuits. The instrumentation proves to be rather simple and some results will be given.

## IV. ANALYSIS OF THE RACING COUNTERS

To lend some physical meaning to the analytical results, the analysis will be accompanied by numerical results for a typical application, namely, transmission of a mastergroup of telephone channels by PCM. A mastergroup carries 600 voice-grade channels frequency-multiplexed together, and its amplitude distribution is very close to Gaussian if the signal load is predominantly message service.[1] With normal busy hour loading the rms value of the signal is approximately ¼ of the system overload voltage. Under extreme conditions the rms may rise to ⅓ of the overload voltage. These figures will be used to calculate the performances of the framing circuits. A nominal sampling rate of $6 \times 10^6$ samples per second is assumed for the mastergroup. This rate will be used to translate misframe rate into misframe interval, the mean time between misframes.

We can consider the two racing counters as a sequential machine having $(N + 1)^2$ possible states. In Fig. 6 the $(N + 1)^2$ states are depicted in a square array $A$; each of its elements $a_{ij}$ represents a state where the upper counter has count $i$ and the lower counter $j$. From $a_{ij}$ transition is possible to 3 adjacent states $a_{i+1,j}$ , $a_{i+1,j+1}$ , or $a_{i,j+1}$ upon receiving as inputs 01, 00, or 10 respectively. In this notation the first digit represents the input to the upper counter and the second digit the input to the lower. Since the counters count only 0's, an input of 11 will not advance the counters and the state will remain at $a_{ij}$ . Starting from the initial state $a_{00}$ , the problems are (a) to find the probability of reaching the bottom row when digits 2 and 3 are connected to the counters (this yields the misframe rate) and (b) to find the probability distribution of the time required to reach either the bottom row or the right-hand column when other pairs of digits are connected to the counters; this leads to the distribution of reframe time when the resulting distributions are convolved.

A convenient technique for finding these probabilities is to use signal

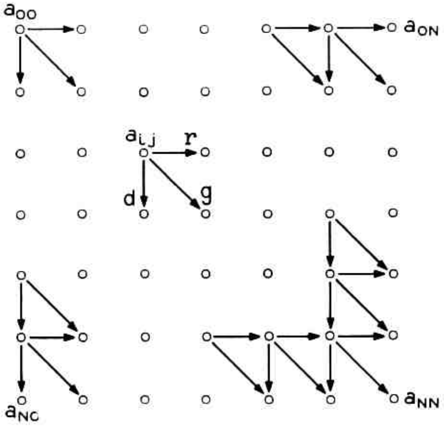


Fig. 6 — State diagram for racing counters.

flow graphs.[2] Using $x = e^{-s}$ as the time delay operator, the transitions indicated in Fig. 6 are as follows:

$$\text{down } d = \frac{xP(01)}{1 - xP(11)}$$
$$\text{diagonal } g = \frac{xP(00)}{1 - xP(11)} \tag{3}$$
$$\text{and to the right } r = \frac{xP(10)}{1 - xP(11)}.$$

The denominator $[1 - xP(11)]$ is due to self-loops at each state when neither counter advances. In principle, this flow graph can be solved for the transmission from the initial state to either the bottom row or the right-hand column as rational functions of the delay operator $x$. From these rational functions the total probability of reaching the bottom row can be calculated by letting $x = 1$, and the probability distribution of the waiting time can be obtained by a power series expansion of the rational functions. However, in a practical situation with counters counting up to 16, the calculations become extremely involved, and even with 20 decimal digits round-off errors become excessive. Approximations are therefore used to estimate the misframe rate and the framing time.

To calculate the average misframe rate, the substitution $x = 1$ can be made before solving the flow graph of Fig. 6. This reduces complexity considerably and one can calculate the probability of reaching the bottom row before the right-hand column. Information about time delay is lost and must be estimated independently.

The flow graph can be solved by observing that

$$Q(i,j) = dQ(i-1,j) + gQ(i-1,j-1) + rQ(i,j-1) \tag{4}$$

for

$$1 \leqq i \leqq N-1 \quad \text{and} \quad 1 \leqq j \leqq N-1$$

where $Q(i,j)$ is the probability that the state $a_{ij}$ is reached at any time starting from $a_{00}$. The $d$, $g$, and $r$ are now numerical quantities calculated from (3) with $x = 1$. The above iteration formula is valid for all states except the border states of the array $A$. To complete the picture we have

$$Q(00) = 1 \tag{5}$$

since $a_{00}$ is the initial state, and going straight down

$$Q(i,0) = dQ(i-1,0) \qquad 1 \leqq i \leqq N. \tag{6}$$

To the right we have

$$Q(0,j) = rQ(0,j-1) \qquad 1 \leqq j \leqq N \tag{7}$$

For the bottom row we have

$$Q(N,j) = dQ(N-1,j) + gQ(N-1,j-1) \qquad 1 \leqq j \leqq N-1 \tag{8}$$

and the rightmost column

$$Q(i,N) = gQ(i-1,N-1) + rQ(i,N-1) \qquad 1 \leqq i \leqq N-1 \tag{9}$$

and, finally, the lower right state has probability

$$Q(N,N) = gQ(N-1,N-1) \tag{10}$$

since it can be reached only by way of $a_{N-1,N-1}$. The special treatment given the bottom row and right-hand column is necessary because they are the end states; from here we start anew at $a_{00}$.

The probability of reaching the bottom row is the sum

$$U = \sum_{j=0}^{N} Q(N,j) \tag{11}$$

which is the probability of an output pulse from the upper counter before the lower counter reaches count $N$. This is the probability of a false out-of-frame signal when digits 2 and 3 are connected to the counters. The recurrence formulas are valid for signals that are independent with respect to the past, so that $d$, $g$, and $r$ are the same for all states. Statistical dependence of the two digit inputs is considered in their joint probabilities. This iterative procedure has been carried out, and some numerical results are presented below.

Assuming a Gaussian distributed input signal the joint probabilities of digits 2 and 3 can be determined for normal loading with an rms input at ¼ of the system overload and for extreme loading with an rms input at ⅓ of the overload. The various probabilities are shown in Table I:

Table I — Probabilities of Digits 2 and 3

| | 01 | 00 | 10 | 11 |
|---|---|---|---|---|
| RMS ¼ overload | 0.0428 | 0.0026 | 0.6826 | 0.2720 |
| RMS ⅓ overload | 0.1092 | 0.0244 | 0.5468 | 0.3196 |

substituting these numbers into (3), we have for $x = 1$ the transition probabilities shown in Table II.

TABLE II — TRANSITION PROBABILITIES

| Transition | Down<br>$d$ | Diagonal<br>$g$ | To the Right<br>$r$ |
|---|---|---|---|
| RMS ¼ overload | 0.0588 | 0.0037 | 0.9375 |
| RMS ⅓ overload | 0.1605 | 0.0359 | 0.8036 |

The strong tendency to go to the right is quite evident here. The probabilities of reaching the bottom row before the right-hand column can be calculated from these data using the iteration formulas developed above. To translate these probabilities into mean time between misframes we proceed as follows. When the signal is in-frame, the lower counter almost always attains full count before the upper. For counters of size $N$, the lower counter resets both counters on the average of every $N/p$ PCM words, where $p$ is the probability of a 0 in digit 3. The mean time between misframes is then $N/pU$. The results are shown graphically in Fig. 7 for various counter sizes. At normal loading and $N = 16$, the mean time between misframes is $1.2 \times 10^{12}$ words which, at a sampling rate of $6 \times 10^6$ per second, amounts to $2 \times 10^5$ seconds or a little more than 2 days. When the rms signal is increased to ⅓ of overload, this mean time deteriorates rapidly to fractions of a second, so that the counter size has to be more than 32 to insure adequate reliability under severe overload conditions.

To complete the picture on the racing counters, the framing time will be estimated. During search for the correct framing we observe that the

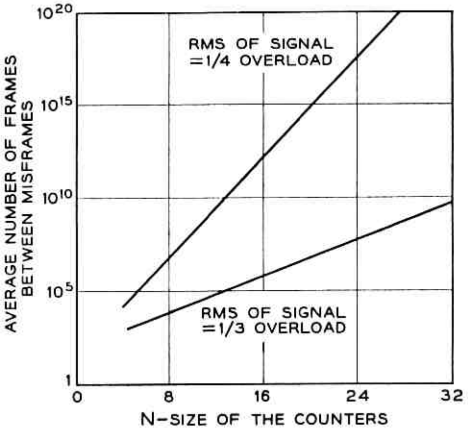


Fig. 7 — Reliability of the racing counters.

upper counter will be advanced, with 0's occurring with probability at least ½, and that it will be able to reach full count without first being reset by the lower with probability at least ½. Thus with $M$ digits in each PCM word and assuming the worst case of searching through all $M - 1$ positions, the counters will be reset on the average of $2(M - 1)$ times. Each reset requires on the average of $2N$ words to either the upper or lower counter. A conservative estimate of the average framing time for the worst case is therefore $4N(M - 1)$ words.

As mentioned earlier, the exact distribution of the framing time is difficult to obtain; however, the variance of this distribution can be estimated. The framing time distribution can be considered as a compound distribution, where the number of times $n$ either counter reaches full count during framing is governed by one distribution and the waiting time $t$ for each reset is governed by another distribution. It is known that such a distribution has mean $E(n)E(t)$ and Variance $E(n)\text{Var}(t) + \text{Var}\ (n)E^2(t)$.[3] The distribution of the number of times either counter reaches full count before the upper counter reaches full count $M - 1$ times is governed by the negative binomial distribution.* With the upper counter having probability ½ of reaching full count, $n$ has average $2(M - 1)$ as mentioned before and variance $2(M - 1)$. The waiting time for each reset is similarly governed by the negative binomial distribution. With probability ½ of receiving a 0, the waiting time $t$ has mean $2N$ and variance $2N$. The variance of the framing time is therefore $2(M - 1)(2N) + 2(M - 1)(2N)^2$; for large $N$ this is approximately $8(M - 1)N^2$.

For a 9-digit PCM system $M = 9$, and if we use $N = 32$, the average framing time for a sampling rate $F_s = 6 \times 10^6$ per second is

$$\frac{8(M - 1)N}{F_s} = \frac{4 \times 8 \times 32}{6 \times 10^6} = 171\ \mu\text{sec}$$

the standard deviation is

$$\frac{[8(M - 1)N^2]^{\frac{1}{2}}}{F_s} = \frac{(8 \times 8)^{\frac{1}{2}} \times 32}{6 \times 10^6} = 43\ \mu\text{sec}.$$

Since the distribution is the result of many convolutions, it can be approximated by a normal distribution; with this assumption we can use three standard deviations as the confidence limit and estimate the maximum framing time as 300 μsec. During out-of-frame conditions the upper

* See Ref. 3, p. 253. Actually the negative binomial distribution governs the number of times the lower counter reaches full count. This average is $M - 1$; the total average waiting time is therefore $(M - 1) + (M - 1) = 2(M - 1)$.

counter actually receives 0's with probability greater than ½, so that the estimates are conservative.

## V. ANALYSIS OF THE REVERSIBLE COUNTER

The use of a reversible counter allows greater reliability without resorting to large-capacity counters as is necessary for the racing counters. The analysis is also simpler, since only one counter is involved. The flow graph for a reversible counter is shown in Fig. 8. The probability of a 0 which increases the count is $p$, and $q = 1 - p$ is the probability of a 1 which decreases the count. The count cannot go below zero. The gain of the graph for any counter size $N$ can be obtained by standard techniques. The result can be expressed conveniently in the form of a recursion formula for the denominator polynomial

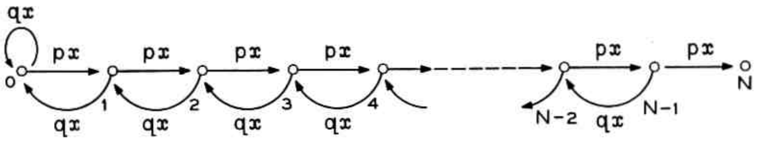


Fig. 8 — Flow graph for reversible counter.

$$D_N(x) = D_{N-1}(x) - pqx^2 D_{N-2}(x)$$

where

$$D_0(x) = 1 \qquad \text{and} \qquad D_1(x) = 1 - qx.$$

The numerator is simply $N_N(x) = p^N x^N$. Some representative results are

$$Q_4(x) = \frac{p^4 x^4}{1 - qx - 3pqx^2 + 2pq^2x^3 + p^2q^2x^4} \tag{12}$$

and

$$Q_8(x) = \frac{p^8 x^8}{\begin{array}{r}1 - qx - 7pqx^2 + 6pq^2x^3 + 15p^2q^2x^4 \\ - 10p^2q^3x^5 - 10p^3q^3x^6 + 4p^3q^4x^7 + p^4q^4x^8\end{array}}. \tag{13}$$

The average time between misframes can be determined from the above by differentiation. Thus*

$$T_{\text{av}} = Q_N'(1) \tag{14}$$

* See Ref. 4.

when $p$ and $q$ are for the second digit of the Gray code. The results for various counter sizes and for system overload at 4 and 3 times rms are shown in Fig. 9. It is seen that with a counter of size 16 and worst-case loading the misframe interval is still sufficiently long, 1000 hours at a 6-mc sampling rate.

One disadvantage of using the reversible counter is the slow reframing process. When the receiver is out of frame the counter can be assumed to receive 1's and 0's with equal probability. Using formulas developed above for $N = 16$ but substituting 0.5 for $p$ and $q$, one obtains an average of 272 words to reach full count. For a 9-digit PCM system sampled at 6 mc, this amounts to 360 $\mu$sec for the average framing time. To shorten the framing time a dual-mode scheme applied frequently in conventional framing circuits can be used. The scheme is described in more detail below.

The framing circuit is designed to have two modes of operation. In the in-frame mode, the counter size is set at 16 for maximum reliability; once the out-of-frame signal is received the counter size is reduced to 8 to secure fast framing. The logic is depicted in Fig. 10.

The flip-flop determines the mode of operation. When in frame, the flip-flop is reset and the counter must reach count 16 excess 0's over 1's of the second-digit Gray code. When the system goes out of frame, the probabilities of 1's and 0's are equal, and an output from the $N = 16$ lead of the binary counter chain sets the flip-flop to the out-of-frame mode. In this mode the output from the $N = 8$ lead of the binary chain is used. At the same time a timer is turned on to reset the flip-flop after

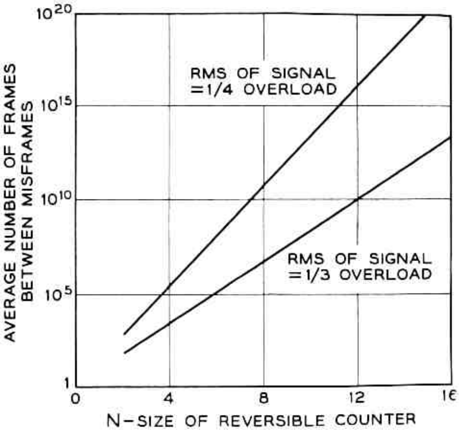


Fig. 9 — Reliability of the reversible counter.

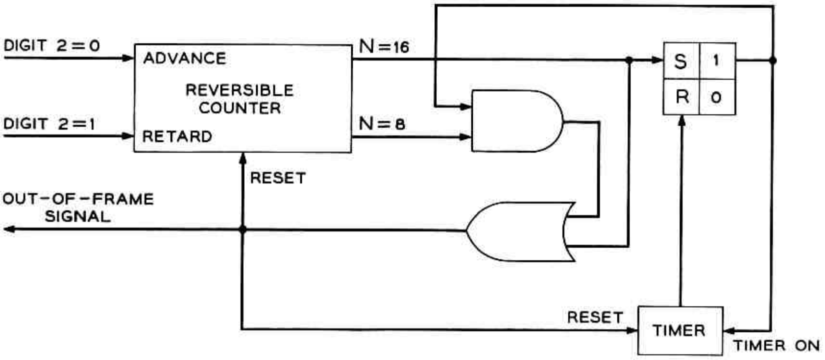


Fig. 10 — Dual-mode reversible counter framer.

a certain elapsed time. This elapsed time is selected such that it is longer than the maximum time required to get an output from the $N = 8$ lead when the system is searching but shorter than the minimum time required to get an output from this same stage when the correct frame is found. "Maximum" and "minimum" are used here in a probabilistic sense to be defined later. Thus during recovery the timer is reset before it reaches the preset time, thereby preventing the flip-flop from resetting back to the in-frame mode. When the system cycles back into frame, the timer will return the system to the in-frame mode. Each time a signal appears at the counter output, the framing counter is inhibited one time slot in order to examine the next bit position; the reversible counter is also reset automatically to zero. With the proper preset time, the system is almost always prevented from cycling past the true in-frame position.

To estimate the framing time for this scheme, we again use the example of a 9-digit PCM system sampled at $6 \times 10^6$ per second. For the worst case of searching through all 9 digits the average framing time is given by

$$T_F = Q_{16}'(1) + 7Q_8'(1). \tag{15}$$

The effect of incorrect decisions by the timer which cause recycling is ignored here. The first term corresponds to detection and the second term corresponds to the search through the next 7 positions. The time spent in verifying that the last position is the correct one is not included, because the system will already be in frame. The above equation is evaluated for $p = q = \frac{1}{2}$ and yields the worst-case average framing time of 130 $\mu$sec. This estimate is again conservative, since 0's occur with probability greater than $\frac{1}{2}$ in some digits of the Gray code.

The exact distribution of the framing time for the worst case may be determined by expanding $Q_{16}(x)Q_8^7(x)$ in a power series. Again, this is difficult to do accurately. To get around this problem an approximation to the inverse transform of $Q_8(x)$ is determined by noting that the decay in the tail of the distribution is dependent mainly on the singularity of $Q_8(x)$ closest to the unit circle (1.01728 in this case). On this basis the inverse transform is approximately

$$q_8(k) = \frac{1.6986 \times 10^{-2}}{(1.01728)^{k-8}} \qquad \text{for} \qquad k \geqq 8$$

and

$$q_8(k) = 0 \qquad 0 \leqq k < 8$$

where $1.6986 \times 10^{-2}$ is selected so that

$$\sum_{k=8}^{\infty} q_8(k) = 1.$$

Using the result and returning to the $x$ domain

$$Q_8(x) \approx \frac{1.6986 \times 10^{-2} x^8}{\left(1 - \dfrac{x}{1.01728}\right)}. \tag{16}$$

We now make the further approximation of replacing $Q_{16}(x)$ by $Q_8(x)$ in the product mentioned above. We can therefore deal with the simple result given by (16) raised to the 8th power. On this basis a somewhat optimistic expression for the distribution of the framing time can be readily obtained:

$$p(n) = \frac{(1.6986 \times 10^{-2})^8 (n - 57)!}{7!(n - 64)!(1.01728)^{n-64}} \qquad \text{for} \qquad n \geqq 64 \tag{17}$$
$$p(n) = 0 \qquad 0 \leqq n < 64.$$

The upper tail of $p(n)$ is shown in Fig. 11.

Taking the $10^{-3}$ point as the confidence limit and multiplying by the sampling period, we get 200 μsec as the maximum framing time. Since the framing process is dominated by the $Q_8^7(x)$ term, the error introduced by the substitution of $Q_8(x)$ for the $Q_{16}(x)$ term should not be significant.

Finally, we note that an optimum time must be chosen for the timer in Fig. 10 to reset the flip-flop back to the in-frame mode. Selection of this time is based on the distributions of waiting times for an output from the $N = 8$ lead of the counter, first under the out-of-frame condi-

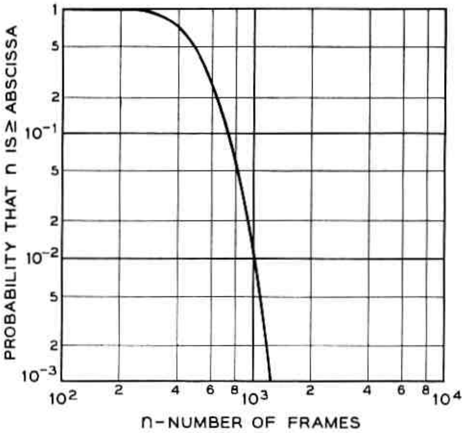


Fig. 11 — Distribution of framing time for reversible counter.

tion and second under the in-frame condition. Summing the first two columns of Table I we obtain 0.13 as the probability of a 0 for the second digit when the rms input is at ⅓ of the system overload. For the other digits a probability of 0.5 is assumed. Expanding $Q_8(x)$ in a power series when $p = 0.5$ and when $p = 0.13$ yields the desired result. This is plotted in Fig. 12. If the time is chosen to be 560 frames, the framing detector will be in the wrong operating mode only 0.01 per cent of the time,

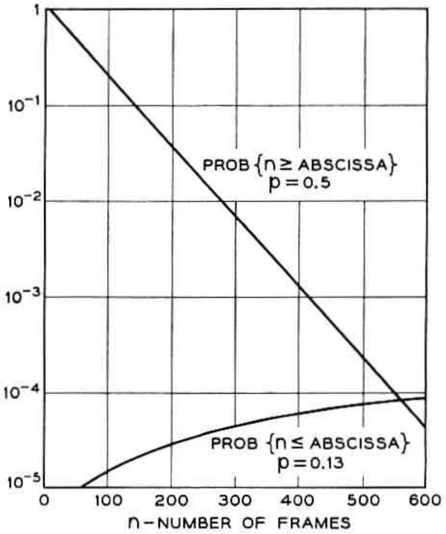


Fig. 12 — Selection of optimum time for the timer.

which means that the framer will seldom cycle past the true frame position during the framing process.

We note for future reference that the distribution of the waiting time in Fig. 12 for $p = 0.5$ is a straight line on semilog paper, which indicates that it has an exponential tail.

## VI. MEASURED FRAMING PERFORMANCE FOR THE *RC* CIRCUIT

We introduce this section by defining the problem. Illustrated in Fig. 5 is a typical input to the *RC* circuit, a random pulse train

$$x(t) = \sum_{n=0}^{\infty} a_n g(t - nT) \tag{18}$$

where $a_n$ is a sequence of independent random variables assuming values 1 or 0 with probabilities $p$ and $(1 - p)$, and $g(t)$ is a rectangular pulse of height $E$ and width $w$. When this pulse train is applied to the circuit of Fig. 5, the capacitor will charge when a pulse is present and discharge otherwise. The charging time constant is

$$\tau_c = \frac{R_1 R_2}{R_1 + R_2} C \tag{19}$$

and the discharge time constant is

$$\tau_d = R_2 C. \tag{20}$$

It is also convenient to refer to the attenuation constant

$$K = \frac{R_2}{R_1 + R_2} = 1 - \frac{\tau_c}{\tau_d}. \tag{21}$$

We are interested in the transient response of the circuit $y(t)$, particularly at times $t = w, t = T + w, \cdots, t = MT + w$ because they are the local maxima. We can proceed step by step:

$$y(w) = a_0 KE[1 - \exp(-w/\tau_c)] \tag{22}$$

$$y(T) = a_0 KE[1 - \exp(-w/\tau_c)] \exp[-(T - w)/\tau_d]; \tag{23}$$

at $t = T + w$, the charge due to $a_1$ is added, the charge due to $a_0$ decays further with a time constant of either $\tau_c$ or $\tau_d$ depending on the value of $a_1$

$$\begin{aligned} y(T + w) = {} & a_1 KE[1 - \exp(-w/\tau_c)] \\ & + \{a_0 KE[1 - \exp(-w/\tau_c)] \exp[-(T - w)\tau_d]\} \quad (24) \\ & [a_1 \exp(-w/\tau_c) + (1 - a_1) \exp(-w/\tau_d)]; \end{aligned}$$

in general

$$y(MT + w) = KE[1 - \exp(-w/\tau_c)] \sum_{n=0}^{M} a_n \exp[-(M - n)(T - w)/\tau_d] \prod_{m=n+1}^{M} [a_m \exp(-w/\tau_c) + (1 - a_m) \exp(-w/\tau_d)]. \tag{25}$$

The framing performance of this circuit is related to the probability distribution of the first time that the output of the circuit exceeds a certain threshold. It is the distribution of the smallest $M$ such that

$$y(MT + w) > \text{threshold}. \tag{26}$$

To find the distribution analytically from (25) appears difficult. Some simplification can be obtained by assuming that the widths of the pulses are small or by assuming that the charge and discharge time constants are the same. Under either of these conditions the product in (25) disappears and the output is essentially of the form

$$z = \sum_{n=0}^{M} a_n \beta^{(M-n)} \qquad 0 < \beta < 1. \tag{27}$$

The behavior of the random variable $z$ when $M \rightarrow \infty$ has received some attention,[4] but the distribution of the first passage time of $z$ with respect to some threshold is still difficult to obtain.

Here the experimental approach is taken; the circuit used is depicted in Fig. 13. The input is derived from an analog-to-digital converter with a Gaussian signal as input. The output of this converter is in Gray code. By adjusting the level of the input signal and by selecting the various digits of the Gray code, a pulse train with any desired pulse density may be obtained. The digital timer measures the waiting time; it is started at the closing of the input switch and stopped by the threshold circuit. The threshold circuit also opens the input switch and signals the recorder to write the timer output on tape. A delay circuit resets the digital timer and initiates the next cycle of measurement after the $RC$ circuit has returned to the rest condition. Each timing and recording operation takes about one msec; about a million measurements were made and recorded in a matter of minutes. A simple computer program reads the data and compiles the cumulative distribution of these data as well as the mean and standard deviation.

Some qualitative results concerning the effects of the various parameters will be given below. First, for all of the combinations of the parame-

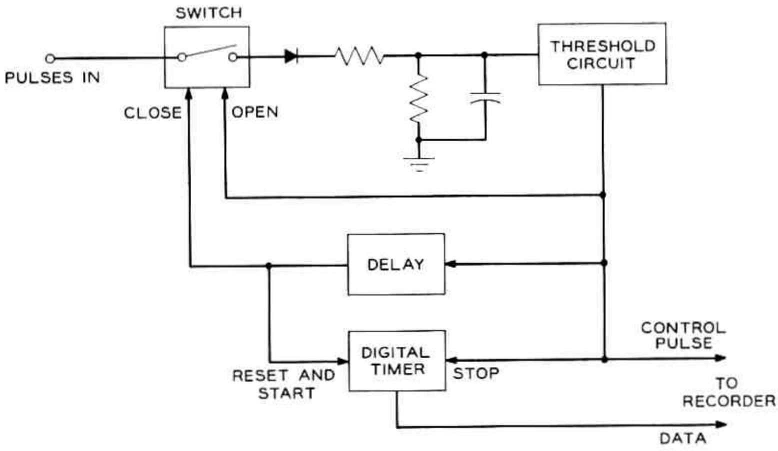


Fig. 13 — Measuring distribution of first passage time.

ters chosen, the measured distributions tend to have an exponential tail; they plot as straight lines on semilog paper (see, for example, Fig. 15 below). An intuitive argument can be given for this result. If we suppose that the threshold is set very low compared to the average output of the circuit, at voltages below this threshold the circuit acts more like an integrator than an *RC* circuit because it charges almost linearly and discharges very little between pulses. The distribution should therefore be similar to the distribution of the waiting times for the $n$th success in a sequence of Bernoulli trials, which has an exponential tail. Now we suppose that the threshold is set high compared to the average output of the circuit. Near this threshold, the circuit decays rapidly between pulses, so that a succession of many pulses in a row is necessary to drive the circuit over the threshold. The problem is now similar to the first occurrence of $n$ consecutive successes in a sequence of Bernoulli trials, which again has an exponential tail. Finally, we can suppose that the threshold is set about equal to the average output of the circuit when the probability of a pulse at the input is 0.5. Near this threshold the decay due to an absence of a pulse is about equal to the charge contributed by a presence of an input pulse. The circuit therefore behaves much like a reversible counter in this region. In the previous section this has been shown to have an exponential tail. All of these arguments are of course approximate, but, lacking a complete theory, they serve to provide some insight. Knowledge that the distribution of the waiting time has an exponential tail enables us to use the techniques developed for the reversible counter to estimate the framing time distribution of this circuit.

The second qualitative result is that the measured distributions of the first passage time for the various circuits are very much the same as long as their composite time constants and relative threshold settings are the same. By "composite time constant" is meant the time required for the output to reach $(1 - e^{-1})$ of the maximum output when all pulses are present at the input. By "relative threshold" is meant the threshold as a fraction of the aforementioned maximum output. The different situations are illustrated in Fig. 14.

The composite time constant and the maximum output can be computed from (25), setting all $a_n$'s to 1.

$$y(MT + w) = KE[1 - \exp(-w/\tau_c)] \sum_{n=0}^{M} \exp - \left[ (M - n) \left( \frac{T - w}{\tau_d} + \frac{w}{\tau_c} \right) \right]. \quad (28)$$

Letting $M$ approach infinity we obtain the maximum output

$$y_{\max} = KE \frac{[1 - \exp(-w/\tau_c)]}{\left[ 1 - \exp - \left( \frac{T - w}{\tau_d} + \frac{w}{\tau_c} \right) \right]}. \quad (29)$$

The expression inside the summation in (28) can be rewritten as

$$\exp - \left[ T(M - n) \left( \frac{1 - w'}{\tau_d} + \frac{w'}{\tau_c} \right) \right] \quad (30)$$

where $w' = T/w$, the duty cycle of the pulses. From this we can see that the composite time constant is

$$\left( \frac{1 - w'}{\tau_d} + \frac{w'}{\tau_c} \right)^{-1}. \quad (31)$$

The third qualitative result is the following. For circuits and threshold settings such that with equal probability of pulses and spaces at the input the distributions of the first passage time are the same, the average first passage time for low probability of input pulses is longer when the relative threshold is higher. Relative threshold is defined as above. This result can be explained by using arguments similar to the first result. At low threshold settings, the circuit acts as an accumulator so that the average first passage time is inversely proportional to the average pulse density. On the other hand, for high threshold settings, the first passage time depends on the occurrence of many consecutive pulses; the probability of this occurrence decreases exponentially with the average pulse density. This result is directly applicable to the framing

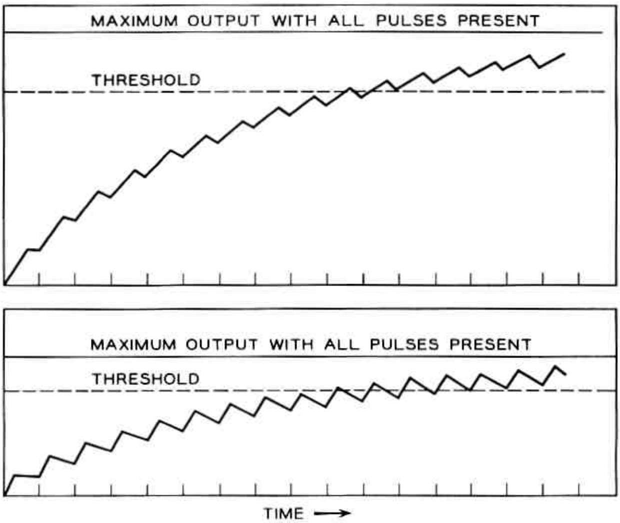

MAXIMUM OUTPUT WITH ALL PULSES PRESENT
THRESHOLD
MAXIMUM OUTPUT WITH ALL PULSES PRESENT
THRESHOLD
TIME ⟶

Fig. 14 — Two situations depicting different parameter settings but with substantially the same distribution of the waiting time to first passage of the threshold voltage.

problem. For the *RC* circuit, the dual-mode operation controlled by a timer used for the reversible counter is not necessary. With appropriate choice of circuit parameters and threshold, one can achieve fast framing and low misframe rate at the same time. To what extent the threshold can be adjusted to improve framing performance depends on the stability of the circuit. When the threshold is set near the level corresponding to all pulses present, a small drift in any of the parameters will cause a large change in reliability.

