THE BELL SYSTEM
TECHNICAL JOURNAL

VOLUME XLVI FEBRUARY 1967 NUMBER 2

Methods of Interpreting Diagnostic Data for Locating Faults in Digital Machines

By H. Y. CHANG and W. THOMIS

(Manuscript received August 29, 1966)

Several techniques for translating the results of diagnostic tests into specific fault identities are described. This translation can be difficult in large and complex machines. The amount of test data required to isolate faults, and the obscure symptoms some faults generate, preclude efficient manual test-by-test interpretations.

The additional observed fact that a significant number of faults yield inconsistent test results from diagnosis to diagnosis demands a flexible interpretation of data. Techniques are described for producing fault dictionaries which can be used by the maintenance craftsman to identify machine faults in a relatively short time. These techniques utilize multidimensional geometric representations of diagnostic results, methods for identifying and ignoring inconsistent tests, pseudo-random mappings, and other procedures for condensing and organizing the information contained in diagnostic test data.

The results of applying these techniques to data obtained from the Bell System's No. 1 Electronic Switching System are also discussed.

I. INTRODUCTION

The problem of locating faults in digital systems is becoming more acute with the increased complexity of these machines and their expanding use in real-time applications. The need for automatic testing techniques for locating faults by automatic programmed diagnostic tests has been realized for some time and is being actively pursued in many areas.[1, 2] This need has particular urgency for a system designed to provide uninterrupted service. Such systems require extensive subsystem duplication as well as facilities for rapid fault isolation, location, and repair. The Bell System's No. 1 Electronic Switching System (No. 1 ESS), designed to control telephone switching functions, is an example.[3]

When machines were smaller and simpler, even the use of diagnostic tests to provide relevant trouble location symptoms was rare. An expert could usually locate the trouble by a quick survey of the behavior of the machine. This may be described as the "eureka" approach.

With large complex machines, however, analysis of symptoms by mere observation is lengthy and costly. Strange behavior sometimes occurs which even the expert is hard put to explain. Even the addition of test points and special diagnostic tools such as programmed testing may not immediately clarify the situation. Further, the sheer quantity of test data necessary to isolate a trouble to one of the myriad of components comprising the machine often demands preprocessing before presentation to the maintenance craftsman.

The techniques described in this paper evolved during the development of No. 1 ESS. They were devised in response to the need for translating the output of diagnostic programs into specific fault identities. Such translation techniques had to use a minimum of real time and memory while being accurate and rapidly applicable by relatively unskilled personnel. Our results were obtained by applying these techniques to the data generated on the No. 1 ESS. This system and particularly its central processor is sufficiently like other digital machines that we believe our conclusions have some general validity. For readers who are not familiar with No. 1 ESS, a brief description of its functions and maintenance plan is provided in Appendix A.

Section I reviews several conventional diagnostic data interpretation techniques—the so-called exact-match presentations, and points out some of their shortcomings. Section II reports some general facts concerning the inconsistency of test results and suggests a number of solutions to the problem. These solutions fall into two general categories: the phase dictionary approach and the cell dictionary approach. Section III describes the phase dictionary approach, its implementation in No. 1 ESS, its advantages and disadvantages, and also discusses some alternate techniques. Section IV introduces the cell dictionary approach and describes how it was implemented on No. 1 ESS. Evaluation results, obtained by applying these techniques to data obtained from No. 1 ESS, are discussed in Section V.

1.1 *Terms and General Background*

A *diagnostic test* consists of the application of special inputs to a machine for the purpose of locating a possible fault. The corresponding responses are termed the *test results* or diagnostic data. A *fault* can be defined as a physical defect in a logical element which will cause incorrect machine operations. Test results are generally processed and interpreted, either automatically by programs or by other manual means, to give field maintenance personnel the necessary information to locate and identify the faulty circuit component or packages. A *circuit package* is the smallest replaceable module of a machine. In No. 1 ESS, it consists of a relatively small number of components mounted on a printed wiring board.[3]

There are at least two types of programmed procedures for fault diagnosis: the "combinational" approach and the "sequential" approach. In the combinational approach a fixed set of tests is applied to the machine, and the results analyzed to identify the fault. The identification process generally utilizes a *fault dictionary* which is a listing of the test results of known faults organized in a fashion convenient for look-up.[2] In the sequential approach, the set of tests applied to the machine is not fixed.[4] The result of each test is used as a basis for determining the next test to be applied. Each fault, or a group of faults giving identical diagnostic results, is then identified by a certain test sequence—no additional data analysis or dictionary look-up is necessary. It is noted that this distinction is to some extent academic. A sequential analysis can be performed on data generated by the combinational approach. In cases where faults cause large numbers of tests to give inconsistent results, this may be advantageous since the sequential approach will be costly in terms of memory storage. Only if the average running time of the sequential approach is significantly less than the combinational approach *and* time is at a premium will the sequential approach be a better choice. For these reasons, the combinational approach is used in No. 1 ESS.

The fault diagnostic information necessary for generating dictionaries can be obtained by two fundamentally different procedures. The two approaches are known as "program simulation" and "hardware simulation." In program simulation the logic description of a machine is compiled into a computer program which is designed to simulate the behavior of the object machine. A particular fault can then be "introduced" into the object machine simply by appropriately changing the program description of the machine's logic. Subsequent logical simulation of the machine under control of its diagnostic program reveals the object machine's actions under test in the presence of the fault. In hardware simulation, faults are physically introduced into a real machine by replacing good circuit components with catastrophic failures such as shorts or opens. The diagnostic tests are performed and results recorded each time a fault is inserted. A fault dictionary can then be generated by processing test results obtained by either method.

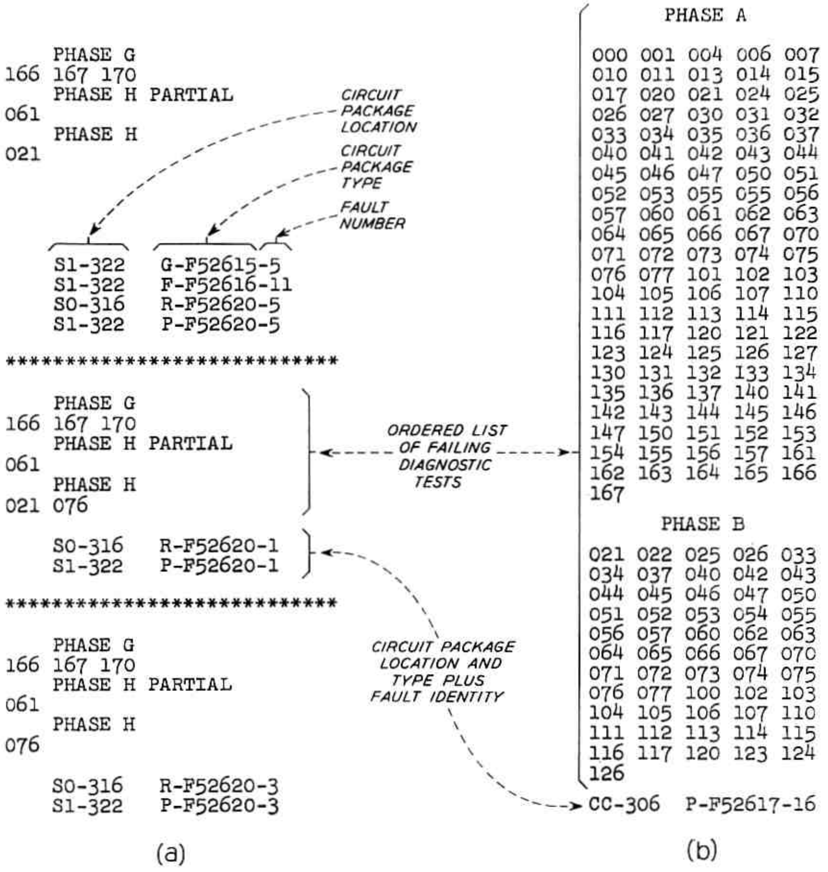
1.2 *Straightforward Dictionary Presentations*

One method of presenting diagnostic data for dictionary use is simply to list for each fault only the failing tests. This method is quite efficient when, on the average, few diagnostic tests fail.[5] A sample page (with added comments) of such a listing appears as Fig. 1 (a).

In this example, the tests were grouped into so-called test phases such as phase A, B, . . . G, H, . . . etc. The tests in each phase are numbered sequentially. Ordering of entries in the dictionary is first by phases in alphabetical order, then by test numbers within the phase. This type of dictionary was employed in the first Electronic Central Office which was in commercial use at Morris, Illinois between 1960 and 1962. A detailed description of its format and implementation is contained in Ref. 5.

As can be seen, such a technique enables a relatively unskilled maintenance man to trouble-shoot by merely searching for matches of any particular pattern with entries in the dictionary. Furthermore, in this representation, the exact configuration of test results is preserved. This feature may be useful when dictionary look-up fails to locate the trouble. The maintenance man may be able to locate faults by a direct examination of the test pattern with the aid of other documents such as diagnostic program listings, logic flow charts, etc.

However, this type of presentation suffers from the disadvantage of bulk when the diagnosis is of any considerable size (i.e., more than about 1000 tests). Further, large numbers of diagnostic tests result in large and complicated patterns, which in turn increase the difficulty of finding matches with dictionary entries [see Fig. 1 (b)]. This technique is one of a class of methods called *exact-match* techniques since an exact match between a test pattern generated by a real fault and a pattern in the dictionary is required to identify the trouble.



```
    PHASE G
166 167 170
    PHASE H PARTIAL
061
    PHASE H
021

    S1-322    G-F52615-5
    S1-322    F-F52616-11
    S0-316    R-F52620-5
    S1-322    P-F52620-5

***************************

    PHASE G
166 167 170
    PHASE H PARTIAL
061
    PHASE H
021 076

    S0-316    R-F52620-1
    S1-322    P-F52620-1

***************************

    PHASE G
166 167 170
    PHASE H PARTIAL
061
    PHASE H
076

    S0-316    R-F52620-3
    S1-322    P-F52620-3
```

```
          PHASE A

000 001 004 006 007
010 011 013 014 015
017 020 021 024 025
026 027 030 031 032
033 034 035 036 037
040 041 042 043 044
045 046 047 050 051
052 053 055 055 056
057 060 061 062 063
064 065 066 067 070
071 072 073 074 075
076 077 101 102 103
104 105 106 107 110
111 112 113 114 115
116 117 120 121 122
123 124 125 126 127
130 131 132 133 134
135 136 137 140 141
142 143 144 145 146
147 150 151 152 153
154 155 156 157 161
162 163 164 165 166
167
          PHASE B

021 022 025 026 033
034 037 040 042 043
044 045 046 047 050
051 052 053 054 055
056 057 060 062 063
064 065 066 067 070
071 072 073 074 075
076 077 100 102 103
104 105 106 107 110
111 112 113 114 115
116 117 120 123 124
126
CC-306   P-F52617-16
```

CIRCUIT PACKAGE LOCATION
CIRCUIT PACKAGE TYPE
FAULT NUMBER
ORDERED LIST OF FAILING DIAGNOSTIC TESTS
CIRCUIT PACKAGE LOCATION AND TYPE PLUS FAULT IDENTITY
(a)
(b)

Fig. 1 — Sample page of fault dictionary.

1.3 *An Improved Exact-Match Technique*

In order to reduce the problem of dictionary bulkiness and the difficulty of manual look-up of large and complicated fault patterns, a "number generation" technique is used in No. 1 ESS. This technique is essentially a "hash" storage technique which effectively "reduces" each diagnostic to a smaller fixed-length decimal number by means of a psuedo-random mapping.[6] Conceptually, the probability of mapping two or more large binary numbers into the same decimal number of relatively small size can be made arbitrarily small by a proper selection of the decimal sample space. A detailed discussion on the particular technique used in No. 1 ESS is incorporated as Appendix B.

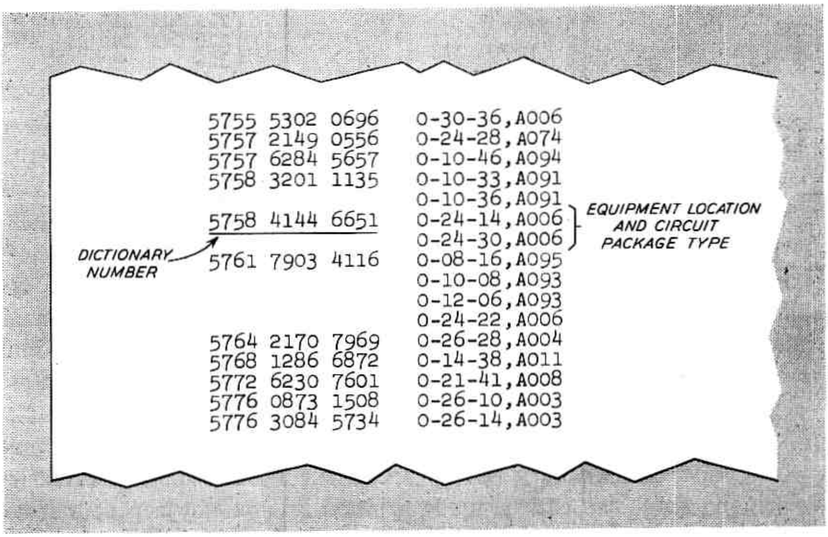
One interesting result is that if the mapping is truly random, the expected number of diagnostics that will map into the same random number is

$$E_r = k - N\left(1 - \left(1 - \frac{1}{N}\right)^k\right),$$

where N is the number of diagnostic test results to be mapped and k denotes the smallest integer greater than or equal to $\log_{10} N$, i.e., the number of digits required in the decimal number. Thus, for No. 1 ESS, in which the largest sample of fault patterns for any unit is no greater than 10^5, a 12-digit representation will suffice to insure that the number of replications, as a result of the data reduction mapping, will be less than one (if, of course, the mapping is truly random).

Fig. 2 shows a sample of a No. 1 ESS exact-match dictionary entries. Each of the 12-digit numbers was derived from a fault pattern by a pseudo-random manipulation of the pattern. As described in Appendix B, the results of this pseudo-random mapping method agreed very well with the theory.

It is estimated that a five-to-one reduction in bulk was thus achieved. In addition, it was found that the time required for look-ups was greatly reduced.



```
5755 5302 0696   0-30-36,A006
5757 2149 0556   0-24-28,A074
5757 6284 5657   0-10-46,A094
5758 3201 1135   0-10-33,A091
                 0-10-36,A091
5758 4144 6651   0-24-14,A006
                 0-24-30,A006
5761 7903 4116   0-08-16,A095
                 0-10-08,A093
                 0-12-06,A093
                 0-24-22,A006
5764 2170 7969   0-26-28,A004
5768 1286 6872   0-14-38,A011
5772 6230 7601   0-21-41,A008
5776 0873 1508   0-26-10,A003
5776 3084 5734   0-26-14,A003
```

DICTIONARY NUMBER
EQUIPMENT LOCATION AND CIRCUIT PACKAGE TYPE

Fig. 2 — Sample format of No. 1 ESS dictionary entries.

1.4 *Shortcomings of "Exact Match" Techniques*

A major problem in using such approaches is that occasionally a diagnostic generated in the field does not exactly match a dictionary entry. This will arise if the results of the field diagnosis of a given fault differ from those yielded by a diagnosis on the same fault which was used to prepare the dictionary, i.e., the test results are *inconsistent.* There are many possible causes of inconsistent test results. Some of the more obvious ones are (*i*) improper machine initialization, i.e., if the effect of the fault is such that it prevents the machine under diagnosis from being properly initialized, the test results may then vary from time to time, depending on the state of memory elements at the time the fault occurs, (*ii*) presence of intermittent or marginal faults, i.e., faults that cause machine malfunction at some times but not at others, and (*iii*) other factors such as the presence of noise and/or variations in circuit component values, etc. Consequently, supplementary techniques are desirable.

II. GENERAL FACTS CONCERNING INCONSISTENT TEST RESULTS

Suppose one had the fault patterns for a number of sample faults that were simulated on a test model of a digital machine for purposes of producing a fault dictionary. Further, suppose that these same faults were introduced into a different but supposedly identical machine and the fault patterns collected. A comparison of the two sets of data is revealing.

Two such sets of data were collected using No. 1 ESS. A sample of faults was inserted in the central processing unit of a No. 1 ESS office in Chase, Maryland and compared with dictionary results obtained from another No. 1 ESS at the laboratory. The sample consisted of 302 faults selected so as to be both well distributed and representative of expected troubles. Of the 302 faults inserted, 58 produced printouts that could not be found in the exact-match dictionary. For various reasons, only 40 of these 58 inconsistent printouts could be analyzed.* This experiment and a comparison of the test results of these 40 faults with their corresponding dictionary patterns show that:

(*i*) About 15 to 20 percent of the data for corresponding faults disagree.

(*ii*) Among the diagnostics that are inconsistent, the majority differ in only a few bits.

* This experiment is reported fully in Section 5.1.

(*iii*) Furthermore, these differences generally cluster, i.e., a relatively small group of adjacent test bits are affected. Several groups may be affected but in general, the groups do not consist of more than about 25 tests (about ½ percent of the overall diagnostic).

(*iv*) Within these clusters the diagnostics produced by the test model tend to have somewhat fewer failing tests.

(*v*) Only about 25 percent of the inconsistent diagnostics have extensive differences.

These observations are qualitative in nature and can only claim to be representative of No. 1 ESS. However, they serve to suggest a number of ways to attack the problem.

Observation (*iii*) above suggests that two differing patterns for the same fault could be made to match if the cluster of differing test bits were masked out. This idea resulted in the Phase Dictionary, the Phase Prime Dictionary, and the Test Group Dictionary (see Section III).

Observations (*ii*) and (*iv*) suggest that the two differing patterns are related and that their relationship could be expressed perhaps by some function of their Hamming distance. This idea resulted in various forms of "Cell" Dictionaries (see Section IV). We shall consider first the Phase Dictionary and some of its relatives.

III. PHASE DICTIONARY APPROACH

3.1 *Characteristics of Diagnostic Data*

The usual technique used to design diagnostic tests for a large and complicated machine is to divide the machine functionally into many small and disjoint (if possible) logic blocks. A logic block can be taken as a group of functionally related circuits, such as an order decoder or an index register, etc., whose input-output terminals are readily accessible. The tests are then designed to pinpoint faults which may exist in each logic block, assuming other logic blocks in the machine are faultless. The overall diagnostic is therefore composed of a concatenation of test results of many so-called *test phases,* each of which consists of tests that are aimed at testing a particular logic block.

Normally, when a fault occurs, it is expected that the fault will be detected by many of the tests that are specially designed to test that part of the circuitry where the fault lies. Thus, one would expect that in each overall diagnostic, the test failures would be roughly distributed over a certain number of test phases, rather than over all test phases. This is, indeed, the case as one examines, for example, the diagnostic data of one unit of the No. 1 ESS central processor complex, the *Central Control*[3] (see Fig. 3). The Central Control has a total of 28 test phases; the overall diagnostic is about 5000 bits long. Out of 102,518 faults simulated, over 97 percent of them yield diagnostics which indicate some-tests-failed in only four or fewer test phases.

3.2 *General Description of Method*

The essential idea of the Phase Dictionary approach lies in the notion of identifying "phase diagnostics," i.e., the failure patterns of individual test phases. That is, in addition to the normal "exact-match" dictionary (as it is described in Section 1.3) one further prepares a supplementary *Phase Dictionary* which is divided into many chapters, each of which is produced by processing each diagnostic phase as if it were the whole diagnostic. Then a fault would be redundantly identified by a number of dictionary entries: one in the exact-match dictionary and many in the phase dictionary, as many as the number of test phases that failed in its overall diagnostic. For example, a fault, f_i, which has failed some tests in test phases 2, 3, 4, and 6 would be identified by an entry in the exact-match dictionary and four additional entries in the phase dictionary, one each in chapters designating test phases 2, 3, 4, and 6.

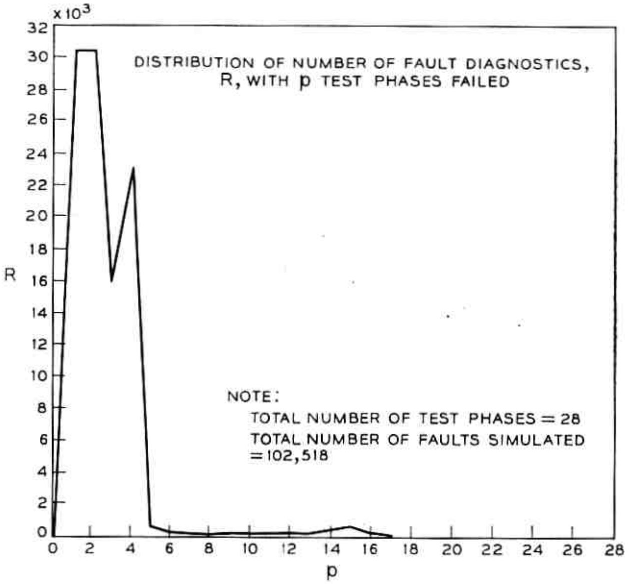


Fig. 3 — No. 1 ESS central control diagnostics.

Now suppose f_i is diagnosed in the field and its failure pattern obtained. Five numbers will then be generated: a normal exact match dictionary number, which is generated by the overall pattern of the diagnostic and four phase numbers, each of which results from manipulating diagnostics of test phases 2, 3, 4, and 6. The maintenance personnel normally will first try to find the overall number in the exact-match dictionary. If there is a match, he can simply replace the circuit package(s) indicated by the dictionary entry. If, on the other hand, he cannot find a match, he will then consult the phase dictionary by matching phase numbers with dictionary entries in chapters 2, 3, 4, and 6. Assume the inconsistency in the diagnostic is small and is confined to, for example, test phase 3 only. The maintenance man will probably discover that he can successfully match phase numbers for test phases 2, 4, and 6 with some entries in chapters 2, 4, and 6. He will not, however, find a match in chapter 3 because the field diagnostic of f_i which generated the phase number for test phase 3 *differs* from the diagnostic which was used to generate the dictionary entry. Nonetheless, the maintenance man can still examine the circuit package(s) indicated by the phase dictionary entries of test phases 2, 4, and 6 where he finds a match to determine which package(s) is to be replaced. Since he knows that the circuit package(s) associated with fault, f_i, ideally would be listed under all of these entries, he should, therefore, select the package(s) that has appeared the greatest number of times. This majority rule approach not only leads one toward replacing the proper circuit package(s) but also reduces the number of unnecessary package replacements.

A slightly more sophisticated form of the phase dictionary can eliminate the need for the somewhat laborious majority rule approach. The phase dictionary consists of entries formed by masking out all failing test bits except those in a single phase. The *Phase Prime Dictionary* consists of entries formed by masking out only a single phase at a time. Thus, if tests in phases 3, 4, and 5 failed during diagnosis, a number for phases 3 and 4, one for 3 and 5, and one for 4 and 5 would be produced. Then if the inconsistencies fall only in say phase 3, only the number produced for phases 4 and 5 could be matched. Note that the majority rule is not needed here as the dictionary for phases 4 and 5 has already performed that function automatically.

The phase prime dictionary, of course, is a little less general than the phase dictionary in that it is useful only if inconsistencies are confined to a single phase.