The framing performance of a typical *RC* circuit will be given here. Again we assume a 9-digit PCM system with 6-mc sampling rate. The parameters are as follows:

$$\begin{aligned}\text{pulse width} &= 50 \text{ per cent duty cycle}\\ \text{charging time constant} &= 0.44\ \mu\text{sec}\\ \text{discharge time constant} &= 1.2\ \mu\text{sec}\\ \text{composite time constant} &= 0.64\ \mu\text{sec}.\end{aligned}$$

With the probability of a pulse set at $\frac{1}{2}$, the variation of the distribution of the waiting time with threshold setting is illustrated in Fig. 15. The variation of the misframe interval and average framing time with threshold setting is illustrated in Fig. 16. If the threshold is chosen such that the misframe interval is $10^5$ seconds (about one day), the average

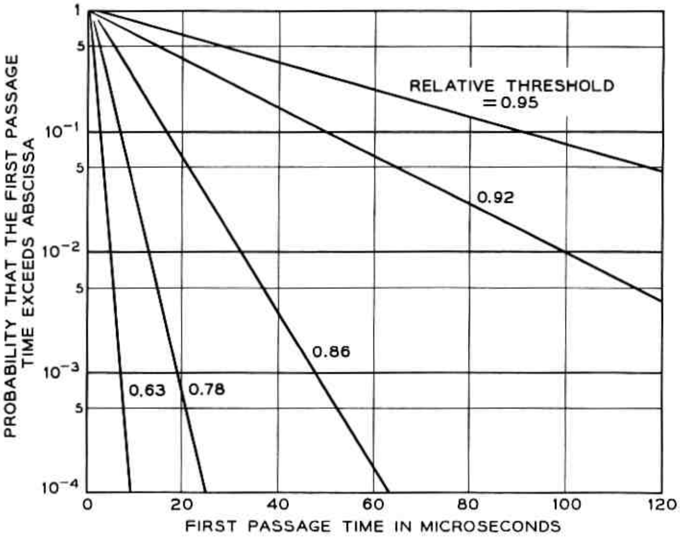


Fig. 15 — Distribution of the first passage time.

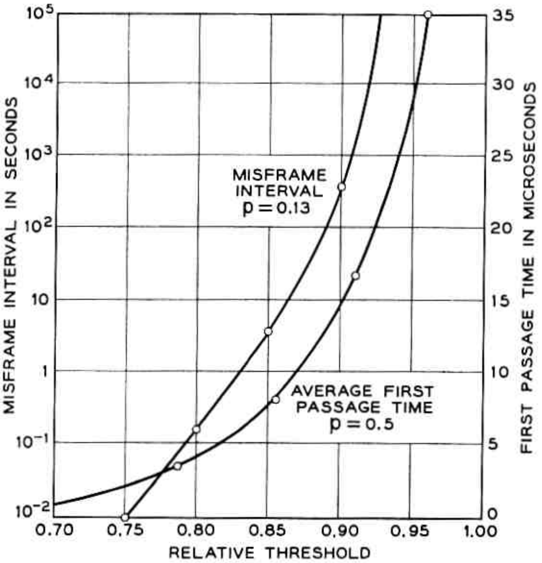


Fig. 16 — Performance variation of the *RC* framer with threshold settings.

first passage time is about 20 μsec. This yields an average framing time of 160 μsec if 8 positions are to be cycled through. Using the results of the reversible counter as a guide, the maximum framing time with 99.9 per cent confidence is about 250 μsec.

## VII. SUMMARY

This paper has considered the possibility of framing a PCM signal by utilizing the statistics of the code digits. Three schemes for testing digit statistics have been proposed and their performances analyzed or measured. Statistical framing is shown to be feasible and effective whenever the signal statistics satisfies certain weak conditions.

The authors wish to acknowledge the help of H. H. Henning and F. P. Rusin for design and construction of equipment used and assistance given in obtaining the experimental results. The encouragement of M. R. Aaron is also appreciated.

## REFERENCES

1. Holbrook, B. D., and Dixon, J. T., Load Rating Theory of Multichannel Amplifiers, B.S.T.J., **18,** Oct., 1939, p. 645.
2. Huggins, W. H., Signal Flow Graphs and Random Signals, Proc. I.R.E., **47,** Jan., 1957, pp. 74–86.
3. Feller, W., *An Introduction to Probability Theory and Its Applications,* John Wiley & Sons, New York, 1957, pp. 268–277.
4. Aaron, M. R., and Gray, J. R., Probability Distributions for the Phase Jitter in Self-Timed Reconstructive Repeaters, B.S.T.J., **41,** Mar., 1962, p. 503.

# Prolate Spheroidal Wave Functions, Fourier Analysis and Uncertainty —IV: Extensions to Many Dimensions; Generalized Prolate Spheroidal Functions

By DAVID SLEPIAN

(Manuscript received June 12, 1964)

*In two earlier papers* in this series, the extent to which a square-integrable function and its Fourier transform can be simultaneously concentrated in their respective domains was considered in detail. The present paper generalizes much of that work to functions of many variables.*

*In treating the case of functions of two variables whose Fourier transforms vanish outside a circle in the two-dimensional frequency plane, we are led to consider the integral equation*

$$\gamma\varphi(x) = \int_0^1 J_N(cxy)\sqrt{cxy}\,\varphi(y)dy. \qquad (i)$$

*It is shown that the solutions are also the bounded eigenfunctions of the differential equation*

$$(1 - x^2)\frac{d^2\varphi}{dx^2} - 2x\frac{d\varphi}{dx} + \left(\chi - c^2x^2 + \frac{\frac{1}{4} - N^2}{x^2}\right)\varphi = 0, \qquad (ii)$$

*a generalization of the equation for the prolate spheroidal wave functions. The functions $\varphi$ (called "generalized prolate spheroidal functions") and the eigenvalues of both (i) and (ii) are studied in detail here, and both analytic and numerical results are presented.*

*Other results include a general perturbation scheme for differential equations and the reduction to two dimensions of the case of functions of $D > 2$ variables restricted in frequency to the $D$ sphere.*

* See Refs. 1 and 2.

## I. INTRODUCTION

In two earlier papers[1,2] in this series, the extent to which a square-integrable function and its Fourier transform can be simultaneously concentrated was considered in detail. In that analysis, the eigenfunctions and eigenvalues of the finite Fourier transform played a key role. These functions, defined for $| x | \leqq 1$ by the integral equation

$$\alpha_j \psi_j(x) = \int_{-1}^{1} e^{icxy} \psi_j(y) dy, \tag{1}$$

can be continued analytically throughout the complex plane. They possess a number of special properties that make them most useful for the study of bandlimited functions. The functions are complete in the class of bandlimited functions; they are orthogonal in $(-1,1)$ and also in $(-\infty, \infty)$; the $\psi_j$ are also the eigenfunctions of the integral equation

$$\begin{aligned} \lambda \psi(x) &= \int_{-1}^{1} \frac{\sin c(x-y)}{\pi(x-y)} \psi(y) dy \\ \lambda &= \frac{c}{2\pi} | \alpha |^2 \end{aligned} \tag{2}$$

whose kernel is positive definite; $\psi_o$, the eigenfunction of (2) belonging to the largest eigenvalue, is in an appropriate sense most concentrated among bandlimited functions of given energy. These and other properties are discussed in detail in Refs. 1 and 2. Some familiarity with these papers will be assumed in the following.

In the present paper we consider certain aspects of the generalization of this earlier work to functions of many variables. Many of the structural results of Refs. 1 and 2 (as was pointed out there) depend only on the fact that the operator defined by the right of (2) is completely continuous and positive definite. The generalizations to $D$ dimensions are perfectly straightforward: we comment briefly on some of them in Section II, but do not belabor them. Our main concern here is with details of the explicit solution of some of the integral equations that are multidimensional generalizations of (1). An unexpected dividend of this work is that one of these equations is of interest in the theory of masers.

In Section III, we point out some general features of the integral equations to be considered. Section IV treats the case of functions of two variables whose Fourier transforms vanish outside a circle in the two-dimensional frequency plane. The analog of (1) is shown to be the integral equation

$$\gamma\varphi(r) = \int_0^1 J_N(crr')\sqrt{crr'}\varphi(r')dr', \qquad 0 \leqq r \leqq 1. \tag{3}$$

This integral equation also describes the modes in a maser interferometer with confocal spherical mirrors of circular cross section (Ref. 3, p. 488). The eigenfunctions of (3) are shown to be the bounded solutions of

$$(1 - x^2)\frac{d^2\varphi}{dx^2} - 2x\frac{d\varphi}{dx} + \left(\chi - c^2x^2 - \frac{\frac{1}{4} - N^2}{x^2}\right)\varphi = 0 \tag{4}$$

that vanish at $x = 0$.

We call the solutions of (4) generalized prolate spheroidal functions. Section V is devoted to their study: 5.1 treats the case of small $c$; 5.2 and 5.3 treat various asymptotic cases.*

In Section VI, the results of Section V are used to discuss the eigenvalues of (3). Various asymptotic forms for these quantities are derived.

Section VII treats the case of functions of $D > 2$ variables whose Fourier transforms vanish outside a sphere in the $D$-dimensional frequency space. It is shown that this more general problem can be reduced completely to the case already treated in Sections IV, V and VI.

Finally, in Section VIII we present some numerical detail about some of the eigenfunctions and eigenvalues encountered. Applications of these results will be presented elsewhere.[9, 10]

## II. GENERALIZATIONS OF EARLIER WORK

We denote points in Euclidean space of $D$ dimensions, $E_D$, by vectors, $\mathbf{x} = (x_1, x_2, \cdots, x_D)$. A square-integrable function of $D$ variables, $f(\mathbf{x})$, is said to be $R$-limited if it can be represented as a Fourier integral

$$f(\mathbf{x}) = (2\pi)^{-D}\int_R \exp\,(i\mathbf{x}\cdot\mathbf{y})F(\mathbf{y})dy \tag{5}$$

over the bounded region $R$. Here $\mathbf{x}\cdot\mathbf{y} = \sum x_iy_i$ is the usual scalar product and we write $dy$ for $\prod dy_i$. If $f$ is of total energy $A$, then by Parseval's theorem

$$A = \int_{E_D} |f(\mathbf{x})|^2dx = (2\pi)^{-D}\int_R |F(\mathbf{y})|^2dy,$$

whereas the energy of $f$ in the bounded region $S$ is

* Some of the results of Sections IV and V have been developed independently by J. C. Huertley,[11] who was led to consider (3) from its laser applications.

$$\int_S |f(\mathbf{z})|^2 dz$$

$$= \int_S dz (2\pi)^{-2D} \int_R dx \exp (i\mathbf{z}\cdot\mathbf{x})F(\mathbf{x}) \int_R dy \exp (-i\mathbf{z}\cdot\mathbf{y})\bar{F}(\mathbf{y})$$

$$= (2\pi)^{-D} \int_R dx \int_R dy\, K_S(\mathbf{x} - \mathbf{y})F(\mathbf{x})\bar{F}(\mathbf{y})$$

where

$$K_S(\mathbf{x} - \mathbf{y}) = (2\pi)^{-D} \int_S \exp\, [i\mathbf{z}\cdot(\mathbf{x} - \mathbf{y})]dz \tag{6}$$

and an overbar denotes complex conjugate. The largest fraction of energy that an $R$-limited function can have in the region $S$ is therefore the maximum value of the fraction

$$\int_R dx \int_R dy\, K_S(\mathbf{x} - \mathbf{y})F(\mathbf{x})\bar{F}(\mathbf{y}) \Big/ \int_R |F(\mathbf{y})|^2 dy$$

taken over all functions $F$ square-integrable through $R$. This maximum is the largest eigenvalue of the integral equation

$$\lambda\psi(\mathbf{x}) = \int_R K_S(\mathbf{x} - \mathbf{y})\psi(\mathbf{y})dy, \qquad \mathbf{x} \in R, \tag{7}$$

which is the analog of (2).

The kernel (6) of (7) is positive definite, since

$$\int_R dx \int_R dy\, K_S(\mathbf{x} - \mathbf{y})f(\mathbf{x})\bar{f}(\mathbf{y})$$

$$= (2\pi)^{-D} \int_S dz \left| \int_R dx \exp (i\mathbf{z}\cdot\mathbf{x})f(\mathbf{x}) \right|^2 > 0$$

whenever

$$\int_R |f(\mathbf{x})|^2 dx > 0.$$

By well-known theorems (see Ref. 4, Chap. 6), the eigenvalues of (7) are real and positive and the eigenfunctions, orthogonal on $R$, are complete in the class of functions square-integrable in $R$. A complete discussion of the simultaneous concentration of square-integrable functions in $E_D$ and their Fourier transforms can be given in terms of the largest eigenvalue of (7) as in Ref. 2, Theorem 2.

The right member of (7) can be used to extend the domain of definition of $\psi$. We define

$$\psi(\mathbf{x}) = \frac{1}{\lambda} \int_R K_S(\mathbf{x} - \mathbf{y})\psi(\mathbf{y})dy, \qquad \mathbf{x} \in E_D .$$

Then for two different eigenfunctions of (7)

$$\int_{E_D} \psi_i(\mathbf{x})\bar{\psi}_j(\mathbf{x})dx$$

$$= \frac{1}{\lambda_i\lambda_j} \int_R dx \int_R dy\ \psi_i(\mathbf{x})\bar{\psi}_j(\mathbf{y}) \int_{E_D} dz\ K_S(\mathbf{z} - \mathbf{x})\bar{K}_S(\mathbf{z} - \mathbf{y}).$$

To evaluate the innermost integral here, we observe from (6) that $K_S$ is given as a Fourier transform, so that from Parseval's theorem,

$$\int_{E_D} K_S(\mathbf{z} - \mathbf{x})\bar{K}_S(\mathbf{z} - \mathbf{y})dz$$

$$= (2\pi)^{-D} \int_S du \exp\left[-i\mathbf{u}\cdot(\mathbf{x} - \mathbf{y})\right] = \bar{K}_S(\mathbf{x} - \mathbf{y}).$$

One then finds

$$\int_{E_D} \psi_i(\mathbf{x})\bar{\psi}_j(\mathbf{x})dx = \frac{1}{\lambda_i\lambda_j} \int_R dx\ \psi_i(\mathbf{x}) \int_R dy\ \bar{K}_S(\mathbf{x} - \mathbf{y})\bar{\psi}_j(\mathbf{y})$$

$$= \frac{1}{\lambda_i} \int_R dx\ \psi_i(\mathbf{x})\bar{\psi}_j(\mathbf{x}).$$

The orthogonality of the $\psi_i$ over $R$ thus implies orthogonality over $E_D$ as well.

Other results of the one-dimensional case extend as easily to $D$ dimensions, but we do not dwell further here on this general structure.

## III. SYMMETRY CONSIDERATIONS

In what follows, we shall be concerned with the explicit solution of a number of instances of (7). Considerable simplification occurs when the region $R$ is symmetric, i.e., when $\mathbf{x} \in R$ implies $-\mathbf{x} \in R$, and when $S$ is a scaled version of $R$. We write $S = cR$ where $\mathbf{x} \in cR$ if and only if $\mathbf{x}/c \in R$ with $c$ a positive constant. We restrict our attention henceforth to this case.

Somewhat simpler than (7) is the integral equation

$$\alpha\psi(\mathbf{x}) = \int_R \exp\,(ic\mathbf{x}\cdot\mathbf{y})\psi(\mathbf{y})dy, \qquad \mathbf{x} \in R \tag{8}$$

which is a natural generalization of (1). We shall show in this section that solution of this equation is completely equivalent to solution of (7) when the symmetries just discussed maintain. We shall accordingly hereafter take (8) as our equation of fundamental concern.

From the symmetry of $R$, it readily follows that if $\psi(\mathbf{x})$ is a solution of (8), so also is $\psi(-\mathbf{x})$, so that both $\psi_e(\mathbf{x}) = \psi(\mathbf{x}) + \psi(-\mathbf{x})$ and $\psi_o(\mathbf{x}) = \psi(\mathbf{x}) - \psi(-\mathbf{x})$ are solutions as well. *The eigenfunctions of* (8) *can be chosen to be either even or odd functions of $x$.*

The complex conjugate of (8) is

$$\bar{\alpha}\bar{\psi}(\mathbf{x}) = \int_R \exp\,(-ic\mathbf{x}\cdot\mathbf{y})\bar{\psi}(\mathbf{y})dy, \qquad \mathbf{x} \in R. \tag{9}$$

Multiply (8) by $\bar{\psi}(\mathbf{x})$ and integrate over $R$. Multiply (9) by $\psi(\mathbf{x})$ and integrate over $R$. Combining these equations, one finds on using the symmetry of $R$ that

$$(\alpha \pm \bar{\alpha}) \int_R \psi(\mathbf{x})\bar{\psi}(\mathbf{x})dx$$
$$= \int_R dx \int_R dy \exp\,(ic\mathbf{x}\cdot\mathbf{y})\bar{\psi}(\mathbf{x})[\psi(\mathbf{y}) \pm \psi(-\mathbf{y})].$$

If then $\psi$ is even, by choosing the negative sign in this equation, one obtains $\alpha - \bar{\alpha} = 0$, whereas if $\psi$ is odd, by choosing the plus sign, one finds $\alpha + \bar{\alpha} = 0$. *The eigenvalues of* (8) *associated with even eigenfunctions are real: the eigenvalues of* (8) *associated with odd eigenfunctions are pure imaginary.* It follows then that (8) is equivalent to the pair of equations

$$\beta_e\psi_e(\mathbf{x}) = \int_R \cos c\mathbf{x}\cdot\mathbf{y}\psi_e(\mathbf{y})dy \tag{10}$$

$$\beta_o\psi_o(\mathbf{x}) = \int_R \sin c\mathbf{x}\cdot\mathbf{y}\psi_o(\mathbf{y})dy \tag{11}$$

in which $\beta_e$ and $\beta_o$ are real. These equations have real symmetric kernels and we can fall back on the extensive theory in the literature treating such equations. We observe that the eigenfunctions of (10) must be even and that $\beta_e = 0$ cannot be an eigenvalue of this equation, for by Fourier theory the only even square-integrable function in $R$ for which

$$\int_R \cos c\mathbf{x}\cdot\mathbf{y}\psi(\mathbf{y})dy = 0, \qquad \mathbf{x} \in R$$

is $\psi(\mathbf{y}) \equiv 0$. It follows then from the theorem on page 234 of Ref. 4 that the eigenfunctions of (10) are complete in the class of even functions square-integrable in $R$. A similar argument shows that the solutions of (11) are complete in the class of odd functions square-integrable in $R$. The solutions of (10) can be chosen real and orthogonal in $R$, as can the solutions of (11). Solutions of (10) are automatically orthogonal to solutions of (11) by symmetry.

We have now shown that the solutions of (8) are complete in the class of functions square-integrable in $R$. The eigenfunctions can be chosen real, orthogonal, and either even (in which case the eigenvalue $\alpha$ is real) or odd (in which case $\alpha$ is pure imaginary). We henceforth assume the $\psi$ so chosen.

By iterating (8), one finds that the $\psi$ also satisfy

$$\lambda\psi(\mathbf{x}) = \int_R K_c(\mathbf{x} - \mathbf{y})\psi(\mathbf{y})dy \tag{12}$$

$$\lambda = \left(\frac{c}{2\pi}\right)^D |\alpha|^2 \tag{13}$$

with

$$K_c(\mathbf{x}) = \left(\frac{c}{2\pi}\right)^D \int_R \exp\,(ic\mathbf{z}\cdot\mathbf{x})\;dz = (2\pi)^{-D} \int_{cR} e^{i\mathbf{z}\cdot\mathbf{x}}\;dz \tag{14}$$

which is (7) in slightly altered notation and is the $D$-dimensional analog of (2). Since the solutions $\psi$ of (8) are complete, it follows that they are also a complete set of solutions of (12). As was asserted, to solve (12), it suffices to solve (8).

The eigenfunctions of (8) can be extended by demanding that equation to hold for all $\mathbf{x} \in E_D$. It is then easy to show that the extended $\psi$ are orthogonal in $E_D$ and that they are complete in the class of $cR$-limited functions.

## IV. THE CASE $D = 2$, $R$ A CIRCLE

We now treat in detail the equation

$$\alpha\psi(x_1, x_2) = \int_R e^{ic(x_1y_1+x_2y_2)}\psi(y_1, y_2)dy_1dy_2 \tag{15}$$

where $R$ is the unit circle $y_1^2 + y_2^2 \leqq 1$. Change to polar coordinates gives

$$\begin{aligned}\alpha\psi(r,\theta) &= \int_0^1 dr'\,r' \int_0^{2\pi} d\theta'\;e^{icrr'\cos(\theta-\theta')}\psi(r',\theta')\\ &= \sum_{-\infty}^{\infty} i^m e^{im\theta} \int_0^1 dr'\,r'J_m(crr') \int_0^{2\pi} d\theta'\;e^{-im\theta'}\,\psi(r',\theta')\end{aligned} \tag{16}$$

on making the usual Bessel function expansion. Here $\psi(r,\theta)$ is exhibited as a Fourier series in $\theta$. A simple argument then gives for the eigenfunctions of (15) and their corresponding eigenvalues

$$\begin{aligned} \psi_{o,n}(r,\theta) &= R_{o,n}(r), & \alpha_{o,n} &= 2\pi\beta_{o,n} \\ \psi_{N,n}(r,\theta) &= R_{N,n}(r)\begin{matrix}\cos N\theta, \\ \sin N\theta,\end{matrix} & \alpha_{N,n} &= 2\pi i^N\beta_{N,n} \\ & N = 1, 2, \dots, & n &= 0, 1, 2, \dots \end{aligned} \tag{17}$$

where

$$\beta_{N,n}R_{N,n}(r) = \int_0^1 J_N(crr')R_{N,n}(r')r'\,dr', \qquad 0 \leqq r \leqq 1, \qquad n, N = 0, 1, 2, \dots. \tag{18}$$

All the eigenvalues of (15), except possibly the $\alpha_{o,n}$ have at least a two-fold degeneracy inherited from the symmetry of the circle.

Our task now is to study the integral equation

$$\beta R(r) = \int_0^1 J_N(crr')R(r')r'\,dr', \qquad 0 \leqq r \leqq 1.$$

It is convenient to make the substitutions

$$\gamma = \sqrt{c}\beta, \qquad \varphi(r) = \sqrt{r}R(r) \tag{19}$$

to obtain the symmetric equation

$$\gamma\varphi(r) = \int_0^1 J_N(crr')\sqrt{crr'}\varphi(r')\,dr', \qquad 0 \leqq r \leqq 1. \tag{20}$$

Note that $\varphi(0) = 0$. We shall show that the eigenfunctions $\varphi_{N,n}(r)$ of (20) can be obtained as the solution of a Sturm-Liouville differential equation.

Let

$$K_N(x) = J_N(x)\sqrt{x} \tag{21}$$

and let the operator $M$ be defined by

$$[M\psi](x) = \int_0^1 K_N(cxy)\psi(y)dy.$$

Denote by $L_x$ the differential operator

$$L_x = \frac{d}{dx}(1 - x^2)\frac{d}{dx} + \left(\frac{\frac{1}{4} - N^2}{x^2} - c^2x^2\right).$$

Consider now

$$\begin{aligned}[ML\psi](x) &= \int_0^1 K_N(cxy)\left[\frac{d}{dy}(1-y^2)\frac{d}{dy}\right. \\ &\qquad \left. + \left(\frac{\frac{1}{4}-N^2}{y^2} - c^2y^2\right)\right]\psi(y)dy \\ &= [K_N(cxy)(1-y^2)\psi'(y) - cx(1-y^2) \\ &\qquad \cdot K_N'(cxy)\psi(y)]_{y=0}^{1} + \int_0^1 \psi(y)\left[c^2x^2(1-y^2)K_N''(cxy)\right. \\ &\qquad \left. - 2cxyK'(cxy) + \left(\frac{\frac{1}{4}-N^2}{y^2} - c^2y^2\right)K_N(cxy)\right]dy \end{aligned} \tag{22}$$

where the right member is obtained by integration by parts. Here primes denote differentiation of the function in question with respect to its argument. The integrated expression vanishes if $\psi(0) = 0$, since from (21), $K_N(0) = 0$. Also from (21) and the differential equation satisfied by Bessel functions, one has the identity

$$K_N''(cxy) = -\left(1 + \frac{\frac{1}{4}-N^2}{c^2x^2y^2}\right)K_N(cxy). \tag{23}$$

Substitute this expression in (22) to yield

$$\begin{aligned}[ML\psi](x) &= \int_0^1 \psi(y)[-2cxyK'(cxy) \\ &\qquad + (\tfrac{1}{4} - N^2 + c^2x^2y^2 - c^2x^2 - c^2y^2)K(cxy)]dy, \\ \psi(0) &= 0. \end{aligned} \tag{24}$$

On the other hand, by direct calculation and use of (23), one has

$$\begin{aligned}[LM\psi](x) &= L_x\int_0^1 K_N(cxy)\psi(y)dy \\ &= \int_0^1 \psi(y)\left[(1-x^2)c^2y^2K_N''(cxy) - 2xcyK_N'(cxy)\right. \\ &\qquad \left. + \left(\frac{\frac{1}{4}-N^2}{x^2} - c^2x^2\right)K_N(cxy)\right]dy \\ &= \int_0^1 \psi(y)\left[-2xcyK_N'(cxy) + \left\{-(1-x^2)c^2y^2\right.\right. \\ &\qquad \left.\left. \cdot\left(1 + \frac{\frac{1}{4}-N^2}{c^2x^2y^2}\right) + \frac{\frac{1}{4}-N^2}{x^2} - c^2x^2\right\}K_N(cxy)\right]dy \end{aligned}$$

$$\begin{aligned}&= \int_0^1 \psi(y)[-2cxyK_N'(cxy) + (\tfrac{1}{4} - N^2 + c^2x^2y^2 \\ &\qquad - c^2x^2 - c^2y^2)K_N(cxy)]dy \\ &= [ML\psi](x)\end{aligned}$$

on comparison with (24).

Let $C$ be the class of functions square-integrable in (0,1) and twice differentiable there that vanish at the origin. Operating on functions in $C$, the operators $M$ and $L$ commute. It follows that solutions of

$$L_x\varphi(x) = -\chi\varphi(x)$$

in $C$ are also solutions of (20). Consequently, we next turn our attention to the differential equation.

$$(1 - x^2)\frac{d^2\varphi}{dx^2} - 2x\frac{d\varphi}{dx} + \left(\frac{\frac{1}{4} - N^2}{x^2} - c^2x^2 + \chi\right)\varphi = 0. \tag{25}$$

V. GENERALIZED PROLATE SPHEROIDAL FUNCTIONS

When $N = \frac{1}{2}$ in (25), this equation reduces to the equation for prolate spheroidal functions of order zero. We shall refer to bounded solutions of (25) for arbitrary values of $N$ as *generalized prolate spheroidal functions*. These functions are similar in many respects to prolate spheroidal functions, as the development that follows shows. Bounded solutions of (25) exist only for discrete values of $\chi$, say $\chi_{N,n}$, $n = 0, 1, 2, \ldots$ which we label so that $\chi_{N,o} \leqq \chi_{N,1} \leqq \chi_{N,2} \leqq \ldots$. We denote the corresponding eigenfunctions by $\varphi_{N,n}(x)$.

### 5.1 *Expansions in Powers of c*

Consider first the case when $c = 0$. Substitution of the series

$$\varphi = \sum_0^\infty a_jx^{\alpha+2j}$$

into (25) shows that we must have $\alpha = \frac{1}{2} \pm N$. If $N \neq 0$, the negative sign leads to solutions having a singularity at $x = 0$. If $N = 0$, a second solution can be found, but it has a logarithmic singularity at $x = 0$. We must have therefore

$$\alpha = \tfrac{1}{2} + N.$$

The coefficients are given by the recurrence

$$a_{j+1} = a_j\frac{(\alpha + 2j)(\alpha + 2j + 1) - \chi}{(\alpha + 2j + 2)(\alpha + 2j + 1) + \frac{1}{4} - N^2}.$$

For large $j$, $a_{j+1}/a_j \to 1$, so unless the series terminates, this solution becomes unbounded as $x \to 1$. Choosing $\chi$ to terminate the series at $x^{\alpha+2n}$, we have

$$\chi = \chi_{N,n}(0) = (N + 2n + \tfrac{1}{2})\,(N + 2n + \tfrac{3}{2}) \tag{26}$$

for the eigenvalues of (25) when $c = 0$. The series solution now becomes (when $a_o$ is set equal to unity)*

$$\begin{aligned} \varphi &= T_{N,n}(x) = x^{N+\frac{1}{2}} R_{N,n}(x) \\ R_{N,n}(x) &= F(-n, n + N + 1; N + 1; x^2) \end{aligned} \tag{27}$$

where

$$F(a,b;\, c;\, z) = 1 + \frac{ab}{c}\frac{z}{1!} + \frac{a(a + 1)b(b + 1)}{c(c + 1)}\frac{z^2}{2!} + \cdots$$

is the usual Gaussian hypergeometric function. The polynomial $R_{N,n}(x)$ is readily expressed in terms of Jacobi polynomials $P_n^{(\alpha,\beta)}(x)$ (Ref. 5, Chap. IV). Adopting the notation of Szegö, we have

$$R_{N,n}(x) = \binom{n + N}{n}^{-1} P_n^{(N,0)}(1 - 2x^2). \tag{28}$$

From (27), (28) and the known properties of the Jacobi polynomials, one finds

$$T_{N,n}(1) = (-1)^n \binom{n + N}{n}^{-1} \tag{29}$$

$$\int_0^1 T_{N,n}(x) T_{N,n'}(x)\, dx = \frac{\delta_{nn'}}{2(2n + N + 1)\binom{n + N}{n}} \tag{30}$$

$$\begin{aligned} &2(n + N + 1)^2(2n + N)R_{N,n+1} \\ &\quad = (2n + N + 1)\,[(2n + N)\,(2n + N + 2)\,(1 - 2x^2) \\ &\qquad + N^2]R_{N,n} - 2n^2(2n + N + 2)R_{N,n-1} \end{aligned}$$

$$\begin{aligned} &(2n + N)x(1 - x^2)\frac{d}{dx}R_{N,n}(x) \\ &\quad = n[(2n + N)(1 - 2x^2) - N]R_{N,n}(x) - 2n^2R_{N,n-1}(x) \end{aligned}$$

$$x^2T_{N,n}(x) = \gamma_{N,n}{}^{1}T_{N,n+1}(x) + \gamma_{N,n}{}^{0}T_{N,n}(x) + \gamma_{N,n}{}^{-1}T_{N,n-1}(x) \tag{31}$$

* It has been called to our attention that our $T_{N,n}(x)$ are closely related to the Zernike polynomials. These latter arise in the diffraction theory of aberrations.[12]

$$\begin{aligned}
\gamma_{N,n}{}^{1} &= -\frac{(n+N+1)^2}{(2n+N+1)(2n+N+2)} \\
\gamma_{N,n}{}^{0} &= \frac{1}{2}\left(1+\frac{N^2}{(2n+N)(2n+N+2)}\right) \qquad (32)\\
\gamma_{N,n}{}^{-1} &= -\frac{n^2}{(2n+N)(2n+N+1)}
\end{aligned}$$

$$| T_{N,n}(x) | \leqq 1 \quad \text{for} \quad 0 \leqq x \leqq 1.$$

The function $T_{N,n}(x)$ has $n$ zeros in (0,1]. We define $T_{N,n}(x) = 0$ if $n < 0$.

Returning now to (25) for arbitrary values of $c$, we attempt a power series solution in $c^2$ by writing

$$\varphi(x) = \varphi_{N,n}(x) = T_{N,n}(x) + \sum_{j=1}^{\infty} c^{2j} Q_j(N,n,x) \qquad (33)$$

$$\chi = \chi_{N,n}(c) = \chi_{N,n}(0) + \sum_{j=1}^{\infty} c^{2j} a_j(N,n), \qquad (34)$$

where the $Q$'s and $a$'s are independent of $c$. When this latter quantity is zero, this solution reduces to that already found. As is shown in Appendix A, the $Q$'s and $a$'s can be determined recursively in an elementary manner. We have

$$Q_j(N,n,x) = \sum_{k=-j}^{j} A_k{}^{j}(N,n) T_{N,n+k}(x) \qquad (35)$$

with

$$a_j(N,n) = \sum_{k=-1}^{1} A_{-k}{}^{j-1}(N,n)\gamma_{N,n-k}{}^{k}, \qquad j = 1, 2, \dots \qquad (36)$$

$$\begin{aligned}
&[\chi_{N,n+m}(0) - \chi_{N,n}(0)]A_m{}^{j}(N,n) \\
&\quad = \sum_{k=1}^{j} a_k(N,n) A_m{}^{j-k}(N,n) - \sum_{k=-1}^{1} A_{-k+m}{}^{j-1}(N,n)\gamma_{N,n-k+m}{}^{k}, \qquad (37)\\
&\qquad\qquad m = -j, -j+1, \dots, j; \quad j = 1, 2, \dots .
\end{aligned}$$

Here $A_k{}^{j}(N,n)$ is defined to be zero if $| k | > j$, or $k < -n$ or $k = 0$ and $j \neq 0$. In addition we have $A_m{}^{0}(N,n) = 0$, $m \neq 0$, $A_o{}^{0}(N,n) = 1$, $a_o(N,n) = 0$. For use in (36) and (37), the $\gamma$'s of (32) must be defined so that for $n < 0$, $\gamma_{N,n}{}^{1} = \gamma_{N,n}{}^{0} = \gamma_{N,n+1}{}^{-1} = 0$.

To terms of order $c^2$ the eigenfunctions and eigenvalues of (25) are explicitly

$$\chi_{N,n}(c) = \left(2n + N + \frac{1}{2}\right)\left(2n + N + \frac{3}{2}\right) + \frac{1}{2}\left(1 + \frac{N^2}{(2n + N)(2n + N + 2)}\right)c^2 + O(c^4) \tag{38}$$

$$\varphi_{N,n}(x) = T_{N,n}(x) + \left(\frac{n^2 T_{N,n-1}(x)}{4(2n + N)^2(2n + N + 1)} - \frac{(n + N + 1)^2 T_{N,n+1}(x)}{4(2n + N + 1)(2n + N + 2)^2}\right)c^2 + O(c^4). \tag{39}$$

In view of (35), the series (33) can be formally regrouped to give

$$\varphi_{N,n}(x) = \sum_0^\infty d_j^{N,n}(c) T_{N,j}(x). \tag{40}$$

Substitution in (25) yields the three-term recurrence

$$\begin{aligned} &c^2\gamma_{N,j-1}{}^1\, d_{j-1}{}^{N,n} \\ &\quad + [c^2\gamma_{N,j}{}^0 + (2j + N + \tfrac{1}{2})(2j + N + \tfrac{3}{2}) - \chi]\, d_j{}^{N,n} \\ &\quad\quad + c^2\gamma_{N,j+1}{}^{-1}\, d_{j+1}{}^{N,n} = 0. \end{aligned} \tag{41}$$

This recurrence can be used to determine the $d_j^{N,n}$ and the eigenvalues in a manner quite parallel to that used in the study of prolate spheroidal wave functions. The method of Bouwkamp[6] can be adopted and used advantageously for the computation of the $d_j^{N,n}$ and the eigenvalues for values of $c$ too large to permit effective use of (33) and (34). The $d$'s can, of course, be expressed in terms of the $A$'s of (35). One has

$$d_{n+j}{}^{N,n}(c) = \sum_{l=|j|}^\infty A_j{}^l(N,n)c^{2l}, \qquad j = -n, -n + 1, \ldots . \tag{42}$$

The series solutions (40) or (33), (35) for the generalized prolate spheroidal function are, of course, valid only for $0 \leqq x \leqq 1$. To obtain a series valid for $x > 1$, we use (20) and the fact (established in Appendix B) that

$$\int_0^1 J_N(cxy)\sqrt{cxy}\, T_{N,n}(y)dy = \binom{N + n}{n}^{-1} \frac{J_{N+2n+1}(cx)}{\sqrt{cx}}. \tag{43}$$

The solution (40) then extends for all $x$ by the series

$$\varphi_{N,n}(x) = \frac{1}{\gamma_{N,n}} \sum_{j=0}^\infty d_j^{N,n} \frac{J_{N+2j+1}(cx)}{\binom{N + j}{j}\sqrt{cx}} \tag{44}$$

which is obtained by inserting (40) in the right of (20) and integrating term by term.

The eigenvalue $\gamma_{N,n}$ can be expressed in terms of the $d_j^{N,n}$. Divide both sides of the equation

$$\gamma_{N,n}\varphi_{N,n}(x) = \int_0^1 J_N(cxx')\sqrt{cxx'}\;\varphi_{N,n}(x')dx' \tag{45}$$

by $x^{N+\frac{1}{2}}$ and take limits as $x \to 0$. From (27) and (40), we see that the left member of (45) becomes

$$\gamma_{N,n}\sum_j d_j^{N,n}R_{N,n}(0) = \gamma_{N,n}\sum_j d_j^{N,n}.$$

Since $J_N(x)\sqrt{x} \sim (x/2)^N\sqrt{x}/\Gamma(N+1)$, the right of (45) becomes

$$\begin{aligned}
&\frac{1}{\Gamma(N+1)2^N}\int_0^1 (cx')^{N+\frac{1}{2}}\varphi_{N,n}(x')dx' \\
&\qquad = \frac{c^{N+\frac{1}{2}}}{\Gamma(N+1)2^N}\sum_{j=0} d_j^{N,n}\int_0^1 x'^{N+\frac{1}{2}}T_{N,j}(x')dx' \\
&\qquad = \frac{c^{N+\frac{1}{2}}}{\Gamma(N+1)2^N}\sum_{j=0} d_j^{N,n}\int_0^1 T_{N,o}(x')T_{N,j}(x')dx \\
&\qquad = \frac{c^{N+\frac{1}{2}}d_o^{N,n}}{\Gamma(N+2)2^{N+1}}
\end{aligned}$$

where we have used successively (40), (27) and (30). The combined result is

$$\gamma_{N,n} = \frac{c^{N+\frac{1}{2}}d_o^{N,n}}{2^{N+1}\Gamma(N+2)\sum_{j=0} d_j^{N,n}}. \tag{46}$$

The integral equation (45) is also useful for obtaining the asymptotic behavior of $\varphi_{N,n}(x)$ as $x \to \infty$. We have

$$\gamma_{N,n}\varphi_{N,n}(x) = \frac{1}{cx}\int_0^{cx} du\; u^{N+1}J_N(u)\,\frac{\varphi_{N,n}(u/cx)}{u^{N+\frac{1}{2}}} \tag{47}$$

on letting $cxx' = u$. Now $(u^{N+1}J_{N+1})' = (u^{N+1}J_N)$, so that (47) can be integrated by parts to yield

$$\begin{aligned}
\gamma_{N,n}\varphi_{N,n}(x) = \frac{1}{cx}\Bigg[& u^{N+1}J_{N+1}(u)\,\frac{\varphi_{N,n}(u/cx)}{u^{N+\frac{1}{2}}}\Bigg|_0^{cx} \\
&- \int_0^{cx} du\; u^{N+1}J_{N+1}(u)\,\frac{d}{du}\,\frac{\varphi_{N,n}(u/cx)}{u^{N+\frac{1}{2}}}\Bigg]
\end{aligned}$$

$$= \frac{1}{cx}\sqrt{cx}\,\varphi_{N,n}(1)J_{N+1}(cx) - R.$$

For large $x$, this becomes

$$\gamma_{N,n}\varphi_{N,n}(x) = \varphi_{N,n}(1)\sqrt{\frac{2}{\pi}}\frac{\cos[cx - (N+1)(\pi/2) - (\pi/4)]}{cx} + O\left(\frac{1}{x^2}\right). \tag{48}$$

This of course is consistent with (44). If now we define $\varphi_{N,n}{}^*(x)$ to be a generalized prolate spheroidal function normalized so that for large $x$

$$\varphi_{N,n}{}^*(x) \sim \frac{\cos[cx - (N+1)(\pi/2) - (\pi/4)]}{cx}, \tag{49}$$

(48) gives us

$$\gamma_{N,n} = \sqrt{\frac{2}{\pi}}\,\varphi_{N,n}{}^*(1) \tag{50}$$

a relation that will be useful to us later.

### 5.2 *Asymptotics for Fixed n and Large c*

The behavior of generalized prolate spheroidal functions for large $c$ can be determined by methods quite parallel to those used in Ref. 7 in discussing the prolate spheroidal functions. Five different asymptotic forms for $\varphi_{N,n}(x)$ are found, depending on the $x$ range under consideration. These are properly joined to furnish a solution for all $x$. For most of these regions, we content ourselves here with writing only the leading term of the asymptotic development.

In (25) we make the substitution $t = x\sqrt{c}$. There results

$$\mathbf{L}\varphi - (1/c)\mathbf{M}\varphi + (\chi/c)\varphi = 0 \tag{51}$$

where the operators are given by

$$\begin{aligned} \mathbf{L} &= \frac{d^2}{dt^2} + \frac{\frac{1}{4} - N^2}{t^2} - t^2 \\ \mathbf{M} &= t^2\frac{d^2}{dt^2} + 2t\frac{d}{dt}. \end{aligned} \tag{52}$$

Now the equation

$$\mathbf{L}U + \lambda U = 0$$

has solutions

$$\begin{aligned} U &= U_{N,n}(t) = e^{-t^2/2}t^{N+\frac{1}{2}}L_n^{(N)}(t^2) \\ \lambda_n &= 4n + 2N + 2, \qquad n = 0, 1, 2, \ldots \end{aligned} \tag{53}$$

(see Ref. 5, p. 99) where $L_n^{(\alpha)}(x)$ is the Laguerre polynomial of degree $n$ in Szegö's notation. The function $U_{N,n}(t)$ has $n$ zeros in $(0, \infty)$. This suggests attempting solution of (51) for large $c$ by the series

$$\varphi_{N,n}{}^{1} = U_{N,n}(t) + \sum_{j=1}^{\infty} (1/c^j) S_j(N,n,t) \tag{54}$$

$$\chi_{N,n}(c)/c = 4n + 2N + 2 + \sum_{j=1}^{\infty} (1/c^j) b_j(N,n). \tag{55}$$

We now note that

$$\mathbf{M}U_{N,n}(t) = \mu_{N,n}{}^{1}U_{N,n+2} + \mu_{N,n}{}^{0}U_{N,n} + \mu_{N,n}{}^{-1}U_{N,n-2} \tag{56}$$

where

$$\begin{aligned} \mu_{N,n}{}^{1} &= (n + 1)(n + 2) \\ \mu_{N,n}{}^{0} &= -[(2n + 1)(n + N + \tfrac{1}{2}) + \tfrac{3}{4}] \\ \mu_{N,n}{}^{-1} &= (n + N)(n + N - 1), \end{aligned} \tag{57}$$

a fact which can be readily derived from (52), (53) and the properties of Laguerre polynomials. The perturbation scheme of Appendix A applies therefore, and we find at once that

$$S_j(N,n,t) = \sum_{k=-j}^{j} B_k{}^{j}(N,n) U_{N,n+2k}(t) \tag{58}$$

where the $B$'s and $b$'s are given by the recurrence

$$\begin{aligned} b_j(N,n) &= \sum_{k=-1}^{1} B_{-k}{}^{j-1}(N,n)\mu_{N,n-2k}{}^{k}, \qquad j = 1, 2, \ldots \\ 8mB_m{}^{j}(N,n) &= \sum_{k=1}^{j} b_k(N,n)B_m{}^{j-k}(N,n) \\ &\qquad - \sum_{k=-1}^{1} B_{-k+m}{}^{j-1}(N,n)\mu_{N,n+2}{}^{k}(m - k) \\ &\qquad m = -j, -j + 1, \ldots, j; \;\; j = 1, 2, \ldots \end{aligned} \tag{59}$$

with the convention $B_k{}^{j}(N,n) \equiv 0$ if $|k| > j$, or $k < -n$ or $k = 0$ and $j \neq 0$. We take $B_m{}^{0}(N,n) = 0$, $m \neq 0$, $B_0{}^{0}(N,n) = 1$, $b_0(N,n) = 0$.

In this manner we obtain explicitly

$$\chi_{N,n}(c) = (4n + 2N + 2)c - [(2n + 1)(n + N + \tfrac{1}{2}) + \tfrac{3}{4}] - \frac{(N + 2n + 1)[2n^2 + 2n(N + 1) + N + 2]}{4c} + O\left(\frac{1}{c^2}\right) \tag{60}$$

which gives the behavior of $\chi_{N,n}$ for large $c$.

We write the solution just found as

$$\varphi_{N,n}{}^1(x) = U_{N,n}(t) + \sum_{j=1}^{\infty} \frac{1}{c^j} \sum_{k=-j}^{j} B_k{}^j(N,n) U_{N,n+2k}(t), \qquad t = x\sqrt{c}. \tag{61}$$

The right side of (61) is ordered in powers of $c^{-1}$ when expressed in terms of the variable $t$. However, if $t = x\sqrt{c}$ is substituted, the terms are no longer so ordered since $U_{N,m+2}(x\sqrt{c})/U_{N,m}(x\sqrt{c}) = O(c)$. The range of $x$ values for which the first few terms of (61) furnish information about $\varphi_{N,n}$ vanishes as $c$ gets large. We shall use (61) only for $0 \leqq x \leqq 1/c^{\frac{1}{4}}$.

To obtain an asymptotic form for $\varphi_{N,n}(x)$ for $c^{-\frac{1}{4}} \leqq x \leqq 1 - (1/c)$ it is convenient to write $\varphi_{N,n}(x) = x^{N+\frac{1}{2}}\psi_{N,n}(x)$ and set $y = \sqrt{1 - x^2}$. Equation (25) now becomes

$$(1 - y^2)\frac{d^2\psi}{dy^2} + \left[\frac{1}{y} - (2N + 3)y\right]\frac{d\psi}{dy} + \left[\chi_{N,n} - c^2 - \left(N + \frac{1}{2}\right)\left(N + \frac{3}{2}\right) + c^2y^2\right]\psi = 0. \tag{62}$$

Into this equation, substitute $\chi_{N,n}(c)$ as given by (60) and set

$$\psi = \frac{e^{cy}(1 - y)^n}{\sqrt{y}(1 + y)^{N+n+1}}\, v.$$

One finds then for $v$,

$$\frac{dv}{dy} + O\left(\frac{1}{c}\right) = 0.$$

Accordingly we write

$$\varphi_{N,n}{}^2(x) \sim \frac{x^{N+\frac{1}{2}}(1 - y)^n e^{cy}}{\sqrt{y}(1 + y)^{N+n+1}} \qquad y = \sqrt{1 - x^2}, \qquad c^{-\frac{1}{4}} \leqq x \leqq 1 - \frac{1}{c}. \tag{63}$$

To obtain an asymptotic form for $\varphi_{N,n}(x)$ valid near $x = 1$, set $y = s/c$ in (62) and again use (60) for $\chi_{N,n}$. There results

$$\frac{d^2\psi}{dy^2} + \frac{1}{s}\frac{d\psi}{ds} - \psi + O\left(\frac{1}{c}\right) = 0.$$

Accordingly we write

$$\begin{aligned} &\varphi_{N,n}{}^3(x) \sim x^{N+\frac{1}{2}} I_o(cy) \\ &\qquad y = \sqrt{1-x^2}, \qquad 1 - (1/c) \leqq x \leqq 1 \end{aligned} \tag{64}$$

where $I_o(x)$ is the modified Bessel function. (See Ref. 8, Vol. II, p. 5).

When $x > 1$, we set $z = \sqrt{x^2 - 1}$, and have $y = iz$. The solutions $\varphi_{N,n}{}^2$ and $\varphi_{N,n}{}^3$ then give rise to two more asymptotic forms. We write

$$\varphi_{N,n}{}^4(x) = x^{N+\frac{1}{2}} J_o(cz), \qquad 1 \leqq x \leqq 1 + \frac{1}{c} \tag{65}$$

$$\varphi_{N,n}{}^5(x) = x^{N+\frac{1}{2}} \operatorname{Re} \frac{e^{icz}(1 - iz)^n}{\sqrt{iz}(1 + iz)^{N+n+1}}, \qquad 1 + \frac{1}{c} \leqq x, \tag{66}$$

$$z = \sqrt{x^2 - 1}.$$

We now determine the joining factors for these five solutions. In $\varphi_{N,n}{}^1$ and $\varphi_{N,n}{}^2$ we set $x = u/c^{\frac{1}{4}}$ and let $c$ become large for fixed $u$. One finds

$$\varphi_{N,n}{}^1(u/c^{\frac{1}{4}}) \sim \frac{(-1)^n c^{(2n+N+\frac{1}{2})/4}}{n!}\, u^{2n+N+\frac{1}{2}} e^{-u^2\sqrt{c}/2}$$

$$\varphi_{N,n}{}^2(x = u/c^{\frac{1}{4}}) \sim \frac{e^c c^{-(2n+N+\frac{1}{2})/4}}{2^{N+2n+1}}\, u^{2n+N+\frac{1}{2}} e^{-u^2\sqrt{c}/2}$$

where we have used the fact that

$$L_n{}^N(u^2\sqrt{c}) \sim (-1)^n u^{2n} c^{n/2}/n!.$$

When $y = v/\sqrt{c}$, one finds for fixed $v$ and large $c$

$$\varphi_{N,n}{}^2(y = v/\sqrt{c}) \sim c^{\frac{1}{4}} e^{v\sqrt{c}}/\sqrt{v}$$

$$\varphi_{N,n}{}^3(y = v/\sqrt{c}) \sim \frac{1}{\sqrt{2\pi c^{\frac{1}{4}}}}\, e^{v\sqrt{c}}/\sqrt{v}$$

where to obtain this last expression we have used the known asymptotic formula $I_o(x) \sim e^x/\sqrt{2\pi x}$ (see Ref. 8, Vol. II, p. 86). Finally, when $z = v/\sqrt{c}$ we find

$$\varphi_{N,n}{}^{4}(z = v/\sqrt{c}) \sim \sqrt{\frac{2}{\pi}}\, c^{-\frac{1}{4}} \frac{\cos\,(v\sqrt{c} - \pi/4)}{\sqrt{v}}$$

$$\varphi_{N,n}{}^{5}(z = v/\sqrt{c}) \sim c^{\frac{1}{4}} \frac{\cos\,(v\sqrt{c} - \pi/4)}{\sqrt{v}}$$

where we have made use of the formula (see Ref. 8, Vol. II, p. 85) $J_o(z) \sim (\pi z/2)^{-\frac{1}{2}} \cos\,(z - \pi/4)$.

All these results can be summarized in the following statement:

$$\hat{\varphi}_{N,n}(x) \sim \begin{cases} e^{-t^2/2} t^{N+\frac{1}{2}} L_n{}^{(N)}(t^2), & 0 \leqq x \leqq c^{-\frac{1}{4}} \\ k_2 \dfrac{x^{N+\frac{1}{2}} e^{cy} (1 - y)^n}{\sqrt{y}(1 + y)^{N+n+1}}, & c^{-\frac{1}{4}} \leqq x \leqq 1 - c^{-1} \\ k_3 x^{N+\frac{1}{2}} I_o(cy), & 1 - c^{-1} \leqq x \leqq 1 \\ k_4 x^{N+\frac{1}{2}} J_o(cz), & 1 \leqq x \leqq 1 + c^{-1} \\ k_5 x^{N+\frac{1}{2}} \operatorname{Re} \dfrac{e^{icz}(1 - iz)^n}{\sqrt{iz}(1 + iz)^{N+n+1}}, & 1 + c^{-1} \leqq x \end{cases} \tag{67}$$

where

$$t = x\sqrt{c}, \qquad y = \sqrt{1 - x^2}, \qquad z = \sqrt{x^2 - 1}$$

$$k_2 = \frac{(-1)^n 2^{N+2n+1} c^{n+N/2+\frac{1}{4}} e^{-c}}{n!}$$

$$k_3 = k_4 = \frac{(-1)^n \sqrt{\pi} 2^{N+2n+3/2} c^{n+N/2+\frac{3}{4}} e^{-c}}{n!}$$

$$k_5 = \frac{(-1)^n 2^{N+2n+2} c^{n+N/2+\frac{1}{4}} e^{-c}}{n!}$$

is the asymptotic form for large $c$ of a bounded continuous solution of (25) belonging to the eigenvalue (60).

We next calculate the normalization constant

$$\frac{1}{N_{N,n}{}^2} = \int_0^1 [\hat{\varphi}_{N,n}(x)]^2\, dx.$$

For the contribution due to $\varphi_{N,n}{}^{1}$ we find

$$\begin{aligned} \int_0^{c^{-\frac{1}{4}}} dx\, [e^{-t^2/2} t^{N+\frac{1}{2}} L_n{}^{(N)}(t^2)]^2 &= \frac{1}{\sqrt{c}} \int_0^{c^{\frac{1}{4}}} dt\, e^{-t^2} t^{2N+1} [L_n{}^{(N)}(t^2)]^2 \\ &= \frac{1}{2\sqrt{c}} \int_0^{\sqrt{c}} du\, e^{-u} u^N [L_n{}^{(N)}(u)]^2 \\ &= \frac{\Gamma(n + N + 1)}{2\sqrt{c}\Gamma(n + 1)} [1 + O(c^{N+2n} e^{-\sqrt{c}})| \end{aligned}$$

where we have used the fact that

$$\int_0^\infty e^{-x} x^\alpha [L_n^{\alpha}(x)]^2 \, dx = \frac{\Gamma(n + \alpha + 1)}{\Gamma(n + 1)}$$

(see Ref. 5, p. 99). It is not hard to show that the contribution to $1/N_{N,n}^{\ 2}$ from integration over the region $c^{-\frac{1}{4}} \leqq x \leqq 1$ is $O(c^p e^{-\sqrt{c}})$ for some $p > 0$. We have then

$$N_{N,n}^{\ 2} \sim \frac{2\sqrt{c}\,\Gamma(n + 1)}{\Gamma(n + N + 1)}. \tag{68}$$

5.3 *Asymptotics for n and c Both Large*

The techniques employed here again follow very closely those used in Ref. 7. We accordingly give a minimum of detail.

We assume that when $n$ and $c$ are both large $\chi$ can be written

$$\chi_{N,n} \sim c^2 + 2\delta c + b_0 + b_1/c + \cdots . \tag{69}$$

The ranges of $n$ and $c$ for which this is valid will appear in the analysis to follow.

In (25) make the substitution $x = t/c$ and replace $\chi$ by (69). One finds

$$\frac{d^2\varphi}{dt^2} + \left(1 + \frac{\frac{1}{4} - N^2}{t^2}\right)\varphi + O\left(\frac{1}{c}\right) = 0$$

and hence for large $c$, $\varphi(t) \sim \sqrt{t}J_N(t)$. We write

$$\varphi_{N,n}^{\ 6}(x) = \sqrt{x}J_N(cx), \qquad 0 \leqq x \leqq \frac{1}{\sqrt{c}}. \tag{70}$$

Returning to (25) with $\chi$ replaced by (69), we observe that the substitution

$$\varphi = \frac{\exp\left[i\left(cx - \frac{\delta}{2}\log\frac{1 - x}{1 + x}\right)\right]}{\sqrt{1 - x^2}}\, v$$

yields $\dfrac{dv}{dx} + O\left(\dfrac{1}{c}\right) = 0$, so that for large $c$, $v$ becomes constant. After multiplying this solution by a complex constant, we take its real part for the next section of $\varphi$. Explicitly we define

$$\varphi_{N,n}^{\ 7}(x) = \sqrt{\frac{2}{\pi c}}\,\frac{\cos\left[cx - \frac{\delta}{2}\log\frac{1 - x}{1 + x} - (N + \frac{1}{2})\frac{\pi}{2}\right]}{\sqrt{1 - x^2}}. \tag{71}$$

Note that when $x = u/\sqrt{c}$ and $c$ is large (71) becomes

$$\varphi_{N,n}{}^{7}\left(\frac{u}{\sqrt{c}}\right) \sim \sqrt{\frac{2}{\pi c}} \cos\left[u\sqrt{c} - (N + \tfrac{1}{2})\frac{\pi}{2}\right].$$

The asymptotic formula for $J_N$ (see Ref. 8, Vol. II, p. 85) shows that

$$\varphi_{N,n}{}^{6}(x = u/\sqrt{c}) \sim \sqrt{\frac{2}{\pi c}} \cos\left[u\sqrt{c} - (N + \tfrac{1}{2})\frac{\pi}{2}\right]$$

also, so that $\varphi_{N,n}{}^{6}$ and $\varphi_{N,n}{}^{7}$ agree for large $c$ in the neighborhood of $x = 1/\sqrt{c}$.

To find an appropriate asymptotic form for $\varphi$ valid near $x = 1$, substitute $\varphi = x^{N+1}e^{ic(1-x)}u$ into (25) with $\chi$ given by (69). Now make the substitution $x = 1 - i\xi/2c$. There results

$$\xi\frac{d^2u}{d\xi^2} + (1 - \xi)\frac{du}{d\xi} - \left(\frac{1}{2} - i\frac{\delta}{2}\right)u + O\left(\frac{1}{c}\right) = 0.$$

Accordingly, we are led to define

$$\varphi_{N,n}{}^{8}(x) = x^{N+\frac{1}{2}}e^{ic(1-x)}\Phi\left[\frac{1}{2} - i\frac{\delta}{2}, 1; -2ic(1 - x)\right] \tag{72}$$

where

$$\Phi(a,b;x) = 1 + \frac{a}{b}\frac{x}{1!} + \frac{a(a + 1)}{b(b + 1)}\frac{x^2}{2!} + \cdots$$

is the confluent hypergeometric function in the notation of Ref. 8, Vol. I, Chap. 6.

The solution (72) is real. Its asymptotic form for large $c$ when $x = 1 \pm v/\sqrt{c}$ can be found from the known† asymptotics for the $\Phi$ function. One finds

$$\begin{aligned}\varphi_{N,n}{}^{8}(x = 1 \pm v/\sqrt{c}) &\sim \frac{\sqrt{2}e^{\pm\delta(\pi/4)}}{\sqrt{v}c^{\frac{1}{4}}R(\delta)} \\ &\qquad \cos\left[v\sqrt{c} \mp \frac{\delta}{2}\log(2v\sqrt{c}) \pm \theta(\delta) - \frac{\pi}{4}\right]\end{aligned} \tag{73}$$

where the real functions $R(\delta)$ and $\theta(\delta)$ are defined by

$$\Gamma\left(\frac{1}{2} + i\frac{\delta}{2}\right) = R(\delta)e^{i\theta(\delta)}. \tag{74}$$

This latter definition is made precise by requiring $\theta(\delta)$ to be continuous with $\theta(0) = 0$.

Now when $x = 1 - v/\sqrt{c}$, (71) shows that

† See Ref. 8, Vol. I, p. 278, Eq. (2).

$$\varphi_{N,n}{}^{7}(x = v/\sqrt{c}) \sim \frac{1}{c^{\frac{1}{4}}\sqrt{2\pi v}} \cos\left[v\sqrt{c} + \frac{\delta}{2}\log\frac{v}{2\sqrt{c}} + (N + \tfrac{1}{2})\frac{\pi}{2} - c\right]. \tag{75}$$

Comparison of this expression with (73) shows that $\varphi_{N,n}{}^{7}$ and $(-1)^{q} \cdot R(\delta)e^{\delta\pi/4}\varphi_{N,n}{}^{8}/2\sqrt{\pi}$ are asymptotically the same for $x = 1 - v/\sqrt{c}$ provided

$$c + \delta\log(2\sqrt{c}) - \theta(\delta) = (N + 1)(\pi/2) + \pi q \tag{76}$$

with $q$ an integer to be determined shortly.