3.3 *Implementation in No.* 1 *ESS*

The general program flow for implementing the phase dictionary is shown in Fig. 4. The diagnostic data obtained through simulation are originally stored on magnetic tapes. Each diagnostic is read into core and is divided into many segments, one segment of each test phase. A phase dictionary number is then generated for each test phase by manipulating diagnostic data in that phase. The phase dictionary number computation is performed for each phase of all fault diagnostics. All phase dictionary numbers of the same test phase are grouped together and sorted. The sorted listings are then printed to form the various chapters of the phase dictionary.

3.4 *Advantages and Disadvantages*

The major advantage of the phase dictionary over the exact-match dictionary is its ability to locate faults whose diagnostics are inconsistent. That is, if the inconsistency of test results is confined to a small number of test phases, the phase dictionary can still locate faults by matching phase dictionary numbers. Further, the use of phase dictionary numbers also offers a possibility for identifying marginal faults. This can be done by repeatedly exercising (on-line) certain phase(s) of the diagnostic tests and then matching the phase numbers, since repeatedly exercising all diagnostic tests could be extremely costly in terms of system real time.

The phase dictionary in general tends to be bulkier than the exact-match dictionary. This is because each fault is multiply listed in the

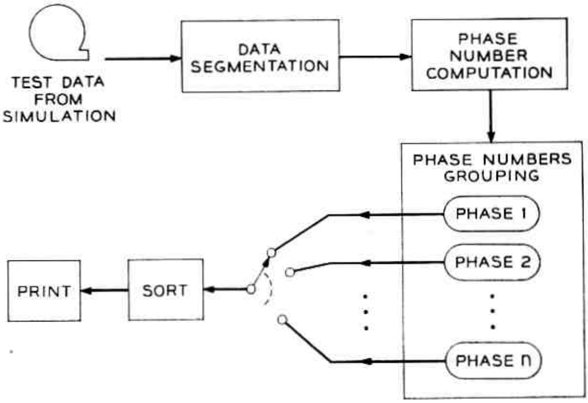


Fig. 4 — Program flow for phase dictionary implementation.

dictionary. For the particular example shown in Fig. 3, a phase dictionary is approximately 2½ times the size of the exact-match dictionary, as can be determined by calculating the average number of failing test phases per diagnostic. Another shortcoming of the phase dictionary is that if the inconsistency of test results affects all test phases or if the phase which the inconsistency affects is the *only* failing phase in the overall diagnostic, the fault cannot be successfully identified with this approach. Although wide variations in test results for the same fault are less likely to arise in a good diagnostic design, they nevertheless represent a problem.

3.5 *The Test Group Dictionary*

A third form of dictionary that uses the idea of masking out inconsistent tests has been investigated. This dictionary was constructed by considering each test of the diagnosis independently and determining what faults caused this test to fail. Those faults would then be grouped and associated with that test. When this was done for all tests, the *Test Group Dictionary* was formed.

It was originally hoped that the list of faults associated with a given test would be small enough on the average so that a majority rule technique could be used to advantage when analyzing a field diagnostic result. Unfortunately, in all of the cases implemented so far, the average listing is well over 30 faults per test. This precludes its use as a manual tool. If stored on tape, however, and searched by machine, it gives promise as being a valuable laboratory tool. It might also be useful if the maintenance facilities for a number of machines were centralized.

IV. CELL DICTIONARY APPROACH

4.1 *Introduction*

The phase dictionary approach is a technique for finding exact matches between patterns by eliminating those portions of the patterns where differences are found. The "cell" dictionary approach, on the other hand, is not concerned with exact matches between patterns but with near matches. In the cell approach, the entire fault pattern is examined and a measure of its "similarity" to other patterns in the dictionary is made. On the basis of this measure, those faults in the dictionary associated with the most "similar" patterns are identified as being more likely to have caused the observed fault pattern.* This

* This assumption can only be justified ultimately by satisfactory results in its application.

then provides the maintenance man with a list in order of probability of faults to repair.

Clearly, the efficacy of this approach depends on (*i*) the way in which "similarity" is defined and measured, and (*ii*) the manner in which this measure is used.

4.2 *"Similarity" and "Dissimilarity" of Diagnostics*

If every test in a diagnosis was of equal significance in the task of isolating faults, one could say that two diagnostic test patterns were dissimilar in proportion to the number of corresponding tests in which the two patterns differ. If the diagnostic results were expressed as binary numbers, their dissimilarity could be expressed as having a direct relationship to their Hamming distance. The greater the Hamming distance, i.e., the more places in which two patterns differ, the less alike the patterns are. Conversely, the smaller the Hamming distance, the more similar two patterns are. Consequently, Hamming distance can be used a measure of both similarity and dissimilarity.

Hamming distance is a readily computable measure. And further, it can be easily modified to take into account the fact that diagnostic tests differ in *significance*. For the purposes of explaining the idea of significance, consider a small sample of faults and just those tests of the diagnosis which fail for at least one fault in that sample. Suppose a number of these tests always pass or fail together as a group for any given fault. Then the results of the entire group could be predicted by examining the result of any one of the tests. Here the information provided by each test in the task of locating a fault is the same but diluted by a factor equal to the number of tests in the group. One would say that the significance of these tests was low. On the other hand, a test whose result cannot be predicted by the results of other tests would have a higher significance. Intuitively speaking, the significance of a test should also be affected by its consistency. Less significance should be attributed to inconsistent tests.

A natural and satisfying way of expressing these considerations quantitatively would be to assign weights to tests in accordance with their significance. Section 4.9 gives a brief discussion of some mathematical techniques available to do this.[8]

Suppose now only that each test has been assigned a weight. Then similarity would be computed by summing the weights of the differing tests in the patterns. Similarity is, thus, now measured by a weighted Hamming distance. Whether or not test weights have been assigned, however, similarity is measured by an easily calculated number.

4.3 *A Problem*

It would be possible to store the dictionary (i.e., all the fault patterns) in a computer memory and to use the similarity measure just described to search the dictionary for patterns similar to some arbitrary pattern when desired. As a practical matter, however, the time required for the search and the storage required for the dictionary are prohibitive. In the case of No. 1 ESS, for example, it is estimated that about 2.5×10^7 36-bit words would be required to store the dictionary for the central control alone. The time required for a single search would be on the order of 10 minutes provided the dictionary were stored in core (rather than on tape or disk) and provided the central control did no other work. Since No. 1 ESS is a time-shared machine, no one job is run for longer than 10 percent of any extended interval. Consequently, a single search would run well over one hour. Clearly, a more sophisticated method is needed.

4.4 *A Geometric Model for the Dictionary*

Consider, for the time being, that diagnostic results are representable by binary numbers and that similarity is measured by simple Hamming distance. It is possible to construct a geometric analog of the binary number system in binary space. In this analog, each binary number represents a unique point in the space and the "distance" between points corresponds to the Hamming distance between the patterns representing those points. Thus, there are exactly 2^N distinct points or patterns in an N-dimensional space.

Suppose that all of the data in the dictionary were placed in some N-dimensional space. For example, the dictionary data for the No. 1 ESS central control consists of about 10^5 different patterns of order 5000 (i.e., N equals 5000). Thus, 10^5 points out of a possible $2^{5000} \approx 10^{1500}$ points would be taken up by dictionary data. (The space is very sparsely populated.) Now suppose some arbitrary pattern (produced by some real fault) is placed at point A in the N-space. If point A is already occupied by a dictionary pattern, then the two patterns match and the real fault is almost certainly the same as the dictionary fault associated with that pattern. If, however, point A is unoccupied, we have (presumably) the case of an inconsistency. Then it would follow, since Hamming distances are preserved in the analog, that those dictionary points that are closest spatially will be the most similar ones according to our definition. The faults associated with these nearby points would then represent those dictionary faults which could most probably have produced the inconsistent pattern.

4.5 *The Cell Dictionary Concept*

Imagine first that a number of points, $c_i(i = 1, 2, \cdots, m)$, in the N-space are selected arbitrarily. Now imagine that every point, x_k, in the space is associated with a c_i such that the Hamming distance $d(c_i, x_k)$ is least. This procedure would then produce a number of "cells" C_i with "centers" c_i, each containing all points x_k, such that $d(c_i, x_k) < d(c_j, x_k)$, for $i \neq j$. Applying this procedure to dictionary data will result in a *Cell Dictionary*. The dictionary would consist of an ordered list of those cells which contained diagnostic results together with the fault identities corresponding to those results. Each cell can be conveniently identified by its c_i. Such a dictionary could be used as follows:

(*i*) Place in memory a list of occupied cells (i.e., cells which have diagnostic results associated with them).

(*ii*) Given an arbitrary pattern, a search would consist of computing the cell containing it and finding the two or three closest occupied cells.

(*iii*) Likely faults could now be found by consulting the printed dictionary.

This method would be practical if such cell lists were relatively small. Unfortunately, No. 1 ESS data does not lend itself to small lists. Therefore, a different form of cell dictionary was adopted.

4.6 *Selection of Practical Cell Center Lists*

In this section, we shall discuss an algorithm which permits the generation of a list of cell centers (c_i's) that, given an arbitrary pattern, can be rapidly searched.

Consider all binary numbers of order N (i.e., having N bits) and a partitioning of the N bits into k equal parts (assuming N is divisible by k). Suppose that all bits in each part, P_i, of the partitioning are assigned like values—either all 0's or all 1's. Then the subset of all numbers of order N meeting the above requirements can be placed into one-to-one correspondence with the set of all binary numbers of order k. This subset can then be taken to form a set of 2^k cell centers which divides the N space into 2^k equal-sized (i.e., containing the same number of points) cells.

To show that every point is contained in some cell and every cell contains the same number of points, imagine an arbitrary pattern of order N is divided into k equal parts. Then all binary numbers having more 0's than 1's within part P_i of this pattern are closer Hamming distance-wise to a cell center whose P_i is all 0's. Similarly, all binary numbers having more 0's (1's) than 1's (0's) within other parts, P_i's, of this pattern are closer to a cell center whose P_i's are all 0's (1's). It follows, therefore, from the definition of a cell that each point is contained in some cell. Moreover, if N/k is odd, each point will be contained in one and only one cell.

Since, in general, there are C_m^n possible binary patterns of order n with m 0's (or 1's),* the number of binary numbers having more 0's than 1's in P_i equals the number of binary numbers having more 1's than 0's. Since this holds for any P_i, and further it holds independently of any other P_i, it holds for the entire partition. Thus, each cell center will have exactly the same number of points closer to it than to any other cell center.

From the nature of the construction of cell centers, a binary number can be assigned to a cell merely by partitioning it and counting either the 0's or 1's in each part. This greatly simplifies and speeds up the assigning of patterns to cells. Furthermore, cell sizes can be varied at will merely by changing the partitioning. Very large cells will be generated if the P_i's are large and vice versa.

4.7 *The Multiple Cell Dictionaries Approach*

When these techniques were applied to No. 1 ESS, it was decided to produce a number of cell dictionaries, each corresponding to a different cell size. This approach evolved because of the problems encountered in apparently simpler implementations. For example, as was discussed in Section 4.5, one way of implementing cell dictionaries in the field would be first to produce only one cell dictionary from the laboratory data. Then, a list of those occupied cells (i.e., cells containing dictionary fault patterns) would be stored in the field ESS machine. This would enable the machine to take a pattern for a field trouble, compute the cell containing it, search the cell list, and print out a number of nearby occupied cells. In the case of No. 1 ESS, however, the list of occupied cells was very large (requiring on the order of 15,000 36-bit words of memory) and search times prohibitively long. The next possibility, which was tested and then discarded, was to print the cell dictionary (i.e., an ordered list of occupied cells) and search it manually. A search consisted of checking whether or not a cell containing a real fault was occupied. If it was not, then a check of the nearby cells was necessary. Thus, the search required that No. 1 ESS machine compute and print the cell containing the real fault and its neighboring cells in N-space. Unfortunately, the number of nearby cells in N-space can be enormous. For

* C_m^n = the number of combinations of n out of a total of m things.

example, there are, in general, C_d^k adjacent cells a distance dN/k from any given cell (N and k are as previously defined and $d = 1, 2, 3, \cdots$). Thus, if $N = 5001$ and $k = 1667$, the closest cells to the given $d = 1$ will be a distance of only 3 away and there will be 1667 of them. For $d = 2$, the number of cells will be about 1.4×10^6. As a result, unless the cell containing the fault pattern was occupied, the time for finding any nearby occupied cell could be extremely long. Consequently, a modification of this idea, the so-called *Multiple Cell Dictionaries* approach, was finally adopted.

Suppose the ESS machine computes the name of the cell that contains a real fault, and a check in the cell dictionary shows that the cell is unoccupied. Now suppose the machine repeats a similar calculation to obtain the name of a larger sized cell. Then, in general, a different cell name will result. In order to check whether or not this cell is occupied, a cell dictionary corresponding to this sized cell would have to be available and searched. Suppose the search was again unsuccessful; then the machine could compute another even larger cell and so on, for as many times as there are available dictionaries. Ultimately, as the cells grow larger, they must become occupied cells for some cell dictionary.

The computation of cells of different sizes for a real fault is relatively simple. In order to use the computed cell names, however, a number of corresponding cell dictionaries must be made available. Thus, the multiple cell dictionaries approach is a trade-off of bulk for search time. The increase in bulk, however, is not so great as might at first be suspected. This is because as the cell sizes increase the number of cells decreases and the number of faults in the "zero" cell* (which is not printed in general) increases. A qualitative evaluation of the results achieved will be presented in Section 5.1.

It should be noted that a partitioning such that each P_i consists of a a single bit will result in an exact-match cell dictionary. Thus, the cell approach can be extended to cover both the case of exact matches as well as the case of inconsistencies.

The form of cell dictionaries is exactly the same as the exact-match dictionary. This is achieved by scrambling cell center coordinates in a fashion similar to the procedure for scrambling diagnostic results for the exact-match dictionary (see Section 1.3 and Appendix B). The scrambled cell center coordinate serves as the cell name when printing the cell dictionary.

* The "zero" cell is one whose center has an all 0's pattern.

4.8 *Some Disadvantages*

Although the major advantages of the cell dictionary are fairly obvious, some of its disadvantages may not be.

First, the smallest practical cell sizes are formed with a partitioning such that each P_i consists of 3 bits. This means that any pattern that never groups at least two 1's within a single P_i will fall in the zero cell. Since the "all zeroes" diagnostic result represents the healthy machine, it could be expected that many faults that cause the machine to be "slightly sick" will fall into the zero cell or its neighbors. This is indeed the case as revealed in the No. 1 ESS diagnostic data. Thus, a fault falling in the smallest zero cell in many cases renders the cell dictionary useless because of its poor resolution. Fortunately, the situation is ameliorated by the fact that most inconsistencies occur with faults which cause many test failures.

A second disadvantage concerns the process of computing larger cells around the real inconsistent fault pattern. It is true that as the partitioning increases, larger cells containing the real fault are examined but two facts should be noted:

(*i*) The real pattern will not necessarily be close to the center of the cell, and

(*ii*) The centers of these cells will not usually coincide.

These facts mean that some faults not necessarily close to the real fault may occasionally be implicated and also that a cell may not be completely included in the next larger sized cell. Eventually, of course, as cell sizes increase, the smaller cells will be completely included but at the cost of a greater number of implicated faults.

Both of these considerations are affected by the algorithm used to select cell centers. A better algorithm might eliminate these problems. Also, an approach such as Kruskal's (see Ref. 8) can reduce if not eliminate them.

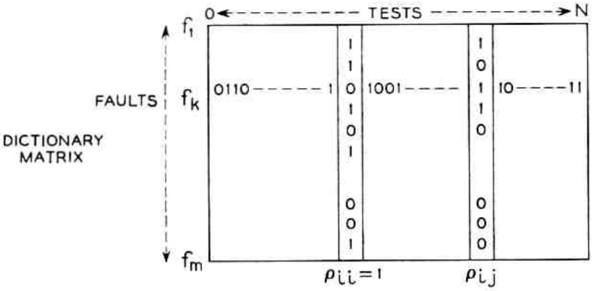
4.9 *Test Weighting*

The modification of similarity measurements by the inclusion of test weights should take the following considerations into account:

(*i*) The significance of a test relative to other tests in the task of isolating faults, and

(*ii*) The consistency of the test, i.e., what is the probability that on multiple diagnoses of the same fault, the test will give the same result as it gave on previous diagnoses.

Item (i) can be derived from the test results obtained from dictionary production. Consider a matrix formed such that each row represents the binary diagnostic result for a fault (see Fig. 5). Then each column represents the results a given test yields for each of the faults. To calculate the significance of a test relative to other tests, first obtain the correlation coefficients ρ_{ij} of column i with every other jth column using standard statistical definitions. The relative significance of the ith test which we will call σ_i is



$$\text{CONSISTENCY FACTOR} = \frac{\text{NO. OF CONSISTENT FAULTS}}{\text{TOTAL NO. OF FAULTS}} = c$$

$$\text{TEST WEIGHT } w_i = \frac{c_i}{\sum_{j=0}^{N} \rho_{ij}{}^2}$$

$$\text{SIMILARITY} \approx \frac{1}{\text{WEIGHTED HAMMING DISTANCE}}$$

Fig. 5 — Test weights.

$$\sigma_i = 1 \Big/ \sum_{j=0}^{N} \rho_{ij}^2 .$$

ρ is squared to obviate sign difficulties and down-weight small correlations.

Item (ii) consists of the ratio of the number of faults for which the test was consistent to the total number of sample faults.* The con-

* This ratio may be obtained practically by repeatedly performing the diagnosis on a given fault and observing the results of the tests. This could be done during physical simulation (at a considerable time cost). Otherwise, field data would have to be collected over a period of time. Kruskal has an idea (yet to be verified) of how to predict the inconsistency of a test theoretically. See Ref. 8.

sistency factor c_i of test i can be defined as:

$$c_i = \frac{\text{total number of faults for which the test } i \text{ was consistent}}{\text{total number of faults}}$$

A test weight w_i then is

$$w_i = f(\sigma_i, c_i).$$

A suggested function is

$$w_i = (1 + c_i \log c_i + (1 - c_i) \log (1 - c_i))\sigma_i .$$

The quantity $[1 + c_i \log c_i + (1 - c_i) \log (1 - c_i)]$ is the usual entropy function. It essentially makes $c_i = \frac{1}{2}$ the zero point (which is the value we would expect if the test was completely inconsistent) and spreads the intermediate factors appropriately.

This formulation of test weights is due to J. B. Kruskal. Ref. 8, Section VII contains an elegant discussion of test weights excluding inconsistency considerations.

Test weighting has not, as yet, been used with No. 1 ESS data. The process of calculating 5000 × 4999 ÷ 2 correlation coefficients (assuming a diagnosis of 5000 related tests in a diagnosis) is very time consuming. However, results obtained without test weights (i.e., every test given the same weight) appear quite satisfactory.

V. RESULTS AND CONCLUSIONS

5.1 *Dictionary Evaluation Results*

An experiment was conducted to evaluate the effectiveness of the No. 1 ESS exact-match dictionary, which is being used in the field, and the phase and cell dictionaries, which are being implemented. A sample of faults was inserted in the central control at a field No. 1 ESS office located in Chase, Maryland. The sample, which consists of 302 central control faults, was selected by persons who were not involved in the dictionary production project so as to reduce the possibility of bias in the selection. The faults selected were well distributed with respect to their types and locations, and were representative of expected troubles. Each fault was inserted when the office was running under a simulated traffic load and a diagnosis was performed. The corresponding diagnostic printout had three possible outcomes: (*i*) a printout that matched the correct dictionary number in the exact-match dictionary, (*ii*) a printout which did not match the dictionary number in the exact-match dictionary, or (*iii*) a printout indicating all-tests-passed.

TABLE I—EXACT–MATCH DICTIONARY EVALUATION OF FAULTS INSERTED IN NO. 1 ESS OFFICE AT CHASE, MARYLAND

Faults inserted		302	
(*i*)	Found in exact-match dictionary	216	(72.2%)
(*ii*)	Produced diagnostic printout, but not found in exact-match dictionary	58	(19.4%)
(*iii*)	Produced all-tests-passed printout	25	(8.4%)
(*iv*)	Produced invalid data for analysis	3	

Of the 302 faults inserted, there were 216 faults that were successfully located by the exact-match dictionary, 58 faults producing a printout not found in the dictionary, 25 faults producing all-tests-passed printouts, and 3 faults whose test results were either invalid or incomplete for this analysis due to errors made in the fault insertion procedure. The evaluation results for the exact-match dictionary are shown in Table I. Since the diagnostic programs were designed under a pressing time schedule with little opportunity for the feedback of evaluation results, we consider these figures quite gratifying.

Further analysis revealed that out of 25 faults producing all-tests-passed printouts, 5 were faults that could not be detected by programs because of inherent circuit redundancy and 20 were faults that were not detected due to test inadequacy. Out of 58 faults producing printouts which did not match any dictionary numbers in the dictionary, 9 were those diagnostic data had not been simulated during the dictionary production process and 9 were faults whose data were incomplete for the purpose of analysis with phase or cell dictionaries. Thus, only the data of the remaining 40 faults were analyzed to illustrate the feasibility and the effectiveness of phase and cell dictionaries. As shown in Table II, only 2 faults could not be located; all other 38 faults were found in either phase or cell dictionaries. In addition, 80 percent of the faults located by the cell dictionary were isolated to 10 or fewer circuit packages. (For this evaluation, only 5 sections of

TABLE II—PHASE AND CELL DICTIONARY EVALUATION OF FAULTS INSERTED IN NO. 1 ESS OFFICE AT CHASE, MARYLAND

Faults producing inconsistent test results		40
(*i*)	Found in phase ictiodnary	27
(*ii*)	Found in cell dictionary	35
(*iii*)	Found in either phase or cell dictionaries	38
(*iv*)	Found nowhere	2

the multiple cell dictionary were constructed. The partitioning of the diagnostic for each section was 3, 5, 11, 21, and 41 bits, respectively.) This is indeed a significant improvement. However, it must also be cautioned that this sample of faults producing inconsistent test results is too small to draw any meaningful quantitative conclusions. What can be said is that significant improvements over present exact-match techniques can be expected and that the cell dictionary approach may be somewhat superior to the phase dictionary approach.

The exact-match dictionary has been in use in No. 1 ESS office in Succasunna, New Jersey, since May 30, 1965. Its effectiveness has been more or less compatible with our evaluation results. The machine can usually be diagnosed and repaired within twenty minutes when the dictionary look-up procedure is successful. The phase and cell dictionaries will be implemented for the No. 1 ESS office in Beverly Hills, California, which will begin service sometime in the fall of 1966. The detailed field performance of all these dictionaries is not covered in the scope of this paper.

5.2 *Concluding Remarks*

The major advantages of these techniques, as a whole, are that they provide the maintenance craftsman with rapid methods for extracting the information from diagnostic test patterns for the purpose of faults location. The techniques require a very modest amount of machine time and memory. They can be quite effective especially if some care is taken during the fault simulation phase of dictionary production.

Each technique has its limitations, however. The "exact-match" dictionary will not handle inconsistencies. The phase dictionary will be of assistance if at least one phase is consistent, but at the cost of resolution and time consumed while manually searching for fault identities using the majority rule approach. The phase prime dictionary will eliminate the manual search but will work only if the inconsistency is confined to a single phase. The cell dictionary is ineffective when the fault falls into the zero cell.