Quite analogous to (71) is the solution for $x > 1$,

$$\varphi_{N,n}{}^{9}(x) = \frac{e^{\delta(\pi/2)}}{\sqrt{\pi c}} \frac{\cos\left[cx - \dfrac{\delta}{2}\log\dfrac{x-1}{x+1} - (N + 1)\dfrac{\pi}{2} - \dfrac{\pi}{4}\right]}{\sqrt{x^{2} - 1}}. \tag{77}$$

When $x = 1 + v/\sqrt{c}$ and $c$ is large, this solution becomes

$$\varphi_{N,n}{}^{9}(x = 1 + v/\sqrt{c}) \sim \frac{e^{\delta(\pi/2)}}{c^{\frac{1}{4}}\sqrt{2\pi v}} \cos\left[c + v\sqrt{c} - \frac{\delta}{2}\log\frac{v}{2\sqrt{c}} - (N + 1)\frac{\pi}{2} - \frac{\pi}{4}\right].$$

Comparison with (73) shows that this is the same as $(-1)^{q}(R(\delta)e^{\delta\pi/4}/2 \cdot \sqrt{\pi})\varphi_{N,n}{}^{8}(x = 1 + v/\sqrt{c})$ when account is taken of (76).

Our results thus far can be summarized as follows:

$$\varphi_{N,n}(x) \sim \begin{cases} \sqrt{x}J_{N}(cx), & 0 \leqq x \leqq c^{-\frac{1}{2}} \\[2ex] \sqrt{\dfrac{2}{\pi c}} \dfrac{\cos\left[cx - \dfrac{\delta}{2}\log\dfrac{1-x}{1+x} - (N + \frac{1}{2})\dfrac{\pi}{2}\right]}{\sqrt{1 - x^{2}}}, & c^{-\frac{1}{2}} \leqq x \leqq 1 - c^{-\frac{1}{2}} \\[2ex] \dfrac{(-1)^{q}R(\delta)e^{\delta(\pi/4)}}{2\sqrt{\pi}} x^{N+\frac{1}{2}}e^{ic(1-x)} \Phi\left[\dfrac{1}{2} - i\dfrac{\delta}{2}, 1; -2ic(1 - x)\right], & |x - 1| \leqq c^{-\frac{1}{2}} \\[2ex] \dfrac{e^{\delta(\pi/2)}}{\sqrt{\pi c}} \dfrac{\cos\left[cx - \dfrac{\delta}{2}\log\dfrac{x-1}{x+1} - (N + 1)\dfrac{\pi}{2} - \dfrac{\pi}{4}\right]}{\sqrt{x^{2} - 1}} & x \geqq 1 + c^{-\frac{1}{2}} \end{cases} \tag{78}$$

is the asymptotic form for large $n$ and $c$ of a continuous solution of (25) provided $\delta$ and $q$ are chosen to satisfy (76) and the requirement that $\varphi$ as given by (78) has $n$ zeros in the open $x$-interval (0,1). The corresponding eigenvalue is given by $\chi_{N,n} \sim c^2 + 2\delta c + O(1)$. Higher-order terms can be found by methods analogous to those presented in Ref. 7.

When $c$ becomes large and $\delta$ remains fixed, i.e., $\delta = O(1)$, the number of zeros of $\varphi_{N,n}(x)$ in $0 < x \leqq 1$ can be estimated roughly from (78). Using the asymptotic expansion for $J_N$, we find that $\varphi_{N,n}{}^6(x)$ contributes

$$z_6 = (\sqrt{c}/\pi) + O(1)$$

zeros as $x$ ranges from zero to $1/\sqrt{c}$. From $\varphi_{N,n}{}^7(x)$ we find

$$z_7 = (1/\pi)[c - 2\sqrt{c} + (\delta/2) \log \sqrt{c}] + O(1)$$

zeros for $1/\sqrt{c} \leqq x \leqq 1 - 1/\sqrt{c}$. Finally, by using the asymptotic form (73) for $\varphi_{N,n}{}^8$, the number of zeros of $\varphi$ for $1 - 1/\sqrt{c} \leqq x \leqq 1$ is estimated as

$$z_8 = (1/\pi)[\sqrt{c} + (\delta/2) \log \sqrt{c}] + O(1).$$

Since we must have $n = z_6 + z_7 + z_8$, the last three equations show that

$$n\pi = c + \delta \log 2\sqrt{c} + O(1).$$

Combined with (76) this implies that as $c \to \infty$,

$$\theta(\delta) + (N + 1)(\pi/2) + \pi q - n\pi = O(1). \tag{79}$$

The equation just established can be used to obtain a limiting result. Let $N$ be fixed and suppose that $n$ grows with $c$ according to

$$n = (1/\pi)[c + b \log (2\sqrt{c})] \tag{80}$$

where $b$ is a fixed number (independent of $c$). Multiply this equation by $\pi$, add to (76) and rearrange to obtain

$$(\delta - b) \log (2\sqrt{c}) = \theta(\delta) + (N + 1)(\pi/2) + \pi q - n\pi = O(1) \tag{81}$$

where the last equality comes from (79). Divide (81) by $\log(2\sqrt{c})$. We then obtain the limit result: *if $n$ grows with $c$ according to* (80), *then*

$$\lim_{c\to\infty} \delta = b. \tag{82}$$

## VI. ASYMPTOTICS OF $\gamma_{N,n}$ AND $\lambda_{N,n}$

### 6.1 *Fixed N and n. Large c*

The asymptotic solution $\hat{\varphi}_{N,n}(x)$ given in (67) has the values

$$\hat{\varphi}_{N,n}(1) = k_3$$

$$\hat{\varphi}_{N,n}(x \to \infty) \sim (-1)^n k_5 \frac{\cos [cx - (N+1)(\pi/2) - (\pi/4)}{x}.$$

On recalling the definition given in (49), we see that for large $c$

$$\varphi_{N,n}^*(x) \sim (-1)^n \hat{\varphi}_{N,n}(x)/ck_5,$$

so that for fixed $n$ and $N$ as $c$ becomes large

$$\varphi_{N,n}^*(1) \sim \frac{(-1)^n k_3}{ck_5} = (-1)^n \sqrt{\frac{\pi}{2c}}. \tag{83}$$

Equation (50) then gives

$$\gamma_{N,n} \sim \frac{(-1)^n}{\sqrt{c}}. \tag{84}$$

We now proceed to use (84) and the useful formula (to be established)

$$\frac{\partial \gamma_{N,n}}{\partial c} = \frac{\gamma_{N,n}}{2c} [\phi_{N,n}^2(1) - 1] \tag{85}$$

where

$$\int_0^1 \varphi_{N,n}^2(x)\, dx = 1 \tag{86}$$

to get a much stronger statement regarding the asymptotic behavior of $\gamma_{N,n}$. First we establish (85)–(86).

For simplicity of notation let us write (45) as

$$\gamma_n \varphi_n(x) = \int_0^1 K(cxx')\varphi_n(x')\, dx' \tag{87}$$

where we have suppressed dependences on $N$. Differentiating, we find

$$\begin{aligned} &\frac{\partial \gamma_n}{\partial c} \varphi_n(x) + \gamma_n \frac{\partial \varphi_n(x)}{\partial c} \\ &\qquad = \int_0^1 xx'K'(cxx')\varphi_n(x')\, dx' + \int_0^1 K(cxx') \frac{\partial \varphi_n(x')}{\partial c}\, dx'. \end{aligned} \tag{88}$$

Differentiating (87) with respect to $x$ gives

$$\gamma_n \varphi_n'(x) = \frac{c}{x} \int_0^1 xx'K'(cxx')\varphi_n(x')\, dx',$$

so that (88) becomes

$$\frac{\partial \gamma_n}{\partial c}\,\varphi_n(x) + \gamma_n\,\frac{\partial \varphi_n(x)}{\partial c} = \frac{x}{c}\,\gamma_n\varphi_n{}'(x) + \int_0^1 K(cxx')\,\frac{\partial \varphi_n(x')}{\partial c}\,dx'.$$

Multiply this equation by $\varphi_n(x)$ and integrate. One finds

$$\frac{\partial \gamma_n}{\partial c}\int_0^1 \varphi_n{}^2(x)\,dx + \gamma_n\int_0^1 \varphi_n(x)\,\frac{\partial \varphi_n(x)}{\partial c}\,dx$$
$$= \gamma_n\int_0^1 \frac{x}{2c}\,\frac{d}{dx}\,\varphi_n{}^2(x)\,dx + \gamma_n\int_0^1 \varphi_n(x')\,\frac{\partial \varphi_n(x')}{\partial c}\,dx',$$

where the last term has been obtained by interchange of orders of integration and use of (87). Equation (85) then follows by integrating the first term on the right by parts and by using (86).

To use effectively (85)–(86) it is convenient to introduce $\kappa_{N,n} \equiv (-1)^n\cdot\sqrt{c}\gamma_{N,n}$. We then have

$$\frac{1}{\kappa_{N,n}}\,\frac{\partial \kappa_{N,n}}{\partial c} = \frac{1}{2c}\,\varphi_{N,n}{}^2(1) \tag{89}$$

$$\lim_{c\to\infty} \kappa_{N,n} = 1 \tag{90}$$

from (85) and (84) respectively. From (67) and (68) we see that

$$\varphi_{N,n}{}^2(1) \sim k_3{}^2 N_{N,n}{}^2 = \frac{\pi 2^{2N+4n+4} c^{N+2n+2} e^{-2c}}{\Gamma(n+1)\Gamma(n+N+1)}.$$

Using this expression in (89) and integrating, we obtain

$$\log \kappa_{N,n}\,\Big|_c^\infty = \frac{\pi 2^{2N+4n+3}}{\Gamma(n+1)\Gamma(n+N+1)}\int_c^\infty t^{N+2n+1} e^{-2t}\,dt.$$

Integrating by parts and using (90), we finally find

$$\gamma_{N,n} = \frac{(-1)^n}{\sqrt{c}} - \frac{(-1)^n \pi 2^{2N+4n+2} c^{N+2n+\frac{1}{2}} e^{-2c}}{\Gamma(n+1)\Gamma(n+N+1)}\left[1 + O\left(\frac{1}{c}\right)\right]. \tag{91}$$

In terms of $\lambda$ of (13), we find from (17) and (19)

$$\lambda_{N,n} = c\gamma_{N,n}{}^2 \tag{92}$$

so that

$$\lambda_{N,n} = 1 - \frac{\pi 2^{2N+4n+3} c^{N+2n+1} e^{-2c}}{\Gamma(n+1)\Gamma(n+N+1)}\left[1 + O\left(\frac{1}{c}\right)\right]. \tag{93}$$

### 6.2 *Fixed N and n. Small c.*

We use (46) to obtain an expression for $\gamma_{N,n}$ for small $c$. From (42), it follows that

$$
\begin{aligned}
d_o^{N,n}(c) &= \sum_{l=n}^{\infty} A_{-n}^{\,l}(N,n)c^{2l} = A_{-n}^{\,n}(N,n)c^{2n}[1 + O(c^2)] \\
&= \frac{(-1)^n\Gamma(n+1)\Gamma(n+N+1)\Gamma(N+2)}{2^{2n}\Gamma(2n+N+1)\Gamma(2n+N+2)}\, c^{2n}[1+O(c^2)]
\end{aligned}
\tag{94}
$$

where we have used (124), (26) and (32). From (42) one has

$$
\sum_{j=0} d_j^{N,n} = d_n^{N,n} + O(c^2) = 1 + O(c^2). \tag{95}
$$

Equations (94), (95) and (46) now give

$$
\gamma_{N,n} = \frac{(-1)^n\Gamma(n+1)\Gamma(n+N+1)}{2^{2n+N+1}\Gamma(2n+N+1)\Gamma(2n+N+2)}\, c^{2n+N+\frac{1}{2}}. \tag{96}
$$

Higher-order terms could be obtained in a similar manner. An alternative route, however, is to use (85) and (86). From (39) and (30), one sees that $[2(2n+N+1)]^{\frac{1}{2}}\binom{n+N}{n}(-1)^n[1+O(c^4)]$ is the normalization factor for (39). Using (39) one then finds for a normalized solution

$$
\begin{aligned}
\varphi_{N,n}(1) &= (-1)^n\sqrt{2(2n+N+1)} \\
&\qquad \left[1 + \frac{N^2c^2}{4(2n+N)^2(2n+N+2)^2}\right] + O(c^4).
\end{aligned}
$$

Inserting this expression in (85) and integrating, we find

$$
\begin{aligned}
\gamma_{N,n} &= \frac{(-1)^n\Gamma(n+1)\Gamma(n+N+1)c^{2n+N+\frac{1}{2}}}{2^{2n+N+1}\Gamma(2n+N+1)\Gamma(2n+N+2)} \\
&\qquad \cdot\left[1 + \frac{N^2c^2}{4(2n+N)^2(2n+N+2)^2} + O(c^4)\right].
\end{aligned}
\tag{97}
$$

### 6.3 *Asymptotics for n and c Both Large*

To obtain an expression for $\gamma_{N,n}$ valid for $n$ and $c$ both large, we use (77) and (49–50). For the asymptotic solution (77) we have

$$
\varphi_{N,n}(1) \sim \frac{R(\delta)e^{\delta(\pi/4)}(-1)^q}{2\sqrt{\pi}} \tag{98}
$$

and for very large $x$

$$
\varphi_{N,n}(x \to \infty) \sim \frac{e^{\delta(\pi/2)}}{\sqrt{\pi c}}\,\frac{\cos[cx - (N+1)(\pi/2) - (\pi/4)]}{x}.
$$

Comparison with (49) shows that $\varphi_{N,n}^{\;*} = \sqrt{\pi/c}\,e^{-\delta(\pi/2)}\varphi_{N,n}$, and (98) and (50) now give

$$\gamma_{N,n} \sim \frac{(-1)^q R(\delta) e^{-\delta(\pi/4)}}{\sqrt{2\pi c}}. \tag{99}$$

Now (91) and (97) show that for large and small $c$ the sign of $\gamma_{N,n}$ is the same as the sign of $(-1)^n$. As $c$ varies, $\gamma_{N,n}$ cannot change sign, for by (92) if $\gamma_{N,n}$ were to vanish for some value of $c \neq 0$, so would $\lambda_{N,n}$. Since, as we have noted in Sections II and III, the kernel $K_c$ of (12) is positive definite, this is impossible. We can therefore replace $q$ by $n$ in (99) and we have

$$q = n (\mathrm{mod}\ 2). \tag{100}$$

From the definition (74) of $R(\delta)$, one has

$$[R(\delta)]^2 = \Gamma\left(\frac{1}{2} + i\,\frac{\delta}{2}\right) \Gamma\left(\frac{1}{2} - i\,\frac{\delta}{2}\right) = \frac{\pi}{\cosh \delta(\pi/2)}. \tag{101}$$

Here we have used the functional relation [Ref. 8, Vol. I, p. 3, Eq. (7)] for the gamma function

$$\Gamma(\tfrac{1}{2} + z)\Gamma(\tfrac{1}{2} - z) = \pi \sec \pi z.$$

Equations (99), (100), (101) and (92) combined are

$$\gamma_{N,n} \sim \frac{(-1)^n}{\sqrt{c(1 + e^{\pi\delta})}}, \qquad \lambda_{N,n} \sim \frac{1}{1 + e^{\pi\delta}}. \tag{102}$$

Finally from (80), (82) and (102) we have the limiting result: if

$$n = [(1/\pi)(c + b \log 2\sqrt{c})]$$

where the brackets denote "largest integer in" and $b$ is a fixed number, then

$$\lim_{c \to \infty} \lambda_{N,n} = \frac{1}{1 + e^{\pi b}}. \tag{103}$$

## VII. THE CASE $D > 2$, $R$ THE UNIT SPHERE

In the previous sections, we have treated the important special case $D = 2$, $R$ the unit circle in considerable detail. Most of the analysis there was concerned with solving the integral equation (20). Fortunately, as we shall now see, the solution of that equation also affords a complete solution of the case $R$ the unit sphere centered at the origin in $E_D$, $D = 3, 4, \ldots$. In treating this general case, we shall draw freely on the theory of $D$-dimensional spherical harmonics as given, for example, in Ref. 8, Vol. II, Chap. XI. We follow the notation of this work and set

$$D = p + 2, \qquad p = 1, 2, \ldots. \tag{104}$$

Let $\mathbf{x} = r\boldsymbol{\xi}$ and $\mathbf{y} = r'\mathbf{n}$ where $\boldsymbol{\xi}$ and $\mathbf{n}$ are unit vectors in $E_{p+2}$. Equation (8) now becomes

$$\alpha\psi(r,\boldsymbol{\xi}) = \int_0^1 dr' r'^{p+1} \int_\Omega \exp\,(icrr'\boldsymbol{\xi}\cdot\mathbf{n})\psi(r',\mathbf{n})\,d\Omega(\mathbf{n}) \tag{105}$$

where $\Omega$ is the surface of the unit sphere in $E_{p+2}$.

Now let

$$h(N,p) = (2N + p)\,\frac{(N + p - 1)!}{p!N!}, \qquad N = 0, 1, 2, \ldots, \tag{106}$$

and let $S_N^{\,l}(\boldsymbol{\xi})$, $l = 1, 2, \cdots, h(N,p)$, be a complete set of orthonormal surface harmonics of degree $N$. The Funk-Hecke theorem (Ref. 8, Vol. II, pp. 247–248) asserts that

$$\int_\Omega \exp\,(icrr'\boldsymbol{\xi}\cdot\mathbf{n})S_N^{\,l}(\mathbf{n})\,d\Omega(\mathbf{n}) = H_N(crr')S_N^{\,l}(\boldsymbol{\xi}) \tag{107}$$

where

$$H_N(crr') = \frac{2\pi^{(p+1)/2}N!\,(p-1)!}{\Gamma\left(\frac{p+1}{2}\right)(N+p-1)!}\int_{-1}^{1} e^{icrr'u}C_N^{\,p/2}(u)\,(1-u^2)^{(p-1)/2}\,du \tag{108}$$

is independent of $l$ and $C_N^{\,\eta}(u)$ is a Gegenbauer polynomial (Ref. 8, Vol. II, p. 235). By expanding $\psi$ in surface harmonics,

$$\psi(r,\boldsymbol{\xi}) = \sum_{N=0}^{\infty}\sum_{l=0}^{h(N,p)} R_{N,l}(r)S_N^{\,l}(\boldsymbol{\xi}),$$

we find from (105) and (107)

$$\alpha_{N,l}R_{N,l}(r) = \int_0^1 dr' r'^{p+1}H_N(crr')R_{N,l}(r'), \tag{109}$$

from which it is seen that $R_{N,l}(r)$ and $\alpha_{N,l}$ are independent of $l$. We have the expected degeneracy of eigenvalues due to spherical symmetry.

Now [Ref. 8, Vol. II, p. 236, Eq. (25)]

$$C_N^{\,p/2}(u) = \frac{(-1)^N}{2^N}\,\frac{(p)_N}{\left(\frac{p+1}{2}\right)_N N!}\,(1-u^2)^{-(p-1)/2}\,\frac{d^N}{du^N}\,(1-u^2)^{N+(p-1)/2},$$

where $(a)_N = a(a+1)\cdots(a+N-1)$, so that from (108)

$$H_N(crr') = \frac{2\pi^{(p+1)/2}(-1)^N}{\Gamma\left(\frac{p+1}{2}\right)2^N\left(\frac{p+1}{2}\right)_N}\int_{-1}^{1} du\, e^{icrr'u}\,\frac{d^N}{du^N}\,(1-u^2)^{N+(p-1)/2}.$$

Integration by parts gives for the integral here

$$(-1)^N \int_{-1}^{1} du(1 - u^2)^{N+(p-1)/2} \frac{d^N}{du^N} e^{icrr'u}$$
$$= (-icrr')^N \int_{-1}^{1} du\, e^{icrr'u}(1 - u^2)^{N+(p-1)/2}$$
$$= (-i)^N \sqrt{\pi}\Gamma\left(N + \frac{p+1}{2}\right) 2^{N+p/2}(crr')^{-p/2} J_{N+p/2}(crr')$$

where we have used the Poisson formula

$$\Gamma(\nu + \tfrac{1}{2})J_\nu(z) = \pi^{-\frac{1}{2}}(z/2)^\nu \int_{-1}^{1} e^{izu}(1 - u^2)^{\nu-\frac{1}{2}}\, du$$

[Ref. 8, Vol. II, p. 81, Eq. (7)]. We have then, finally

$$H_N(crr') = i^N(2\pi)^{1+p/2} J_{N+p/2}(crr')/(crr')^{p/2}.$$

We see now from (109) that the eigenfunctions and eigenvalues of (105) are

$$\begin{aligned} \psi_{N,l,n}(r,\xi) &= R_{N,n}(r)S_N{}^l(\xi), \qquad l = 1, 2, \dots, h(N,p) \\ \alpha_{N,n} &= i^N(2\pi)^{1+p/2}\beta_{N,n} \\ N,n &= 0, 1, 2, \dots \end{aligned} \tag{110}$$

where

$$\beta_{N,n}R_{N,n}(r) = \int_0^1 \frac{J_{N+p/2}(crr')}{(crr')^{p/2}} r'^{p+1} R_{N,n}(r')\, dr'. \tag{111}$$

These equations are the analogues of (16), (17) and (18). Set

$$\gamma = \beta c^{(p+1)/2}, \qquad \varphi = r^{(p+1)/2}R. \tag{112}$$

Equation (111) becomes

$$\gamma\varphi(r) = \int_0^1 J_{N+p/2}(crr')\sqrt{crr'}\varphi(r')\, dr'. \tag{113}$$

This, however, is (20) with $N$ replaced by $N + p/2$. The formulae of Section IV for the solutions of (20) can be taken over exactly replacing $N$ by $N + p/2$ throughout. (Expressions involving factorials must be replaced by the appropriate ones in terms of $\Gamma$ functions when $p$ is an odd integer.) Together with (110), (111) and (112), they provide solution of (105) for all $D \geqq 2$.

It is interesting to note that the one-dimensional case treated in Refs. 1 and 2 can be obtained as a special case of the present theory by appropriate interpretation. The parameter $N$ of this section is the degree of the homogeneous polynomial solution to Laplace's equation in $D$ dimension afforded by the spherical harmonic $S_N^{\,l}$ when expressed in rectangular coordinates. When $D = 1$, Laplace's equation $d^2\psi/dx^2 = 0$ has only two homogeneous solutions, $\psi = k$ and $\psi = x$, respectively of degrees zero and one. For $D = 1$, i.e., $p = -1$ from (104), we have

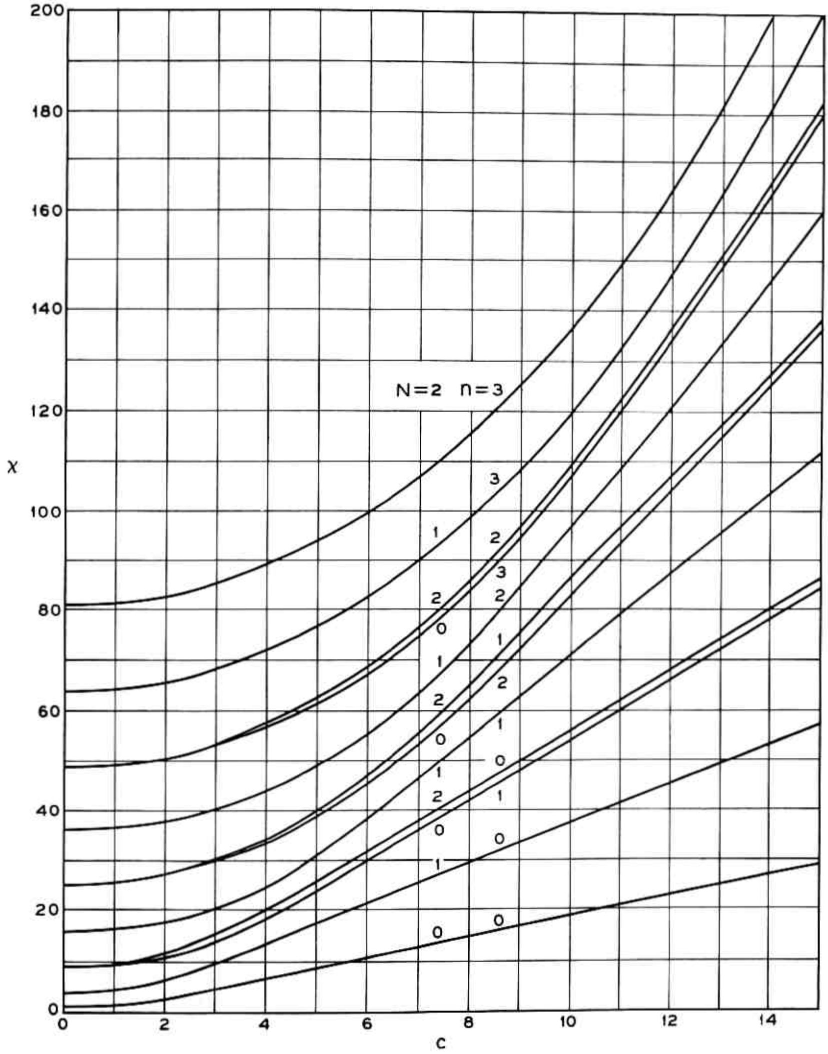


Fig. 1 — Curves of $\chi_{N,n}$ of (25) vs $c$.

only two allowed values, $N = 0$ and $N = 1$. The quantity $N + p/2$ occurring in (113) then has values $-\frac{1}{2}$ and $\frac{1}{2}$. The kernel becomes $\sqrt{2/\pi} \cos crr'$ and $\sqrt{2/\pi} \sin crr'$ respectively in these two cases, and we retrieve the integral equations for the even and odd prolate spheroidal functions of zero order. Note that when $N = \pm\frac{1}{2}$, (25) reduces to the prolate spheroidal equation.

## VIII. NUMERICAL RESULTS

A program for the IBM 7090 has been written to compute generalized prolate spheroidal functions using formulae (40) and (44). Trial values for the $\chi_{N,n}$ were obtained from (34) and (55) and the recurrences (36)–(37) and (59). The method of Bouwkamp[6] was then used to correct these estimates and obtain the $d_j^{N,n}$. Values of $\gamma_{N,n}$ were obtained from (46) and these were converted to values of $\lambda$ by $\lambda_{N,n} = c\gamma_{N,n}^2$.

Fig. 1 shows plots of $\chi_{N,n}$ versus $c$. Fig. 2 gives the behavior of the first few $\lambda_{N,n}$. By definition of the labels, $\chi_{N,n+1} \geqq \chi_{N,n}$ for $N,n = 0, 1, \ldots$ and if $c > 0$ the inequality is strict. From Sturmian theory, it follows that $\chi_{N+1,n} > \chi_{N,n}$. For the $\lambda$'s, one can show correspondingly that $\lambda_{N,n+1} < \lambda_{N,n}$ and $\lambda_{N+1,n} < \lambda_{N,n}$ for $N,n = 0, 1, \ldots$. The problem of ordering the $\lambda$'s and $\chi$'s for all $N$ and $n$ appears to be a difficult one. Some values are listed on Table I.

Figs. 3 and 4 show plots of $\varphi_{N,n}(x)$ versus $x$ for $N = 0,2$, $n = 0,1,2,3$ and $c = 2,10$. Values of the $\varphi_{N,n}$ for a larger set of parameter values are given in Table II. Normalization is as in (86).

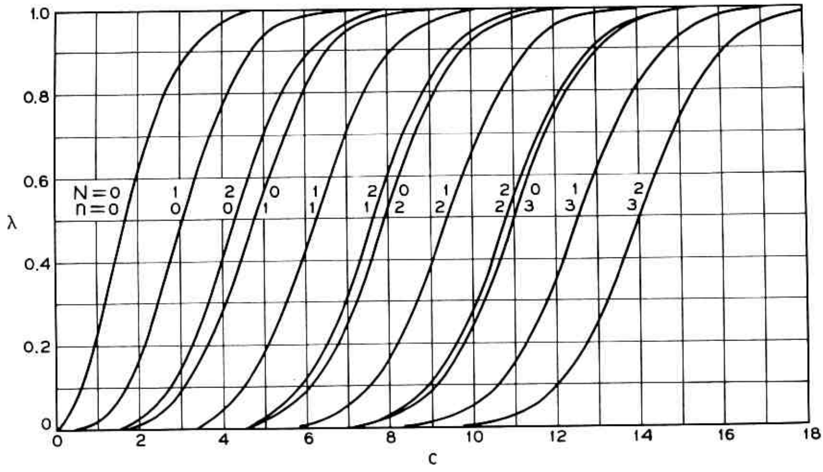


Fig. 2 — Curves of $\lambda_{N,n}$ of (13) and (15) vs $c$.

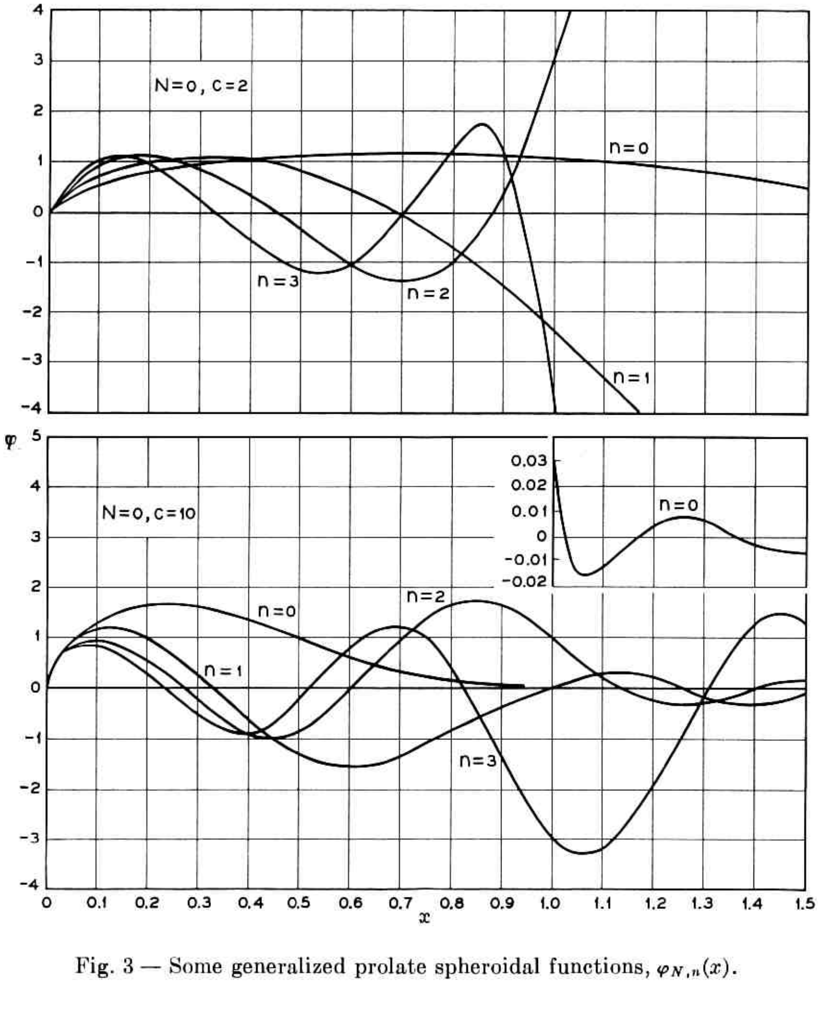


Fig. 3 — Some generalized prolate spheroidal functions, $\varphi_{N,n}(x)$.