We think that the results obtained will be fairly typical of what can be expected when implementing a maintenance dictionary approach on a digital machine. The combination of techniques is not perfect but is one of the most powerful for locating faults in real-time systems.

A logical continuation of this work would probably involve:

(*i*) A study of why tests are inconsistent. This would permit a better technique for eliminating inconsistencies from test patterns when generating phase dictionaries. It would also permit modifications of the measure of "similarity" between fault patterns.

(*ii*) Development of better cell center algorithms for the particular form of cell dictionary described. An "ultimate" cell dictionary is probably an implementation of Kruskal's ideas.[8]

(*iii*) Develop criteria for establishing figures of merit for dictionary techniques which take into account: (*a*) the percentage of real troubles located (as compared to simulated faults), (*b*) the resolution, (*c*) the speed or facility with which dictionary look-ups can be made, and (*d*) the machine time and memory processing requirements.

Some progress has already been made on items (*ii*) and (*iii*) but considerably more is needed.

VI. ACKNOWLEDGMENTS

The authors would especially like to acknowledge the work of J. B. Kruskal whose ideas were the stimulus for much of the work on the cell dictionary. We are also indebted to F. M. Goetz for the initial suggestions on "pseudo-random" mapping methods and the phase prime dictionary. The programming skills of R. E. Archer and J. L. Kodner were invaluable, as well as the comments of R. L. Campbell, R. W. Downing, L. S. Tuomenoksa, and many of our colleagues.

APPENDIX A

No. 1 ESS System Organization and Maintenance Plan

No. 1 ESS is a general purpose electronic telephone switching machine which employs for its control a time-shared multiple-program computer operating in real time.[3] Functionally, the system can be divided into a central processor and a peripheral system (see Fig. 6). The *central processor*, which operates with $5\frac{1}{2}$ microsecond cycle time, provides the data processing facility for telephone, maintenance and administrative functions. It consists of program stores, call stores, and a central control. The *program store*, which is a read-only type of semi-permanent memory, contains the stored program and translation information that are needed to switch calls and provide services as well as maintenance programs. The *call store,* which is a temporary ferrite sheet memory, stores all transient information for processing calls, such as the digits dialed by the subscriber or the busy-idle states of lines and trunks. The *central control* consists mainly of wired logic. Its duty is to coordinate and command all system operations.

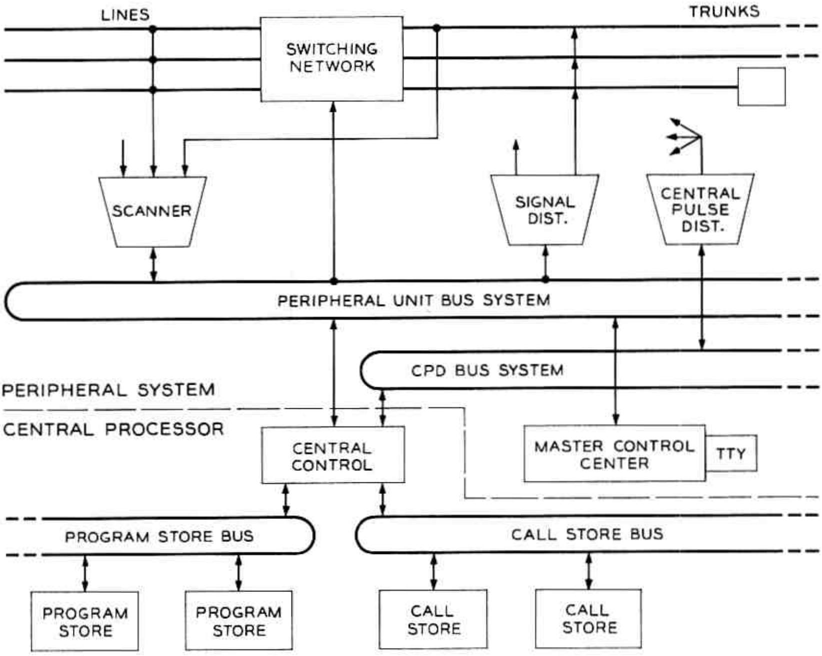


Fig. 6 — No. 1 ESS block diagram.

The *peripheral system* consists of a ferreed switching network, scanners and distributors, and a master control center. The *switching network* provides the connections between lines, trunks, and service circuits (i.e., auxiliary devices such as tone sources, signal receivers, and signal transmitters). The *scanners* are used to collect information from lines, trunks, and points internal to the system. The *distributor* is made up of two units: the *central pulse distributor,* which operates at machine cycle speed, is used for signaling logic circuits, whereas, the *signal distributor* is used for controlling slower devices such as relays in trunk circuits. The various subsystems are interconnected with balanced ac coupled bus systems. The *master control center* includes a teletypewriter for input-output, a panel for manual testing of lines and trunks, and some controls and displays.

The No. 1 ESS was designed for a high degree of maintainability and dependability. These objectives require that telephone service is not interrupted even in the presence of internal component failures.[2] To achieve these objectives, the major units of the central processor are duplicated, and circuit and program facilities are provided for detecting troubles, locating the faulty subsystems and re-establishing an operational configuration without interrupting telephone service. Once a faulty subsystem is identified, the diagnostic programs are called in to analyze the trouble. These programs are executed (on-line) by the operational configuration; they are segmented and interleaved with the main call processing program to avoid interference with the normal system operation. The diagnostic programs carry out a fixed sequence of tests by observing the normal outputs of the faulty subsystem or by monitoring some special test points strategically embedded within the unit. The combinational testing approach is used to conserve program storage space and to simplify the processing of test results. The pass or fail test data are recorded and processed using the exact-match techniques discussed previously (in Section 1.3 and Appendix B) to produce a compact diagnostic printout on the teletypewriter. The translation of a diagnostic printout into the location of the replaceable faulty circuit package(s) is then accomplished with the aid of a dictionary. The data for the dictionary are generated through hardware simulation, i.e., by actually inserting almost every possible simple hard fault sequentially into the unit and then recording its diagnostic. The general approach of dictionary production is similar to the one on Morris machine[5] and therefore, will not be described here. A sample format of exact-match dictionary entries is illustrated in Fig. 2.

APPENDIX B

Pseudo-Random Number Generation

In No. 1 ESS, the diagnostic test results of each fault is represented by an n-bit binary number where each bit or a sequence of bits designates the pass or fail result of a particular test(s). The number of bits n is usually very large, e.g., $n \approx 5000$. The "hash" technique used to reduce No. 1 ESS diagnostic data to a smaller fixed length number employs a "pseudo-random" function which manipulates an arbitrary and large pattern of test results to produce a number with relatively few digits. The reduction procedure is pseudo-random in the sense that the mapping is deterministic but approximates the process of assigning one truly random number to each fault pattern.

This process is analogous to the problem of selecting numbers at random from an urn. Assume the urn has N distinct numbers. A total of k numbers are selected, one at a time with replacement, from the urn. The probability P that all these k numbers are distinct is

$$P = \prod_{i=1}^{k-1} \left(1 - \frac{i}{N}\right)$$

$$\therefore \quad P \cong 1 - \frac{\sum_{i=1}^{k-1} i}{N} = 1 - \frac{k(k-1)}{2N}, \quad \text{for} \quad k \ll N.$$

Thus, suppose k denotes the total number of distinct fault patterns and $d = [\log_{10} N]$ denotes the number of decimal digits in the diagnostic printout (the symbol $[x]$ denotes the smallest integer greater than or equal to x). The probability of generating, at least conceptually, *all distinct* d-digit numbers from a pseudo-random number generator can be made arbitrarily large by increasing d. However, in practice the number of digits d in the printout should be kept reasonably small so as to simplify the look-up process and to reduce the dictionary's bulkiness. Hence, for a sample of 10^5 distinct fault patterns, a 12-digit representation would probably suffice, since it yields a probability of 0.995 that all resultant numbers will be distinct.

When the ratio of k to N is not exceptionally small, a few duplicated (and replicated) numbers could result. Thus, it is also necessary to compute the expected number, E_r, of replicated pseudo-random numbers when k fault patterns are assigned random values from a sample space N. Suppose E_e represent the expected number of distinct numbers generated, then the probability that a particular number is selected, at least once, from the urn in k selections is $1 - (1 - 1/N)^k$. This probability is also equal to E_e/N. Thus,

$$E_e = N\left(1 - \left(1 - \frac{1}{N}\right)^k\right).$$

The expected number of replicated numbers becomes,

$$E_r = k - E_e = \sum_{i=1}^{k-1} (-1)^{i+1} C_{i+1}^{k} \left(\frac{1}{N}\right)^i.$$

Hence, for a sample of 10^5 distinct fault patterns, the expected number of replicated numbers in a 6-digit representation is about 4837 whereas the expected number of replicated numbers in a 12-digit representation is less than one.

An experiment was performed to verify the hypothesis that this method of diagnostic data reduction is analogous to the problem of selecting numbers at random from an urn. A sample of 964 fault patterns each consisting of approximately 1000 bits of test results was used. Initially, the reduction process just used the addition and rotation instructions to produce a 6-digit number, nine duplicated resulted. On the subsequent attempts, shift and other peculiar instructions were added to better scramble the data; the number of duplicates finally decreased to zero. From the urn analog, the expected number of duplicates in assigning a 6-digit pseudo-random number to 964 distinct patterns is 0.5, and the probability of no duplication is 0.61.

The experiment demonstrates the feasibility of number generation schemes. An effective method can readily reduce each pattern of a collection of diagnostics to a smaller fixed-length number with very little loss in resolution. Since the greatest possible total number of fault patterns for any No. 1 ESS subsystem is about 10^5, the diagnostic printout uses a 12-digit representation. As mentioned earlier, this representation is amall enough so that the dictionary look-up process is easy, yet large enough so that the probability of all generated numbers being distinct is quite high and the expected number of duplicates is very small.

Basically, each fault pattern undergoes two stages of reduction process. In the first stage, each binary fault pattern of n bits is ANDed (bit-wise) with each member of a set of m preselected reference vectors $R_1, R_2, \cdots R_m$, and the resultant "bit-sums" $S_1, S_2, \cdots, S_m$ are collected ($m \ll n$). That is, suppose the binary fault pattern of fault F is $T_1T_2 \cdots T_n(T_i = 1 \text{ or } 0)$, and the pattern of reference vector R_j is $r_1^jr_2^j \cdots r_n^j(r_i^j = 1 \text{ or } 0)$. Then the bit-wise ANDing operation will generate a "bit-sum" S_j where,

$$S_j = \sum_{i=1}^{n} T_i \cdot r_i^j$$

and $j = 1, 2, \cdots, m$. To further reduce the size of the fault pattern, each set of bit-sums $S_1, S_2, \cdots, S_m$ undergoes three independent stages of "data scrambling" manipulation, each resulting in a 4-digit number. Each stage is a pseudo-random number generation procedure based on the shift, rotation, and addition orders.

The final diagnostic printout is, therefore, a 12-digit number, formed by a concatenation of three 4-digit numbers. Fig. 7 shows the general program flow of the final reduction process.

The final three stage reduction process used was evolved through experimentation. The resolution of the dictionary so generated was quite high, high enough so that most entries in the dictionary associate with only four or fewer circuit packages. For example, in the case of the central pulse distributor,[7] which has 3312 simulated faults and whose

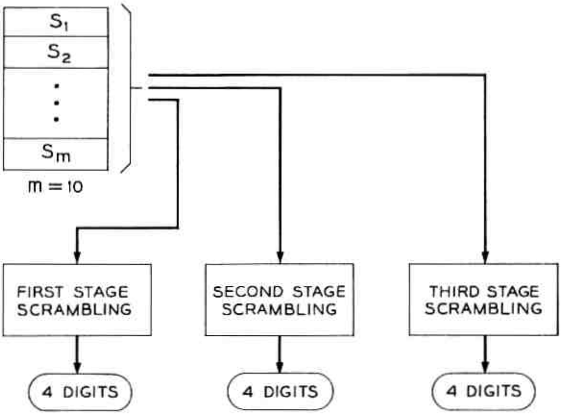


Fig. 7 — Data reduction.

TABLE III — DICTIONARY RESOLUTION STUDY (CENTRAL PULSE DISTRIBUTOR)

i	$N(i)$		
	1st stage	1st & 2nd stage	All three stages
1	436	910	924
2	300	653	663
3	101	68	67
4	65	26	26
5	18	4	3
6	49	12	10
7	5	4	2
8	54	1	1
9	8	1	1
10	7	1	1
11	3	1	1
12	3	0	0
13	0	0	0
14	1	0	0
15	0	0	0
16	1	1	1
17 AND UP	0	0	0

$N(i)$ = Number of dictionary numbers having i associated packages.

fault pattern has 1150 bits, 15 reference vectors were used. Initially, only the *first stage* of the final reduction process (rotation and addition) was used to produce a 4-digit number. In the resultant dictionary only 59 percent of the entries were associated with four or fewer circuit packages and the mean was 2.56 packages per entry (see Table III). As the *second stage* was added to better scramble the data, 90 percent of the entries in the dictionary associated with four or fewer circuit packages and the mean was improved to 1.62 packages per entry. The *third stage* was added and the resultant dictionary had 95 percent of its entries associating with four or fewer circuit packages and a mean of 1.58 packages per entry. Further experimentation with the addition of a fourth stage reduction process did not provide significant improvement in resolvability. Thus, a three-stage reduction process was adopted. The dictionary is rather compact; it has only 52 (8-½ by 11) pages (for the central pulse distributor). A sample of the format printout is shown in Fig. 2. Sample evaluation results of this type of dictionary are discussed in Section V.

REFERENCES

1. Doyle, R. H., Meyer, R. A., and Pedowtiz, R. P., Automatic Failure Recovery in Digital Data Processing System, IBM J. Res. Dev., *3,* January, 1959, pp. 2–12.
2. Downing, R. W., Nowak, J. S., and Tuomenoksa, L. S., No. 1 ESS Maintenance Plan, B.S.T.J., *43,* September, 1964, pp. 1961–2020.
3. Keister, W., Ketchledge, R. W., and Vaughan, H. E., No. 1 ESS: System Organization and Objectives, B.S.T.J., *43,* September, 1964, pp. 1831–1844.
4. Brule, J. D., Johnson, R. A., and Kletsky, E., Diagnosis of Equipment Failures, IRE Trans. Rel. Qual. Contr., *RQC-9,* April, 1960, pp. 23–24.
5. Tsiang, S. H. and Ulrich, W., Automatic Trouble Diagnosis of Complex Logic Circuits, B.S.T.J., *41,* July, 1962, pp. 1177–1200.
6. McIlroy, M. D., A Variant Method of File Searching, Commun. ACM, March, 1963, p. 101.
7. Freimanis, L., Guercio, A. M., and May, H. F., No. 1 ESS Scanner, Signal Distributor and Central Pulse Distributor, B.S.T.J., *43,* September, 1964, pp. 2255–2282.
8. Kruskal, J. B. and Hart, R. E., A Geometric Interpretation of Diagnostic Data from a Digital Machine, Based on a Study of the Morris, Illinois Electronic Central Office, B.S.T.J., *45,* October, 1966, pp. 1299–1338.

Dynamic Response of Systems of Mutually Synchronized Oscillators

By M. B. BRILLIANT

(Manuscript received September 8, 1966)

Recent studies have been concerned with conditions for the stability of synchronized systems and expressions for equilibrium frequency. This paper describes the transient response of special configurations of synchronized systems of arbitrary size, as well as frequency response and jitter response for a few cases. Tentative extrapolations to more general configurations are suggested.

I. INTRODUCTION

Recent studies have established the sufficiency of certain rather broad conditions for the stability of linear synchronized networks,[1] and have shown that valid expressions for the equilibrium frequency of such systems can be obtained if initial conditions are taken into account.[2] Description of the transient response of such systems is of interest, but results for general configurations have not been obtained. This paper describes the results of studies of special configurations of systems of arbitrary size, and some tentative conclusions about more general configurations are suggested.

II. SYSTEM EQUATIONS

The equations for the synchronized system will be taken in the form used by Gersho and Karafin[1] in their (9):

$$p'_n(t) = v_n(t) + h_n(t) * \sum_{m=1}^{N} a_{nm}[p_m(t - \tau_{nm}) - p_n(t)], \qquad n = 1, \cdots, N \qquad (1)$$

(where the star denotes convolution). In Laplace transformed form, assuming zero initial conditions,

$$sP_n(s) = V_n(s) + H_n(s) \sum_{m=1}^{N} a_{nm}e^{-s\tau_{nm}}P_m(s) - H_n(s)P_n(s), \qquad n = 1, \cdots, N. \tag{2}$$

In these equations, $p_n(t)$ is the phase of the oscillator at the nth station, $v_n(t)$ is the free-running frequency of that oscillator with the effects of local disturbances added, $h_n(t)$ is the impulse response of a control filter at the nth station, τ_{nm} is the delay on the transmission link from the mth station to the nth, and a_{nm} is an averaging coefficient associated with that link. The coefficients are normalized so that

$$\sum_{m=1}^{N} a_{nm} = 1. \tag{3}$$

Normally, a_{nn} is zero. The filter gain $H_n(s)$ has the dimensions of inverse time; its zero-frequency value, assumed to be nonnegative, is

$$H_n(0) = \lambda_n . \tag{4}$$

These equations, as pointed out by Gersho and Karafin, are conformable with the linear equations used by Karnaugh[2] if $v_n(t)$ is understood to include not only the free-running frequency of the oscillator at the nth station but also the sum of the transient disturbances at that station as well as some initial condition terms.

The assumption of zero initial conditions in (2) depends on the following simplifying procedure. Since only dynamic responses are to be studied, and since the equations are linear, the steady-state solution can be subtracted from the total response. Thus, $v_n(t)$ and $p_n(t)$ will be taken to represent only the disturbance component. Where the disturbance is transient, the $v_n(t)$ will be assumed to have zero values before the disturbance begins, and the initial phases will be taken as zero. The result of this procedure shows only the response to the disturbance, to which the steady-state solution would have to be added to determine the total frequencies and phases.

Although formal results for arbitrary filters and arbitrary delays will be obtained in a few cases, emphasis will be placed on the simple case of flat filters $H_n(s) \equiv \lambda_n$ (in effect, no filters) and zero transmission delays. In this case, the filter gains λ_n determine the time scale of the response. There seems to be no compelling practical reason to make the λ_n much greater than the reciprocal of a second. The response time can then be assumed to be large compared with the transmission delays expected in most cases as well as large enough so that the response would not be severely affected by the inherent low-pass filter effects of ordinary electromechanical control elements. The results are sufficiently encouraging, from a practical standpoint, to suggest that it may not be necessary to incorporate filtering by design, so that the simple case appears to have some practical value.

III. AN ELECTRICAL ANALOG; RECIPROCITY

Although explicit transient responses have been derived only for specific system configurations, it is possible to state, for systems of arbitrary configuration, a condition sufficient to guarantee that the transient response is not oscillatory. This condition is a reciprocity condition derived from the properties of a passive electrical network analog.

3.1 *Case 1:* $\tau_{nm} \equiv 0$, $H_n(s) \equiv \lambda_n$

Consider an electrical network as shown in Fig. 1, having N nodes in addition to a ground node. A capacitor C_n is connected from the nth node, $n = 1, \cdots, N$, to ground, and a resistor $R_{nm} = R_{mn}$ is connected between some, not necessarily all, pairs of nodes n, m. A current source delivers current from ground into each node. The Laplace-transformed node equations are

$$sE_n(s) = \frac{1}{C_n} I_n(s) + \sum_{m=1}^{N} \frac{1}{R_{nm}C_n} [E_m(s) - E_n(s)]. \tag{5}$$

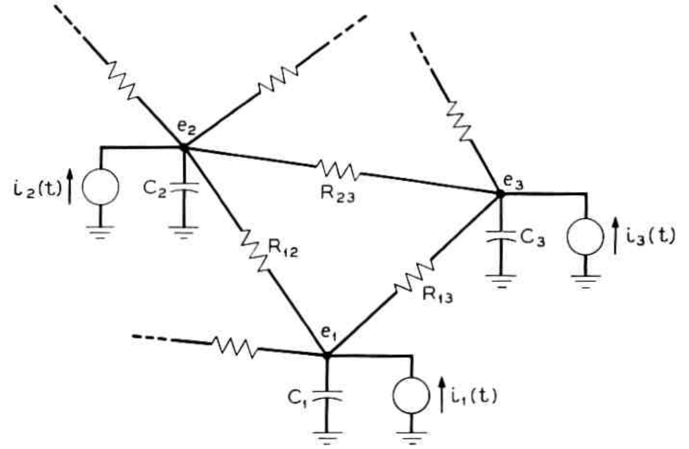


Fig. 1 — Part of the electrical analog of a reciprocal system with flat filters and zero delays.

These equations are similar in form to (2) representing a synchronized system, and can be identified with them if the nth node of the electrical network is identified with the nth station of the synchronized system, and

$$E_n(s) = P_n(s), \tag{6a}$$

$$I_n(s) = C_n V_n(s), \tag{6b}$$

$$R_{nm} = \frac{1}{a_{nm}\lambda_n C_n}, \tag{6c}$$

$$\tau_{nm} = 0, \tag{6d}$$

$$H_n(s) = \lambda_n . \tag{6e}$$

Note that $V_n(s)$ is not a voltage, but a reference frequency.

The reciprocity condition

$$R_{nm} = R_{mn} \tag{7}$$

imposes a condition on the averaging coefficients and filter gains in addition to the simplifying conditions of flat filters and zero delays. This condition immediately gives, from (6c),

$$a_{nm}\lambda_n C_n = a_{mn}\lambda_m C_m . \tag{8}$$

The capacitances C_n are to a certain extent arbitrary, in that if a system has an analog with given C_n, equivalent analogs can be formed by multiplying all the C_n by any common factor and rescaling the other elements. The capacitance at the node corresponding to any one selected station can therefore be chosen arbitrarily; (8) shows how the capacitances for stations to which it is connected can then be derived using only parameters of the synchronized system:

$$C_m = \frac{a_{nm}\lambda_n}{a_{mn}\lambda_m} C_n . \tag{9}$$

For a station that is connected to the selected one by a path of M links, via $M - 1$ intermediate stations, iteration gives a formula of the form

$$C_{n_M} = \frac{a_{n_0 n_1}}{a_{n_1 n_0}} \cdot \frac{a_{n_1 n_2}}{a_{n_2 n_1}} \cdots \frac{a_{n_{M-1} n_M}}{a_{n_M n_{M-1}}} \cdot \frac{\lambda_{n_0} C_{n_0}}{\lambda_{n_M}}, \tag{10}$$

where n is the index of the selected station and n_k is the index of the kth station in sequence along the path.