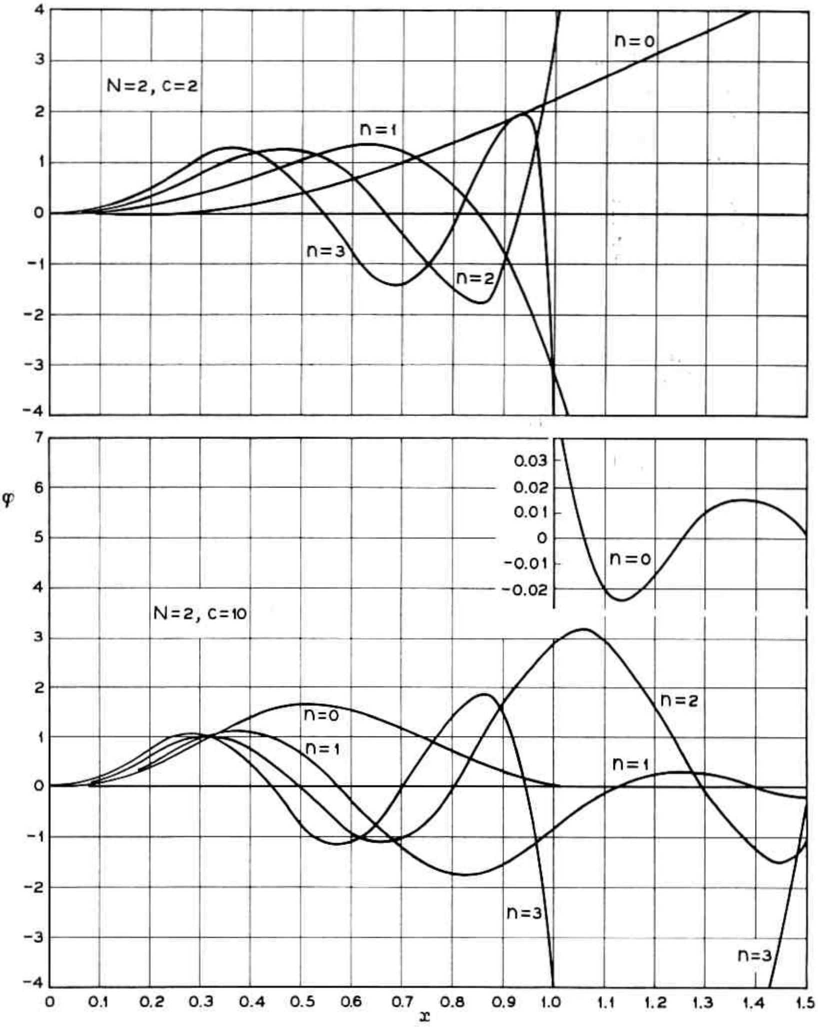


Fig. 4 — More generalized prolate spheroidal functions, $\varphi_{N,n}(x)$.

TABLE I — NUMERICAL VALUES OF $\chi_{N,n}$ AND $\lambda_{N,n}$

| $c$ | $\chi$ | $\lambda$ | $c$ | $\chi$ | $\lambda$ |
|---|---|---|---|---|---|
| | $N = 0 \quad n = 0$ | | | $N = 1 \quad n = 0$ | |
| 0.1 | 7.5499895 − 1 | 2.4968775 − 3 | 0.5 | 3.9163765 + 0 | 9.4982658 − 4 |
| 0.5 | 8.7434899 − 1 | 6.0585348 − 2 | 1.0 | 4.4119661 + 0 | 1.3986168 − 2 |
| 1.0 | 1.2395933 + 0 | 2.2111487 − 1 | 2.0 | 6.3394615 + 0 | 1.6123183 − 1 |
| 1.5 | 1.8225178 + 0 | 4.2951906 − 1 | 3.0 | 9.3427678 + 0 | 4.8326866 − 1 |
| 2.0 | 2.5857968 + 0 | 6.2963045 − 1 | 4.0 | 1.3086855 + 1 | 7.8473505 − 1 |
| 3.0 | 4.4622709 + 0 | 8.8705036 − 1 | 5.0 | 1.7170130 + 1 | 9.3671678 − 1 |
| 4.0 | 6.5208586 + 0 | 9.7495117 − 1 | 6.0 | 2.1310500 + 1 | 9.8534266 − 1 |
| 5.0 | 8.5869176 + 0 | 9.9534230 − 1 | 10.0 | 3.7555900 + 1 | 9.9998314 − 1 |
| 10.0 | 1.8690110 + 1 | 9.9999957 − 1 | | | |
| | $N = 0 \quad n = 1$ | | | $N = 1 \quad n = 1$ | |
| 1 | 9.2562398 + 0 | 1.0829815 − 4 | 1 | 1.6255011 + 1 | 7.4672551 − 7 |
| 2 | 1.0847476 + 1 | 6.7214485 − 3 | 2 | 1.7912353 + 1 | 1.8549511 − 4 |
| 3 | 1.3698728 + 1 | 6.6745424 − 2 | 4 | 2.4832293 + 1 | 3.8313651 − 2 |
| 4 | 1.7898720 + 1 | 2.6742780 − 1 | 5 | 3.0401459 + 1 | 1.6818804 − 1 |
| 5 | 2.3241561 + 1 | 5.7877057 − 1 | 6 | 3.7326440 + 1 | 4.2912557 − 1 |
| 6 | 2.9277622 + 1 | 8.3060712 − 1 | 7 | 4.5219234 + 1 | 7.1473948 − 1 |
| 7 | 3.5550580 + 1 | 9.4973850 − 1 | 8 | 5.3565692 + 1 | 8.9618892 − 1 |
| 8 | 4.1805821 + 1 | 9.8782700 − 1 | 9 | 6.1976089 + 1 | 9.7041388 − 1 |
| 9 | 4.7985976 + 1 | 9.9738554 − 1 | 10 | 7.0297509 + 1 | 9.9279210 − 1 |
| 10 | 5.4108072 + 1 | 9.9947801 − 1 | | | |
| | $N = 0 \quad n = 2$ | | | $N = 1 \quad n = 2$ | |
| 1 | 2.5751488 + 1 | 1.8834675 − 9 | 1 | 3.6265101 + 1 | 4.6976877 − 12 |
| 2 | 2.6773866 + 1 | 1.9235204 − 6 | 2 | 3.7820310 + 1 | 1.9082396 − 8 |
| 5 | 3.8241737 + 1 | 1.6017987 − 2 | 5 | 4.9160037 + 1 | 1.0305314 − 3 |
| 6 | 4.4846367 + 1 | 8.1254764 − 2 | 6 | 5.5464880 + 1 | 8.3652966 − 3 |
| 7 | 5.3021146 + 1 | 2.5847455 − 1 | 7 | 6.3286568 + 1 | 4.4641026 − 2 |
| 8 | 6.2527715 + 1 | 5.3544699 − 1 | 8 | 7.2759605 + 1 | 1.6080875 − 1 |
| 9 | 7.2854528 + 1 | 7.8635574 − 1 | 9 | 8.3789365 + 1 | 3.9082845 − 1 |
| 10 | 8.3461406 + 1 | 9.2600949 − 1 | 10 | 9.5955815 + 1 | 6.6399691 − 1 |
| 11 | 9.4019226 + 1 | 9.7915064 − 1 | 11 | 1.0867089 + 2 | 8.6120123 − 1 |
| 12 | 1.0443896 + 2 | 9.9484586 − 1 | 12 | 1.2145589 + 2 | 9.5495434 − 1 |
| 13 | 1.1474313 + 2 | 9.9882732 − 1 | 13 | 1.3409696 + 2 | 9.8759852 − 1 |
| 14 | 1.2496987 + 2 | 9.9974793 − 1 | 14 | 1.4657506 + 2 | 9.9692509 − 1 |
| 15 | 1.3514611 + 2 | 9.9994820 − 1 | | | |
| 16 | 1.4528810 + 2 | 9.9998984 − 1 | | | |
| | $N = 0 \quad n = 3$ | | | $N = 1 \quad n = 3$ | |
| 1 | 4.9250694 + 1 | 6.0066949 − 15 | 1 | 6.4258409 + 1 | 7.6540787 − 18 |
| 2 | 5.0761114 + 1 | 9.8333952 − 11 | 2 | 6.5789319 + 1 | 4.9988893 − 13 |
| 5 | 6.1688709 + 1 | 3.5422330 − 5 | 5 | 7.6749767 + 1 | 1.1190662 − 6 |
| 7 | 7.4995083 + 1 | 3.5278392 − 3 | 9 | 1.0834214 + 2 | 1.0298512 − 2 |
| 8 | 8.3823340 + 1 | 2.0130790 − 2 | 10 | 1.2001776 + 2 | 4.5573797 − 2 |
| 9 | 9.4336396 + 1 | 8.2918248 − 2 | 11 | 1.3350648 + 2 | 1.5017502 − 1 |
| 10 | 1.0659367 + 2 | 2.4212641 − 1 | 12 | 1.4872078 + 2 | 3.5789707 − 1 |
| 11 | 1.2034708 + 2 | 4.9658387 − 1 | 13 | 1.6522672 + 2 | 6.1969392 − 1 |
| 12 | 1.3504432 + 2 | 7.4660703 − 1 | 14 | 1.8237982 + 2 | 8.2818634 − 1 |
| 13 | 1.5007176 + 2 | 9.0244395 − 1 | 15 | 1.9961900 + 2 | 9.3866345 − 1 |
| 14 | 1.6502439 + 2 | 9.6944048 − 1 | 16 | 2.1666412 + 2 | 9.8150785 − 1 |
| 15 | 1.7977291 + 2 | 9.9165294 − 1 | 17 | 2.3347382 + 2 | 9.9500835 − 1 |
| 16 | 1.9433894 + 2 | 9.9791376 − 1 | | | |

TABLE I — *Continued*

| $c$ | $\chi$ | $\lambda$ | $c$ | $\chi$ | $\lambda$ |
|---|---|---|---|---|---|
|  | $N = 2$ $n = 0$ |  |  | $N = 2$ $n = 2$ |  |
| 1 | 9.4976317 + 0 | 3.9517707 − 4 | 1 | 4.9292042 + 1 | 1.0614368 − 14 |
| 2 | 1.1710916 + 1 | 1.9088335 − 2 | 2 | 5.0922802 + 1 | 1.7074109 − 10 |
| 3 | 1.5291960 + 1 | 1.3627864 − 1 | 5 | 6.2566028 + 1 | 5.5396960 − 5 |
| 4 | 2.0048498 + 1 | 4.0298411 − 1 | 7 | 7.6509011 + 1 | 5.0165514 − 3 |
| 5 | 2.5667098 + 1 | 7.0317221 − 1 | 8 | 8.5676381 + 1 | 2.7077307 − 2 |
| 6 | 3.1747966 + 1 | 8.9417071 − 1 | 9 | 9.6519700 + 1 | 1.0455606 − 1 |
| 7 | 3.7952889 + 1 | 9.7092917 − 1 | 10 | 1.0904728 + 2 | 2.8443961 − 1 |
| 8 | 4.4125829 + 1 | 9.9324606 − 1 | 11 | 1.2295597 + 2 | 5.4626700 − 1 |
| 10 | 5.6324064 + 1 | 9.9972026 − 1 | 12 | 1.3768257 + 2 | 7.8216583 − 1 |
|  |  |  | 13 | 1.5265435 + 2 | 9.1919613 − 1 |
|  |  |  | 14 | 1.6752489 + 2 | 9.7528766 − 1 |
|  |  |  | 15 | 1.8220066 + 2 | 9.9334500 − 1 |
|  | $N = 2$ $n = 1$ |  |  | $N = 2$ $n = 3$ |  |
| 1 | 2.5333581 + 1 | 4.1659113 − 9 | 1 | 8.1275289 + 1 | 9.2115325 − 21 |
| 2 | 2.7088321 + 1 | 4.0530517 − 6 | 2 | 8.2854667 + 1 | 2.3940665 − 15 |
| 5 | 3.9788041 + 1 | 2.5787154 − 2 | 5 | 9.4065006 + 1 | 3.2594658 − 8 |
| 6 | 4.6842565 + 1 | 1.1592915 − 1 | 10 | 1.3679420 + 2 | 5.8064778 − 3 |
| 7 | 5.5371030 + 1 | 3.2573165 − 1 | 11 | 1.4973956 + 2 | 2.6682390 − 2 |
| 8 | 6.5067655 + 1 | 6.0754452 − 1 | 12 | 1.6450899 + 2 | 9.5057491 − 2 |
| 9 | 7.5425367 + 1 | 8.3161633 − 1 | 13 | 1.8113152 + 2 | 2.5380401 − 1 |
| 10 | 8.5969115 + 1 | 9.4455397 − 1 | 14 | 1.9931616 + 2 | 4.9785798 − 1 |
| 11 | 9.6442775 + 1 | 9.8484078 − 1 | 15 | 2.1846670 + 2 | 7.3845601 − 1 |
|  |  |  | 16 | 2.3793070 + 2 | 8.9389256 − 1 |
|  |  |  | 17 | 2.5727411 + 2 | 9.6462917 − 1 |
|  |  |  | 18 | 2.7635393 + 2 | 9.8968676 − 1 |

TABLE II — VALUES OF $\varphi_{N,n}(x)$

| | | $N = 0$ $n = 0$ | | |
|---|---|---|---|---|
| $x$ | $c = 1$ | $c = 2$ | $c = 5$ | $c = 10$ |
| 0.1 | 4.74638 − 1 | 5.55421 − 1 | 9.15662 − 1 | 1.31455 + 0 |
| 0.2 | 6.68776 − 1 | 7.74706 − 1 | 1.22032 + 0 | 1.62247 + 0 |
| 0.3 | 8.14070 − 1 | 9.27095 − 1 | 1.35165 + 0 | 1.57689 + 0 |
| 0.4 | 9.31948 − 1 | 1.03607 + 0 | 1.35103 + 0 | 1.30428 + 0 |
| 0.5 | 1.03044 + 0 | 1.11011 + 0 | 1.24626 + 0 | 9.31637 − 1 |
| 0.6 | 1.11351 + 0 | 1.15353 + 0 | 1.06660 + 0 | 5.70325 − 1 |
| 0.7 | 1.18341 + 0 | 1.16921 + 0 | 8.43474 − 1 | 2.91331 − 1 |
| 0.8 | 1.24157 + 0 | 1.15957 + 0 | 6.07845 − 1 | 1.17077 − 1 |
| 0.9 | 1.28896 + 0 | 1.12695 + 0 | 3.86969 − 1 | 3.19741 − 2 |
| 1.0 | 1.32627 + 0 | 1.07383 + 0 | 2.01532 − 1 | 3.00159 − 3 |
| 1.1 | 1.35405 + 0 | 1.00285 + 0 | 6.38588 − 2 | −1.09501 − 3 |
| 1.2 | 1.37278 + 0 | 9.16840 − 1 | −2.24980 − 2 | 4.23236 − 4 |
| 1.3 | 1.38285 + 0 | 8.18791 − 1 | −6.18395 − 2 | 6.58696 − 4 |
| 1.4 | 1.38464 + 0 | 7.11797 − 1 | −6.43066 − 2 | −2.21883 − 4 |
| 1.5 | 1.37853 + 0 | 5.98995 − 1 | −4.32142 − 2 | −5.95391 − 4 |
| 1.6 | 1.36489 + 0 | 4.83499 − 1 | −1.22744 − 2 | −1.21194 − 4 |
| 1.7 | 1.34410 + 0 | 3.68328 − 1 | 1.68342 − 2 | 4.20886 − 4 |
| 1.8 | 1.31655 + 0 | 2.56332 − 1 | 3.61314 − 2 | 3.79898 − 4 |
| 1.9 | 1.28264 + 0 | 1.50130 − 1 | 4.20554 − 2 | −8.00341 − 5 |

TABLE II — *Continued*

$N = 0 \quad n = 0$

| $x$ | $c = 1$ | $c = 2$ | $c = 5$ | $c = 10$ |
|---|---|---|---|---|
| 2.0 | 1.24281 + 0 | 5.20429 − 2 | 3.52268 − 2 | −3.67577 − 4 |
| 2.1 | 1.19748 + 0 | −3.59472 − 2 | 1.94309 − 2 | −2.42458 − 4 |
| 2.2 | 1.14713 + 0 | −1.12245 − 1 | 1.30129 − 4 | 1.44876 − 4 |
| 2.3 | 1.09221 + 0 | −1.75666 − 1 | −1.71274 − 2 | 3.08117 − 4 |
| 2.4 | 1.03322 + 0 | −2.25459 − 1 | −2.80628 − 2 | 1.76068 − 4 |
| 2.5 | 9.70655 − 1 | −2.61300 − 1 | −3.05323 − 2 | −1.11428 − 4 |
| 2.6 | 9.05010 − 1 | −2.83289 − 1 | −2.47697 − 2 | −3.05239 − 4 |
| 2.7 | 8.36800 − 1 | −2.91927 − 1 | −1.30522 − 2 | −6.14275 − 5 |
| 2.8 | 7.66537 − 1 | −2.88082 − 1 | 1.07933 − 3 | 1.53741 − 4 |
| 2.9 | 6.94735 − 1 | −2.72956 − 1 | 1.38275 − 2 | 6.59394 − 5 |
| 3.0 | 6.21906 − 1 | −2.48005 − 1 | 2.20781 − 2 | 1.15331 − 4 |

$N = 0 \quad n = 1$

| $x$ | $c = 1$ | $c = 2$ | $c = 5$ | $c = 10$ |
|---|---|---|---|---|
| 0.1 | 7.57682 − 1 | 7.49125 − 1 | 7.51850 − 1 | 1.11517 + 0 |
| 0.2 | 1.00189 + 0 | 9.78062 − 1 | 8.84348 − 1 | 9.64135 − 1 |
| 0.3 | 1.08562 + 0 | 1.03485 + 0 | 7.51864 − 1 | 2.39045 − 1 |
| 0.4 | 1.02660 + 0 | 9.38690 − 1 | 4.08540 − 1 | −6.68891 − 1 |
| 0.5 | 8.24694 − 1 | 6.94825 − 1 | −7.50154 − 2 | −1.35733 + 0 |
| 0.6 | 4.76162 − 1 | 3.08610 − 1 | −6.10130 − 1 | −1.58691 + 0 |
| 0.7 | −2.29984 − 2 | −2.11235 − 1 | −1.10302 + 0 | −1.36404 + 0 |
| 0.8 | −6.75867 − 1 | −8.51754 − 1 | −1.47075 + 0 | −8.87374 − 1 |
| 0.9 | −1.48414 + 0 | −1.59571 + 0 | −1.65550 + 0 | −4.07419 − 1 |
| 1.0 | −2.44790 + 0 | −2.41297 + 0 | −1.63388 + 0 | −9.25172 − 2 |
| 1.1 | −3.56556 + 0 | −3.30607 + 0 | −1.41956 + 0 | 2.63841 − 2 |
| 1.2 | −4.83382 + 0 | −4.22107 + 0 | −1.05901 + 0 | 2.11679 − 2 |
| 1.3 | −6.24769 + 0 | −5.13836 + 0 | −6.20954 − 1 | −1.29202 − 2 |
| 1.4 | −7.80051 + 0 | −6.02868 + 0 | −1.82563 − 1 | −2.10428 − 2 |
| 1.5 | −9.48401 + 0 | −6.86308 + 0 | 1.85537 − 1 | −2.91179 − 3 |
| 1.6 | −1.12884 + 1 | −7.61390 + 0 | 4.31871 − 1 | 1.47146 − 2 |
| 1.7 | −1.32026 + 1 | −8.25574 + 0 | 5.32289 − 1 | 1.36618 − 2 |
| 1.8 | −1.52139 + 1 | −8.76628 + 0 | 4.92313 − 1 | −1.33175 − 3 |
| 1.9 | −1.73086 + 1 | −9.12711 + 0 | 3.43349 − 1 | −1.28433 − 2 |
| 2.0 | −1.94720 + 1 | −9.32424 + 0 | 1.34214 − 1 | −9.24495 − 3 |
| 2.1 | −2.16882 + 1 | −9.34869 + 0 | −8.02848 − 2 | 2.21756 − 3 |
| 2.2 | −2.39406 + 1 | −9.19674 + 0 | −2.51019 − 1 | 1.09934 − 2 |
| 2.3 | −2.62120 + 1 | −8.87007 + 0 | −3.44099 − 1 | 7.19680 − 3 |
| 2.4 | −2.84846 + 1 | −8.37588 + 0 | −3.46476 − 1 | −2.21224 − 3 |
| 2.5 | −3.07405 + 1 | −7.72616 + 0 | −2.66729 − 1 | −6.65003 − 3 |
| 2.6 | −3.29612 + 1 | −6.93789 + 0 | −1.31077 − 1 | −7.12403 − 3 |
| 2.7 | −3.51286 + 1 | −6.03209 + 0 | 2.39279 − 2 | 8.00041 − 4 |
| 2.8 | −3.72245 + 1 | −5.03334 + 0 | 1.60443 − 1 | 8.27644 − 3 |
| 2.9 | −3.92310 + 1 | −3.96899 + 0 | 2.48208 − 1 | 1.58822 − 3 |
| 3.0 | −4.11308 + 1 | −2.86785 + 0 | 2.70290 − 1 | −1.20582 − 3 |

$N = 0 \quad n = 2$

| $x$ | $c = 1$ | $c = 2$ | $c = 5$ | $c = 10$ |
|---|---|---|---|---|
| 0.1 | 9.39351 − 1 | 9.35161 − 1 | 8.84844 − 1 | 9.05937 − 1 |
| 0.2 | 1.08208 + 0 | 1.06262 + 0 | 9.01678 − 1 | 5.16232 − 1 |

TABLE II — *Continued*

$N = 0 \quad n = 2$

| $x$ | $c = 1$ | $c = 2$ | $c = 5$ | $c = 10$ |
|---|---|---|---|---|
| 0.3 | 8.66349 − 1 | 8.22287 − 1 | 5.05390 − 1 | −3.42126 − 1 |
| 0.4 | 3.64401 − 1 | 2.96652 − 1 | −1.35886 − 1 | −9.82679 − 1 |
| 0.5 | −3.05506 − 1 | −3.80932 − 1 | −7.81957 − 1 | −8.93819 − 1 |
| 0.6 | −9.56032 − 1 | −1.00914 + 0 | −1.16120 + 0 | −7.44970 − 2 |
| 0.7 | −1.32240 + 0 | −1.31928 + 0 | −1.03567 + 0 | 9.85253 − 1 |
| 0.8 | −1.05832 + 0 | −9.83017 − 1 | −2.64525 − 1 | 1.65961 + 0 |
| 0.9 | 2.65717 − 1 | 3.74968 − 1 | 1.15147 + 0 | 1.60911 + 0 |
| 1.0 | 3.16217 + 0 | 3.16046 + 0 | 3.05638 + 0 | 9.70106 − 1 |
| 1.1 | 8.22453 + 0 | 7.79020 + 0 | 5.16083 + 0 | 2.03097 − 1 |
| 1.2 | 1.61238 + 1 | 1.46703 + 1 | 7.09600 + 0 | −2.55403 − 1 |
| 1.3 | 2.76035 + 1 | 2.41735 + 1 | 8.48753 + 0 | −2.68864 − 1 |
| 1.4 | 4.34731 + 1 | 3.66166 + 1 | 9.03589 + 0 | −2.22585 − 2 |
| 1.5 | 6.46001 + 1 | 5.22386 + 1 | 8.58154 + 0 | 1.83653 − 1 |
| 1.6 | 9.19010 + 1 | 7.11812 + 1 | 7.14579 + 0 | 1.76198 − 1 |
| 1.7 | 1.26331 + 2 | 9.34716 + 1 | 4.93120 + 0 | 9.12279 − 3 |
| 1.8 | 1.68874 + 2 | 1.19010 + 2 | 2.28768 + 0 | −1.36988 − 1 |
| 1.9 | 2.20527 + 2 | 1.47564 + 2 | −3.55354 − 1 | −1.39402 − 1 |
| 2.0 | 2.82291 + 2 | 1.78757 + 2 | −2.57167 + 0 | −1.18170 − 2 |
| 2.1 | 3.55155 + 2 | 2.12083 + 2 | −4.02160 + 0 | 1.05884 − 1 |
| 2.2 | 4.40084 + 2 | 2.46904 + 2 | −4.51700 + 0 | 1.18832 − 1 |
| 2.3 | 5.38002 + 2 | 2.82468 + 2 | −4.05482 + 0 | 1.72440 − 2 |
| 2.4 | 6.49779 + 2 | 3.17933 + 2 | −2.81239 + 0 | −8.73616 − 2 |
| 2.5 | 7.76218 + 2 | 3.52375 + 2 | −1.10632 + 0 | −7.36407 − 2 |
| 2.6 | 9.18038 + 2 | 3.84832 + 2 | 6.76904 − 1 | −2.63630 − 2 |
| 2.7 | 1.07586 + 3 | 4.14324 + 2 | 2.16111 + 0 | 3.13871 − 2 |
| 2.8 | 1.25021 + 3 | 4.39892 + 2 | 3.05658 + 0 | 9.47047 − 2 |
| 2.9 | 1.44146 + 3 | 4.60641 + 2 | 3.21902 + 0 | 7.91312 − 3 |
| 3.0 | 1.64988 + 3 | 4.75731 + 2 | 2.66304 + 0 | −5.44166 − 2 |

$N = 0 \quad n = 3$

| $x$ | $c = 1$ | $c = 2$ | $c = 5$ | $c = 10$ |
|---|---|---|---|---|
| 0.1 | 1.04335 + 0 | 1.03904 + 0 | 1.00396 + 0 | 8.16002 − 1 |
| 0.2 | 9.41806 − 1 | 9.22225 − 1 | 7.82101 − 1 | 2.80879 − 1 |
| 0.3 | 2.91804 − 1 | 2.54987 − 1 | 1.09411 − 2 | −5.85742 − 1 |
| 0.4 | −5.66503 − 1 | −6.04220 − 1 | −8.19616 − 1 | −9.01249 − 1 |
| 0.5 | −1.16122 + 0 | −1.17035 + 0 | −1.15387 + 0 | −2.96465 − 1 |
| 0.6 | −1.04537 + 0 | −1.00493 + 0 | −6.55699 − 1 | 7.15663 − 1 |
| 0.7 | −6.94405 − 2 | 3.87818 − 3 | 5.00576 − 1 | 1.12898 + 0 |
| 0.8 | 1.23639 + 0 | 1.27745 + 0 | 1.45426 + 0 | 2.99477 − 1 |
| 0.9 | 1.17030 + 0 | 1.12044 + 0 | 6.85065 − 1 | −1.44767 + 0 |
| 1.0 | −3.74163 + 0 | −3.74119 + 0 | −3.72277 + 0 | −2.98825 + 0 |
| 1.1 | −1.94508 + 1 | −1.86263 + 1 | −1.36118 + 1 | −3.23654 + 0 |
| 1.2 | −5.51898 + 1 | −5.13331 + 1 | −3.01767 + 1 | −1.98417 + 0 |
| 1.3 | −1.24365 + 2 | −1.12506 + 2 | −5.34231 + 1 | −6.65948 − 2 |
| 1.4 | −2.45530 + 2 | −2.15895 + 2 | −8.18229 + 1 | 1.27315 + 0 |
| 1.5 | −4.43417 + 2 | −3.78480 + 2 | −1.12269 + 2 | 1.31011 + 0 |
| 1.6 | −7.50023 + 2 | −6.20420 + 2 | −1.40431 + 2 | 3.14866 − 1 |
| 1.7 | −1.20572 + 3 | −9.64812 + 2 | −1.61419 + 2 | −7.44769 − 1 |
| 1.8 | −1.86039 + 3 | −1.43724 + 3 | −1.70733 + 2 | −1.00264 + 0 |
| 1.9 | −2.77451 + 3 | −2.06513 + 3 | −1.65204 + 2 | −4.10716 − 1 |
| 2.0 | −4.02027 + 3 | −2.87683 + 3 | −1.43828 + 2 | 4.24001 − 1 |
| 2.1 | −5.68257 + 3 | −3.90062 + 3 | −1.08200 + 2 | 8.04224 − 1 |

TABLE II — *Continued*

$N = 0 \quad n = 3$

| $x$ | $c = 1$ | $c = 2$ | $c = 5$ | $c = 10$ |
|---|---|---|---|---|
| 2.2 | −7.85998 + 3 | −5.16340 + 3 | −6.24706 + 1 | 4.53989 − 1 |
| 2.3 | −1.06656 + 4 | −6.68939 + 3 | −1.27832 + 1 | −2.40332 − 1 |
| 2.4 | −1.42276 + 4 | −8.49881 + 3 | 3.37407 + 1 | −6.90334 − 1 |
| 2.5 | −1.86902 + 4 | −1.06060 + 4 | 7.02754 + 1 | −3.30234 − 1 |
| 2.6 | −2.42135 + 4 | −1.30186 + 4 | 9.15596 + 1 | 1.36373 − 1 |
| 2.7 | −3.09743 + 4 | −1.57357 + 4 | 9.49226 + 1 | 2.58923 − 1 |
| 2.8 | −3.91655 + 4 | −1.87469 + 4 | 8.08334 + 1 | 4.73302 − 1 |
| 2.9 | −4.89962 + 4 | −2.20322 + 4 | 5.29160 + 1 | −7.99752 − 3 |
| 3.0 | −6.06911 + 4 | −2.55591 + 4 | 1.70562 + 1 | −4.71948 − 1 |

$N = 1 \quad n = 0$

| $x$ | $c = 1$ | $c = 2$ | $c = 5$ | $c = 10$ |
|---|---|---|---|---|
| 0.1 | 6.67799 − 2 | 7.82376 − 2 | 1.75066 − 1 | 3.92683 − 1 |
| 0.2 | 1.88413 − 1 | 2.19147 − 1 | 4.70668 − 1 | 9.77051 − 1 |
| 0.3 | 3.44707 − 1 | 3.96102 − 1 | 7.93680 − 1 | 1.44397 + 0 |
| 0.4 | 5.27637 − 1 | 5.96043 − 1 | 1.08116 + 0 | 1.62479 + 0 |
| 0.5 | 7.31901 − 1 | 8.08692 − 1 | 1.28498 + 0 | 1.49159 + 0 |
| 0.6 | 9.53337 − 1 | 1.02496 + 0 | 1.37488 + 0 | 1.13769 + 0 |
| 0.7 | 1.18837 + 0 | 1.23655 + 0 | 1.34097 + 0 | 7.13592 − 1 |
| 0.8 | 1.43380 + 0 | 1.43586 + 0 | 1.19359 + 0 | 3.52409 − 1 |
| 0.9 | 1.68661 + 0 | 1.61605 + 0 | 9.60016 − 1 | 1.21545 − 1 |
| 1.0 | 1.94398 + 0 | 1.77112 + 0 | 6.78651 − 1 | 1.73276 − 2 |
| 1.1 | 2.20321 + 0 | 1.89606 + 0 | 3.91685 − 1 | −6.46094 − 3 |
| 1.2 | 2.46172 + 0 | 1.98689 + 0 | 1.37762 − 1 | −6.17021 − 4 |
| 1.3 | 2.71702 + 0 | 2.04079 + 0 | −5.42516 − 2 | 4.36117 − 3 |
| 1.4 | 2.96675 + 0 | 2.05612 + 0 | −1.69153 − 1 | 1.77204 − 3 |
| 1.5 | 3.20863 + 0 | 2.03244 + 0 | −2.06531 − 1 | −2.50040 − 3 |
| 1.6 | 3.44053 + 0 | 1.97051 + 0 | −1.79111 − 1 | −2.92887 − 3 |
| 1.7 | 3.66041 + 0 | 1.87221 + 0 | −1.08794 − 1 | 5.03281 − 5 |
| 1.8 | 3.86640 + 0 | 1.74049 + 0 | −2.14572 − 2 | 2.47164 − 3 |
| 1.9 | 4.05675 + 0 | 1.57927 + 0 | 5.85005 − 2 | 1.94403 − 3 |
| 2.0 | 4.22984 + 0 | 1.39322 + 0 | 1.12680 − 1 | −5.40072 − 4 |
| 2.1 | 4.38426 + 0 | 1.18771 + 0 | 1.31527 − 1 | −2.13282 − 3 |
| 2.2 | 4.51871 + 0 | 9.68543 − 1 | 1.15060 − 1 | −1.41516 − 3 |
| 2.3 | 4.63210 + 0 | 7.41814 − 1 | 7.15483 − 2 | 6.26047 − 4 |
| 2.4 | 4.72350 + 0 | 5.13706 − 1 | 1.46127 − 2 | 1.92158 − 3 |
| 2.5 | 4.79216 + 0 | 2.90272 − 1 | −4.05001 − 2 | 7.95172 − 4 |
| 2.6 | 4.83753 + 0 | 7.72847 − 2 | −8.05860 − 2 | −6.86022 − 4 |
| 2.7 | 4.85923 + 0 | −1.19954 − 1 | −9.73081 − 2 | −7.63503 − 4 |
| 2.8 | 4.85708 + 0 | −2.96762 − 1 | −8.86612 − 2 | −9.86627 − 4 |
| 2.9 | 4.83107 + 0 | −4.49226 − 1 | −5.89386 − 2 | 1.47764 − 4 |
| 3.0 | 4.78140 + 0 | −5.74265 − 1 | −1.69513 − 2 | 1.50302 − 3 |

$N = 1 \quad n = 1$

| $x$ | $c = 1$ | $c = 2$ | $c = 5$ | $c = 10$ |
|---|---|---|---|---|
| 0.1 | 1.78797 − 1 | 1.86209 − 1 | 2.26456 − 1 | 4.69291 − 1 |
| 0.2 | 4.81623 − 1 | 4.98465 − 1 | 5.77313 − 1 | 1.01396 + 0 |

TABLE II — *Continued*

$N = 1 \quad n = 1$

| $x$ | $c = 1$ | $c = 2$ | $c = 5$ | $c = 10$ |
|---|---|---|---|---|
| 0.3 | 8.11322 − 1 | 8.30625 − 1 | 8.81124 − 1 | 1.10226 + 0 |
| 0.4 | 1.09167 + 0 | 1.09943 + 0 | 1.01214 + 0 | 5.69582 − 1 |
| 0.5 | 1.24498 + 0 | 1.22371 + 0 | 8.87650 − 1 | −3.54754 − 1 |
| 0.6 | 1.19004 + 0 | 1.12484 + 0 | 4.82831 − 1 | −1.21477 + 0 |
| 0.7 | 8.42479 − 1 | 7.29666 − 1 | −1.63619 − 1 | −1.61645 + 0 |
| 0.8 | 1.15744 − 1 | −2.61055 − 2 | −9.56512 − 1 | −1.44879 + 0 |
| 0.9 | −1.07772 + 0 | −1.19460 + 0 | −1.76308 + 0 | −9.06520 − 1 |
| 1.0 | −2.82516 + 0 | −2.81285 + 0 | −2.43994 + 0 | −3.26420 − 1 |
| 1.1 | −5.21224 + 0 | −4.90037 + 0 | −2.86233 + 0 | 3.12685 − 2 |
| 1.2 | −8.32175 + 0 | −7.45742 + 0 | −2.94990 + 0 | 1.08983 − 1 |
| 1.3 | −1.22323 + 1 | −1.04640 + 1 | −2.68284 + 0 | 2.38728 − 2 |
| 1.4 | −1.70172 + 1 | −1.38798 + 1 | −2.10636 + 0 | −6.19490 − 2 |
| 1.5 | −2.27431 + 1 | −1.76444 + 1 | −1.32097 + 0 | −6.25964 − 2 |
| 1.6 | −2.94693 + 1 | −2.16792 + 1 | −4.62707 − 1 | −1.89305 − 4 |
| 1.7 | −3.72461 + 1 | −2.58893 + 1 | 3.24280 − 1 | 5.11365 − 2 |
| 1.8 | −4.61145 + 1 | −3.01663 + 1 | 9.15552 − 1 | 4.39191 − 2 |
| 1.9 | −5.61051 + 1 | −3.43921 + 1 | 1.22972 + 0 | −3.52397 − 3 |
| 2.0 | −6.72374 + 1 | −3.84424 + 1 | 1.24232 + 0 | −3.87077 − 2 |
| 2.1 | −7.95192 + 1 | −4.21912 + 1 | 9.87598 − 1 | −3.49613 − 2 |
| 2.2 | −9.29459 + 1 | −4.55151 + 1 | 5.48435 − 1 | 2.81473 − 3 |
| 2.3 | −1.07501 + 2 | −4.82978 + 1 | 3.67863 − 2 | 3.09193 − 2 |
| 2.4 | −1.23154 + 2 | −5.04353 + 1 | −4.30799 − 1 | 3.14921 − 2 |
| 2.5 | −1.39863 + 2 | −5.18371 + 1 | −7.57370 − 1 | −6.17863 − 4 |
| 2.6 | −1.57573 + 2 | −5.24330 + 1 | −8.84163 − 1 | −2.95268 − 2 |
| 2.7 | −1.76214 + 2 | −5.21739 + 1 | −8.00250 − 1 | −1.28822 − 2 |
| 2.8 | −1.95708 + 2 | −5.10350 + 1 | −5.41499 − 1 | −1.34150 − 3 |
| 2.9 | −2.15961 + 2 | −4.90184 + 1 | −1.79916 − 1 | 6.07715 − 3 |
| 3.0 | −2.36871 + 2 | −4.61486 + 1 | 1.95783 − 1 | 2.63666 − 2 |

$N = 1 \quad n = 2$

| $x$ | $c = 1$ | $c = 2$ | $c = 5$ | $c = 10$ |
|---|---|---|---|---|
| 0.1 | 3.17572 − 1 | 3.23375 − 1 | 3.58452 − 1 | 4.61327 − 1 |
| 0.2 | 7.89038 − 1 | 7.98456 − 1 | 8.45034 − 1 | 8.84405 − 1 |
| 0.3 | 1.13929 + 0 | 1.13924 + 0 | 1.10027 + 0 | 6.93291 − 1 |
| 0.4 | 1.16268 + 0 | 1.13700 + 0 | 9.09831 − 1 | −8.13386 − 2 |
| 0.5 | 7.46367 − 1 | 6.88194 − 1 | 2.60402 − 1 | −8.60192 − 1 |
| 0.6 | −6.41935 − 2 | −1.40156 − 1 | −6.18349 − 1 | −9.72981 − 1 |
| 0.7 | −9.90830 − 1 | −1.04501 + 0 | −1.27959 + 0 | −2.01002 − 1 |
| 0.8 | −1.44413 + 0 | −1.42980 + 0 | −1.15587 + 0 | 1.02436 + 0 |
| 0.9 | −4.43333 − 1 | −3.60476 − 1 | 2.92752 − 1 | 1.92471 + 0 |
| 1.0 | 3.46336 + 0 | 3.46067 + 0 | 3.41429 + 0 | 1.94059 + 0 |
| 1.1 | 1.22761 + 1 | 1.16998 + 1 | 8.22613 + 0 | 1.12202 + 0 |
| 1.2 | 2.86165 + 1 | 2.63729 + 1 | 1.43419 + 1 | 5.54149 − 2 |
| 1.3 | 5.57952 + 1 | 4.98022 + 1 | 2.09872 + 1 | −5.92374 − 1 |
| 1.4 | 9.78715 + 1 | 8.45483 + 1 | 2.71147 + 1 | −5.45072 − 1 |
| 1.5 | 1.59705 + 2 | 1.33319 + 2 | 3.15908 + 1 | −5.23392 − 2 |
| 1.6 | 2.46994 + 2 | 1.98856 + 2 | 3.34313 + 1 | 3.78477 − 1 |
| 1.7 | 3.66308 + 2 | 2.83810 + 2 | 3.20166 + 1 | 4.06284 − 1 |
| 1.8 | 5.25100 + 2 | 3.90597 + 2 | 2.72591 + 1 | 7.68751 − 2 |
| 1.9 | 7.31713 + 2 | 5.21256 + 2 | 1.96583 + 1 | −2.68863 − 1 |
| 2.0 | 9.95363 + 2 | 6.77296 + 2 | 1.02468 + 1 | −3.08683 − 1 |

TABLE II — *Continued*

$N = 1 \quad n = 2$

| $x$ | $c = 1$ | $c = 2$ | $c = 5$ | $c = 10$ |
|---|---|---|---|---|
| 2.1 | 1.32612 + 3 | 8.59566 + 2 | 4.17984 − 1 | −9.67834 − 2 |
| 2.2 | 1.73484 + 3 | 1.06812 + 3 | −8.32684 + 0 | 1.95130 − 1 |
| 2.3 | 2.23316 + 3 | 1.30213 + 3 | −1.46613 + 1 | 2.53304 − 1 |
| 2.4 | 2.83336 + 3 | 1.55981 + 3 | −1.76896 + 1 | 1.16918 − 1 |
| 2.5 | 3.54832 + 3 | 1.83831 + 3 | −1.71224 + 1 | −9.81521 − 2 |
| 2.6 | 4.39138 + 3 | 2.13379 + 3 | −1.33267 + 1 | −2.49959 − 1 |
| 2.7 | 5.37625 + 3 | 2.44140 + 3 | −7.25156 + 0 | −5.82649 − 2 |
| 2.8 | 6.51688 + 3 | 2.75538 + 3 | −2.34462 − 1 | 9.65074 − 2 |
| 2.9 | 7.82726 + 3 | 3.06921 + 3 | 6.27542 + 0 | 5.32562 − 2 |
| 3.0 | 9.32135 + 3 | 3.37554 + 3 | 1.10027 + 1 | 1.36992 − 1 |

$N = 1 \quad n = 3$

| $x$ | $c = 1$ | $c = 2$ | $c = 5$ | $c = 10$ |
|---|---|---|---|---|
| 0.1 | 4.70292 − 1 | 4.74792 − 1 | 5.03663 − 1 | 5.60653 − 1 |
| 0.2 | 1.03640 + 0 | 1.03944 + 0 | 1.05094 + 0 | 9.52509 − 1 |
| 0.3 | 1.15261 + 0 | 1.13873 + 0 | 1.02612 + 0 | 4.75116 − 1 |
| 0.4 | 5.87098 − 1 | 5.49119 − 1 | 2.80849 − 1 | −5.29321 − 1 |
| 0.5 | −4.34815 − 1 | −4.78943 − 1 | −7.50431 − 1 | −1.03082 + 0 |
| 0.6 | −1.22557 + 0 | −1.23658 + 0 | −1.23879 + 0 | −3.74929 − 1 |
| 0.7 | −9.49593 − 1 | −9.03408 − 1 | −5.28507 − 1 | 8.87232 − 1 |
| 0.8 | 5.95761 − 1 | 6.56832 − 1 | 1.04428 + 0 | 1.26903 + 0 |
| 0.9 | 1.58393 + 0 | 1.56393 + 0 | 1.35005 + 0 | −3.80514 − 1 |
| 1.0 | −3.99973 + 0 | −3.99876 + 0 | −3.98415 + 0 | −3.66768 + 0 |
| 1.1 | −2.94218 + 1 | −2.82967 + 1 | −2.13760 + 1 | −6.63805 + 0 |
| 1.2 | −9.93028 + 1 | −9.31025 + 1 | −5.82859 + 1 | −7.12958 + 0 |
| 1.3 | −2.54877 + 2 | −2.33198 + 2 | −1.21470 + 2 | −4.48170 + 0 |
| 1.4 | −5.60512 + 2 | −5.00166 + 2 | −2.14807 + 2 | −1.91044 − 1 |
| 1.5 | −1.11157 + 3 | −9.66261 + 2 | −3.37164 + 2 | 3.09018 + 0 |
| 1.6 | −2.04369 + 3 | −1.72812 + 3 | −4.80966 + 2 | 3.47603 + 0 |
| 1.7 | −3.54344 + 3 | −2.91000 + 3 | −6.31861 + 2 | 1.24407 + 0 |
| 1.8 | −5.86049 + 3 | −4.66630 + 3 | −7.70049 + 2 | −1.54042 + 0 |
| 1.9 | −9.32112 + 3 | −7.18302 + 3 | −8.72930 + 2 | −2.73125 + 0 |
| 2.0 | −1.43431 + 4 | −1.06778 + 4 | −9.19104 + 2 | −1.43583 + 0 |
| 2.1 | −2.14517 + 4 | −1.53984 + 4 | −8.92696 + 2 | 7.02031 − 1 |
| 2.2 | −3.12971 + 4 | −2.16194 + 4 | −7.87269 + 2 | 2.14894 + 0 |
| 2.3 | −4.46725 + 4 | −2.96368 + 4 | −6.08384 + 2 | 1.44329 + 0 |
| 2.4 | −6.25326 + 4 | −3.97614 + 4 | −3.74068 + 2 | −1.90745 − 1 |
| 2.5 | −8.60136 + 4 | −5.23076 + 4 | −1.12896 + 2 | −1.17683 + 0 |
| 2.6 | −1.16452 + 5 | −6.75841 + 4 | 1.40286 + 2 | −1.61055 + 0 |
| 2.7 | −1.55406 + 5 | −8.58796 + 4 | 3.50030 + 2 | −7.92157 − 2 |
| 2.8 | −2.04672 + 5 | −1.07450 + 5 | 4.86619 + 2 | 1.29119 + 0 |
| 2.9 | −2.66305 + 5 | −1.32504 + 5 | 5.32282 + 2 | 3.73378 − 1 |
| 3.0 | −3.42633 + 5 | −1.61182 + 5 | 4.83375 + 2 | 4.32692 − 1 |

$N = 2 \quad n = 0$

| $x$ | $c = 1$ | $c = 2$ | $c = 5$ | $c = 10$ |
|---|---|---|---|---|
| 0.1 | 8.11214 − 3 | 9.30928 − 3 | 2.21477 − 2 | 8.00048 − 2 |
| 0.2 | 4.58035 − 2 | 5.22724 − 2 | 1.20067 − 1 | 4.01321 − 1 |

TABLE II — *Continued*

$N = 2 \quad n = 0$

| $x$ | $c = 1$ | $c = 2$ | $c = 5$ | $c = 10$ |
|---|---|---|---|---|
| 0.3 | 1.25827 − 1 | 1.42274 − 1 | 3.07992 − 1 | 9.01985 − 1 |
| 0.4 | 2.57171 − 1 | 2.87029 − 1 | 5.71015 − 1 | 1.38094 + 0 |
| 0.5 | 4.46741 − 1 | 4.90296 − 1 | 8.72575 − 1 | 1.62971 + 0 |
| 0.6 | 6.99877 − 1 | 7.52394 − 1 | 1.16320 + 0 | 1.54932 + 0 |
| 0.7 | 1.02059 + 0 | 1.07049 + 0 | 1.39112 + 0 | 1.19392 + 0 |
| 0.8 | 1.41171 + 0 | 1.43889 + 0 | 1.51294 + 0 | 7.25292 − 1 |
| 0.9 | 1.87493 + 0 | 1.84932 + 0 | 1.50259 + 0 | 3.16554 − 1 |
| 1.0 | 2.41089 + 0 | 2.29129 + 0 | 1.35656 + 0 | 6.79148 − 2 |
| 1.1 | 3.01924 + 0 | 2.75254 + 0 | 1.09462 + 0 | −2.02873 − 2 |
| 1.2 | 3.69863 + 0 | 3.21944 + 0 | 7.56039 − 1 | −1.46045 − 2 |
| 1.3 | 4.44678 + 0 | 3.67758 + 0 | 3.91913 − 1 | 1.04085 − 2 |
| 1.4 | 5.26054 + 0 | 4.11222 + 0 | 5.55434 − 2 | 1.52034 − 2 |
| 1.5 | 6.13588 + 0 | 4.50887 + 0 | −2.07467 − 1 | 1.29918 − 3 |
| 1.6 | 7.06802 + 0 | 4.85377 + 0 | −3.67156 − 1 | −1.11917 − 2 |
| 1.7 | 8.05139 + 0 | 5.13443 + 0 | −4.13815 − 1 | −9.67213 − 3 |
| 1.8 | 9.07977 + 0 | 5.34008 + 0 | −3.58514 − 1 | 1.55354 − 3 |
| 1.9 | 1.01463 + 1 | 5.46205 + 0 | −2.29434 − 1 | 9.62330 − 3 |
| 2.0 | 1.12436 + 1 | 5.49406 + 0 | −6.51947 − 2 | 6.50298 − 3 |
| 2.1 | 1.23637 + 1 | 5.43250 + 0 | 9.35338 − 2 | −2.03792 − 3 |
| 2.2 | 1.34985 + 1 | 5.27657 + 0 | 2.12220 − 1 | −8.19124 − 3 |
| 2.3 | 1.46391 + 1 | 5.02833 + 0 | 2.68832 − 1 | −5.06774 − 3 |
| 2.4 | 1.57769 + 1 | 4.69273 + 0 | 2.57171 − 1 | 1.94735 − 3 |
| 2.5 | 1.69026 + 1 | 4.27721 + 0 | 1.86697 − 1 | 4.94597 − 3 |
| 2.6 | 1.80072 + 1 | 3.79179 + 0 | 7.90086 − 2 | 5.04030 − 3 |
| 2.7 | 1.90816 + 1 | 3.24856 + 0 | −3.78705 − 2 | −7.03146 − 4 |
| 2.8 | 2.01167 + 1 | 2.66139 + 0 | −1.36199 − 1 | −6.15552 − 3 |
| 2.9 | 2.11036 + 1 | 2.04550 + 0 | −1.94874 − 1 | −1.13042 − 3 |
| 3.0 | 2.20337 + 1 | 1.41675 + 0 | −2.03218 − 1 | 1.09442 − 3 |

$N = 2 \quad n = 1$

| $x$ | $c = 1$ | $c = 2$ | $c = 5$ | $c = 10$ |
|---|---|---|---|---|
| 0.1 | 3.01525 − 2 | 3.18266 − 2 | 4.36114 − 2 | 1.10168 − 1 |
| 0.2 | 1.63412 − 1 | 1.71721 − 1 | 2.27802 − 1 | 5.09403 − 1 |
| 0.3 | 4.17552 − 1 | 4.35481 − 1 | 5.46215 − 1 | 9.75159 − 1 |
| 0.4 | 7.63468 − 1 | 7.87547 − 1 | 9.08386 − 1 | 1.10096 + 0 |
| 0.5 | 1.12468 + 0 | 1.14274 + 0 | 1.16775 + 0 | 6.37139 − 1 |
| 0.6 | 1.37439 + 0 | 1.36734 + 0 | 1.16137 + 0 | −2.87398 − 1 |
| 0.7 | 1.33292 + 0 | 1.28261 + 0 | 7.56169 − 1 | −1.22293 + 0 |
| 0.8 | 7.66020 − 1 | 6.71043 − 1 | −1.09012 − 1 | −1.69212 + 0 |
| 0.9 | −6.15776 − 1 | −7.14755 − 1 | −1.39595 + 0 | −1.50113 + 0 |
| 1.0 | −3.15683 + 0 | −3.14077 + 0 | −2.96182 + 0 | −8.45332 − 1 |
| 1.1 | −7.25498 + 0 | −6.87847 + 0 | −4.57844 + 0 | −1.47679 − 1 |
| 1.2 | −1.33591 + 1 | −1.21904 + 1 | −5.97349 + 0 | 2.35668 − 1 |
| 1.3 | −2.19659 + 1 | −1.93150 + 1 | −6.88402 + 0 | 2.24248 − 1 |
| 1.4 | −3.36151 + 1 | −2.84518 + 1 | −7.11308 + 0 | 3.10851 − 3 |
| 1.5 | −4.88847 + 1 | −3.97466 + 1 | −6.57336 + 0 | −1.65861 − 1 |
| 1.6 | −6.83847 + 1 | −5.32791 + 1 | −5.31243 + 0 | −1.46461 − 1 |
| 1.7 | −9.27502 + 1 | −6.90514 + 1 | −3.50808 + 0 | 2.53438 − 3 |
| 1.8 | −1.22634 + 2 | −8.69800 + 1 | −1.43922 + 0 | 1.23063 − 1 |
| 1.9 | −1.58698 + 2 | −1.06891 + 2 | 5.66677 − 1 | 1.16626 − 1 |
| 2.0 | −2.01606 + 2 | −1.28518 + 2 | 2.19548 + 0 | 3.46739 − 3 |
| 2.1 | −2.52012 + 2 | −1.51503 + 2 | 3.20680 + 0 | −9.50870 − 2 |

Table II — *Continued*

$N = 2 \quad n = 1$

| $x$ | $c = 1$ | $c = 2$ | $c = 5$ | $c = 10$ |
|---|---|---|---|---|
| 2.2 | −3.10555 + 2 | −1.75404 + 2 | 3.47937 + 0 | −1.00215 − 1 |
| 2.3 | −3.77844 + 2 | −1.99703 + 2 | 3.03214 + 0 | −1.01271 − 2 |
| 2.4 | −4.54453 + 2 | −2.23824 + 2 | 2.01698 + 0 | 7.86142 − 2 |
| 2.5 | −5.40910 + 2 | −2.47136 + 2 | 6.84586 − 1 | 6.25620 − 2 |
| 2.6 | −6.37687 + 2 | −2.68987 + 2 | −6.69192 − 1 | 1.86542 − 2 |
| 2.7 | −7.45191 + 2 | −2.88715 + 2 | −1.76387 + 0 | −2.83514 − 2 |
| 2.8 | −8.63756 + 2 | −3.05673 + 2 | −2.39029 + 0 | −8.09642 − 2 |
| 2.9 | −9.93633 + 2 | −3.19263 + 2 | −2.45263 + 0 | −6.05693 − 3 |
| 3.0 | −1.13498 + 3 | −3.28923 + 2 | −1.97549 + 0 | 4.94337 − 2 |

$N = 2 \quad n = 2$

| $x$ | $c = 1$ | $c = 2$ | $c = 5$ | $c = 10$ |
|---|---|---|---|---|
| 0.1 | 6.92988 − 2 | 7.12673 − 2 | 8.51196 − 2 | 1.31062 − 1 |
| 0.2 | 3.52496 − 1 | 3.60980 − 1 | 4.18221 − 1 | 5.67604 − 1 |
| 0.3 | 8.01544 − 1 | 8.14659 − 1 | 8.93117 − 1 | 9.45356 − 1 |
| 0.4 | 1.20746 + 0 | 1.21237 + 0 | 1.21221 + 0 | 7.67342 − 1 |
| 0.5 | 1.27474 + 0 | 1.25311 + 0 | 1.05255 + 0 | −3.76902 − 2 |
| 0.6 | 7.57995 − 1 | 7.02242 − 1 | 2.84685 − 1 | −9.13196 − 1 |
| 0.7 | −3.27092 − 1 | −3.94312 − 1 | −8.22806 − 1 | −1.05173 + 0 |
| 0.8 | −1.41364 + 0 | −1.43709 + 0 | −1.49256 + 0 | −7.12440 − 2 |
| 0.9 | −1.01239 + 0 | −9.55456 − 1 | −4.82515 − 1 | 1.57998 + 0 |
| 1.0 | 3.74003 + 0 | 3.73507 + 0 | 3.69332 + 0 | 2.86776 + 0 |
| 1.1 | 1.76108 + 1 | 1.68770 + 1 | 1.23963 + 1 | 2.91426 + 0 |
| 1.2 | 4.78738 + 1 | 4.45904 + 1 | 2.64247 + 1 | 1.66265 + 0 |
| 1.3 | 1.04979 + 2 | 9.51393 + 1 | 4.56130 + 1 | −6.78936 − 2 |
| 1.4 | 2.03280 + 2 | 1.79119 + 2 | 6.85921 + 1 | −1.18788 + 0 |
| 1.5 | 3.61809 + 2 | 3.09537 + 2 | 9.27854 + 1 | −1.12986 + 0 |
| 1.6 | 6.05086 + 2 | 5.01768 + 2 | 1.14715 + 2 | −2.05881 − 1 |
| 1.7 | 9.63954 + 2 | 7.73357 + 2 | 1.30538 + 2 | 7.08003 − 1 |
| 1.8 | 1.47641 + 3 | 1.14368 + 3 | 1.36797 + 2 | 8.79453 − 1 |
| 1.9 | 2.18846 + 3 | 1.63343 + 3 | 1.31147 + 2 | 3.17527 − 1 |
| 2.0 | 3.15489 + 3 | 2.26396 + 3 | 1.12984 + 2 | −4.11104 − 1 |
| 2.1 | 4.44006 + 3 | 3.05647 + 3 | 8.37540 + 1 | −7.13268 − 1 |
| 2.2 | 6.11858 + 3 | 4.03107 + 3 | 4.68923 + 1 | −3.71699 − 1 |
| 2.3 | 8.27598 + 3 | 5.20576 + 3 | 7.34290 + 0 | 2.40288 − 1 |
| 2.4 | 1.10092 + 4 | 6.59547 + 3 | −2.92519 + 1 | 6.17233 − 1 |
| 2.5 | 1.44270 + 4 | 8.21064 + 3 | −5.75607 + 1 | 2.78319 − 1 |
| 2.6 | 1.86502 + 4 | 1.00565 + 4 | −7.35509 + 1 | −1.46009 − 1 |
| 2.7 | 2.38122 + 4 | 1.21320 + 4 | −7.52754 + 1 | −2.33031 − 1 |
| 2.8 | 3.00582 + 4 | 1.44287 + 4 | −6.32751 + 1 | −4.05877 − 1 |
| 2.9 | 3.75458 + 4 | 1.69309 + 4 | −4.05587 + 1 | 1.20902 − 2 |
| 3.0 | 4.64442 + 4 | 1.96136 + 4 | −1.18748 + 1 | 4.27288 − 1 |

$N = 2 \quad n = 3$

| $x$ | $c = 1$ | $c = 2$ | $c = 5$ | $c = 10$ |
|---|---|---|---|---|
| 0.1 | 1.26968 − 1 | 1.29109 − 1 | 1.44081 − 1 | 1.94166 − 1 |
| 0.2 | 5.91772 − 1 | 5.99200 − 1 | 6.48662 − 1 | 7.75163 − 1 |
| 0.3 | 1.13308 + 0 | 1.13787 + 0 | 1.15991 + 0 | 1.06828 + 0 |

TABLE II — *Continued*

| $N = 2 \quad n = 3$ | | | | |
|---|---|---|---|---|
| $x$ | $c = 1$ | $c = 2$ | $c = 5$ | $c = 10$ |
| 0.4 | 1.22242 + 0 | 1.20735 + 0 | 1.08268 + 0 | 4.42410 − 1 |
| 0.5 | 5.02125 − 1 | 4.61442 − 1 | 1.72834 − 1 | −7.08002 − 1 |
| 0.6 | −7.28650 − 1 | −7.67417 − 1 | −9.99672 − 1 | −1.12281 + 0 |
| 0.7 | −1.34994 + 0 | −1.33947 + 0 | −1.20442 + 0 | −6.35200 − 2 |
| 0.8 | −1.55815 − 1 | −9.78895 − 2 | 3.13440 − 1 | 1.38983 + 0 |
| 0.9 | 1.72855 + 0 | 1.73119 + 0 | 1.69507 + 0 | 6.83573 − 1 |
| 1.0 | −4.24197 + 0 | −4.23990 + 0 | −4.22216 + 0 | −4.05639 + 0 |
| 1.1 | −4.29825 + 1 | −4.14973 + 1 | −3.22563 + 1 | −1.17483 + 1 |
| 1.2 | −1.69966 + 2 | −1.60436 + 2 | −1.05765 + 2 | −1.81766 + 1 |
| 1.3 | −4.92189 + 2 | −4.54612 + 2 | −2.55016 + 2 | −1.86221 + 1 |
| 1.4 | −1.19761 + 3 | −1.08177 + 3 | −5.12503 + 2 | −1.15388 + 1 |
| 1.5 | −2.59483 + 3 | −2.28970 + 3 | −9.05339 + 2 | −3.41475 − 1 |
| 1.6 | −5.16494 + 3 | −4.44652 + 3 | −1.44608 + 3 | 8.52760 + 0 |
| 1.7 | −9.62735 + 3 | −8.07479 + 3 | −2.12354 + 3 | 1.01259 + 1 |
| 1.8 | −1.70216 + 4 | −1.38880 + 4 | −2.89670 + 3 | 4.31901 + 0 |
| 1.9 | −2.88067 + 4 | −2.28281 + 4 | −3.69313 + 3 | −3.63918 + 0 |
| 2.0 | −4.69805 + 4 | −3.61016 + 4 | −4.41450 + 3 | −7.22580 + 0 |
| 2.1 | −7.42193 + 4 | −5.52126 + 4 | −4.94883 + 3 | −5.31249 + 0 |
| 2.2 | −1.14040 + 5 | −8.19885 + 4 | −5.18833 + 3 | 8.57079 − 1 |
| 2.3 | −1.70988 + 5 | −1.18598 + 5 | −5.05014 + 3 | 5.14940 + 0 |
| 2.4 | −2.50841 + 5 | −1.67557 + 5 | −4.49592 + 3 | 5.60742 + 0 |
| 2.5 | −3.60853 + 5 | −2.31714 + 5 | −3.54624 + 3 | 5.87060 − 1 |
| 2.6 | −5.10005 + 5 | −3.14232 + 5 | −2.28432 + 3 | −4.18090 + 0 |
| 2.7 | −7.09289 + 5 | −4.18531 + 5 | −8.50402 + 2 | −2.39443 + 0 |
| 2.8 | −9.72020 + 5 | −5.48229 + 5 | 5.76599 + 2 | −1.90159 + 0 |
| 2.9 | −1.31415 + 6 | −7.07072 + 5 | 1.80814 + 3 | 6.87950 − 1 |
| 3.0 | −1.75464 + 6 | −8.98762 + 5 | 2.67521 + 3 | 4.80391 + 0 |

I am indebted to Mrs. E. Sonnenblick for programming and carrying out the computations reported here.