Unambiguous determination of the C_n requires that if two or more paths exist between two stations, the formula (10) should give the same result on all paths. This is equivalent to the condition that the product of the averaging coefficients taken counterclockwise around any closed loop must equal the product of the averaging coefficients taken clockwise around the loop:

$$a_{n_1n_2}a_{n_2n_3} \cdots a_{n_Mn_1} = a_{n_1n_M} \cdots a_{n_3n_2}a_{n_2n_1} \, . \tag{11}$$

This condition will be called the reciprocity condition for synchronized systems; a system that satisfies this condition will sometimes be called a reciprocal system. It is easily seen that the reciprocity condition is both necessary and sufficient for the existence of a passive electrical analog of the form of Fig. 1, assuming that the conditions of flat filters and zero delays are also satisfied.

Since the poles of an RC network response function are all simple and lie on the negative real s-axis,[3] its transient response consists entirely of real exponential components. It follows immediately that a reciprocal system with flat filters and zero delays cannot have an oscillatory transient response. Moreover, errors in parameter values that cause small departures from reciprocity cannot immediately result in the appearance of oscillatory components. Such components are represented by conjugate pairs of complex poles; since the pole locations are continuous functions of the parameter values, no pole can move off the real axis until it has first moved along the axis and joined another real-axis pole to form a double pole, assuming that the departure from the reciprocal ideal is not of such form as to add new poles.

3.2 *Case 2: τ_{nm} small, $H_n(s)$ nearly flat*

This conclusion is strictly true only for zero delays and flat filters. However, it may be expected that delays much smaller than the system response time, or filters that are nearly flat up to frequencies much larger than the reciprocal of the response time, will have little effect on the transient response. In fact, it can be shown in specific cases that the addition of any delay, however small, introduces an infinite number of oscillatory components, which nevertheless are small in amplitude and rapidly damped so that their total effect is small. It may be assumed that the omission of delays and high frequency cutoffs is comparable to the neglect of the same parameters in ordinary circuit analysis.

It is not necessary that the filters be flat in order that the system have an electrical analog. The resistors can be replaced by any 2-terminal networks so as to simulate any filter that has a "positive real" frequency response function. If the transfer function of the filter can be synthesized as the admittance of a network of resistors and capacitors only, the analog will still be an RC network and the transient response obviously remains nonoscillatory.

While this study is nominally confined to dynamic behavior, the Appendix shows how the reciprocity condition simplifies the steady-state analysis.

V. E. Beneš has pointed out that if the a_{nm} are considered as the transition probabilities of a Markov process, as in his original study (unpublished work, 1959) of stability and equilibrium frequency, the reciprocity condition introduced here is equivalent to the condition of reversibility of the Markov process, which in turn is related to detailed balance in statistical mechanics.

IV. TWO-STATION SYSTEMS

The analysis of a system of two stations offers not only an introduction to the techniques of analysis but also an example of the behavior of small systems for comparison with the behavior of the large systems to be described in later sections.

An impulse disturbance of frequency is assumed to occur at one of the stations, which we then designate (without loss of generality) as station 1. This form of disturbance can be interpreted as a brief rise in frequency which is almost immediately corrected, leaving a residual phase error of one unit of phase. Alternatively, it could represent any disturbance that gives rise to the sudden appearance of a phase error.

The system equations, from (2), are

$$\begin{aligned} sP_1(s) &= 1 + H_1(s)e^{-s\tau_{12}}P_2(s) - H_1(s)P_1(s), \\ sP_2(s) &= H_2(s)e^{-s\tau_{21}}P_1(s) - H_2(s)P_2(s). \end{aligned} \tag{12}$$

These equations are easily solved to give

$$\begin{aligned} P_1(s) &= \frac{s + H_2(s)}{s^2 + s[H_1(s) + H_2(s)] + H_1(s)H_2(s)[1 - e^{-s(\tau_{12}+\tau_{21})}]}, \\ P_2(s) &= \frac{H_2(s)e^{-s\tau_{21}}}{s^2 + s[H_1(s) + H_2(s)] + H_1(s)H_2(s)[1 - e^{-s(\tau_{12}+\tau_{21})}]}. \end{aligned} \tag{13}$$

The final value theorem gives

$$p_1(\infty) = p_2(\infty) = \frac{\lambda_2}{\lambda_1 + \lambda_2 + \lambda_1\lambda_2(\tau_{12} + \tau_{21})} \tag{14}$$

as the ultimate displacement in phase caused by the disturbance. The equality of the two final values, signifying no net change in the phase difference between the two stations, is a necessary consequence of the uniqueness of the steady-state solution.

4.1 *Case 1:* $H_n(s) \equiv \lambda_n$, $\tau_{nm} \equiv 0$.

More explicit results for the transient response are obtained in the special case of flat filters and zero delays. The transforms become simple enough for inversion by inspection; the result in the time domain is

$$
\begin{aligned}
p_1(t) &= \frac{\lambda_2}{\lambda_1 + \lambda_2} + \frac{\lambda_1}{\lambda_1 + \lambda_2} e^{-(\lambda_1+\lambda_2)t}, \\
p_2(t) &= \frac{\lambda_2}{\lambda_1 + \lambda_2} [1 - e^{-(\lambda_1+\lambda_2)t}].
\end{aligned}
\tag{15}
$$

These equations indicate a simple exponential approach to the final value, starting with initial phases [immediately after the impulse in $v_1(t)$] of 1 at the first station and 0 at the second. Such behavior appears satisfactory for a practical system.

4.2 *Case 2:* $H_n(s) \equiv \lambda$, $\tau_{nm} \equiv \tau$.

To determine the effect of delays, the system will be made as simple as possible in other respects. The filters will be assumed flat with equal gains, and the delays will be assumed equal. In this case, the solution (13) becomes

$$
\begin{aligned}
P_1(s) &= \frac{s + \lambda}{s^2 + 2\lambda s + \lambda^2 - \lambda^2 e^{-2s\tau}}, \\
P_2(s) &= \frac{\lambda e^{-s\tau}}{s^2 + 2\lambda s + \lambda^2 - \lambda^2 e^{-2s\tau}}.
\end{aligned}
\tag{16}
$$

The denominator can be factored, and a partial expansion in partial fractions gives

$$
\begin{aligned}
P_1(s) &= \tfrac{1}{2}[Q_1(s) + Q_2(s)], \\
P_2(s) &= \tfrac{1}{2}[Q_1(s) - Q_2(s)],
\end{aligned}
\tag{17}
$$

where

$$
\begin{aligned}
Q_1(s) &= \frac{1}{s + \lambda - \lambda e^{-s\tau}}, \\
Q_2(s) &= \frac{1}{s + \lambda + \lambda e^{-s\tau}}.
\end{aligned}
\tag{18}
$$

One approach to the inversion of these transforms is to divide numerator and denominator by $s + \lambda$ and treat the result as the summation of a geometric series. Expansion of the series gives

$$
\begin{aligned}
Q_1(s) &= \sum_{m=0}^{\infty} \frac{\lambda^m e^{-ms\tau}}{(s+\lambda)^{m+1}}, \\
Q_2(s) &= \sum_{m=0}^{\infty} \frac{(-\lambda)^m e^{-ms\tau}}{(s+\lambda)^{m+1}},
\end{aligned}
\tag{19}
$$

from which, by (17),

$$
\begin{aligned}
P_1(s) &= \sum_{k=0}^{\infty} \frac{\lambda^{2k} e^{-2ks\tau}}{(s+\lambda)^{2k+1}}, \\
P_2(s) &= \sum_{k=0}^{\infty} \frac{\lambda^{2k+1} e^{-(2k+1)s\tau}}{(s+\lambda)^{2k+2}}.
\end{aligned}
\tag{20}
$$

Inversion term by term gives

$$
\begin{aligned}
p_1(t) &= \sum_{k=0}^{[t/2\tau]} \frac{\lambda^{2k}(t-2k\tau)^{2k} e^{-\lambda(t-2k\tau)}}{2k!}, \\
p_2(t) &= \sum_{k=0}^{[(t-\tau)/2\tau]} \frac{\lambda^{2k+1}[t-(2k+1)\tau]^{2k+1} e^{-\lambda[t-(2k+1)\tau]}}{(2k+1)!},
\end{aligned}
\tag{21}
$$

where the square bracket in the limit of summation (but only there) denotes the integer part of the enclosed expression. This result can be numerically evaluated term by term if the product $\lambda\tau$ is known. It gives an exact result (for the assumed model) up to a time depending on the number of terms evaluated. Fig. 2 shows a graph of the calculated results for $\lambda\tau = 0.1$, that is, delay equal to one-tenth of the reciprocal of the filter gain.

The interpretation of this result is that the response of each station to changes in phase at the other is delayed for a time equal to the link delay τ. Thus, from $t = 0$ to $t = \tau$, station 2 is completely undisturbed. Meanwhile, from $t = 0$ to $t = 2\tau$, station 1 observes no change in the frequency received from station 2 and therefore, its response is exponential with time constant $1/\lambda$. Therefore, from $t = \tau$ to $t = 3\tau$ station 2 responds to the exponential response received from station 1, and so on. The result (21) could in fact have been derived by tracing out the response of the system in this manner.

A second approach, inherently inexact but more useful for times that are long compared to the transmission delay, is to complete the partial-fraction expansion of (18). This requires in principle determina-

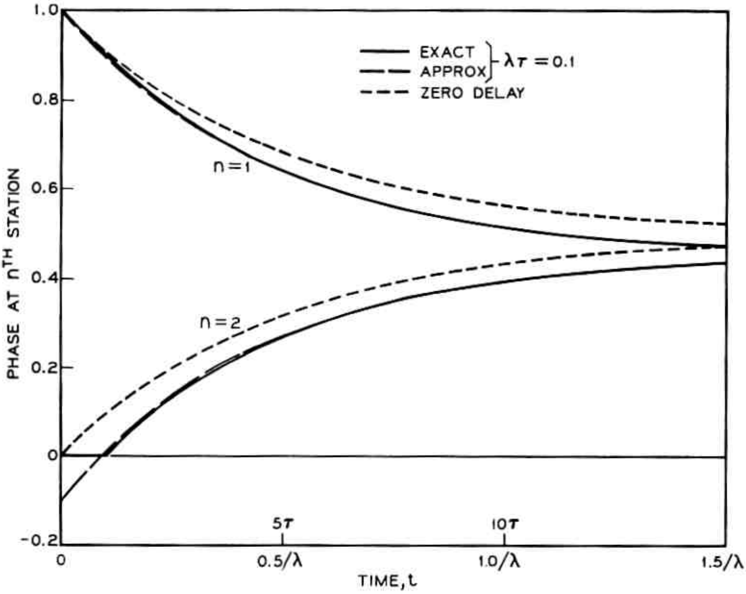


Fig. 2 — Impulse response of a two-station system.

tion of the locations of all the poles, which are determined by the transcendental equation

$$s + \lambda = \pm\lambda e^{-s\tau}, \tag{22}$$

where the "plus" sign refers to $Q_1(s)$ and the "minus" sign to $Q_2(s)$. This equation has, in general, an infinity of solutions. However, if $\lambda\tau$ is a small number, the most important components will be those due to poles of the order of magnitude of λ. The exponent in (22) is then small, so that the exponential can be approximated as

$$e^{-s\tau} \approx 1 - s\tau. \tag{23}$$

Using this approximation in (18) gives a form which is easily inverted to give, finally, from (17),

$$\begin{aligned} p_1(t) &\approx \frac{1}{2}\left[\frac{1}{1+\lambda\tau} + \frac{1}{1-\lambda\tau}\, e^{-2\lambda t/(1-\lambda\tau)}\right] \\ p_2(t) &\approx \frac{1}{2}\left[\frac{1}{1+\lambda\tau} - \frac{1}{1-\lambda\tau}\, e^{-2\lambda t/(1-\lambda\tau)}\right]. \end{aligned} \tag{24}$$

This obviously has an error at $t = 0$, which is small if $\lambda\tau$ is small, but has the correct final value determined by (14). This result is compared with the exact response, as well as with the response of a zero-delay system, in Fig. 2, which illustrates the case of $\lambda\tau = 0.1$. The approximation is better for smaller $\lambda\tau$.

V. LARGE FULLY INTERCONNECTED SYSTEMS

Next to be considered is a network of N identical stations in which all stations transmit via identical direct links to all others. All coefficients a_{nm} are assumed equal:

$$a_{nm} = \frac{1}{N-1}, \qquad n = 1, \cdots, N, \qquad m \neq n. \tag{25}$$

If an impulse disturbance occurs at the first station, all the other stations display identical responses, so that

$$p_2(t) = p_3(t) = \cdots = p_N(t). \tag{26}$$

The system response can, therefore, be described in terms of two equations in $P_1(s)$ and $P_2(s)$:

$$\begin{aligned} sP_1(s) &= 1 + H(s)e^{-s\tau}P_2(s) - H(s)P_1(s) \\ sP_2(s) &= \frac{H(s)}{N-1}e^{-s\tau}[P_1(s) + (N-2)P_2(s)] - H(s)P_2(s). \end{aligned} \tag{27}$$

These equations can be formally solved to give

$$\begin{aligned} P_1(s) &= \frac{s + H(s) - \left(\frac{N-2}{N-1}\right)H(s)e^{-s\tau}}{\Delta} \\ P_2(s) &= \frac{H(s)e^{-s\tau}}{(N-1)\Delta}, \end{aligned} \tag{28}$$

where

$$\begin{aligned} \Delta = s^2 &+ sH(s)\left[2 - \left(\frac{N-2}{N-1}\right)e^{-s\tau}\right] \\ &+ H^2(s)\left[1 - \left(\frac{N-2}{N-1}\right)e^{-s\tau} - \left(\frac{1}{N-1}\right)e^{-2s\tau}\right]. \end{aligned} \tag{29}$$

The final value theorem gives

$$p_1(\infty) = p_2(\infty) = \frac{1}{N(1 + \lambda\tau)}. \tag{30}$$

5.1 *Case 1:* $H_n(s) \equiv \lambda$, $\tau_{nm} \equiv 0$.

In the special case of flat filters and no delays, inversion of the Laplace transforms gives

$$
\begin{aligned}
p_1(t) &= \frac{1}{N} + \left(1 - \frac{1}{N}\right)e^{n\lambda t/(N-1)} \\
p_2(t) &= \frac{1}{N}\,[1 - e^{-n\lambda t/(N-1)}].
\end{aligned}
\tag{31}
$$

In this case, the response is an exponential approach to equilibrium as in the 2-station system. When one station is disturbed, the other stations respond in unison, as one.

5.2 *Case 2:* $H_n(s) \equiv \lambda$, $\tau_{nm} \equiv \tau$

In the case of equal positive delays and flat filters, the solution (28) in transformed form can be partially expanded in partial fractions to give

$$
\begin{aligned}
P_1(s) &= \frac{1}{N}\,[Q_1(s) + (N - 1)Q_2(s)], \\
P_2(s) &= \frac{1}{N}\,[Q_1(s) - Q_2(s)],
\end{aligned}
\tag{32}
$$

where

$$
\begin{aligned}
Q_1(s) &= \frac{1}{s + \lambda - \lambda e^{-s\tau}}, \\
Q_2(s) &= \frac{1}{s + \lambda + \dfrac{\lambda}{N - 1}\,e^{-s\tau}}.
\end{aligned}
\tag{33}
$$

This is similar in form to (17) and (18), and the same methods can be used to evaluate the transient response. The principal difference between this and the 2-station case is that the conditions for cancellation of odd or even terms in the series of delayed responses (21) do not hold in the many-station case, and the antisymmetric component, $q_2(t)$, is more rapidly damped than the symmetric component $q_1(t)$. The results for the zero-delay case and for the case of $\lambda\tau = 0.1$ are shown for a 6-station system in Fig. 3.

The simplicity of both the analysis and the result can be attributed to the condition that all stations and all paths are identical. Although the effects of slight departures from this condition may be of practical

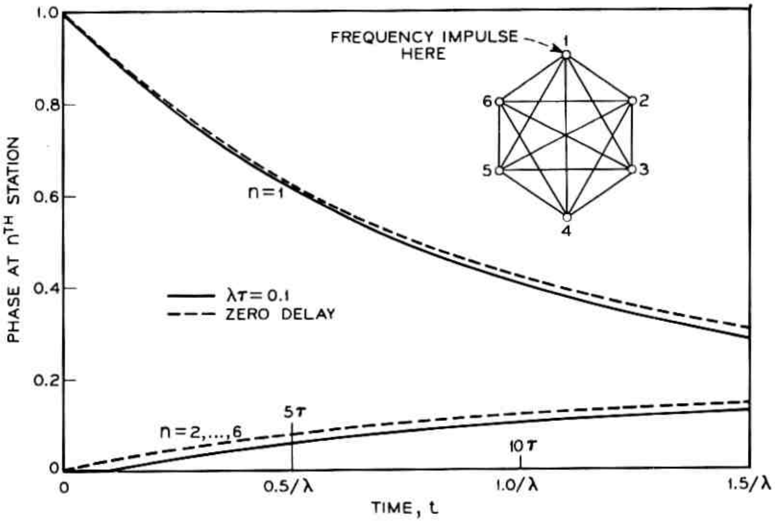


Fig. 3 — Impulse response of a fully interconnected 6-station system.

interest, the slightest departure will destroy the symmetry and vastly complicate the analysis. As a guess, it may be supposed that the effect of a slight dissimilarity among paths will be smaller than the effect of removing some of the paths. When all but N paths have been removed, in such a way that the system forms a ring in which each station receives only from its two nearest neighbors, a new form of symmetry appears, which will be used in the next section.

VI. THE BILATERAL RING

A bilateral ring is defined as a ring of N identical stations, with $2N$ identical one-way links forming N two-way links by which each station sends to, and receives from, its two nearest neighbors, one on each side. This may be viewed as the opposite extreme to the fully interconnected system, providing the longest possible indirect paths in a system of N identical stations. (Longer paths are possible in a chain, but the stations cannot be identical because each end station has only one neighbor.)

The equations of the bilateral ring, in transform form, are

$$sP_n(s) = V_n(s) + H(s)\{\tfrac{1}{2}[P_{n+1}(s) + P_{n-1}(s)]e^{-s\tau} - P_n(s)\},$$
$$n = 1, 2, \cdots, N, \qquad (34)$$

where addition in the index n is performed modulo N, so that $P_{N+1}(s)$ is $P_1(s)$, $P_0(s)$ is $P_N(s)$, and the $(N - m)$th station can be alternatively designated as the $(-m)$th. This system of equations will be simplified by a form of Fourier analysis. We define

$$Q_k(s) = \sum_{n=1}^{N} P_n(s)e^{-i2\pi nk/N}, \qquad k = 1, \cdots, N, \tag{35}$$

where j is the imaginary unit. It can then be shown by direct substitution that

$$P_n(s) = \frac{1}{N}\sum_{k=1}^{N} Q_k(s)e^{i2\pi nk/N}, \qquad n = 1, \cdots, N. \tag{36}$$

Similarly, variables $U_k(s)$ will be defined by transformation of the $V_n(s)$ as in (35), with inversion as in (36). The linearity of the Laplace transformation implies similar relations among the variables in the time domain. All these relations remain unaffected if any n or k is changed by adding or subtracting N.

Let the nth equation in (34) be multiplied by $e^{-i2\pi nk/N}$, and the equation summed over all n. The result is

$$sQ_k(s) = U_k(s) + H(s)[\tfrac{1}{2}(e^{i2\pi k/N} + e^{-i2\pi k/N})e^{-s\tau} - 1]Q_k(s), \qquad k = 1, \cdots, N. \tag{37}$$

This can be solved immediately to give

$$Q_k(s) = \frac{U_k(s)}{s + H(s)[1 - e^{-s\tau}\cos(2\pi k/N)]}. \tag{38}$$

Given a set of transient frequency disturbances $v_n(t)$, one may find their Laplace transforms $V_n(s)$, find the $U_k(s)$ using (35), find the $Q_k(s)$ using (38), use (36) to obtain $P_n(s)$, and find the phase disturbances $p_n(t)$ by inverse transformation.

In the case of an isolated impulse in frequency at the Nth station, we have

$$V_n(s) = 0, \qquad n = 1, 2, \cdots, N-1; \qquad V_N(s) = 1. \tag{39}$$

By using (35) we get

$$U_k(s) = 1, \qquad k = 1, \cdots, N. \tag{40}$$

Explicit solutions will be obtained here only for cases in which the filters are flat. Under these conditions,

$$Q_k(s) = \frac{1}{s + \lambda[1 - e^{-s\tau}\cos(2\pi k/N)]}. \tag{41}$$

The complexity of the result depends on whether the delay τ is assumed zero or positive.

6.1 *Case 1:* $H_n(s) \equiv \lambda,\ \tau_{nm} \equiv 0,\ N < \infty$.

If the delay is zero, (41) can be inverted immediately to give

$$q_k(t) = \exp\{-\lambda[1 - \cos(2\pi k/N)]t\} \tag{42}$$

and the phase disturbances, using (36), are

$$p_n(t) = \frac{1}{N}\sum_{k=1}^{N} e^{j2\pi nk/N} \exp\{-\lambda[1 - \cos(2\pi k/N)]t\} \tag{43}$$

The Nth term in this sum is real, as is the $(N/2)$th term if N is even. For all other k, the kth term is the complex conjugate of the $(N - k)$th term, so that the sum is real, and may be expressed as the sum of the the real parts of the individual terms:

$$p_n(t) = \frac{1}{N}\sum_{k=1}^{N} \cos(2\pi nk/N) \exp\{-\lambda[1 - \cos(2\pi k/N)]t\}. \tag{44}$$

The Nth term in this sum is a constant term, which applies equally to all stations and does not affect the phase differences between stations. All other terms are real exponentials approaching zero with increasing time. The dashed curves in Figs. 4, 5, and 6 show the response of a 6-station ring calculated from (44).

6.2 *Case 2:* $H_n(s) \equiv \lambda,\ \tau_{nm} \equiv 0,\ N = \infty$.

This result can be extended to rings of indefinitely large size in two different ways, so as to specify the response either a given number of stations away from the source of the disturbance, or a given fraction of the circumference away from the source. For the first approach, which gives an exact result for an infinite ring, let

$$\theta_k = 2\pi k/N \tag{45}$$

and let N increase without limit (approach infinity). Then the limit of (43) defines the integral

$$p_n(t) = \frac{e^{-\lambda t}}{2\pi}\int_0^{2\pi} e^{jn\theta} \exp(\lambda t \cos\theta)\, d\theta, \tag{46}$$

which is related to a known integral form[4] for the modified Bessel function of the first kind, order n, and gives

$$\begin{aligned} p_n(t) &= e^{-\lambda t} I_n(\lambda t), \qquad n = \cdots -1, 0, 1, \cdots \\ &= p_{-n}(t). \end{aligned} \tag{47}$$

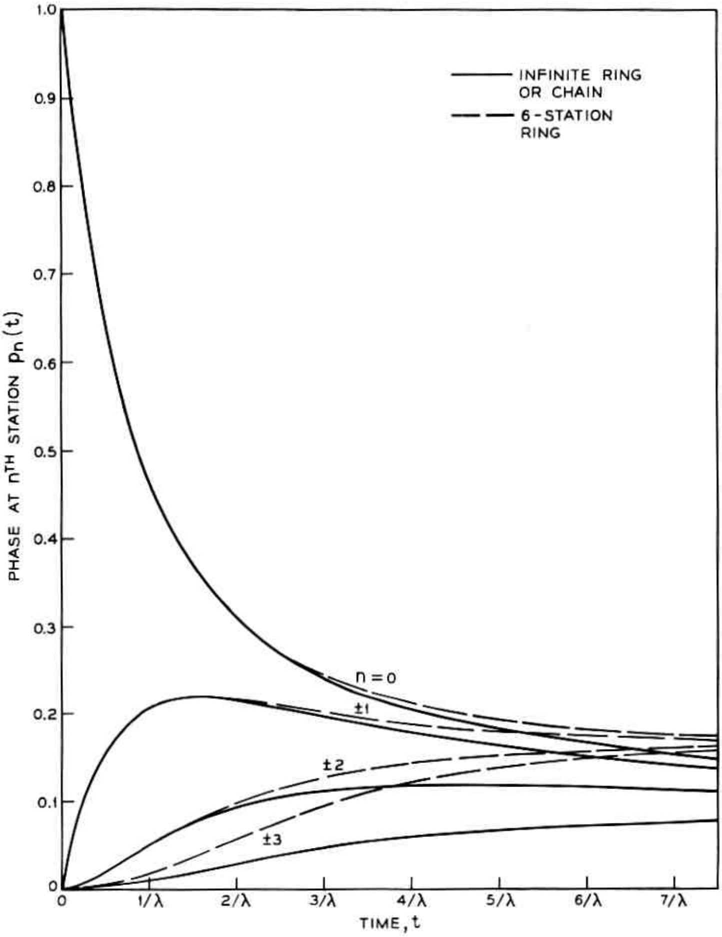


Fig. 4 — Impulse response of bilateral rings with zero delays.

Curves calculated from this equation are shown as the solid curves in Fig. 4.

Full exploitation of this result requires that the station at which the disturbance originates be called the zeroth, and that neighboring stations be indexed with positive integers to one side and negative integers to the other side. At any time t, the largest phase disturbance is at the station at which the original disturbance occurred. The asymptotic approximation for large x,

$$I_n(x) \approx \frac{e^x}{\sqrt{2\pi x}} \tag{48}$$

shows that the phase disturbance decreases with increasing time roughly as

$$p_n(t) \approx \frac{1}{\sqrt{2\pi\lambda t}}. \tag{49}$$

Although this result gives the wrong limit for a finite system, it gives a clear picture of the early behavior while the response is still substantially localized.

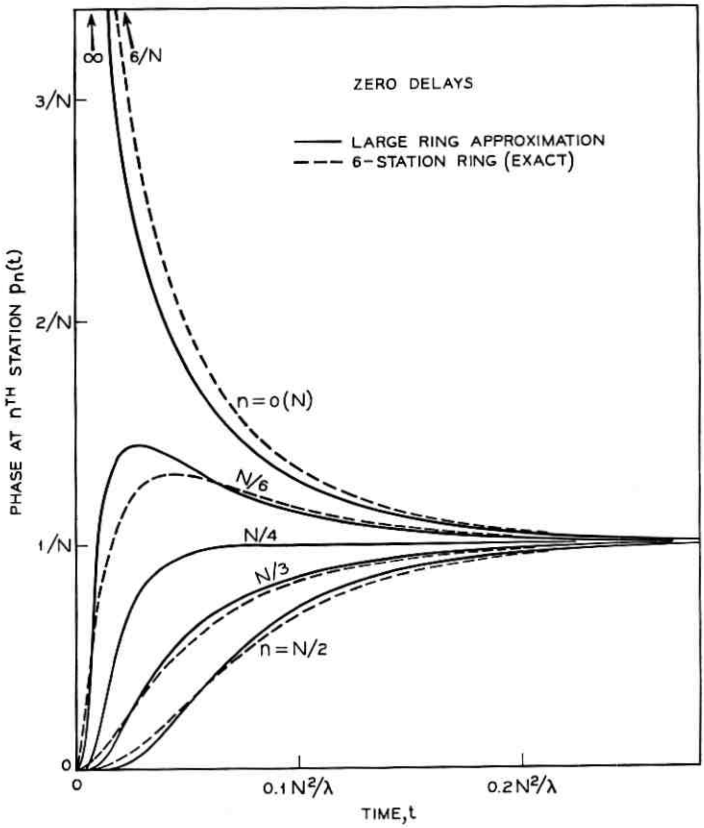


Fig. 5 — Large-t approximation for a large bilateral ring.

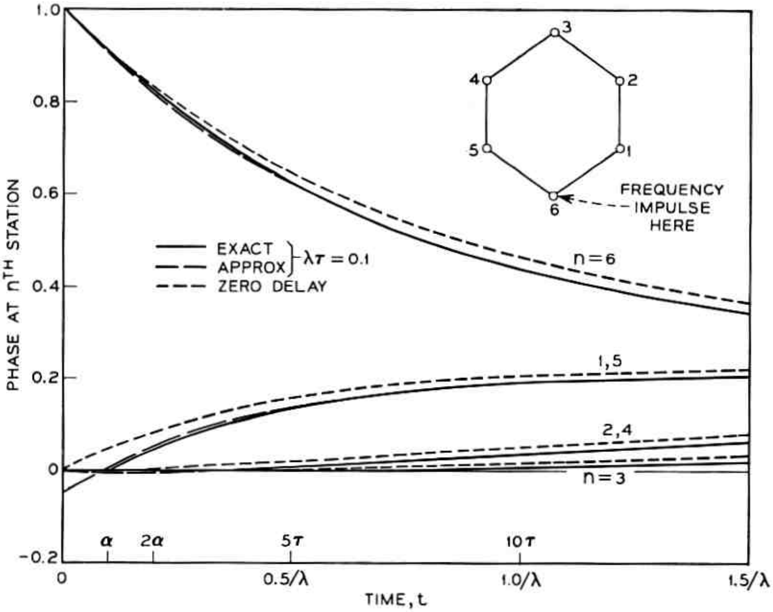


Fig. 6 — Impulse response of a 6-station bilateral ring.

6.3 *Case 3:* $H_n(s) \equiv \lambda$, $\tau_{nm} \equiv 0$, *N large, t large.*

The alternative approach provides a better approximation for large t after the disturbance has spread around the ring. When t is large, the dominant terms in (44) are those in which $\cos(2\pi k/N)$ is closest to unity, including not only those in which k is small but also those in which k is close to N, or, equivalently, k is small and negative. Since the kth term and the $(N - k)$th or $(-k)$th term are equal, the latter terms can be effectively included by doubling each term for small k. For large N, the approximation

$$\cos x \approx 1 - \frac{x^2}{2} \tag{50}$$

can be used for these terms. The Nth or zeroth term is a constant $1/N$. The other terms, which are small, can be omitted or included as convenient; since it is difficult to specify in advance which terms are negligible, it seems safest to include them all, at least formally. Thus, approximately, for large N,

$$p_n(t) \approx \frac{1}{N} + \frac{2}{N} \sum_{k=1}^{\infty} \cos(2\pi nk/N) \exp\left(\frac{-2\pi^2 k^2}{N^2} \lambda t\right). \tag{51}$$

The time constants are proportional to the square of the number of stations in the ring. The components have sinusoidal spatial distributions around the ring, and the time constant is inversely proportional to the square of the spatial frequency. Some curves calculated from (51) are shown in Fig. 5, compared with the response of a 6-station ring.

6.4 *Case 4:* $H_n(s) \equiv \lambda,\ \tau_{nm} \equiv \tau,\ N < \infty.$

If the delays are positive but all equal, the methods used in the 2-station system can be applied to the inversion of (41). The exact result is

$$q_k(t) = \sum_{m=0}^{[t/\tau]} \frac{\lambda^m \cos^m (2\pi k/N)(t - m\tau)^m e^{-\lambda(t-m\tau)}}{m!}. \tag{52}$$

The approximation based on (23) gives

$$q_k(t) \approx \frac{1}{1 + \lambda\tau \cos (2\pi k/N)} \exp \left\{ -\lambda t \left[\frac{1 - \cos (2\pi k/N)}{1 + \lambda\tau \cos (2\pi k/N)} \right] \right\} \tag{53}$$

which may be compared with (42). Curves calculated from these equations for a 6-station ring with $\lambda\tau = 0.1$ are shown in Fig. 6 and compared with the zero-delay case.

VII. BILATERAL CHAINS

It has been mentioned previously that a chain lacks the simplicity of a ring because of the exceptional nature of the end stations. However, given any chain of N stations, an analogous ring can be formed by duplicating all stations except the end stations so as to form a second chain between the end stations as shown in Fig. 7, and taking the value $\frac{1}{2}$ for each of the two averaging coefficients at each end station, leaving all other parameters unchanged. The response of the chain to a disturbance at any station can be found by applying the same disturbance at the corresponding station or stations in the analogous ring; the response of each half of the ring will be the same as the response of the original chain.

A bilateral ring, as studied in the preceding section, will result if the stations in the chain all have equal filter gains and if all averaging coefficients (except at the end stations) equal $\frac{1}{2}$. Such a chain will be called a bilateral chain. Thus, in particular, the response shown for 6-station rings in Figs. 4, 5, and 6 will also be observed in 4-station bilateral chains disturbed by an impulse at an end station. The response to a disturbance at any other station may be obtained by superposition of two station responses calculated from the ring; the responses to be

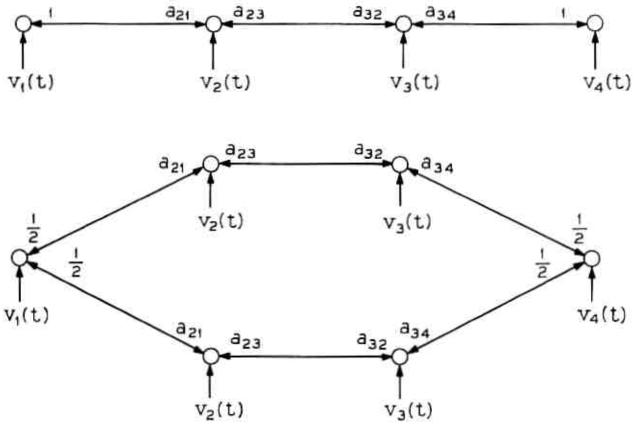


Fig. 7 — A chain and its analogous ring.

superposed may be identified by supposing that the disturbance is propagated in both directions from the source and is reflected at either end of the chain.

Alternatively, in principle, the response of either the ring or the chain might be determined by superposition of an infinite number of terms of the infinite-ring response determined in the preceding section. The response of an infinite ring is the same as that of an infinite chain extending in both directions from the source of the disturbance, since the two networks are indistinguishable. The response of a finite ring could be calculated by supposing the disturbance to propagate around the ring an unlimited number of times in both directions. For a finite chain, the disturbance could be considered to spread in both directions (except when the disturbance originates at an end station) and to be reflected whenever it reaches an end station. This method may be useful in large chains or rings as a refinement of the simple approximation of a large chain or ring as an infinite one.

VIII. UNILATERAL RINGS AND CHAINS

All the networks studied in previous sections have satisfied the reciprocity condition, and in consequence all components of the response have been nonoscillatory: strictly so in the zero-delay case, and approximately in the case of small $\lambda\tau$. In this section, the opposite extreme is studied. In the unilateral ring, the product of the averaging coefficients in one direction is positive, while every coefficient in the other direction is zero.

8.1 *Rings*

To define a unilateral ring, we assume a ring of N identical stations and assign a positive direction around the ring. Each station transmits only to its nearest neighbor in the positive direction. Each station then receives from only one other station and thus has only one averaging coefficient equal to unity. All links are identical. The system equations are

$$sP_n(s) = V_n(s) + H(s)[P_{n-1}(s)e^{-s\tau} - P_n(s)], \qquad n = 1, \cdots, N. \qquad (54)$$

The transformation defined by (35) and (36) may be applied to this network also. In place of (41), assuming the same impulsive disturbance as given in (39), we get

$$Q_k(s) = \frac{1}{s + \lambda(1 - e^{-s\tau - j2\pi k/N})}. \qquad (55)$$

8.1.1 *Case 1: $H_n(s) \equiv \lambda$, $\tau_{nm} \equiv 0$, $N < \infty$*

Only the case of zero delays has been studied in detail. In this case,

$$q_k(t) = \exp\{-\lambda[1 - \cos(2\pi k/N)]t\} \exp[-j\lambda t \sin(2\pi k/N)]. \qquad (56)$$

Hence,

$$p_n(t) = \frac{1}{N}\sum_{k=1}^{N} \exp\{j[2\pi nk/N - \lambda t \sin(2\pi k/N)]\} \cdot \exp\{-\lambda[1 - \cos(2\pi k/N)]t\}. \qquad (57)$$

The sum is real and may be alternatively expressed as

$$p_n(t) = \frac{1}{N}\sum_{k=1}^{N} \cos[2\pi nk/N - \lambda t \sin(2\pi k/N)] \cdot \exp\{-\lambda[1 - \cos(2\pi k/N)]t\}. \qquad (58)$$

The components are not real exponentials, but exponentially damped sinusoids.

8.1.2 *Case 2: $H_n(s) \equiv \lambda$, $\tau_{nm} \equiv 0$, $N = \infty$*

In the infinite unilateral ring, using (45) in (57) and passing to the limit,

$$p_n(t) = \frac{e^{-\lambda t}}{2\pi}\int_0^{2\pi} e^{jn\theta} \exp(\lambda t e^{-j\theta})\, d\theta. \qquad (59)$$

Expanding $\exp(\lambda t e^{-j\theta})$ as a power series in $\lambda t e^{-j\theta}$ and integrating term by term gives

$$\begin{aligned} p_n(t) &= \frac{(\lambda t)^n e^{-\lambda t}}{n!}, \qquad n = 0, 1, \cdots \\ &= 0, \qquad\qquad\quad n = -1, -2, \cdots . \end{aligned} \tag{60}$$

This result is plotted in Fig. 8. The phase disturbances at adjacent stations are in the ratio

$$\frac{p_n(t)}{p_{n-1}(t)} = \frac{\lambda t}{n}, \tag{61}$$

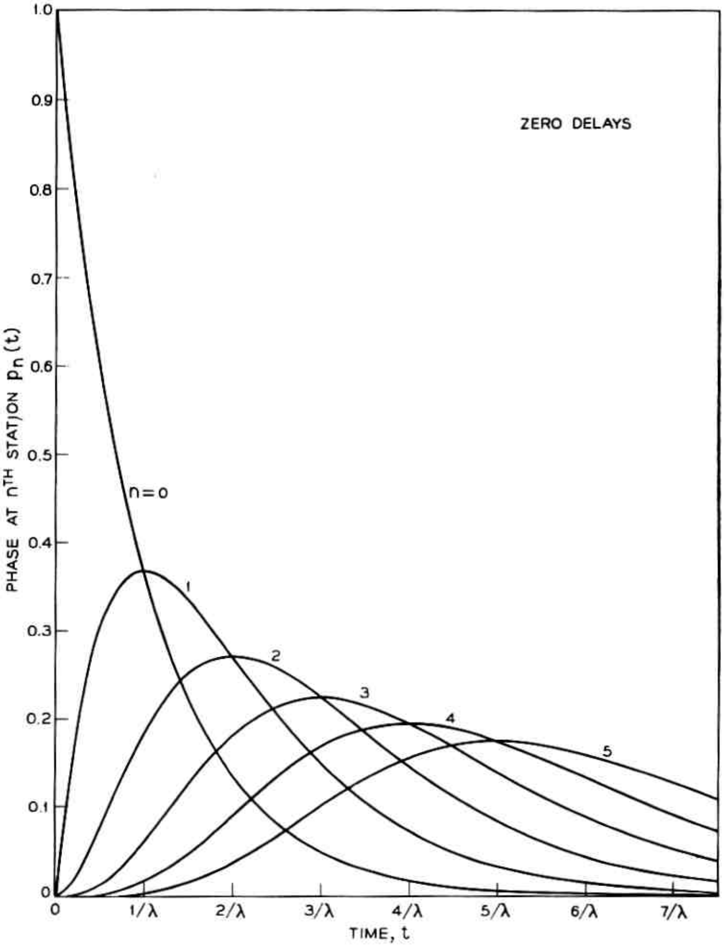


Fig. 8—Impulse response of an infinite unilateral ring.

so that for fixed t, and increasing n, $p_n(t)$ increases until $n \geqq t$. Therefore, at any time t the largest disturbance is at the mth station, where

$$\lambda t - 1 \leqq m \leqq \lambda t. \tag{62}$$

For large t, the magnitude of the largest phase disturbance, obtained with the aid of Stirling's approximation for the factorial, is asymptotically

$$p_m(t) \sim \frac{1}{\sqrt{2\pi\lambda t}}. \tag{63}$$

This is the same as the asymptotic value (49) obtained for the infinite bilateral ring, except that in the unilateral case the peak precesses at the rate of λ stations per unit time.

8.1.3 *Case 3: $H_n(s) \equiv \lambda$, $\tau_{nm} \equiv 0$, N large, t large*

Application of the approximation (50) together with

$$\sin x \approx x \tag{64}$$

gives, for large t in a large ring,

$$p_n(t) \approx \frac{1}{N} + \frac{2}{N} \sum_{k=1}^{\infty} \cos\left[2\pi k(n - \lambda t)/N\right] \exp\left(\frac{-2\pi^2 k^2}{N^2} \lambda t\right). \tag{65}$$

Compared with (51) this shows a response that resembles that of a bilateral ring except that it precesses around the ring in the positive direction at the rate of λ stations per unit time. The oscillatory nature of the response is associated with the progression of the disturbance around the ring.

8.2 *Chains*

A finite unilateral chain is a system with one master station and $N - 1$ slaves. If such a system is disturbed at one of the slave stations, the response of each station following it in the positive direction is the same as that of the corresponding station in an infinite unilateral ring. An impulse disturbance at the master station, however, does not correspond directly to any situation in a unilateral ring. A permanent phase shift of one unit occurs in the master station output. The effect at the second station is the same as that of a step of magnitude λ in the free-running frequency of the second station, and, a step being the integral of an impulse, the response of the entire chain can be inferred by integrating the response to an impulse at the second station. Each station can thus be shown to approach its new equilibrium phase monotonically. If the nominal time of response of each station is defined as the time of maximum rate of change of phase (maximum frequency shift), each station responds with a delay of $1/\lambda$ after the preceding station. The effect of positive link delays in the unilateral chain is to further delay the response without changing its form.

The unilateral ring is not the analog of any chain in the sense of the preceding section.

IX. RECTANGULAR ARRAYS

A rectangular array, in which each station is connected to four nearest neighbors, can be considered as intermediate between a fully interconnected system and a chain or ring, and may be more appropriate than either as a model of a network of stations on the surface of the earth. A rectangular network with no edges or corners can be laid out on the surface of a toroid as in Fig. 9. This network can be analyzed by methods similar to those used for rings.

The stations are most conveniently indexed with double subscripts, $m = 1, \cdots, M_1$, and $n = 1, \cdots, M_2$; the number of stations is $N = M_1M_2$. Assuming equal filters and equal delays, the system equations are

$$\begin{aligned} sP_{mn}(s) &= V_{mn}(s) - H(s)P_{mn}(s) \\ &\quad + \frac{H(s)}{4} e^{-s\tau}[P_{m,n-1}(s) + P_{m,n+1}(s) + P_{m-1,n}(s) + P_{m+1,n}(s)], \qquad (66) \end{aligned}$$

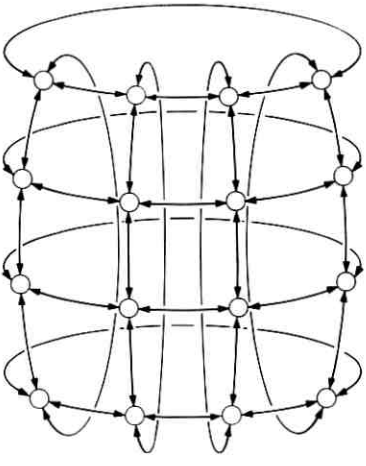


Fig. 9 — A toroidally-connected rectangular array.

assuming addition modulo M_1 in the first index, and modulo M_2 in the second. Defining

$$Q_{kl}(s) = \sum_{m=1}^{M_1} \sum_{n=1}^{M_2} P_{mn}(s) \exp\left[-j2\pi\left(\frac{mk}{M_1} + \frac{nl}{M_2}\right)\right], \qquad k = 1, \cdots, M_1, \quad l = 1, \cdots, M_2 \tag{67}$$

and proceeding as with the bilateral ring, we obtain, in the case of flat filters and zero delays, with an impulse disturbance at the M_1, M_2th (or zero-zeroth) station,

$$p_{mn}(t) = \frac{1}{N} \sum_{k=1}^{M_1} \sum_{l=1}^{M_2} \exp\left[j2\pi\left(\frac{mk}{M_1} + \frac{nl}{M_2}\right)\right] \cdot \exp\left[-\lambda\left(1 - \tfrac{1}{2}\cos\frac{2\pi k}{M_1} - \tfrac{1}{2}\cos\frac{2\pi l}{M_2}\right)t\right]. \tag{68}$$

Comparison with the bilateral ring is most convenient in the limiting cases of large systems. For the infinite array,

$$p_{mn}(t) = e^{-\lambda\tau} I_m\left(\frac{\lambda t}{2}\right) I_n\left(\frac{\lambda t}{2}\right) \tag{69}$$

which has the asymptotic form

$$p_{mn}(t) \sim \frac{1}{\pi\lambda t} \tag{70}$$

indicating a more rapid approach to the final value in the rectangular array than in the ring. The approximation (50) for large t in large arrays gives

$$p_{mn}(t) \approx \frac{1}{N} \sum_{k=-\infty}^{\infty} \sum_{l=-\infty}^{\infty} \cos\left[2\pi\left(\frac{mk}{M_1} + \frac{nl}{M_2}\right)\right] \cdot \exp\left[-\pi^2\lambda t\left(\frac{k^2}{M_1^2} + \frac{l^2}{M_2^2}\right)\right]. \tag{71}$$

The longest time constant is shorter for a rectangular array than for a ring with the same number of stations. Fig. 10 shows some curves calculated from (69).