## APPENDIX A

### *A Perturbation Scheme*

We treat briefly the following problem. Eigenfunctions $u_n$ and eigenvalues $\lambda_n$ of an operator $\mathbf{L}$ are assumed known. That is, we have

$$\mathbf{L}u_n + \lambda_n u_n = 0, \qquad n = 0, 1, 2, \ldots . \tag{114}$$

It is desired to find eigenfunctions $\psi_n$ and eigenvalues $\chi_n$ of the perturbed equation

$$(\mathbf{L} - \epsilon\mathbf{M})\psi + \chi\psi = 0. \tag{115}$$

It is assumed that the $u_n$ satisfy the boundary condition to be imposed on the $\psi$'s and that the $u_n$ are complete in some appropriate sense. We proceed further in a purely formal manner.

Substitute the series

$$\psi_n = u_n + \sum_{j=1} \epsilon^j Q_j \tag{116}$$

$$\chi_n = \lambda_n + \sum_{j=1} \epsilon^j a_j \tag{117}$$

into (115). Here in the notation we have suppressed the dependence of the $Q_j$ and $a_j$ on $n$. By equating to zero the coefficients of distinct powers of $\epsilon$, we find

$$\mathbf{L}u_n + \lambda_n u_n = 0 \tag{118}$$

$$\mathbf{L}Q_j + \lambda_n Q_j = \mathbf{M}Q_{j-1} - \sum_{k=1}^{j} a_k Q_{j-k}\,, \qquad j = 1, 2, \ldots \tag{119}$$

where we define $Q_o = u_n$.

Now it frequently happens that the operator $\mathbf{M}$ is such that $\mathbf{M}u_n$ can be expressed as a finite linear combination of the $u$'s with constant coefficients. We assume this to be the case and write

$$\mathbf{M}u_n = \sum_{i=-l}^{l} \gamma_n^{\,i} u_{n+i\alpha}\,, \qquad n = 0, 1, 2, \ldots\,. \tag{120}$$

Here $\alpha$ is a positive integer, $l$ a nonnegative integer, and the superscript $i$ on $\gamma$ is not an exponent, but a label.

If the $u_n$ are linearly independent, formal solution of system (119) is now straightforward. Set

$$\begin{gathered} Q_j = \sum_{k=-jl}^{jl} A_k^{\,j} u_{n+k\alpha} \\ A_o^{\,j} = 0 \\ j = 1, 2, \ldots\,. \end{gathered} \tag{121}$$

The $A$'s of course depend on $n$, but for simplicity we have suppressed this fact in the notation. Again the superscript is a label, not an exponent. Substitute (121) and (120) into (119). Setting the coefficient of $u_n$ equal to zero in the resultant expression yields

$$a_j = \sum_{k=-l}^{l} A_{-k}^{\;j-1} \gamma_{n-k\alpha}^{\;k}, \qquad j = 1, 2, \dots . \tag{122}$$

Requiring the coefficient of $u_{n+m\alpha}$ to vanish gives

$$\begin{gathered}(\lambda_{n+m\alpha} - \lambda_n) A_m^{\;j} = \sum_{k=1}^{j} a_k A_m^{\;j-k} - \sum_{k=-l}^{l} A_{m-k}^{\;j-1} \gamma_{n+(m-k)\alpha}^{\;k} \\ m = -jl, -jl+1, \dots, jl; \quad j = 1, 2, \dots .\end{gathered} \tag{123}$$

Here we have adopted the conventions

$$A_k^{\;j} \equiv 0$$

if either

$$|k| > jl, \quad \text{or} \quad \alpha k < -n, \quad \text{or} \quad k = 0 \quad \text{and} \quad j = 1, 2, \dots$$
$$A_o^{\;0} = 1, \qquad A_k^{\;0} = 0, \qquad k \neq 0, a_o = 0.$$

Equations (122) and (123) together with these conventions permit successive determination of the $a$'s and $A$'s. The case $l = 1$ occurs frequently. The first few coefficients for this case are given below where we have set

$$\begin{aligned}
h_j &= [\lambda_{n+j} - \lambda_n]^{-1}. \\
a_1 &= \gamma_n^{\;0} \\
A_{-1}^{\;1} &= -h_{-\alpha} \gamma_n^{\;-1} \\
A_1^{\;1} &= -h_\alpha \gamma_n^{\;1} \\
a_2 &= -[h_\alpha \gamma_n^{\;1} \gamma_{n+\alpha}^{\;-1} + h_{-\alpha} \gamma_n^{\;-1} \gamma_{n-\alpha}^{\;1}] \\
A_{-2}^{\;2} &= h_{-2\alpha} h_{-\alpha} \gamma_n^{\;-1} \gamma_{n-\alpha}^{\;-1} \\
A_{-1}^{\;2} &= (h_{-\alpha})^2 \gamma_n^{\;-1} [-\gamma_n^{\;0} + \gamma_{n-\alpha}^{\;0}] \\
A_1^{\;2} &= (h_\alpha)^2 \gamma_n^{\;1} [-\gamma_n^{\;0} + \gamma_{n+\alpha}^{\;0}] \\
A_2^{\;2} &= h_{2\alpha} h_\alpha \gamma_n^{\;1} \gamma_{n+\alpha}^{\;1} \\
a_3 &= (h_\alpha)^2 \gamma_n^{\;1} \gamma_{n+\alpha}^{\;-1} (-\gamma_n^{\;0} + \gamma_{n+1}^{\;0}) + (h_{-\alpha})^2 \gamma_n^{\;-1} \gamma_{n-\alpha}^{\;1} (-\gamma_n^{\;0} + \gamma_{n-\alpha}^{\;0}) \\
A_1^{\;3} &= h_\alpha [A_1^{\;2} (\gamma_n^{\;0} - \gamma_{n+\alpha}^{\;0}) + a_2 A_1^{\;1} - \gamma_{n+2\alpha}^{\;-1} A_2^{\;2}] \\
A_{-1}^{\;3} &= h_{-\alpha} [A_{-\alpha}^{\;2} (\gamma_n^{\;0} - \gamma_{n-\alpha}^{\;0}) + a_2 A_{-1}^{\;2} - \gamma_{n-2\alpha}^{\;1} A_{-2}^{\;2}] \\
a_4 &= A_1^{\;3} \gamma_{n+\alpha}^{\;-1} + A_{-1}^{\;3} \gamma_{n-\alpha}^{\;1}.
\end{aligned}$$

More generally for this case one finds

$$
\begin{aligned}
A_{\pm j}{}^{j} &= (-1)^{j} \prod_{k=1}^{j} h_{\pm k\alpha} \gamma_{n\pm(k-1)\alpha}{}^{\pm 1}, && j = 1, 2, \ldots \\
A_{\pm(j-1)}{}^{j} &= A_{\pm(j-1)}{}^{j-1} \sum_{k=1}^{j-1} h_{\pm k\alpha}[a_1 - \gamma_{n\pm k\alpha}{}^{0}], && j = 2, 3, \ldots \\
A_{\pm(j-2)}{}^{j} &= A_{\pm(j-2)}{}^{j-2} \sum_{l=1}^{j-2} h_{\pm l\alpha} && (124) \\
&\qquad \times [a_2 + h_{\pm(l+1)\alpha} \gamma_{n\pm(l+1)\alpha}{}^{\mp 1} \gamma_{n\pm l\alpha}{}^{\pm 1} \\
&\qquad + (a_1 - \gamma_{n\pm l\alpha}{}^{0}) \sum_{k=1}^{l} h_{\pm k\alpha}(a_1 - \gamma_{n\pm k\alpha}{}^{0})], \\
& && j = 3, 4, \ldots .
\end{aligned}
$$

APPENDIX B

### *Evaluation of an Integral*

We establish here the formula (43). Let

$$
\begin{aligned}
F_{N,n}(x) &= \int_0^1 J_N(xy)\sqrt{xy}\,T_{N,n}(y)\,dy \\
&= \int_0^1 K_N(xy)\,T_{N,n}(y)\,dy
\end{aligned}
\tag{125}
$$

on using the notation of (21). Then

$$
\begin{aligned}
&\left[x^2 \frac{d^2}{dx^2} + 2x \frac{d}{dx} + (x^2 - \chi)\right] F_{N\ n}(x) \\
&= \int_0^1 T_{N,n}(y)[x^2y^2K_N''(xy) + 2xyK_N'(xy) + (x^2 - \chi)K_N(xy)]dy \\
&= \int_0^1 T_{N,n}(y)[(-x^2y^2 - \tfrac{1}{4} + N^2 + x^2 \\
&\qquad\qquad - x)K_N(xy) + 2xyK_N'(xy)]dy
\end{aligned}
\tag{126}
$$

by (23). Here primes denote differentiation with respect to the argument indicated.

Now

$$
\frac{d}{dy}(1 - y^2)\frac{dT_{N,n}(y)}{dy} + \left(\frac{\frac{1}{4} - N^2}{y^2} + \chi\right) T_{N,n}(y) = 0
$$

with $\chi$ given by (26). Multiply this equation by $K_N(xy)$ and integrate from zero to one. There results

$$0 = \int_0^1 K_N(xy)\left[\frac{d}{dy}(1-y^2)\frac{dT_{N,n}}{dy} + \left(\frac{\frac{1}{4} - N^2}{y^2} + \chi\right)T_{N,n}\right]dy$$

$$= \int_0^1 T_{N,n}(y)\left[\frac{d}{dy}(1-y^2)\frac{d}{dy}K_N(xy) + \left(\frac{\frac{1}{4} - N^2}{y^2} + \chi\right)K_N(xy)\right]dy$$

where we have integrated by parts and made use of the fact that $K_N(0) = T_{N,n}(0) = 0$. Carrying out the indicated differentiation, we find

$$0 = \int_0^1 T_{N,n}(y)\Big[(1-y^2)x^2K_N''(xy) - 2xyK_N'(xy) + \left(\frac{\frac{1}{4} - N^2}{y^2} + \chi\right)K_N(xy)\Big]dy \tag{127}$$

$$= -\int_0^1 T_{N,n}(y)[(-x^2y^2 - \tfrac{1}{4} + N^2 + x^2 - \chi)K_N(xy) + 2xyK_N'(xy)]dy.$$

Equations (126) and (127) give

$$\left[x^2\frac{d^2}{dx^2} + 2x\frac{d}{dx} + \left\{x^2 - \left(N + 2n + \frac{1}{2}\right)\left(N + 2n + \frac{3}{2}\right)\right\}\right]F_{N,n}(x) = 0.$$

The only solution of this equation that vanishes for $x = 0$ is

$$F_{N,n}(x) = k\,\frac{J_{N+2n+1}(x)}{\sqrt{x}}.$$

Using (125), we then have

$$k\,\frac{J_{N+2n+1}(x)}{\sqrt{x}} = \int_0^1 J_N(xy)\sqrt{xy}\,T_{N,n}(y)\,dy. \tag{128}$$

To determine $k$, we have only to compare the coefficients of $x^{N+2n+\frac{1}{2}}$ on both sides of (128). In this way we find

$$\frac{k}{2^{N+2n+1}\Gamma(N+2n+2)} = \frac{(-1)^n}{2^{N+2n}\Gamma(N+n+1)n!}\int_0^1 y^{N+2n+\frac{1}{2}}T_{N,n}(y)\,dy. \tag{129}$$

The integral here can be evaluated by using (27), (28) and known properties of the Jacobi polynomials. We have

$$\int_0^1 y^{N+2n+\frac{1}{2}} T_{N,n}(y)dy$$

$$= \binom{n+N}{n}^{-1} \int_0^1 y^{2N+2n+1} P_n^{(N,0)}(1-2y^2)dy$$

$$= 2^{-(N+n+2)} \binom{n+N}{n}^{-1} \int_{-1}^1 (1-u)^N (1-u)^n P_n^{(N,0)}(u)du.$$

Now the coefficient of $u^n$ in $(1-u)^n$ is $(-1)^n$ and the coefficient of $u^n$ in $P_n^{(N,0)}(u)$ is $\binom{2n+N}{n} \Big/ 2^n$ [Ref. 8, Vol. II, p. 169, Eq. (5)], so

$$(1-u)^n = \frac{(-1)^n 2^n}{\binom{2n+N}{n}} P_n^{(N,0)}(u) + \sum_{j=0}^{n-1} A_j P_j^{(N,0)}(u).$$

It follows, then, that

$$\int_1^1 (1-u)^N (1-u)^n P_n^{(N,0)}(u)du$$

$$= \frac{(-1)^n 2^n}{\binom{2n+N}{n}} \int_{-1}^1 (1-u)^N P_n^{(N,0)}(u) P_n^{(N,0)}(u)du$$

$$= \frac{(-1)^n 2^{n+N+1}}{\binom{2n+N}{n}(2n+N+1)}$$

where we have used the orthogonality of the Jacobi polynomials and the known normalization integral [see Ref. 5, page 68, Eq. (4.3.3), for example]. Combining these results, (129) yields

$$k = \binom{n+N}{n}^{-1}$$

and together with (128) this establishes (43).

## REFERENCES

1. Slepian, D., and Pollak, H. O., Prolate Spheroidal Wave Functions, Fourier Analysis and Uncertainty — I, B.S.T.J., **40,** Jan., 1961, pp. 43–64.
2. Landau, H. J., and Pollak, H. O., Prolate Spheroidal Wave Functions, Fourier Analysis and Uncertainty — II, B.S.T.J., **40,** January, 1961, pp. 65–84.

3. Fox, A. G., and Li, Tingye, Resonant Modes in a Maser Interferometer, B.S.T.J., **40**, March, 1961, pp. 453–488.
4. Riesz, F., and Sz.-Nagy, B., *Functional Analysis*, Frederick Ungar Co., New York, 1955.
5. Szegö, G., *Orthogonal Polynomials*, Am. Math. Soc. Colloquium Publication 23, New York, 1959.
6. Bouwkamp, C. J., On Spheroidal Wave Functions of Order Zero, J. Math. Phys., **26**, 1957, pp. 79–92.
7. Slepian, D., Some Asymptotic Expansions for Prolate Spheroidal Wave Functions, to be published.
8. Erdelyi, A., *Higher Transcendental Functions*, McGraw-Hill Book Co., New York, 1953.
9. Kogelnik, H., Modes in Optical Resonators, Section 5.1, chapter in *Advances in Lasers*, ed. Levine, A. K., Dekker Publishers, New York — to appear.
10. Slepian, D., Analytic Solution of Two Apodization Problems, to be published.
11. Heurtley, J. C., Hyperspheroidal Functions — Optical Resonators with Circular Mirrors, to appear in Proceedings of the Polytechnic Inst. of Brooklyn Symposium on Quasi-Optics, June 8–10, 1964.
12. Born, M., and Wolf, E., *Principles of Optics*, Pergamon Press, 1959.

# Contributors to This Issue

R. D. Barnard, B.E.E., 1952, and M.E.E., 1955, Polytechnic Institute of Brooklyn; Ph.D., 1959, Case Institute of Technology; Bell Telephone Laboratories, 1959–61; faculty, Wayne State University, 1961–62; Bell Telephone Laboratories, 1962—. Presently, he is primarily concerned with theoretical problems in signal theory and control. Member, IEEE and American Physical Society, Sigma Xi, Eta Kappa Nu and Tau Beta Pi.

Vaclav E. Beneš, A.B., 1950, Harvard College; M.A. and Ph.D. 1953, Princeton University; Bell Telephone Laboratories, 1953—. Mr. Beneš has been engaged in mathematical research on stochastic processes, traffic theory, and servomechanisms. In 1959–60 he was visiting lecturer in mathematics at Dartmouth College. He is the author of *General Stochastic Process in the Theory of Queues* (Addison-Wesley, 1963). Member, American Mathematical Society, Association for Symbolic Logic, Institute of Mathematical Statistics, SIAM, Mind Association and Phi Beta Kappa.

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 systems: dc telegraph, multichannel AM and FM 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.

James P. Gordon, B.S., Massachusetts Institute of Technology, 1949; M.A., 1951, and Ph.D., 1955, Columbia University; Bell Telephone Laboratories, 1955—.Visiting Professor of Applied Physics, University of California, San Diego, 1962–1963. His research in quantum electronics has involved work on paramagnetic resonance, masers, and the quantum mechanical aspects of communication theory. He has written several technical articles in the field of quantum electronics. Member, AAAS, American Physical Society and Sigma Xi.

J. R. GRAY, B.S.E.E., 1954, M.S.E.E., 1955, University of Florida; Bell Telephone Laboratories, 1955—. He was first engaged in repeater design for a PCM exchange area system, and more recently has been concerned with studies of signal impairment in PCM terminals. He is presently responsible for a group concerned with systems design of a high-speed PCM terminal, working with such problems as framing, synchronization, and signal deterioration.

D. C. HOGG, B.Sc., 1949, University of Western Ontario; M.Sc., 1950, and Ph.D., 1953, McGill University; Bell Telephone Laboratories, 1953—. His work has included studies of artificial dielectrics for microwaves, diffraction of microwaves, and over-the-horizon and millimeter wave propagation. He has been concerned with evaluation of sky noise, analysis of performance characteristics of microwave antennas and, most recently, propagation of optical waves. Senior member, IEEE; member, Commission 2, U.R.S.I., and Sigma Xi.

T. T. KADOTA, B.S., 1953, Yokahama National University (Japan); M.S., 1956, and Ph.D., 1960, University of California (Berkeley); Bell Telephone Laboratories, 1960—. He has been engaged in the study of noise theory with application to optimum detection theory. Member, Sigma Xi and SIAM.

HERWIG KOGELNIK, Dipl.-Ing., 1955, Dr. techn., 1958, Technische Hoschschule Wien, Austria; D. Phil., 1960, Oxford University, England; Bell Telephone Laboratories, 1961—. He is engaged in optical maser research. Member, American Physical Society, IEEE, Elektrotechnischer Verein Österreichs (Austria).

JOHN A. LEWIS, B.S., 1944, Worchester Polytechnic Institute; Sc.M., 1948, Brown University; Ph.D., 1950, Brown University; Bell Telephone Laboratories, May, 1951—. A member of the mathematical physics department, he has been engaged in theoretical investigation of problems of fluid dynamics, piezoelectric vibrations, heat transfer, and satellite attitude control. He is currently studying problems of hypersonic flow. Member, American Mathematical Society, SIAM, l'Unione Matematica Italiana and the Society for Natural Philosophy.

E. A. J. MARCATILI, Aeronautical Engineer, 1947, and E.E., 1948, University of Cordoba (Argentina); research staff, University of Cordoba, 1947–1954; Bell Telephone Laboratories, 1954—. He has been engaged in theory and design of filters in multimode waveguides and in waveguide systems research. More recently he has concentrated in the study of optical transmission media. Member, IEEE.

JOSEPH F. OSSANNA, JR., B.S.E.E., 1952, Wayne State University; Bell Telephone Laboratories, 1952—. He has been concerned with low-noise amplifier research, feedback amplifier theory and design, satellite position prediction in Project Echo, mobile radio fading studies, and data processing. Currently he is involved in the operation of the Murray Hill Computation Center. Member, IEEE, Sigma Xi and Tau Beta Pi.

JOHN W. PAN, B.S. in E.E., 1955, University of Cincinnati; Sc.D. in E.E., 1962, Massachusetts Institute of Technology; Bell Telephone Laboratories, 1955—. He received the Communications Development Training Fellowship for study at MIT, 1958–1962. He has been concerned with the process of pulse code modulation and is currently in charge of a group engaged in design of PCM terminals. Member, Sigma Xi and IEEE.

IRWIN W. SANDBERG, B.E.E., 1955, M.E.E., 1956, and D.E.E., 1958, Polytechnic Institute of Brooklyn; Bell Telephone Laboratories, 1958—. He has been concerned with analysis of military systems, particularly radar systems, and with synthesis and analysis of active and time-varying networks. He is currently involved in a study of the signal-theoretic properties of nonlinear systems. Member, IEEE, SIAM, Eta Kappa Nu, Sigma Xi and Tau Beta Pi.

R. A. SEMPLAK, B.S., 1961, Monmouth College; Bell Telephone Laboratories, 1955—. He has been engaged in beyond-the-horizon radio propagation and three satellite communications projects: Project Echo, Telstar I and Telstar II. He has also participated in studies of the effects of rain on sky noise temperatures at 6-gc frequency and has recently completed an experimental study of the near-field Cassegrainian antenna. He is currently engaged in measuring the scattered radiation from various surfaces at 0.6-micron wavelength.

W. C. SLAUSON, A.B. and B.S., 1927, Hamilton College; Bell Telephone Laboratories, 1927—. His first work consisted of design and development of polarized relays and signals. Later he became concerned with U and Y type relays, making the original capability studies and establishing the original requirements for U relays. During World War II he was in charge of specifications and drawings for permanent magnets for magnetron tubes, and designed special solenoids for electrically controlled torpedoes and radar scanning devices. He also developed watertight sealed relays for submarine detection apparatus. During the war and afterward he took part in the development of hydrogen annealing processes for stabilizing the magnetic characteristics of relays. More recently he has been responsible for the design and development of cost reduction items for relays, one of which is described in this issue. He is a Professional Engineer of the State of New York and holds three Bell System patents.

David Slepian, University of Michigan, 1941–43; M.A., 1947, and Ph.D., 1949, Harvard University; Bell Telephone Laboratories, 1950—. He has been engaged in mathematical research in communication theory, switching theory, and theory of noise, as well as various aspects of applied mathematics. He has been mathematical consultant on a number of Bell Laboratories projects. During the academic year 1958–59, he was Visiting Mackay Professor of Electrical Engineering at the University of California at Berkeley. Member AAAS, American Mathematical Society, Institute of Mathematical Statistics, IEEE, SIAM and U.R.S.I. Commission 6.

Richard L. Townsend, B.M.E., 1954, Cornell University; M.S.E.E., 1959, and E.E., 1960, Massachusetts Institute of Technology; Bell Telephone Laboratories, 1960—. He has been engaged mainly in studies of data transmission over telephone facilities and in the analysis of error control schemes. Member, IEEE, American Mathematical Society, ACM and Sigma Xi.

Robert N. Watts, B.S.E.E., 1955, University of Miami; S.M., 1957, and E.E., 1960, Massachusetts Institute of Technology; Bell Telephone Laboratories, 1960—. His initial assignment was development of methods for switching multiaddress data traffic. He also has participated in the development of automatic calling units and supervised a group engaged in the exploratory development and analysis of several error control systems. Mr. Watts presently supervises a group engaged in the development of DATA-PHONE systems.

Joseph H. Weber, B.E.E., 1952, Rensselaer Polytechnic Institute; M.S.E., 1959, George Washington University; Bell Telephone Laboratories, 1956—. He has been engaged in traffic analysis of communications systems, with particular emphasis on computer techniques such as stochastic simulation. He has also been concerned with the systems engineering of electronic switching systems for commercial and military applications. He presently supervises a Traffic Studies Center group primarily responsible for the analysis of military and civilian communications networks.

E. E. Zajac, B.M.E., 1950, Cornell University; M.S.E., 1952, Princeton University; Ph.D., 1954, Stanford University; Bell Telephone Laboratories, 1954—. A specialist in applied mathematics and mechanics, he has worked on the dynamics of submarine cable laying and recovery, elastic wave propagation in solids, theory of elastic stability, and theory of dynamical systems. More recently, he has been concerned with satellite attitude control studies and computer-made perspective movies. Member, ASME, American Mathematical Society, Sigma Xi, Phi Kappa Phi, Tau Beta Pi and Pi Tau Sigma.

# B.S.T.J. BRIEFS

## A Note on a Signal Recovery Problem

By I. W. SANDBERG

(Manuscript received July 17, 1964)

In a recent study[1] of the recoverability of square-integrable bandlimited signals that are distorted by a frequency-selective time-varying nonlinear system and subsequently are bandlimited to the original bands, certain assumptions were made concerning three of the four functions of frequency that characterize the linear time-invariant part of the system. These assumptions, which are stated in Section 3.4 of Ref. 1, are satisfied in most, but not all, cases of engineering interest. The purpose of this note is to report on an extension of Theorem I of Ref. 1 that covers cases in which the conditions of Section 3.4 of Ref. 1 are not met. More specifically, a proof of the following result is outlined.

*Theorem:* Let $\mathcal{L}_{2R}$ and $\mathcal{B}(\Omega)$ be as defined in Section II of Ref. 1. Let **A**, **B**, **C**, **D**, $\alpha$, $\psi(\cdot,\cdot)$, and **P** be as defined in Sections 3.1, 3.2 and 3.3 of Ref. 1, and, for any $f \varepsilon \mathcal{L}_{2R}$, let $\psi[f]$ denote the function with values $\psi(f(t),t)$ $(-\infty < t < \infty)$.

With $\Omega^*$ the complement of $\Omega$ with respect to $(-\infty, \infty)$, and

$$\begin{aligned} E(\omega) &= D(\omega) \qquad \text{for} \quad \omega \varepsilon \Omega \\ &= 1 \qquad\qquad \text{for} \quad \omega \varepsilon \Omega^*, \end{aligned}$$

let

$$\operatorname*{ess\,sup}_{-\infty<\omega<\infty} \left| [E(C-1) - PAB]^{-1} \right| < \infty,$$

and

$$(1-\alpha) \operatorname*{ess\,sup}_{-\infty<\omega<\infty} \left| E[E(C-1) - PAB]^{-1} \right| < 1.$$

Let $s_3$ be an arbitrary element of $\mathcal{B}(\Omega)$. Then $\mathcal{B}(\Omega)$ contains a unique element $s_1$, and $\mathcal{L}_{2R}$ contains unique elements $w$, $v$, and $s_2$ such that

$$v = \mathbf{A}s_1 + \mathbf{C}w, \qquad s_2 = \mathbf{D}s_1 + \mathbf{B}w,$$

$$s_3 = \mathbf{P}s_2, \qquad \text{and} \quad v = \psi[w]$$

[i.e., such that (1), (2), (3), and (4) of Ref. 1 are satisfied]. Furthermore, there exists a positive constant $k$, that depends only on **A**, **B**, **C**,

**D**, and $\alpha$, such that if

$$\bar{v} = \mathbf{A}\bar{s}_1 + \mathbf{C}\bar{w}, \qquad \bar{s}_2 = \mathbf{D}\bar{s}_1 + \mathbf{B}\bar{w},$$

$$\bar{s}_3 = \mathbf{P}\bar{s}_2\,, \qquad \text{and} \quad \bar{v} = \psi[\bar{w}]$$

where $\bar{w}$, $\bar{v}$, $\bar{s}_2 \; \varepsilon \; \mathfrak{L}_{2R}$ and $\bar{s}_1$, $\bar{s}_3 \; \varepsilon \; \mathcal{B}(\Omega)$, then

$$\| s_1 - \bar{s}_1 \| \leqq k \| s_3 - \bar{s}_3 \|.$$

*Outline of Proof:*

Let the mapping of $\mathfrak{L}_{2R}$ into itself represented in the frequency domain by multiplication by $E(\omega)$ be denoted by **E**, and let $\mathcal{K}(\Omega)$ denote the Banach space of two-vector-valued functions of $t$ belonging to

$$\mathfrak{L}_{2R} \times \mathcal{B}(\Omega),$$

with norm $\| \cdot \|'$ defined by

$$\| f \|' = \left( \int_{-\infty}^{\infty} | f_1(t) |^2 \, dt + \int_{-\infty}^{\infty} | f_2(t) |^{\frac{1}{2}} \, dt \right)^{\frac{1}{2}}, \qquad f = \begin{bmatrix} f_1 \\ f_2 \end{bmatrix} \varepsilon \; \mathfrak{L}_{2R} \times \mathcal{B}(\Omega).$$

Assume that the hypotheses of the theorem are satisfied. To prove the first part of the theorem, it clearly suffices to show that $\mathcal{B}(\Omega)$ contains a unique element $s_1$, and $\mathfrak{L}_{2R}$ contains a unique element $w$, such that

$$\psi[w] = \mathbf{A}s_1 + \mathbf{C}w, \tag{1}$$

and

$$s_3 = \mathbf{D}s_1 + \mathbf{PB}w, \tag{2}$$

in which **P** is defined in Section 3.2 of Ref. 1. For this purpose, we may replace **D** in (2) by **E** and write (1) and (2) as

$$\begin{bmatrix} 0 \\ s_3 \end{bmatrix} = \begin{bmatrix} (\mathbf{C} - \mathbf{I}) & \zeta\mathbf{A} \\ \mathbf{PB} & \zeta\mathbf{E} \end{bmatrix} \begin{bmatrix} w \\ \zeta^{-1} s_1 \end{bmatrix} - \begin{bmatrix} \tilde{\psi}[w] \\ 0 \end{bmatrix} \tag{3}$$

in which $\tilde{\psi}[w] = \tilde{\psi}[w] - w$, $\zeta$ is an arbitrary positive constant, and **I** is the identity operator.