A bounded rectangular array in a plane is more complicated than a toroidally connected array because of the exceptional edge and corner stations. However, a bounded M_1 by M_2 array can be analyzed in terms of an analogous $2M_1 - 2$ by $2M_2 - 2$ toroidal array as shown in Fig. 11. All columns except the first and last are duplicated and connected as shown by the solid lines to form a cylindrical array, and

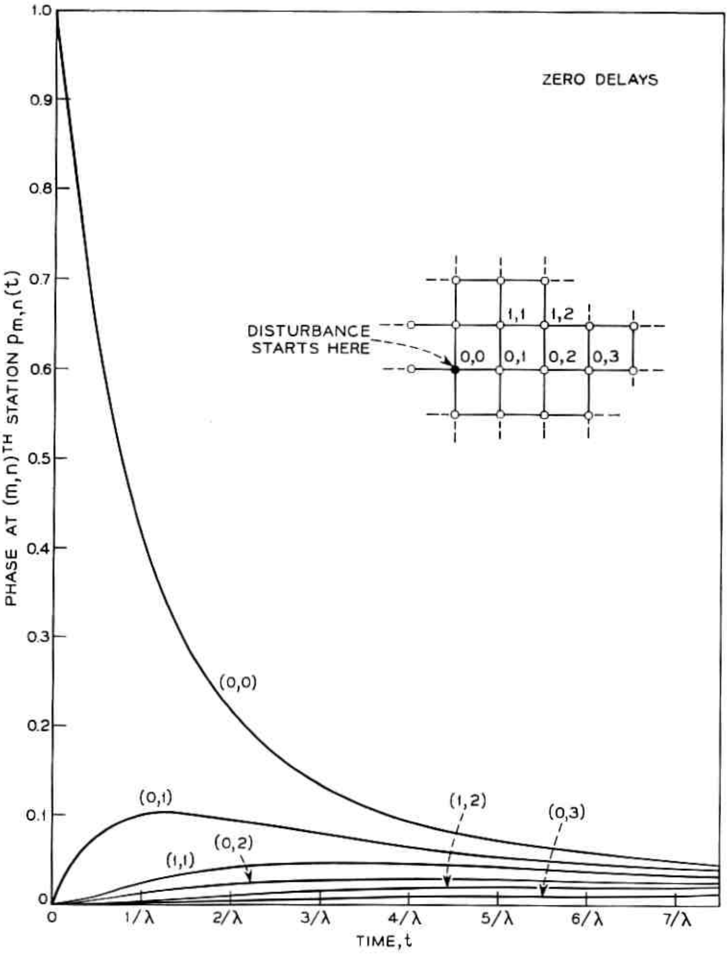


Fig. 10—Impulse response of an infinite rectangular array.

then all rows except the first and last are duplicated and connected as shown by the dashed lines; averaging coefficients are divided by two whenever a station receives from duplicate stations. The toroidal array has one station corresponding to each original corner station, two for each edge station, and four for each interior station. The response of the original bounded array to a disturbance at any station is identical with the response of the corresponding part of the toroid when the original disturbance is applied to corresponding stations.

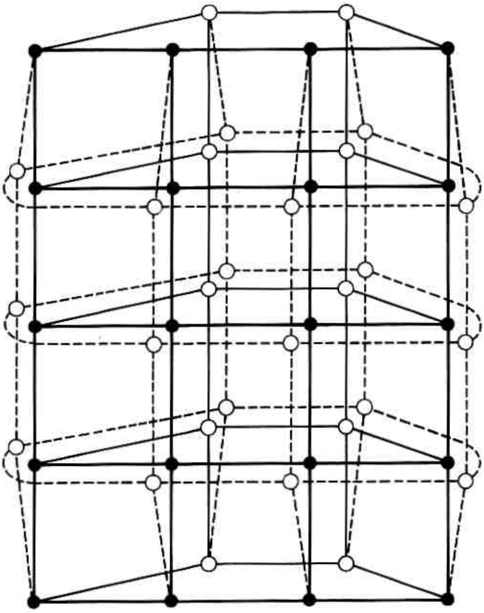


Fig. 11 — The toroidally-connected array analogous to a bounded rectangular array.

Alternatively, in principle, the response of a finite toroidal array can be determined from that of the infinite array by considering the response to propagate around a toroidal array in the four cardinal directions, or to be reflected from the sides of a bounded array.

X. RESPONSE TO SINUSOIDAL DISTURBANCES

The steady-state response of a linear system to a sinusoidal disturbance is sinusoidal, and the phase difference between the response and the input disturbance, together with the ratio of the amplitudes, is given by the frequency response function as a function of frequency. The impulse response is equivalent in principle to the frequency response function as a specification of dynamic properties, since either can be expressed in terms of the other through Fourier or Laplace transformation. The frequency response functions of a bilateral ring, in particular, are the functions $P_n(j\omega)$, which are the $P_n(s)$ evaluated along the "real frequency axis" $s = j\omega$, for real ω.

The frequency response will be determined in this section for infinite rings, both bilateral and unilateral, in the case of arbitrary equal filters and arbitrary equal delays. Although the expressions are more complicated than the impulse response expression in the case of flat filters and zero delays, they do not become very much more complicated in the more general case, for which closed form expressions for the impulse response have not been obtained.

10.1 *Case 1: Bilateral Ring, $N = \infty$*

For the bilateral ring, an expression for $Q_k(s)$ is obtained from (38) using (40), and $P_n(s)$ is obtained using (36). Using the substitution (45) and passing to the limit of infinite N gives

$$P_n(s) = \frac{1}{2\pi} \int_0^{2\pi} \frac{e^{in\theta}\, d\theta}{s + H(s)(1 - e^{s\tau} \cos \theta)}. \tag{72}$$

To evaluate this by contour integration, let

$$z = e^{i\theta}. \tag{73}$$

Then

$$P_n(s) = \frac{1}{j2\pi} \oint \frac{z^{n-1}\, dz}{s + H(s)\left[1 - \frac{e^{-s\tau}}{2}\left(z + \frac{1}{z}\right)\right]}, \tag{74}$$

integrated counterclockwise around the unit circle centered at the origin in the z-plane.

When $n \geqq 0$ the integrand has two poles in the z-plane, located at the roots of the quadratic equation

$$z^2 - 2e^{s\tau}\left[1 + \frac{s}{H(s)}\right]z + 1 = 0. \tag{75}$$

Since the denominator of the integrand is symmetric in z and $1/z$, one root is the reciprocal of the other. We defer consideration of the case where both roots have unit magnitude; then one pole will lie inside the path of integration and the other outside. Denote the root inside the contour by

$$z_1 = e^{s\tau}\left[1 + \frac{s}{H(s)}\right] - \sqrt{e^{2s\tau}\left[1 + \frac{s}{H(s)}\right]^2 - 1}, \tag{76}$$

where it is understood that the square root is to be taken to have whichever sign gives z_1 the smaller magnitude.

For convenience, let

$$\beta(s) = \frac{H(s)}{s + H(s)}, \tag{77}$$

this incidentally being the quantity whose magnitude is required to be less than unity in the sufficient condition for stability given by Gersho and Karafin.[1] Then

$$z_1 = \frac{1 - \sqrt{1 - \beta^2(s)e^{-2s\tau}}}{\beta(s)e^{-s\tau}} = \frac{\beta(s)e^{-s\tau}}{1 + \sqrt{1 - \beta^2(s)e^{-2s\tau}}}, \tag{78}$$

where the second form can be obtained by rationalizing the numerator. The integral around the contour is $2\pi j$ times the residue at this pole, so that $P_n(s)$ is equal to the residue, which can be put in the alternative forms

$$P_n(s) = \frac{[1 - \sqrt{1 - \beta^2(s)e^{-2s\tau}}\,]^n}{H(s)\beta^{n-1}(s)e^{-ns\tau}\sqrt{1 - \beta^2(s)e^{-2s\tau}}} = \frac{\beta^{n+1}(s)e^{-ns\tau}}{H(s)\sqrt{1 - \beta^2(s)e^{-2s\tau}}\,[1 + \sqrt{1 - \beta^2(s)e^{-2s\tau}}\,]^n}. \tag{79}$$

For negative n, (74) can be transformed into an integral in $y = 1/z$ to show that

$$P_n(s) = P_{-n}(s). \tag{80}$$

The deferred case in which z_1 has unit magnitude will now be briefly considered. In this case, the quadratic equation (75) has two conjugate roots of unit magnitude, or double roots at 1 or -1, and it is easily shown that this occurs when $\beta(s)e^{-s\tau}$ is real and has magnitude 1 or greater. If the sufficient condition for stability mentioned earlier is satisfied, this cannot occur in the left half s-plane or on the real frequency axis except at zero frequency, where a singularity is expected to occur in any system configuration.

Where $\beta(s)e^{-s\tau}$ is ± 1, $P_n(s)$ is infinite and will ordinarily have a branch point. This always occurs at $s = 0$, and occurs for other values depending on the filters and delays. Where $\beta(s)e^{-s\tau}$ is real and has magnitude greater than unity, $P_n(s)$ will be finite but will have a step discontinuity, because as s passes through a value at which $\beta(s)e^{-s\tau}$ is real, z_1 crosses the unit circle and must immediately be redefined as z_2, and the square roots in (79) abruptly change sign. The function $P_n(s)$ is thus defined as a single-valued function in the s-plane with line discontinuities where it might be expected to have branch cuts.

If the system is stable, these discontinuities are confined to the interior of the left half s-plane except for $s = 0$.

Thus, (79) defines $P_n(j\omega)$ as a continuous single-valued function except at $\omega = 0$. In the case of flat filters and zero delays, $P_n(s)$ is the Laplace transform of (47). Fig. 12 shows the magnitude of $P_n(j\omega)$ for this case and for the case of $\lambda\tau = 0.1$.

10.2 *Case 2: Unilateral Ring, $N = \infty$*

For an infinite unilateral ring, a similar procedure gives

$$P_n(s) = \frac{1}{j2\pi} \oint \frac{z^n \, dz}{[s + H(s)]z - H(s)e^{-s\tau}} \tag{81}$$

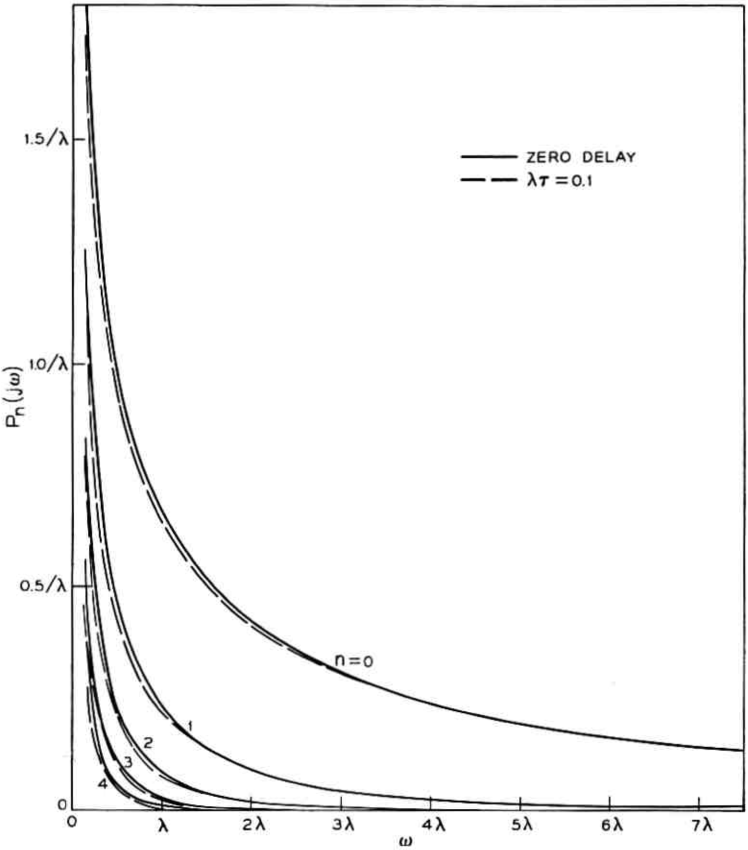


Fig. 12—Frequency response of an infinite bilateral ring.

to be integrated over the same path as (74). When $n \geqq 0$, the integrand has a single pole at $z = \beta(s)e^{-s\tau}$. If the magnitude of $\beta(s)e^{-s\tau}$ is less than 1, the pole is inside the unit circle, and for nonnegative n

$$P_n(s) = \frac{\beta^{n+1}(s)e^{-ns\tau}}{H(s)}, \qquad n = 0, 1, 2, \cdots, \tag{82}$$

while for negative n under the same conditions, the substitution $y = 1/z$ puts all poles outside the unit circle and

$$P_n(s) = 0, \qquad n = -1, -2, \cdots . \tag{83}$$

Fig. 13 shows the magnitude of $P_n(j\omega)$ graphically. As the magnitude of $\beta(s)e^{-s\tau}$ becomes greater than 1, the pole crosses the unit circle and there is a step discontinuity in $P_n(s)$ for all n. However, the sufficient condition for stability mentioned previously is both necessary and sufficient, in the unilateral ring, for these discontinuities to be confined to the left half-plane.

The finite value of $P_n(0)$, where a singularity should occur, is attributable to the fact that every station in the infinite unilateral ring is a slave station, and no finite change at any given station can alter the equilibrium frequency. The infinite unilateral ring is in this sense a pathological limiting case of the unilateral chain in which the master station recedes to infinity and becomes inaccessible.

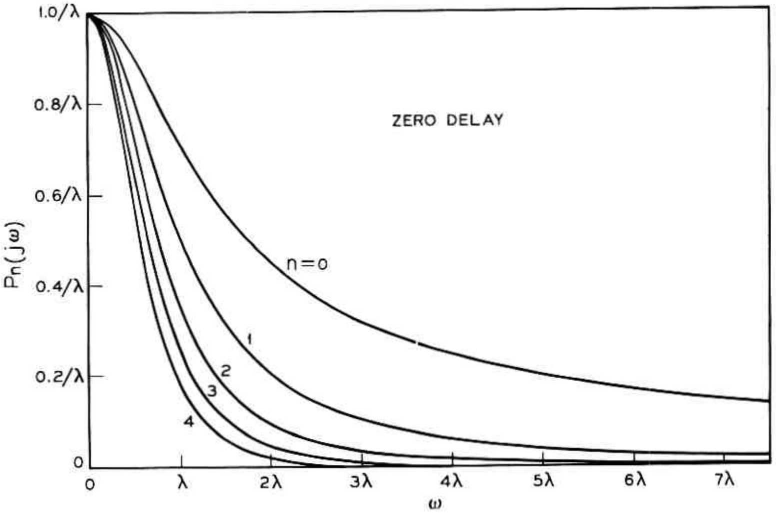


Fig. 13 — Frequency response of an infinite unilateral ring.

XI. JITTER RESPONSE

Jitter denotes random variations in the phase of a signal. In a digital signal, jitter can occur as a result of pattern-induced retiming errors in regenerative repeaters.[5] Jitter reducers[6] can reduce the high-frequency components of jitter, but, because the jitter reducer output frequency is slaved to the input frequency, the low-frequency components cannot be reduced.

In a mutually synchronized system, the low-frequency components of jitter will affect the observed phase differences used to control the clocks. Even if the variations in the received phase, after jitter reduction, are not themselves objectionable, they might cause objectionable variations in clock phases through the cumulative effects of each clock acting on the next. To simplify the analysis, only the effects on the clock phases are considered; the immediate effects of jitter are omitted.

It is assumed that the effect of jitter on the link from the mth station to the nth is to add a random component $\mu_{nm}(t)$ to the phase $p_m(t - \tau_{nm})$ that would be received without jitter. This random component is assumed to have the properties of white Gaussian noise and to be independent on different links. Assuming that a jitter reducer can be designed that will compensate the immediate effects of jitter, we determine only the cumulative effect of jitter propagating through the system as a result of its effect on the station clocks.

The autocorrelation function assumed for $\mu_{nm}(t)$ is

$$E[\mu_{nm}^*(t)\mu_{nm}(t + x)] = K\delta(x). \tag{84}$$

Here "E" stands for the "expectation" or mean value, the star denotes complex conjugation (immaterial here since μ_{nm} is real), and $\delta(t)$ is the Dirac delta function. K represents the noise power density, assumed to be the same for every link in the systems to be considered.

11.1 *Case 1: Phase-Locked Oscillator*

As a standard of comparison, consider the effect of this jitter on a simple phase-locked loop of gain λ, in which an oscillator is controlled by the signal received from an unperturbed source over a jittered link. The equation for the output phase $p(t)$ in this system is

$$p'(t) = F_1 + \lambda(F_0 t + \mu(t) - p(t)), \tag{85}$$

where F_1 is the free-running frequency of the controlled oscillator, F_0 the frequency of the master source, and the link delay is assumed zero. Since the system is linear, and we are interested only in the random component of the output, we may set $F_1 = F_0 = 0$ without loss of generality. Thus, the Laplace-transformed system equation becomes

$$sP(s) = \lambda M(s) - \lambda P(s) \tag{86}$$

with solution

$$P(s) = \frac{\lambda}{s + \lambda} M(s). \tag{87}$$

We obtain the mean-square value of $p(t)$ from its autocorrelation function $\varphi(x)$ evaluated at $x = 0$. This is determined from the power-density spectrum

$$\Phi(\omega) = \frac{1}{2\pi} \int_{-\infty}^{\infty} \varphi(x) e^{-i\omega x} \, dx \tag{88}$$

by means of the inverse transformation

$$\varphi(x) = \int_{-\infty}^{\infty} \Phi(\omega) e^{i\omega x} \, d\omega. \tag{89}$$

The power-density spectrum of the input $\mu(t)$, obtained from (84) and an integral of the form of (88), is flat, equal to $K/2\pi$, for all ω. The output power density is obtained by multiplying this by the square of the magnitude of the frequency response, obtained from (87):

$$\Phi(\omega) = \frac{\lambda^2 K}{2\pi \mid j\omega + \lambda \mid^2}. \tag{90}$$

The inversion integral (89) is evaluated by means of a partial-fraction expansion. The analytic continuation of (90) in the s-plane, $s = j\omega$, has poles in both the right and left half-planes. Since (89) is a Fourier (not Laplace) inversion, terms due to poles in the right half-plane will be zero for positive x; thus, for positive x we need only consider the left half-plane. We obtain

$$\varphi(x) = \frac{\lambda K e^{-\lambda x}}{2} \tag{91}$$

and, as a limit,

$$\varphi(0) = \frac{\lambda K}{2} \tag{92}$$

is the mean-square value of $p(t)$. The rms phase error is of course the square root of this.

11.2 *Case 2: Bilateral Ring*

We now consider a bilateral ring with flat filters and small delays. Each station receives two inputs, each with jitter with the autocorrelation function (84), and therefore power density $K/2\pi$. Each input is multiplied by $\lambda/2$, to produce an error-signal component with power density $(\lambda/2)^2K/2\pi$, and when two independent components are added, their power densities add to produce a total power density of $2(\lambda/2)^2K/2\pi$, or $\lambda^2K/4\pi$. The effect of the two jitter components is thus equal to the effect of a white noise component of power density $\lambda^2K/4\pi$ added to the free-running frequency. We ignore the steady-state components and consider this to be the only input at each station. Thus, we assume

$$E[v_n^*(t)v_n(t + x)] = \frac{\lambda^2K}{2}\,\delta(x). \tag{93}$$

The variables $u_k(t)$ are derived from the $v_n(t)$ as in (84); direct evaluation gives

$$E[u_k^*(t)u_l(t + x)] = \begin{cases} \dfrac{\lambda^2K}{2N}\,\delta(x), & k = l; \\ 0, & k \neq l. \end{cases} \tag{94}$$

This shows that the $u_k(t)$ are uncorrelated, and therefore (since we have assumed Gaussian distributions) independent, each with power density $\lambda^2K/4\pi N$. It follows, since each $q_k(t)$ depends only on the corresponding $u_k(t)$ as in (38), that the $q_k(t)$ are independent. Denoting their autocorrelation functions by $\psi_k(x)$, we obtain their power-density spectra, using the frequency response given by (38), as

$$\Psi_k(\omega) = \frac{\lambda^2K/4\pi N}{[\lambda + j\omega - \lambda e^{-j\omega\tau}\cos\theta_k][\lambda - j\omega - \lambda e^{j\omega\tau}\cos\theta_k]}, \tag{95}$$

where the substitution (45) is used as an abbreviation.

When $k = N$, (95) indicates infinite power density at zero frequency. The autocorrelation function $\psi_N(x)$ is consequently infinite for all x, and in particular the mean-square value of $q_N(t)$ is infinite, so that the mean-square value of each $p_n(t)$ is infinite. This occurs because the random variations that the jitter induces in the system frequency cause the system phase to execute a random walk. However, since $q_N(t)$ contributes equally to every $p_n(t)$, it does not affect the phase differences between clocks, and all other $q_k(t)$ have finite mean-square values. It follows that while the phase of each clock tends to deviate indefinitely far from that of an unperturbed clock of the same frequency, the deviation between clocks in the same system tends to remain bounded.

We are primarily interested in the phase difference between the clock at each station and the delayed signal received from an adjacent station. The mean-square value of this phase difference will be denoted by

$$\Phi_{n,n\pm 1} = E\{[p_n(t) - p_{n\pm 1}(t - \tau)]^2\}. \tag{96}$$

We express the phases in terms of $q_k(t)$ using (36), writing the square of the real sum of these complex quantities as the product of the sum by its conjugate so that expansion of the product gives terms of the form of the left side of (84); thus,

$$\Phi_{n,n\pm 1} = 2 \sum_{k=1}^{N-1} \{\psi_k(0) - \text{Re}\,[e^{\mp i\theta_k}\psi_k(\tau)]\}. \tag{97}$$

Therefore, (95) should be used in an integral of the form of (89) to determine $\psi_k(0)$ and $\psi_k(\tau)$. The analytic continuation of (95) in the s-plane has, in the left half-plane, all the poles of (41), and in addition the reflections of these poles in the right half-plane. We continue to use (23) as an approximation when $\lambda\tau$ is small. The result is

$$\psi_k(x) \approx \frac{\lambda K \exp\left[-\lambda x\left(\dfrac{1 - \cos\theta_k}{1 + \lambda\tau\cos\theta_k}\right)\right]}{4N(1 - \cos\theta_k)(1 + \lambda\tau\cos\theta_k)}, \qquad x \geqq 0. \tag{98}$$

In particular, to the first order in $\lambda\tau$,

$$\psi_k(0) \approx \frac{\lambda K(1 - \lambda\tau\cos\theta_k)}{4N(1 - \cos\theta_k)} \tag{99}$$

and, again using the linear approximation to the exponential,

$$\psi_k(\tau) \approx \frac{\lambda K(1 - \lambda\tau)}{4N(1 - \cos\theta_k)}. \tag{100}$$

We now find, from (97), that

$$\Phi_{n,n\pm 1} \approx \left(\frac{N - 1}{N}\right)\frac{\lambda K}{2}. \tag{101}$$

The mean-square phase discrepancy observed in received signals is thus substantially independent of the size of the system and substantially unaffected by small link delays. It is roughly equal to the mean-square phase error, given by (92), that would be induced, by the jitter in a single link, in a simple phase-locked oscillator with control gain λ.

11.3 *Case 3: Unilateral Ring*

In a unilateral ring, each station receives only one input, so that the equivalent $v_n(t)$ has power density $\lambda^2 K/2\pi$ and its autocorrelation function has twice the value given in (93). The appropriate frequency response is given by (55), so that instead of (95) we get

$$\Psi_k(\omega) = \frac{\lambda^2 K/2\pi N}{[\lambda + j\omega - \lambda e^{-i\omega\tau - i\theta_k}][\lambda - j\omega - \lambda e^{i\omega\tau + i\theta_k}]}. \tag{102}$$

We continue to use (23) to determine a simple approximation. The result is

$$\psi_k(x) \approx \frac{\lambda K \exp\left[-\lambda x\left(\dfrac{1 - e^{-i\theta_k}}{1 + \lambda\tau e^{-i\theta_k}}\right)\right]}{2N(1 - \lambda\tau)(1 - \cos\theta_k)}. \tag{103}$$

In particular,

$$\psi_k(0) \approx \frac{\lambda K(1 + \lambda\tau)}{2N(1 - \cos\theta_k)} \tag{104}$$

and

$$\psi_k(\tau) \approx \frac{\lambda K[1 + \lambda\tau e^{-i\theta_k}]}{2N(1 - \cos\theta_k)}. \tag{105}$$

Equation (97) is equally valid for the unilateral ring as for the bilateral ring, giving

$$\begin{aligned} \Phi_{n,n-1} &\approx \lambda K\left(\frac{N-1}{N}\right) \\ \Phi_{n,n+1} &\approx \lambda K\left[\left(\frac{N-1}{N}\right) + 2\lambda\tau\left(\frac{N-2}{N}\right)\right]. \end{aligned} \tag{106}$$

The mean-square phase discrepancy is essentially twice that which occurs in the bilateral ring. The link delay has a first-order effect on the signal received at each station from the station to which it transmits timing control because of the round-trip delay.

XII. SUMMARY AND CONCLUSIONS

In this section, I propose to extrapolate the specific results of the preceding sections to general conclusions that, although not strictly proven, seem quite likely to be true from a practical standpoint.

It was shown in Section III that a system that satisfies the reciprocity condition and has flat filters and no delays will have a nonoscillatory transient response. The response was described more specifically in later sections for specific configurations: 2-station systems, fully interconnected systems, and bilateral rings and chains, all of which met these conditions. These configurations appear to span the extremes of practical systems.

The effect of delays was determined specifically only for these special configurations and for the special case of equal delays and flat filters. The effect was shown graphically for $\lambda\tau = 0.1$; it appeared to be small and unobjectionable. Fig. 6 shows that at all stations, over the time range shown, the response to a transient disturbance is actually smaller when delays exist. At the zeroth (or Nth) station, where the disturbance originates, this can be attributed to the short period after the disturbance during which the neighboring stations remain undisturbed and are, therefore, reliable indicators of the original state. At other stations, the appearance of smaller disturbances is due in part to the delayed peak of the response.

I propose to conjecture that the dynamic effect of delay in any reciprocal system with flat filters will be equally unobjectionable as long as the product of the largest filter gain and the largest single link delay is less than 0.1. This would be a unjustified extrapolation from a purely mathematical standpoint, but it seems reasonable in the light of the physical interpretation suggested in the preceding paragraph.

The effect of filters with other than flat frequency response has not been shown at all in terms of transient response. Two aspects of this question appear important. In the first place, it may be possible to obtain some improvement in transient response by appropriate filter design, but further analysis appears necessary to answer this question. In the second place, assuming that the flat filter gives a satisfactory response, the effect of high-frequency cutoff, which is inevitable in a practical system, must be estimated. A tentative answer to this question can be obtained by examination of the expressions for frequency response $P(s)$ developed for specific configurations. In all these expressions, the system response is substantially the same as in the flat-filter case as long as the filters $H_n(s)$ remain substantially flat until the frequency s becomes large compared with the zero-frequency filter gains λ_n. This condition establishes an approximate bandwidth requirement for the filters. The extrapolation to arbitrary configurations is proposed in this case also.

The effect of departure from the reciprocity conditions is illustrated in only one case: the unilateral ring. Here, although the departure from reciprocity is the greatest possible, the effect on the transient response is mild. The magnitude of the response, and its rate of subsidence, are substantially unchanged; the principal effect is the precession of the disturbance around the ring. The oscillatory components in the response can be associated with this precession.

Extrapolation of this result appears uncertain. The reciprocity condition can be stated in terms of the equality of the products of averaging coefficients in opposite directions around any loop. It can easily be conjectured that if the product of the averaging coefficients around any loop is much larger in one direction than the other, there will be a tendency for disturbances to precess around the loop in this direction, thereby generating oscillatory components in the response. On the other hand, it is hard to imagine pure precession in a multiloop network. A possible answer is suggested by the argument in Section III, in terms of pole loci, suggesting that a considerable departure from reciprocity could be tolerated before oscillatory components began to appear.

This extrapolation is suggested only for the case of flat or nearly flat filters and zero or small delays. For other cases, departures from reciprocity may give rise to a stability problem. This is suggested by the analysis in Section VIII of the discontinuities in the frequency response of an infinite unilateral ring, which showed that the stability condition that has been shown in the general case only to be a sufficient condition is in this case not merely sufficient but necessary. The latitude for filter shaping may be smaller in the nonreciprocal case, limited not simply by instability but by the deterioration of transient response that generally accompanies an approach to instability.

The analysis of jitter response shows that in certain representative cases the effect of jitter does not accumulate in a large system. This gives a definite negative answer to the question of whether cumulative jitter necessarily occurs in a large system. It seems reasonable to conjecture that this conclusion is independent of configuration, and remains true for substantially flat filters and small delays, but less reasonable to suppose that it will remain true for arbitrary filters.

Nothing in this study should be construed to indicate a preferred configuration for a practical system. Full or nearly full interconnections, nearest-neighbor connections, branching networks, or other forms may be appropriate. In particular, the apparent superiority of the fully interconnected network from the standpoint of transient response must be tempered by the practical considerations against setting up a large number of very long connections.

XIII. ACKNOWLEDGMENTS

This article would be incomplete without mention of J. P. Runyon, whose pioneering contributions[7] and unceasing encouragement were a necessary (if not sufficient) condition for the work herein reported.

APPENDIX

Reciprocal Systems in the Steady State

The assumption of zero initial conditions, used in the study of transient behavior, must now be dropped. Thus, the transformed (2) are no longer valid, but the original equations (1) may be used. In the steady state, the rate of change of phase at every station is equal to the common system frequency f,

$$p_n'(t) = f, \qquad n = 1, \cdots, N \tag{107}$$

so that

$$p_n(t) = ft + \psi_n, \qquad -\infty < t < \infty. \tag{108}$$

Thus, in the steady state, the $v_n(t)$ being constant, the system equations (1) become

$$f = v_n + \lambda_n \sum_{m=1}^{N} a_{nm}(\psi_m - \psi_n - f\tau_{nm}), \qquad n = 1, \cdots, N. \tag{109}$$

The general solution to these equations is the expression given by Gersho and Karafin[1] in terms of cofactors of a matrix derived from the a_{nm}. In the reciprocal case, let the nth equation in (109) be multiplied by C_n and the equations be assumed over all n; when the reciprocity condition in the form of (8) is applied, all the terms in the phases ψ_n drop out and one gets

$$f \sum_{n=1}^{N} C_n = \sum_{n=1}^{N} C_n v_n - f \sum_{n=1}^{N} \lambda_n C_n \sum_{m=1}^{N} a_{nm}\tau_{nm}. \tag{110}$$

This can be solved immediately for f, the expression being similar in form to the solution reported by Gersho and Karafin, except that the C_n, which are easily determined by (9), replace the matrix cofactors.

REFERENCES

1. Gersho, A. and Karafin, B. J., Mutual Synchronization of Geographically Separated Oscillators, B.S.T.J., *45*, December, 1966, pp. 1689–1704.
2. Karnaugh, M., A Model for the Organic Synchronization of Communications Systems, B.S.T.J., *45*, December, 1966, pp. 1705–1735.
3. Guillemin, E. A., Synthesis of Passive Networks, John Wiley and Sons, New York, 1957, p. 64.
4. Churchill, R. V., Fourier Series and Boundary Value Problems, McGraw-Hill Book Co., Inc., New York, 1963, pp. 171–174.
5. Byrne, C. J., Karafin, B. J., and Robinson, D. B., Jr., Systematic Jitter in a Chain of Digital Regenerators, B.S.T.J., *42*, November, 1963, pp. 2679–2714.
6. Witt, F. J., An Experimental 224 Mb/s Digital Multiplexer-Demultiplexer Using Pulse Stuffing Synchronization, B.S.T.J., *44*, November, 1965, pp. 1843–1885 (see pp. 1852–1856).
7. Runyon, J. P., Reciprocal Timing of Time-Division Switching Centers, U. S. Patent No. 3,050,586, August 21, 1962.

Deformation of Gas Lenses by Gravity

By D. GLOGE

(Manuscript received September 14, 1966)

Gravity forces cause distortions in tubular gas lenses. A theory is derived here which yields excellent quantitative agreement with measured distortions for various tube lengths, diameters, and gases. It is shown that in a gas lens of optimum design the displacement of the optical center has a maximum at the end of the lens. The amount of displacement increases with the fourth power of the tube diameter and with the square of the gas pressure.

I. INTRODUCTION

If a cool gas is blown into a hot tube (Fig. 1), the gas heats up first at the wall of the tube and remains cool longer at its center. The density therefore, is higher in the center of the tube and decreases toward the wall. The increase in density is accompanied by an increase in dielectric constant. In this way the gas acts as a positive lens.[1,2]

At the same time, however, the cooler gas tends to sink down because of gravity, thus causing an asymmetric density profile in a horizontal tube.[3] Though a simple approach already gives an estimate of this effect,[4] a more rigorous theory is derived here using a perturbation calculation which determines the transverse convection currents from the unperturbed temperature profile and then uses the currents to correct the temperature profile.

II. TRANSVERSE CONVECTION CURRENTS

The tube walls are at a temperature T_W and ΔT degrees warmer than the entering gas. Heat diffuses toward the axis and determines the temperature field. Using the coordinate system shown in Fig. 1, the temperature field may be approximated by[2]

$$T = T_W - \Delta T\left[1 - 2\,\frac{x^2 + y^2}{a^2} + \left(\frac{x^2 + y^2}{a^2}\right)^2\right]e^{-z/s}, \qquad (1)$$

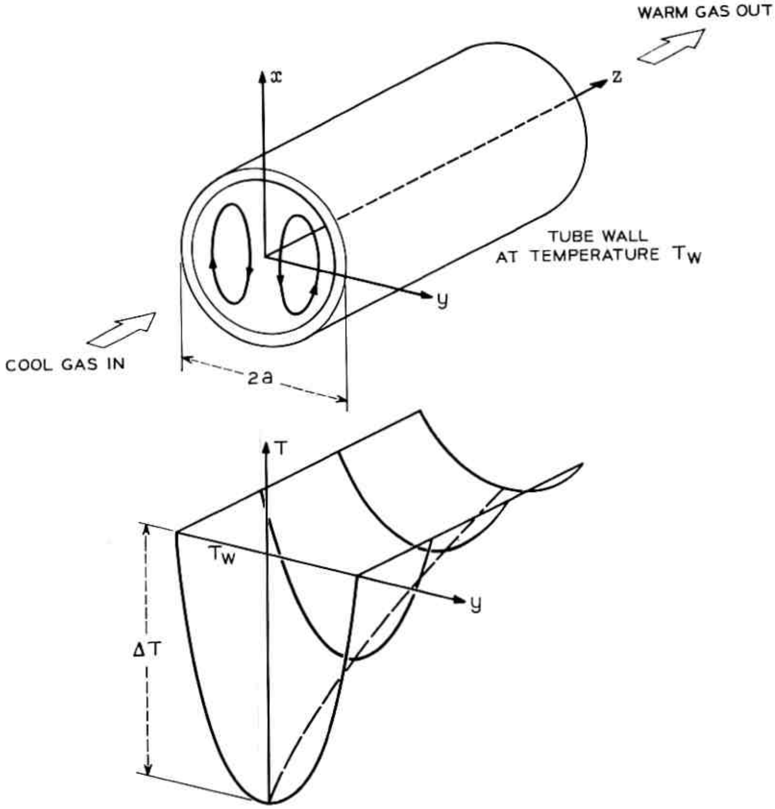


Fig. 1 — Convection currents and temperature distribution in a gas lens.

where a is the tube radius and s a decay length given by the formula

$$s = \frac{a^2 v_{zo}}{7.3\alpha}. \tag{2}$$

v_{zo} is the gas velocity along the axis and α the thermal diffusivity defined as the ratio of heat conductivity κ to heat capacity:

$$\alpha = \frac{\kappa}{\rho c_p}. \tag{3}$$

The heat capacity is written here as the product of density ρ and specific heat at constant pressure.

The temperature is related to the density ρ and the pressure p by the gas equation

$$p = R\rho T. \tag{4}$$

The density determines the gravitational forces $\mathbf{g}\rho$ which drive the gas particles in the transverse direction. The transverse components of the velocity field $\mathbf{v}(x, y, z)$ can be found from Newton's Law

$$\mathbf{g}\rho = \text{grad } p - \nu\rho\nabla^2\mathbf{v} + \rho \frac{d\mathbf{v}}{dt}, \tag{5}$$

where ν is the kinematic viscosity determining the frictional forces. The acceleration is described by the total differential $d\mathbf{v}/dt$ and for the steady state takes the form

$$\frac{d\mathbf{v}}{dt} = (\mathbf{v} \text{ grad})\mathbf{v}. \tag{6}$$

In the problem under consideration the gas may be treated as a quasi-incompressible (Boussinesq) fluid. That means that variations of density may be neglected, except insofar as they modify the action of gravity. Forming the curl of (5) therefore, yields

$$\text{curl } (\mathbf{g}\rho) = -\nu\rho \text{ curl } \nabla^2\mathbf{v} + \rho \text{ curl} \frac{d\mathbf{v}}{dt}. \tag{7}$$

Using (4) and rearranging (7) one finds

$$\mathbf{g} \times \left(\frac{\text{grad } T}{T} - \frac{\text{grad } p}{p}\right) = \nu \text{ curl curl curl } \mathbf{v} + \text{curl} \frac{d\mathbf{v}}{dt}. \tag{8}$$

Here grad p/p can be neglected compared with grad T/T, and T in the denominator will be replaced by the mean (absolute) temperature T_o. Finally, by inserting (6) one finds

$$\frac{1}{T_o}(\mathbf{g} \times \text{grad } T) = \nu \text{ curl curl curl } \mathbf{v} + \text{curl } (\mathbf{v} \text{ grad})\mathbf{v}. \tag{9}$$

To solve this equation, a tentative velocity distribution is introduced which represents the flow lines shown in Fig. 1. The unknown coefficients are chosen in such a way that the equation

$$\text{div } \mathbf{v} = 0 \tag{10}$$

is fulfilled, which assumes that the gas is incompressible. Then the velocity components

$$\begin{aligned} v_x &= -\frac{v_o}{a^4}(a^2 - x^2 - y^2)(a^2 - x^2 - 5y^2) \\ v_y &= -4\frac{v_o}{a^4}xy(a^2 - x^2 - y^2) \\ v_z &= \frac{v_{zo}}{a^2}(a^2 - x^2 - y^2) \end{aligned} \tag{11}$$

result which leaves only the coefficient v_o unknown, since the velocity v_{zo} is determined by the forced laminar flow in the tube. v_o is the vertical gas velocity at the tube center caused by the gravitational forces. It may be assumed to be much smaller than the longitudinal velocity v_{zo}. Though v_o is a function of z the variation of $\mathbf{v}$ in the z-direction is negligible compared to its variation in the cross-sectional plane and has to be considered only in the acceleration term where $\partial v_o/\partial z$ occurs multiplied with the velocity v_{oz}.

With these approximations, v_o can be determined by inserting (1) and (11) into (9) which yields

$$4g\frac{\Delta T}{T_o}e^{-z/s}y = 192\frac{v_o}{a^2}y + 18v_{zo}\frac{\partial v_o}{\partial z}. \tag{12}$$

Third- and higher-order products of x and y are neglected in this equation since they are only important at the wall of the tube and contribute little at the tube center.

Equation (12) is a linear inhomogeneous differential equation in z with the solution

$$v_o = \frac{g}{\nu}\frac{\Delta T}{T_o}\frac{a^2}{48}\frac{s}{s-q}(e^{-z/s} - e^{-z/q}), \tag{13}$$

where

$$q = \frac{3}{32}\frac{a^2}{\nu}v_{zo}. \tag{14}$$

A discussion of (13) is postponed in order to proceed with the calculation of the lens disturbance by using the derived convection currents to correct the temperature profile which, in turn, gives the density distribution and the lens profile.

III. DISPLACEMENT OF THE OPTICAL CENTER

The gravitational forces cause a continuous flow of cool gas toward the bottom of the pipe, which distorts the temperature profile more and more in the way shown in Fig. 2. The growing temperature gradient at the bottom, however, will increase the heat diffusion toward the center and counteract the convection effect. The equality of both effects is expressed by the equation

$$\alpha \nabla^2 T = \mathbf{v} \operatorname{grad} T \tag{15}$$

which determines the actual temperature profile under the boundary condition that $T = T_W$ at the tube wall.

Considering that the temperature function for axial direction is much less curved than the radial one, $\partial^2 T/\partial z^2$ may be neglected and (15) separated with respect to z.[6] This yields

$$\alpha\left(\frac{\partial^2 T}{\partial x^2} + \frac{\partial^2 T}{\partial y^2}\right) = v_x \frac{\partial T}{\partial x} + v_z \frac{\partial T}{\partial y} + v_z \frac{T - T_W}{s}, \tag{16}$$

where $T - T_W$ is an exponential function of z as already introduced by (1) for the undisturbed temperature profile.

No straightforward solution of (16) is known. Assuming, however, that the gravity effect, to first order, tilts the temperature profile in the x-direction as shown in Fig. 2, the amount of this disturbance can be calculated. The assumption implies that by transforming $T(x, y, z)$ into new coordinates

$$\xi = x - \delta(T_W - T); \qquad \eta = y; \qquad \zeta = z \tag{17}$$

the undisturbed profile can be regained, which in the following is denoted by $\theta(\xi, \eta, \zeta)$. Since this is symmetric with respect to ξ, the corresponding transformation in (16) must generate a differential equation for θ which contains only even terms in ξ. The requirement that the odd terms

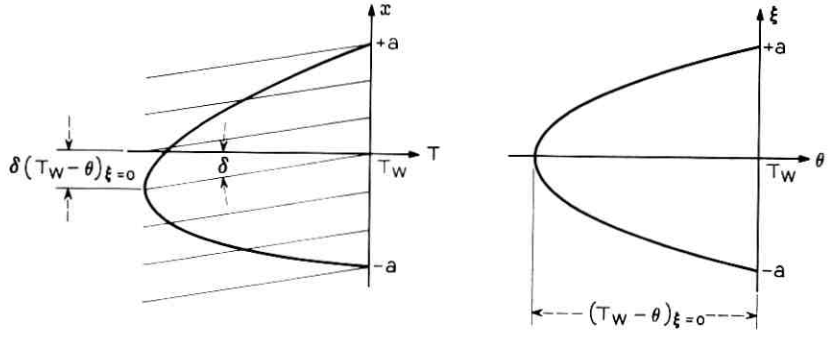


Fig. 2 — The temperature function $T(x)$ and its transformation into a symmetric function $\theta(x - \delta T)$.

cancel yields the following equation for δ:

$$2\alpha\delta\left(\frac{\partial\theta}{\partial\xi}\right)^2\frac{1}{\xi} = v_x\frac{\partial\theta}{\partial\xi} + 2\delta v_z\frac{\theta - T_W}{s}\frac{\partial\theta}{\partial\xi}. \tag{18}$$

The locus of the minimum of the temperature $T(x)$ is of particular interest, for this is the optical center of the distorted lens profile. Fig. 2 shows that this center occurs at a distance

$$d = \delta(T_W - \theta)_{\xi=0} \tag{19}$$

below the tube axis. Using for θ the undisturbed temperature profile given in (1) and solving (18) for $\delta(T_W - \theta)$ at $\xi = 0$ yields

$$d = \frac{v_o}{2\frac{v_{zo}}{s} + 8\frac{\alpha}{a^2}}. \tag{20}$$

By inserting (2) and (3) into (20) one finally finds

$$d = \frac{1}{750}\frac{ga^4}{\alpha^2}\frac{\Delta T}{T_o}\frac{q}{s - q}(e^{-z/s} - e^{-z/q}). \tag{21}$$

The diffusivity α and the viscosity ν for perfect gases are related by Eukens formula[7]

$$\frac{\alpha}{\nu} = \frac{1}{4}\left(9 - 5\frac{c_v}{c_p}\right), \tag{22}$$

c_v being the specific heat at constant volume. As Table I shows, the decay lengths s and q given by (2) and (14) differ very little. Since (21) is not defined for $s = q$ it is more convenient to use the following approximation for (21):

$$d = \frac{1}{750}\frac{ga^4}{\alpha^2}\frac{\Delta T}{T_o}\frac{z}{s}e^{-z/s}, \tag{23}$$

which is valid for $z < 2sq\,|q - s|$.

In Fig. 3 the displacement of the center of the lens profile is plotted

TABLE I

	c_p/c_v	α/ν	q/s
He	1.66	1.55	1.03
N_2	1.41	1.35	0.93
CO_2	1.31	1.30	0.89
CH_4	1.31	1.30	0.89

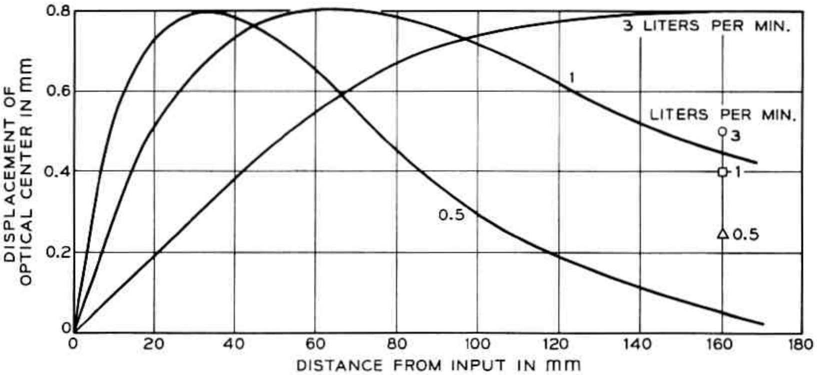


Fig. 3 — Displacement of the profile center in a tubular gas lens of $\frac{3}{8}$-inch i. d. for flow rates of 0.5, 1, and 3 liters per minute using air. ($\alpha = 0.22$ cm²/s). Measured data by DeGano.[8]

versus the distance from the tube input for flow rates of 0.5, 1, and 3 liters per minute. A tube of $\frac{3}{8}$-inch diameter and 100°C wall temperature is assumed. The gas enters at room temperature. The mean temperature during the process is assumed to be $T_o = 50°$C. The gas is air with a diffusivity $\alpha = 0.27 \text{ cm}^2/s$.

All curves show a linear increase of the displacement at the tube input, determined by the transverse acceleration of the gas. Further from the input the displacement follows the exponential decay of the temperature profile. The maximum displacement occurs at $z = s$. Measurements at the end of a 16-cm gas lens using the mentioned parameters are in fair agreement with the theory.[8]

In Fig. 4 the displacement is shown for a tube of $\frac{1}{4}$-inch diameter and two different gases: CO_2 with $\alpha = 0.125 \text{ cm}^2/s$ and N_2 with $\alpha = 0.25 \text{ cm}^2/s$. The temperatures are the same as in Fig. 3. The flow rate is 1 liter per minute. In this case, data are available for various tube lengths.[3] They show an excellent agreement with the predicted behavior of d versus z.