The operator

$$\mathbf{L} = \begin{bmatrix} (\mathbf{C} - \mathbf{I}) & \zeta\mathbf{A} \\ \mathbf{PB} & \zeta\mathbf{E} \end{bmatrix}$$

is a bounded mapping of $\mathcal{K}(\Omega)$ into itself. In view of the first inequality of the theorem, it possesses an inverse on $\mathcal{K}(\Omega)$, and $\mathbf{L}^{-1}$ can be represented in the frequency domain by the matrix-valued function

$$L^{-1}(\omega) = \frac{1}{E(C-1) - PAB} \begin{bmatrix} E & -A \\ -\zeta^{-1}PB & \zeta^{-1}(C-1) \end{bmatrix}.$$

In particular, (3) can be written as

$$\begin{bmatrix} w \\ \zeta^{-1}s_1 \end{bmatrix} = \mathbf{L}^{-1}\mathbf{N}\begin{bmatrix} w \\ \zeta^{-1}s_1 \end{bmatrix} + \mathbf{L}^{-1}\begin{bmatrix} 0 \\ s_3 \end{bmatrix} \tag{5}$$

in which $\mathbf{N}$ is the operator defined on $\mathcal{K}(\Omega)$ by

$$\mathbf{N}f = \begin{bmatrix} \tilde{\psi}[f_1] \\ 0 \end{bmatrix}, \qquad f = \begin{bmatrix} f_1 \\ f_2 \end{bmatrix} \varepsilon\ \mathcal{K}(\Omega).$$

The second inequality of the theorem implies that there exists a positive number $\zeta_0$ such that $\mathbf{L}^{-1}\mathbf{N}$ is a contraction mapping of $\mathcal{K}(\Omega)$ into itself for all $\zeta > \zeta_0$. In fact, using Parseval's identity and the frequency domain representation of $\mathbf{L}^{-1}$, we find that for all $f, g\ \varepsilon\ \mathcal{K}(\Omega)$,

$$\begin{aligned} \| \mathbf{L}^{-1}\mathbf{N}f - \mathbf{L}^{-1}\mathbf{N}g \|' &\leqq \max(c_1, c_2) \| \mathbf{N}f - \mathbf{N}g \|' \\ &\leqq (1 - \alpha) \max(c_1, c_2) \| f - g \|' \end{aligned}$$

in which

$$c_1 = \operatorname*{ess\,sup}_{\omega} | E[E(C - 1) - PAB]^{-1} |,$$

and

$$c_2 = \zeta^{-1} \operatorname*{ess\,sup}_{\omega} | PB[E(C - 1) - PAB]^{-1} |.$$

In view of the contraction-mapping fixed-point theorem, this establishes the existence and uniqueness of the functions $w$ and $s_1$ (as well as the important fact that these functions can be determined by an iteration procedure that converges at a geometric rate).

The second part of the theorem follows directly from (5), the relation

$$\begin{bmatrix} \bar{w} \\ \zeta^{-1}\bar{s}_1 \end{bmatrix} = \mathbf{L}^{-1}\mathbf{N}\begin{bmatrix} \bar{w} \\ \zeta^{-1}\bar{s}_1 \end{bmatrix} + \mathbf{L}^{-1}\begin{bmatrix} 0 \\ \bar{s}_3 \end{bmatrix},$$

and the fact that $\mathbf{L}^{-1}\mathbf{N}$ is a contraction for $\zeta > \zeta_0$.

## REFERENCE

1. Sandberg, I. W., On the Properties of Some Systems that Distort Signals — II, BSTJ, **43**, Jan., 1964, p. 91.

# Detection of Weakly Modulated Light at Microwave Frequencies

By M. G. COHEN and E. I. GORDON

(Manuscript received August 10, 1964)

Studies of the photoelastic, electro-optic or magnetic-optic properties of materials at high frequencies often require the detection of microwave modulated light. In many cases, the modulation depth is sufficiently small that quantitative measurement becomes difficult if not impossible. The purpose of this brief is to describe a homodyne-superheterodyne technique which allows measurement of modulation depths of considerably less than $10^{-6}$.

At such small modulation depths, the light-associated shot noise can be large compared to the modulation signal. Under these circumstances, it is customary to use synchronous detection techniques following a sensitive superheterodyne receiver and to chop the modulation signal at some low frequency. This requires extremely good RF shielding between the modulation source and the receiver to avoid pickup. The variations in amplitude and phase of the pickup produce an unsteady output signal. Alternately, one can chop both the light and the modulation and perform the synchronous detection at a sum or difference frequency. In any case, the limiting sensitivity is determined by noise originating in the photodetector and the receiver.

The technique described here is considerably simpler and more sensitive. Fig. 1 indicates the usual synchronous detection scheme using a photodetector feeding a microwave receiver with a 30-mcs IF strip. A reference signal for the synchronous detector is derived from the light chopping wheel. The added feature is the injection into the line incident on the receiver of a small fraction of the CW modulating signal, taken from the input line to the interaction region or modulator and passed through a variable attenuator and phase shifter. The amplitude of the injected signal is kept at least 30 db larger than that of the pickup. Thus the amplitude and phase of the total injected signal, including pickup, are essentially independent of fluctuations in the phase and amplitude of the pickup. The mixer and IF strip are operated in their linear regions. Thus the signal output $v_o$ from the final stage of the IF amplifier, which is an envelope detector, can be written

$$v_o = | v_i(1 + \tilde{\Gamma}) + \tilde{v}_m + \tilde{v}_s + v_r |_{\text{time average}} . \qquad (1)$$

Here $v_i$ is the injected signal and $\tilde{\Gamma} <<< 1$ represents that part of the

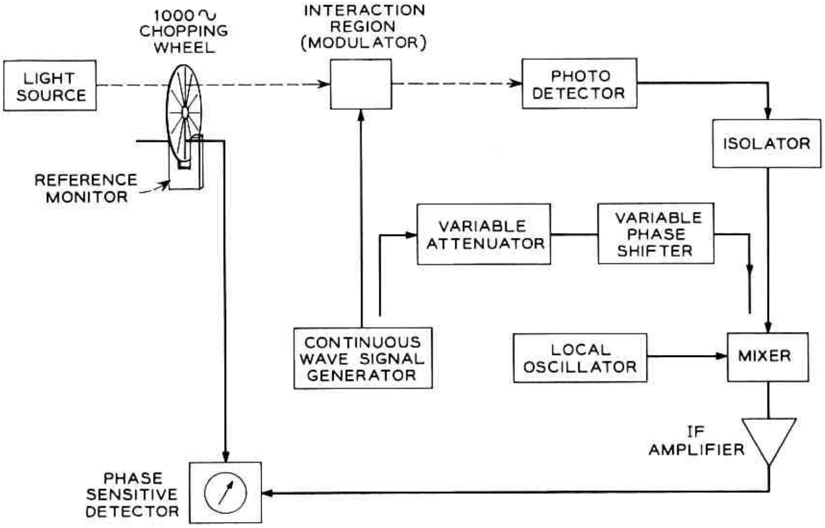


Fig. 1 — Block diagram of the homodyne-superheterodyne detection scheme.

injected signal which, because of imperfect isolation, is incident on the photodetector and is reflected with a component at the chopping frequency (because the photodetector RF impedance is dependent on the light intensity as, for example, in a photodiode) $\tilde{v}_m$ is the modulation signal from the light, $\tilde{v}_s$ is the light-associated shot noise from the photodetector and $v_r$ is the receiver equivalent input noise. The tilde over some of the quantities indicates that they are chopped at the reference frequency of the synchronous detector. By far the largest signal is $v_i$ , so $v_o$ can be written to a very good approximation

$$\begin{aligned} v_o &= | v_i | [1 + 2 \operatorname{Re} \tilde{\Gamma} + 2( | \tilde{v}_m |/| v_i | ) \cos \theta + \text{terms of order} \\ &\qquad | \tilde{v}_s |^2/| v_i |^2, | v_r |^2/| v_i |^2 \text{ and higher}] \\ &= | v_i | (1 + 2 \operatorname{Re} \tilde{\Gamma}) + 2 | \tilde{v}_m | \cos \theta + \text{terms of order} \\ &\qquad | \tilde{v}_s |^2/| v_i |, | v_r |^2/| v_i | \text{ etc.} \end{aligned} \tag{2}$$

in which $\theta$ is the phase angle between $\tilde{v}_m$ and $v_i$ , and Re indicates the real part. All other cross terms have a time average of zero; the only signals which are coherently related are $v_i$ and $\tilde{v}_m$ . Thus all noise terms can be made arbitrarily small compared to $| \tilde{v}_m |$ by making $| v_i |$ large. The contribution from the chopped term containing $\Gamma$ can be made arbitrarily small by using more isolation in the line immediately following the photodetector. In the case of a photomultiplier, this is not normally necessary. Even when these terms are not completely negligible compared to $\bar{v}_m \cos \theta$, their effect can be eliminated by varying the phase of the injected signal so that $\cos \theta$ takes on the value $\pm 1$. The only output which depends on $\theta$ is the desired modulation signal. Thus one need only take the algebraic difference between the extreme deflections of the synchronous detector as $\theta$ is varied. The fact that there is a synchronous detector deflection which depends on the phase of the injected signal is an unambiguous indication of microwave modulation on the light.

All aspects of (2) have been verified in the course of photoelastic and electro-optic modulation experiments above 150 mc by placing variable attenuators in various parts of the circuit to see if the variation of each term had the proper dependence. Modulation depths of $10^{-6}$ could be easily and accurately determined with integration times following the synchronous detector of less than one second. No special shielding was required. It should also be noted that the output of the synchronous detector is proportional to the RF amplitude rather than the square of the amplitude as in most other radiometer detection schemes. Thus, the output is proportional to the amplitude of the light modulation rather than its square.

# An Improved Error Bound for Gaussian Channels

By A. D. WYNER

(Manuscript received August 18, 1964)

## I. INTRODUCTION

The problem considered here is that of coding for the time-discrete amplitude-continuous memoryless channel with additive Gaussian noise, the code words lying on the surface of an $n$-dimensional hypersphere with center at the origin and radius $\sqrt{nP}$.

We define a *code* as a set of $M$ real $n$-vectors $\bar{x} = (x_1, x_2, \cdots, x_n)$ satisfying the ("energy") constraint,

$$\sum_{k=1}^{n} x_k^2 = nP. \tag{1}$$

The *transmission rate* $R$ is defined by $M = e^{nR}$, so that $R = (1/n) \ln M$.

## Detection of Weakly Modulated Light at Microwave Frequencies

By M. G. COHEN and E. I. GORDON

(Manuscript received August 10, 1964)

Studies of the photoelastic, electro-optic or magnetic-optic properties of materials at high frequencies often require the detection of microwave modulated light. In many cases, the modulation depth is sufficiently small that quantitative measurement becomes difficult if not impossible. The purpose of this brief is to describe a homodyne-superheterodyne technique which allows measurement of modulation depths of considerably less than $10^{-6}$.

At such small modulation depths, the light-associated shot noise can be large compared to the modulation signal. Under these circumstances, it is customary to use synchronous detection techniques following a sensitive superheterodyne receiver and to chop the modulation signal at some low frequency. This requires extremely good RF shielding between the modulation source and the receiver to avoid pickup. The variations in amplitude and phase of the pickup produce an unsteady output signal. Alternately, one can chop both the light and the modulation and perform the synchronous detection at a sum or difference frequency. In any case, the limiting sensitivity is determined by noise originating in the photodetector and the receiver.

The technique described here is considerably simpler and more sensitive. Fig. 1 indicates the usual synchronous detection scheme using a photodetector feeding a microwave receiver with a 30-mcs IF strip. A reference signal for the synchronous detector is derived from the light chopping wheel. The added feature is the injection into the line incident on the receiver of a small fraction of the CW modulating signal, taken from the input line to the interaction region or modulator and passed through a variable attenuator and phase shifter. The amplitude of the injected signal is kept at least 30 db larger than that of the pickup. Thus the amplitude and phase of the total injected signal, including pickup, are essentially independent of fluctuations in the phase and amplitude of the pickup. The mixer and IF strip are operated in their linear regions. Thus the signal output $v_o$ from the final stage of the IF amplifier, which is an envelope detector, can be written

$$v_o = \left| v_i(1 + \bar{\Gamma}) + \bar{v}_n + \bar{v}_s + v_p \right|_{\text{time average}}. \tag{1}$$

Here $v_i$ is the injected signal and $\bar{\Gamma} <<< 1$ represents that part of the

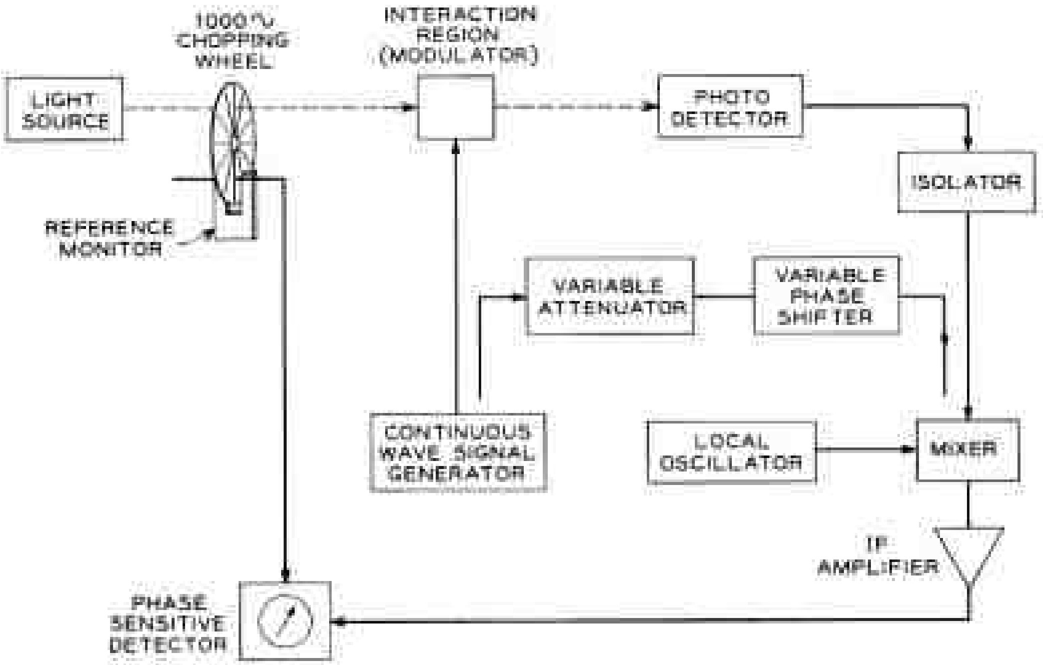


Fig. 1 — Block diagram of the homodyne-superheterodyne detection scheme.

injected signal which, because of imperfect isolation, is incident on the photodetector and is reflected with a component at the chopping frequency (because the photodetector RF impedance is dependent on the light intensity as, for example, in a photodiode) $\tilde{v}_m$ is the modulation signal from the light, $\tilde{v}_s$ is the light-associated shot noise from the photodetector and $v_r$ is the receiver equivalent input noise. The tilde over some of the quantities indicates that they are chopped at the reference frequency of the synchronous detector. By far the largest signal is $v_i$, so $v_o$ can be written to a very good approximation

$$\begin{aligned} v_o &= |v_i|[1 + 2\,\mathrm{Re}\,\tilde{\Gamma} + 2(|\tilde{v}_m|/|v_i|)\cos\theta + \text{terms of order} \\ &\qquad |\tilde{v}_s|^2/|v_i|^2,\ |v_r|^2/|v_i|^2 \text{ and higher}] \\ &= |v_i|(1 + 2\,\mathrm{Re}\,\tilde{\Gamma}) + 2|\tilde{v}_m|\cos\theta + \text{terms of order} \\ &\qquad |\tilde{v}_s|^2/|v_i|,\ |v_r|^2/|v_i| \text{ etc.} \end{aligned} \tag{2}$$

in which $\theta$ is the phase angle between $\tilde{v}_m$ and $v_i$, and Re indicates the real part. All other cross terms have a time average of zero; the only signals which are coherently related are $v_i$ and $\tilde{v}_m$. Thus all noise terms can be made arbitrarily small compared to $|\tilde{v}_m|$ by making $|v_i|$ large. The contribution from the chopped term containing $\Gamma$ can be made arbitrarily small by using more isolation in the line immediately following the photodetector. In the case of a photomultiplier, this is not normally necessary. Even when these terms are not completely negligible compared to $e_m \cos \theta$, their effect can be eliminated by varying the phase of the injected signal so that $\cos \theta$ takes on the value $\pm 1$. The only output which depends on $\theta$ is the desired modulation signal. Thus one need only take the algebraic difference between the extreme deflections of the synchronous detector as $\theta$ is varied. The fact that there is a synchronous detector deflection which depends on the phase of the injected signal is an unambiguous indication of microwave modulation on the light.

All aspects of (2) have been verified in the course of photoelastic and electro-optic modulation experiments above 150 mc by placing variable attenuators in various parts of the circuit to see if the variation of each term had the proper dependence. Modulation depths of $10^{-5}$ could be easily and accurately determined with integration times following the synchronous detector of less than one second. No special shielding was required. It should also be noted that the output of the synchronous detector is proportional to the RF amplitude rather than the square of the amplitude as in most other radiometer detection schemes. Thus, the output is proportional to the amplitude of the light modulation rather than its square.

# An Improved Error Bound for Gaussian Channels

By A. D. WYNER

(Manuscript received August 18, 1964)

## I. INTRODUCTION

The problem considered here is that of coding for the time-discrete amplitude-continuous memoryless channel with additive Gaussian noise, the code words lying on the surface of an $n$-dimensional hypersphere with center at the origin and radius $\sqrt{nP}$.

We define a *code* as a set of $M$ real $n$-vectors $\bar{x} = (x_1, x_2, \cdots, x_n)$ satisfying the ("energy") constraint,

$$\sum_{k=1}^{n} x_k^2 = nP. \tag{1}$$

The *transmission rate* $R$ is defined by $M = e^{nR}$, so that $R = (1/n) \ln M$.

ing the photodetector. In the case of a photomultiplier, this is not normally necessary. Even when these terms are not completely negligible compared to $\bar{v}_m \cos \theta$, their effect can be eliminated by varying the phase of the injected signal so that $\cos \theta$ takes on the value $\pm 1$. The only output which depends on $\theta$ is the desired modulation signal. Thus one need only take the algebraic difference between the extreme deflections of the synchronous detector as $\theta$ is varied. The fact that there is a synchronous detector deflection which depends on the phase of the injected signal is an unambiguous indication of microwave modulation on the light.

All aspects of (2) have been verified in the course of photoelastic and electro-optic modulation experiments above 150 mc by placing variable attenuators in various parts of the circuit to see if the variation of each term had the proper dependence. Modulation depths of $10^{-6}$ could be easily and accurately determined with integration times following the synchronous detector of less than one second. No special shielding was required. It should also be noted that the output of the synchronous detector is proportional to the RF amplitude rather than the square of the amplitude as in most other radiometer detection schemes. Thus, the output is proportional to the amplitude of the light modulation rather than its square.

# An Improved Error Bound for Gaussian Channels

By A. D. WYNER

(Manuscript received August 18, 1964)

## I. INTRODUCTION

The problem considered here is that of coding for the time-discrete amplitude-continuous memoryless channel with additive Gaussian noise, the code words lying on the surface of an $n$-dimensional hypersphere with center at the origin and radius $\sqrt{nP}$.

We define a *code* as a set of $M$ real $n$-vectors $\bar{x} = (x_1, x_2, \cdots, x_n)$ satisfying the ("energy") constraint,

$$\sum_{k=1}^{n} x_k^2 = nP. \qquad (1)$$

The *transmission rate* $R$ is defined by $M = e^{nR}$, so that $R = (1/n) \ln M$.

The code words are transmitted through a channel in which they are corrupted by noise, the received word $\bar{y} = (y_1, y_2, \cdots, y_n)$ being the vector sum of the transmitted word $\bar{x}$ and a noise vector $\bar{z}$, i.e.,

$$\bar{y} = (y_1, y_2, \cdots, y_n) = (x_1 + z_1, x_2 + z_2, \cdots, x_n + z_n) = \bar{x} + \bar{z}. \quad (2)$$

The components of the noise vector $z_k (k = 1, 2, \cdots, n)$ are assumed to be statistically independent Gaussian random variables with mean zero and variance $N$.

The signal "energy" is $\sum_{k=1}^{n} x_k^2 = nP$, and the expected noise "energy" is $E[\sum_k z_k^2] = nN$, so that the signal-to-noise energy ratio is $P/N$. This quantity is also the signal-to-noise "average power."

It is the task of the decoder to examine the received vector $\bar{y}$ and decide which code word $\bar{x}$ was actually transmitted. If $P_{ei}$ is the probability that the decoder makes an incorrect choice when code word $i$ is transmitted $(i = 1, 2, 3, \cdots, M)$, and if each of the $M$ code words is equally likely to be transmitted, then the over-all probability of a decoding error is

$$P_e = \frac{1}{M} \sum_{i=1}^{M} P_{ei}. \quad (3)$$

It is not hard to show that the decoding scheme which minimizes $P_e$ for a given code is the *minimum-distance decoder*, where the decoder selects that code word which has smallest Euclidean distance from the received vector and announces that word as the one which was transmitted. Thus if $\bar{y} = (y_1, y_2, \cdots, y_n)$ is the received vector, the decoder announces that code word $\bar{x}$ which minimizes (with respect to $\bar{x}$)

$$d(\bar{x},\bar{y}) = \sum_{k=1}^{n} (x_k - y_k)^2 = \sum_k x_k^2 + \sum_k y_k^2 - 2 \sum_k x_k y_k.$$

Since $\sum_k x_k^2 = nP$, $d(\bar{x},\bar{y})$ is minimized when $\sum_k x_k y_k$ is maximized. Hence minimum-distance decoding is equivalent to selection of that code word $\bar{x}$ which minimizes the angle in $n$ space $a(\bar{x},\bar{y})$ between $\bar{x}$ and $\bar{y}$, where

$$\cos a(\bar{x},\bar{y}) = \frac{\sum_k x_k y_k}{(\sum_k x_k^2)^{\frac{1}{2}} (\sum_k y_k^2)^{\frac{1}{2}}}. \quad (4)$$

The behavior of codes for this channel has been investigated in detail by Shannon,[1,2] who has shown the following:

*Fundamental Coding Theorem: Let R be any number such that*

$$R < C = \tfrac{1}{2}ln\,[1 + (P/N)].$$

*For each n, there exists an n-dimensional code with rate* $R(M = e^{nR})$ *such that the error probability is*

$$P_e = e^{-nE(R)+o(n)}, \tag{5}$$

*where the exponent* $E(R)$ *(called the "reliability") is positive when* $R < C$ *(so that* $P_e \xrightarrow{n} 0$*).*

Shannon[2] also obtained estimates of the best possible exponent

$$E(R) = \lim_{n\to\infty} -(1/n)\ln P_e\,.$$

In this note we establish the following upper bound on $E(R)$ (i.e., a lower bound on $P_e$):

$$E(R) \leqq \frac{P}{4N}\,e^{-2R} \tag{6}$$

For small rates $R$, (6) is sharper than the bounds of Ref. 2. Inequality (6) is plotted together with the estimates on $E(R)$ in Ref. 2 in Fig. 1.

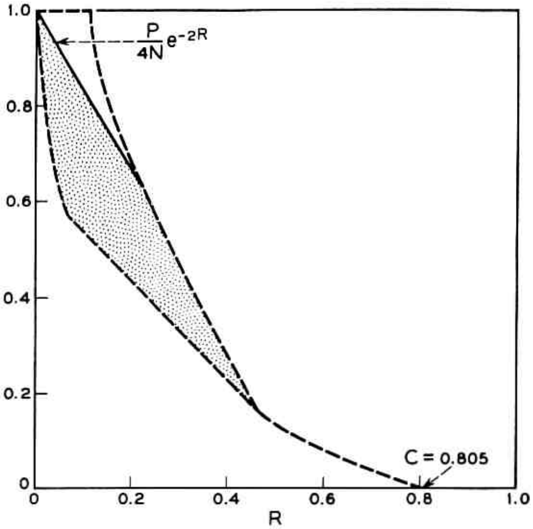


Fig. 1 — New upper bound on $E(R)$ vs $R$ for $P/N = 4$ (solid line). The bounds on $E(R)$ of Ref. 2 are in dotted lines. $E(R)$ lies in the shaded area.

## II. DERIVATION OF THE BOUND

Consider an $n$-dimensional code with $M$ code words $\bar{x}_1, \bar{x}_2, \cdots, \bar{x}_M$. Let $\theta$ be the minimum angle between pairs of code words $a(\bar{x}_i, \bar{x}_j)$ $(i \neq j)$. Denote by $\theta_n(M)$ the largest possible minimum angle $\theta$ in an $n$-dimensional code with $M$ code words, and by

$$s_n(M) = 2\sqrt{nP} \sin [\theta_n(M)/2],$$

the largest possible minimum distance between pairs of code words in an $n$-dimensional code with $M$ code words. Paralleling an argument of Shannon [Ref. 2, pp. 647–648] it is not hard to show that the error probability satisfies

$$P_e \geqq \tfrac{1}{2}\Phi\left(-\sqrt{\frac{nP}{N}} \sin \frac{\theta_n(M/2)}{2}\right), \tag{7}$$

where

$$\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-u^2/2} du$$

is the cumulative error function.

We now employ the following result of Rankin[3] to obtain an upper bound on $\theta_n(M)$:

$$M \leqq \frac{\pi^{\frac{1}{2}}\Gamma\left(\frac{n-1}{2}\right) \sin \beta \tan \beta}{2\Gamma\left(\frac{n}{2}\right) \int_0^\beta (\sin \varphi)^{n-2}(\cos \varphi - \cos \beta) d\varphi}, \tag{8}$$

where $\beta = \sin^{-1} \sqrt{2} \sin (\theta/2)$, and $\theta$ is the minimum angle in an $n$-dimensional code with $M$ code words. Taking logarithms of (8) yields

$$R = \frac{1}{n} \ln M \leqq \frac{1}{n} \ln \frac{\pi^{\frac{1}{2}}}{2} \sin \beta \tan \beta + \frac{1}{n} \ln \frac{\Gamma\left(\frac{n-1}{2}\right)}{\Gamma\left(\frac{n}{2}\right)} - \frac{1}{n} \ln \int_0^\beta (\sin \varphi)^{n-2}(\cos \varphi - \cos \beta) d\varphi. \tag{9}$$

It is shown in the appendix that for large $n$ we may approximate the upper bound of (9) by $-\ln \sqrt{2} \sin (\theta/2)$, yielding

$$\sin \frac{\theta}{2} \leqq \frac{1}{\sqrt{2}} e^{-R}. \tag{10}$$

Since for large $n$, a code with $M/2$ points has the same rate as one with $M$ points (10) and (7) yield (for large $n$)

$$P_e \geqq \tfrac{1}{2}\Phi\left(-\sqrt{\frac{nP}{N}}\,\frac{e^{-R}}{\sqrt{2}}\right). \tag{11}$$

Using the well known asymptotic form of the cumulative error function $\Phi(-x) \sim (1/x\sqrt{2\pi})e^{-x^2/2}$ (large $x$) we obtain

$$E(R) = \lim_{n\to\infty} -\frac{1}{n}\ln P_e \leqq \frac{P}{4N}\,e^{-2R}. \tag{12}$$

## APPENDIX

We must show that the limit of the right-hand member of inequality (9) as $n$ tends to infinity is $-\ln\sqrt{2}\sin(\theta/2)$. The first two terms of this quantity both tend to zero as $n$ becomes large, so that we must show the following:

Let

$$I_n = \int_0^\beta \sin^{n-2}\varphi\,(\cos\varphi - \cos\beta)d\varphi\,,$$

then

$$E = \lim_{n\to\infty}\frac{1}{n}\ln I_n = \ln\sin\beta\,.$$

*Proof:*

$$\text{(a)}\;\; I_n \leqq \int_0^\beta \sin^{n-2}\beta\,(\cos\varphi - \cos\beta)d\varphi = \sin^{n-2}\beta\,[\sin\beta - \beta\cos\beta],$$

so that

$$\frac{1}{n}\ln I_n \leqq \frac{n-2}{n}\ln\sin\beta + \frac{1}{n}\ln[\sin\beta - \beta\cos\beta] \xrightarrow{n} \ln\sin\beta\,.$$

$$\begin{aligned}\text{(b)}\qquad I_n &\geqq \int_{\beta-(\beta/n)}^{\beta} \sin^{n-2}\varphi\,(\cos\varphi - \cos\beta)d\varphi \\ &\geqq \sin^{n-2}\left(\beta - \frac{\beta}{n}\right)\int_{\beta-(\beta/n)}^{\beta}(\cos\varphi - \cos\beta)d\varphi\,.\end{aligned} \tag{13}$$

Now

$$I = \int_{\beta-(\beta/n)}^{\beta} (\cos\varphi - \cos\beta)d\varphi = \sin\beta - \sin\left(\beta - \frac{\beta}{n}\right) - \frac{\beta}{n}\cos\beta$$

$$= \sin\beta - \sin\beta\cos\frac{\beta}{n} + \cos\beta\sin\frac{\beta}{n} - \frac{\beta}{n}\cos\beta.$$

Expanding sin $(\beta/n)$ and cos $(\beta/n)$ into power series in $(\beta/n)$, we obtain

$$I = \sin\beta\left[\frac{\beta^2}{2n^2} + o\left(\frac{1}{n^2}\right)\right] = \frac{\beta^2}{2n^2}\sin\beta[1 + o(1)].$$

Thus

$$\frac{1}{n}\ln I = \frac{1}{n}\ln\frac{\beta^2}{2n^2}\sin\beta + \frac{1}{n}\ln(1 + o(1)) \xrightarrow{n} 0.$$

From (13) we have

$$\frac{1}{n}\ln I_n \geqq \frac{n-2}{n}\ln\sin\left(\beta - \frac{\beta}{n}\right) + \frac{1}{n}\ln I \xrightarrow{n} \ln\sin\beta.$$

Therefore $E = \ln\sin\beta$, which completes the proof.

REFERENCES

1. Shannon, C. E., A Mathematical Theory of Communication, B.S.T.J., **27**, July, 1948, pp. 379–424.
2. Shannon, C. E., Probability of Error for Optimal Codes in a Gaussian Channel B.S.T.J., **38**, May, 1959, pp. 611–656.
3. Rankin, R. A., The Closest Packing of Spherical Caps in $n$ Dimensions, Proc. Glasgow Math. Soc., **2**, 1955, pp. 139–144.