The focal length of the tubular gas lens has a minimum if the flow rate is chosen in such a way that s equals the tube length. The maximum displacement occurs at the end of such a lens and has the value

$$d_{\max} = \frac{1}{2040} \frac{ga^4}{\alpha^2} \frac{\Delta T}{T}. \tag{24}$$

A more useful measure for the gravity effect is the distance D by which a light beam has to be displaced off the tube axis to pass the lens without deflection. Integrating at $x = D$ over the tube length L one finds D from the requirement that the total deflection cancels:

$$\int_0^L \left.\frac{\partial T}{\partial x}\right|_{x=D} dz = 0. \tag{25}$$

The development of the (disturbed) temperature field T about the axis yields for small distortion

$$\int_0^L [D - d(z)]e^{-z/s}\, dz = 0; \tag{26}$$

and finally, by using (23) one has

$$D = \frac{1}{3000}\frac{ga^4}{\alpha^2}\frac{1 - \left(\frac{2L}{s} + 1\right)e^{-2L/s}}{1 - e^{-L/s}}. \tag{27}$$

In Fig. 5 the displacement D is plotted versus the flow rate for CO_2 in a 7-inch tube assuming the same temperatures as in Figs. 3 and 4. Data measured by Steier[3] show good agreement with the theory. For $L > s$

$$D \approx \frac{1}{3000}\frac{ga^4}{\alpha^2} \tag{28}$$

is a good approximation. According to this formula, the optical center of a CO_2 lens of optimum design would occur outside the tube if the tube diameter is larger than 1 cm.

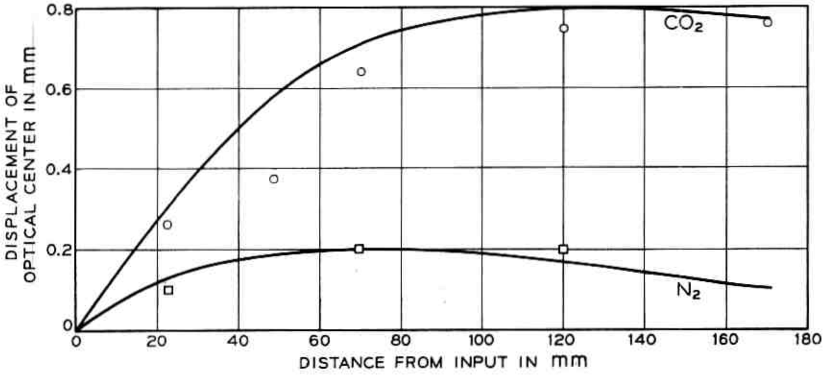


Fig. 4 — Displacement of the profile center in a tubular gas lens of ¼-inch i. d. for a flow rate of 1 liter per minute using CO_2 ($\alpha = 0.1$ cm^2/s) or N_2 ($\alpha = 0.2$ cm^2/s). Measured data by Steier.[3]

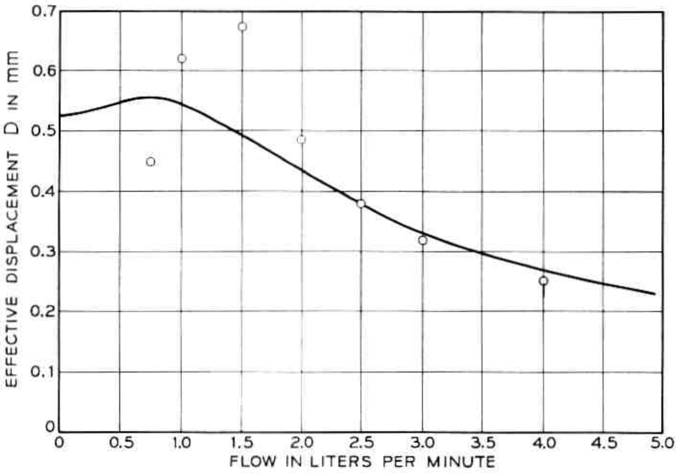


Fig. 5 — Displacement of the optical center of a tubular gas lens using CO_2 and a tube of 7 inches long and ¼-inch diameter. Measured data by Steier.[3]

IV. CONCLUSIONS

The calculations show that the temperature distribution in a gas-filled tube undergoes a distortion which increases with the fourth power of the tube radius. A square law dependence on pressure is predicted for the range of 0.05 to 50 atmospheres where the thermal conductivity is independent of the pressure and therefore, the diffusivity $\alpha \propto 1/p$.

As a measure of the distortion, the displacement of the effective optical center in a tubular gas lens is calculated. Using CO_2 at room temperature and a tube of 10-mm diameter at 100°C wall temperature the optical center occurs at the bottom of the tube.

REFERENCES

1. Berreman, D. W., A Lens or Light Guide Using Convectively Distorted Thermal Gradients in Gases, B.S.T.J., *43,* July, 1964, pp. 1476–1479.
2. Marcuse, D. and Miller, S. E., Analysis of a Tubular Gas Lens, B.S.T.J., *43,* July, 1964, pp. 1759–1782.
3. Steier, W. H., Measurements on a Thermal Gradient Gas Lens, IEEE Trans., *MTT-136,* November, 1965, pp. 740–748.
4. Miller, S. E., unpublished work.
5. Jacob, M., *Heat Transfer,* Vol. 1, John Wiley & Sons, New York, N. Y., 1949, p. 17.
6. Ibid., p. 433.
7. Kennard, E. H., *Kinetic Theory of Gases,* McGraw-Hill Book Co., Inc., New York, N. Y., 1938, p. 182.
8. Degano, A., private communication.

Acoustic Light Modulators Using Optical Heterodyne Mixing

By R. W. DIXON and E. I. GORDON

(Manuscript received September 21, 1966)

Acoustic light modulators are discussed in which the optical carrier is reinserted in the diffracted, frequency-shifted light beam. Reinsertion is accomplished in a novel fashion using a Kösters prism. In contrast to the usual acoustic modulator, the diffracted light is power modulated at the acoustic frequency. Modulation bandwidth and depth are each calculated as a function of the optical and acoustic beam parameters, assuming a Gaussian optical beam and rectangular acoustic beam. It is shown that the modulation bandwidth increases proportional to the optical beam diffraction angle and equals the inverse of the transit time of the sound across the waist of the optical beam. The optimum modulation depth, for a given acoustic power, corresponds to approximate equality of the optical and acoustic diffraction angles. Confirming experiments in the frequency range 250/350 MHz are described.

I. INTRODUCTION

Detection of optical radiation using heterodyne mixing was pioneered in the classic experiments of Forrester, Gudmundsen, and Johnson.[1] They successfully detected the microwave beat between two Zeeman components of a mercury arc. With coherent light sources the technique was utilized initially in the investigation of the mode structure and frequency stability of the helium-neon laser.[2] Subsequently, optical heterodyne mixing has been used as a sensitive, high-resolution detector of frequency shifts in the study of Brillouin scattering[3] and of the frequency broadening of Rayleigh scattered light.[4]

It is well known that under the correct circumstances an optical beam, passing through a transparent material containing a traveling acoustic wave, has part of its energy diffracted by the refractive index variations associated with the acoustic wave. In the proper range of parameters, known as the Bragg region, the diffracted light is confined around a single direction. The system acts as a single-sideband suppressed-carrier modulator in which the diffracted light is intensity modulated with the envelope of the modulation subcarrier. Depending on the relative directions of the light and sound the diffracted light frequency is increased or decreased by the frequency of the modulation subcarrier.

One may arrange to have the diffracted light fall into the same solid angle as a portion of the original light, and thereby allow detection of the diffracted beam by heterodyne mixing with the undiffracted beam. Heterodyne mixing of light in which the signal frequency is shifted by diffraction from sound in a liquid cell ultrasonic modulator has been demonstrated by Cummins and Knable,[5] and the increased sensitivity of this technique has been briefly pointed out by Goodwin and Pedinoff.[6]

Optical heterodyne procedures in combination with Bragg scattering can also be used as a sensitive detector of sound. For example, Lastovska and Benedek[3] have shown that thermal sound (thermal Brillouin scattering) may be detected in this way. In fact, in some applications—e.g., at very high frequencies—this technique can be a more sensitive detector of sound than the best available transducer. The advantages of optical techniques for investigating sound beam intensity profiles, angular distributions, etc. at any point in the medium in which the acoustic wave propagates have been pointed out earlier[7] and may now be supplemented by the increased sensitivity and spectral range which the utilization of optical heterodyne detection affords. An additional advantage of this method of optical detection is that solid-state photodetectors may be used instead of photomultipliers, without compromising sensitivity, whenever sufficient light is available that the shot noise associated with the optical local oscillator limits detector sensitivity.[8]

This paper is concerned with the detailed properties of an acoustic modulator when optical heterodyne detection of the modulated light is employed. A coherent optical source is assumed. The range of useful modulation frequencies for this technique extends at present well into the microwave region. Present limitations of efficient thin-film transducers limit operation to below 10 GHz. Acoustic loss for some applications becomes important at lower frequencies than this. The frequency response of commercial photodiodes extends to about 30 GHz.

The analysis includes a discussion of modulation bandwidth and optimum modulation conditions and concludes with the prediction that very large dynamic bandwidths may be obtained using optical heterodyne mixing in conjunction with an acoustic modulator. A series of experiments involving a novel beam splitter and various modulating materials confirm this prediction. It is shown that bandwidth is related to the diffraction angle of the optical beam and may be varied over a large range by changing this angle. Experiments have been restricted to solid modulating materials, but the results are applicable without modification to liquids if their higher acoustic loss can be accepted. It is concluded that large modulation depths should be practical at modulation frequencies well into the microwave region and at optical frequencies throughout the visible and infrared. It is also pointed out that for small diffracted light intensities, proportionately much larger modulation depths are possible using this technique than if the transmitted light beam alone were monitored.

The relation of modulation depth and bandwidth for a given acoustic power is discussed, and experimental confirmation of the conclusions is presented. In addition, it is shown that with these acoustic light modulators frequency and phase information, as well as power modulation, may be transferred to the light with large dynamic bandwidth.

II. DISCUSSION

2.1 *Optical Beam Geometry*

In order to obtain two beams for use in optical heterodyne mixing experiments, it has been common practice to use a beam splitter and mirror assembly similar to that shown schematically in Fig. 1. The configuration shown would be appropriate for coherent detection of light modulated by Bragg diffraction from an acoustic wave.[5] In the experiments described here it has been found very useful to replace the mirror and beam splitter of such an experiment with a Kösters prism.[9] These prisms have been commonly used in Michelson interferometers and similar apparatus where ease of alignment is desired.[10] Probably this prism has not received the attention it deserves for applications in modern physical experiments.

Fig. 2 shows the general construction of a Kösters prism. Two accurately constructed 30-60-90° prisms are carefully cemented together with a dielectric beam splitter between them. Because the two exit beams travel symmetrical paths to their intersection with any plane perpendicular to the beam splitter, they are optically identical.

The experimental apparatus used in the present series of experiments is shown schematically in Fig. 3. The two beams from the Kösters

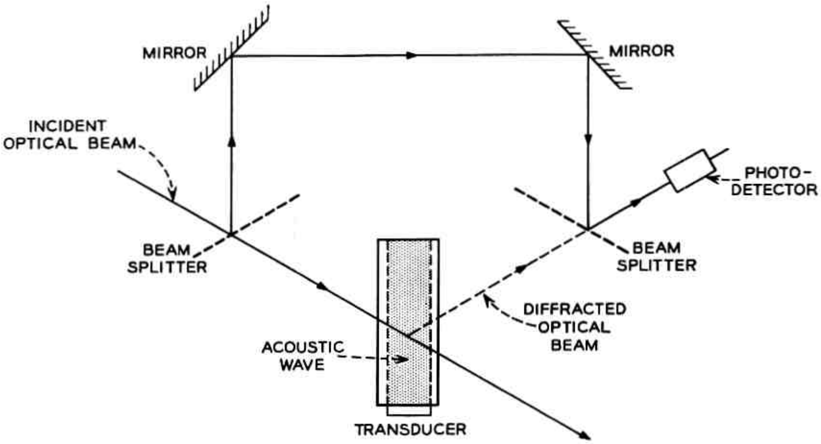


Fig. 1 — Mirrors and beam splitters arranged for optical heterodyne detection of Bragg diffracted light (after Cummins and Knable, Ref. 5).

prism are set to intersect within the acoustic beam at an angle equal to twice the Bragg angle so that Bragg diffracted light from each of the incident beams falls into the diffraction angle of the other beam. The Kösters prism assures symmetrical paths for the diffracted and undiffracted light and thereby makes alignment very easy. A lens of appropriate focal length is positioned so that the optical beam waists intersect with the desired convergence angle in the center of the modulator's acoustic beam.

Consider, in Fig. 3, an acoustic traveling wave originating from the transducer. The frequency of the light diffracted from beam 1 into beam 2 is increased by the acoustic frequency while the light diffracted

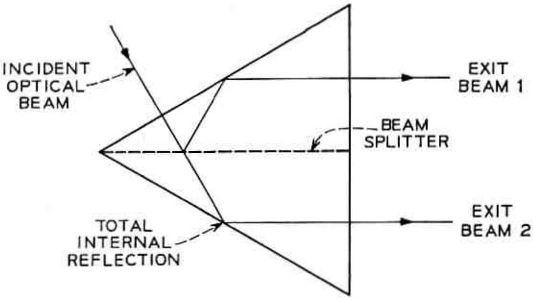


Fig. 2 — A Kösters prism or double image beam splitter.

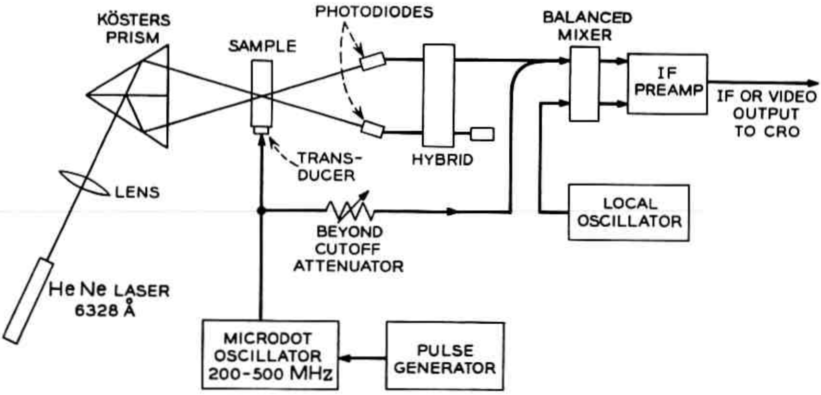


Fig. 3 — Schematic diagram of experimental apparatus.

from beam 2 into beam 1 is decreased by the same amount. The diffracted and transmitted beams interfere at the photodiodes after which the difference frequencies are amplified and detected. Since the difference frequency signal from diode 1 is 180 degrees out of phase with that from diode 2, the two signals may be fed, if desired, into a hybrid and will add at the port for which the output is the difference between the two inputs (the push-pull output). This has the advantage, which was not important in the present experiments but which might be for some applications, that amplitude noise on the incident optical beam (such as the beating of various laser modes) does not appear at the same output as the detected signal if the two sides of the detection system are balanced. In some of the experiments discussed only one diode was used, in others the balanced system worked very well and was no more difficult to align than the system employing a single diode. Modulation frequencies were normally near 300 MHz.

2.2 *Modulator Bandwidth*

In order to appreciate the limits on modulation bandwidth, it is instructive to consider qualitatively several special cases. For simplicity, assume that only photodiode 2 is used and that only an outgoing acoustic wave is present. The Kösters prism is positioned so that the two light beams of frequency ν which are incident on the acoustic modulator have their point of intersection at the center of the acoustic beam. They intersect at twice the Bragg angle, $2\theta_o$, [$\sin \theta_o = \frac{1}{2} f_o \lambda / v$], which is a function of the desired modulator center frequency f_o, the acoustic velocity v, and the optical wavelength in the medium λ. The optical

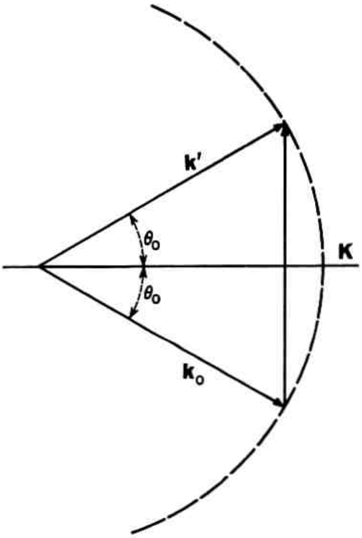


Fig. 4 — A k-vector diagram of the Bragg diffraction process in which optical and acoustic diffraction angles are assumed negligibly small.

velocity is c'. The wave vector relation (momentum conservation) among the three waves may be written $\mathbf{k}_o + \mathbf{K} = \mathbf{k}'$, where $|\mathbf{k}_o| = 2\pi\nu/c'$ and $|\mathbf{k}'| = 2\pi(\nu + f_o)/c'$ are the incident and scattered optical k-vectors, and $\mathbf{K}$ is the k-vector of the ultrasonic wave. Since to order $|f_o/\nu| \ll 1$, $|\mathbf{k}_o| = |\mathbf{k}'|$ it is possible to make the elementary but useful construction shown in Fig. 4. The dotted circle has radius $|\mathbf{k}_o|$ and defines the locus of allowed $|\mathbf{k}'|$. Only a phonon of precisely the correct $|\mathbf{K}|$ will scatter $\mathbf{k}_o$ if one assumes that neither the optical nor acoustic beam has any angular width.

In order to appreciate the effect of nonzero diffraction angle, consider the limiting case in which the diffraction angle of the acoustic beam is large compared with the diffraction angle of the optical beam (Fig. 5); $\mathbf{K}$ has a well-defined magnitude but an angular width $\Delta\theta$. Only those acoustic k-vectors near the direction of $\mathbf{K}_o$ scatter light into spatial coherence with the heterodyning beam $\mathbf{k}_o'$. Thus, the detected signal amplitude is lower than if the same acoustic power were confined to a smaller diffraction angle.

If the acoustic frequency is increased to a new value, $|\mathbf{K}|$ is increased and the construction shown dashed is appropriate. No signal will be observed on the photodiode because no $\mathbf{K}$ can scatter into $\mathbf{k}_o'$. This modulator therefore, not only produces low modulation intensity but also possesses small bandwidth.

Now consider the other limiting case in which the diffraction angle of the optical beam is much larger than that of the acoustic beam (Fig. 6). In this case, only that portion of the optical energy near the center of the diffraction angle can be scattered by the acoustic wave of vector $\mathbf{K}_o$ and heterodyned with the other optical wave. Therefore, the scattered intensity is much less than it would be for the same optical and acoustic powers, if the diffraction angle of the light were decreased.

If the frequency f_o is changed slightly, a new $\mathbf{K}$ is defined which is slightly different in magnitude but which has the same direction as $\mathbf{K}_o$. In this case, however, the detected optical signal at the photodiode is essentially unchanged as long as the deviation ΔK is such that $K_o + \Delta K$ still connects two points on the dashed circle which are within the diffraction angles of the incident and hetrodyning optical beams. By varying f one traces out the angular profile of the *optical* beam. A bandwidth may be approximately defined by the condition

$$\{ \mid K_{\max} \mid - \mid K_{\min} \mid \} \approx \sqrt{2}\, \Delta\theta_o k_o \cos\theta_o \tag{1}$$

from which it follows that the bandwidth Δf_B is given by

$$\Delta f_B \approx \sqrt{2}\, (v/\lambda)\, \Delta\theta_o \cos\theta_o \tag{2}$$

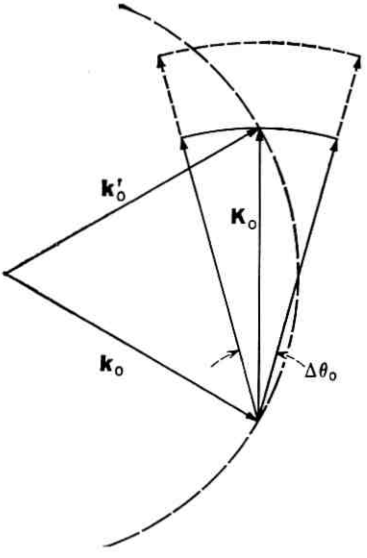


Fig. 5 — A k-vector diagram of the Bragg diffraction process in which the acoustic diffraction angle is large compared with the optical diffraction angle.

in which λ is the optical wavelength. The center frequency f_o is defined by $\sin \theta_o = \frac{1}{2} f_o v/\lambda$. The optical beam diffraction angle $\Delta\theta_o$ may be expressed in terms of the beam waist diameter w_o,[11] corresponding to a Gaussian beam for which w_o is the full width at half intensity,

$$w_o \, \Delta\theta_o \approx \lambda\left(\frac{2 \ln 2}{\pi}\right). \tag{3}$$

Substituting for $\Delta\theta_o$ yields for the bandwidth

$$\Delta f_B \approx \left(\frac{2\sqrt{2} \ln 2}{\pi}\right) \frac{v \cos \theta_o}{w_o} \tag{4}$$

which approximates the reciprocal of the transit time of the sound across the waist of the gaussian light beam. This important result shows how acoustic modulators with large bandwidths are made possible by increasing the diffraction angle of the optical beam. The fact that the diffraction angle $\Delta\theta_o$ of the optical beam must be less than the Bragg angle requires that the fractional bandwidth of the modulator obeys the inequality, using (2) and (3),

$$\frac{\Delta f_B}{f_o} < \frac{1}{\sqrt{2}}. \tag{5}$$

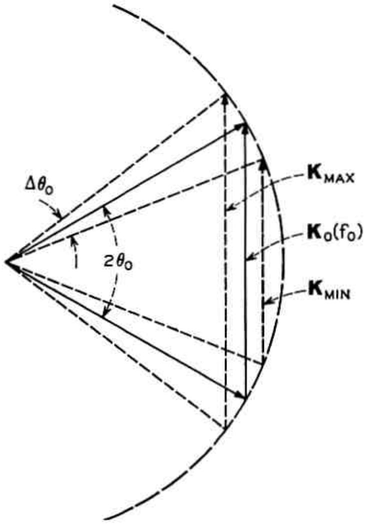


Fig. 6 — A k-vector diagram of the Bragg diffraction process showing the origin of acoustic modulator bandwidth; the optical diffraction angle is large compared with the acoustic diffraction angle.

Thus, fractional bandwidths as large as 70 percent are possible if transducer bandwidths can be made compatable with this requirement.

It should be clear now for the situation depicted in Fig. 6 that increasing the acoustic diffraction angle would increase the detected signal without affecting the bandwidth. Likewise, from Fig. 5, in the other extreme the beat signal increases but the bandwidth is unchanged when the acoustic diffraction angle is decreased. It is therefore, plausible on the basis of the foregoing comments that the choice of *acoustic diffraction angle approximately equal to optical diffraction angle* is the optimum choice for a given bandwidth. The bandwidth is determined entirely by the diffraction angle of the optical beam. This conclusion is given quantitative expression in Appendix A.

III. THEORY AND COMPARISON WITH EXPERIMENT

In order to check the preceding ideas quantitatively, the photodiode current which would be expected with the geometry shown in Fig. 3 was calculated for the case of light beams with identical Gaussian profiles and an acoustic beam with rectangular cross section. It was known that these beam profiles could be accurately approximated experimentally. The lens was positioned in the incident optical beam so that the optical beam waist occurred at the center of the modulator. The diode photocurrent was computed as the far-field interference integral of the product of the acoustically diffracted optical amplitude and the heterodyning optical amplitude. Details of the calculation are presented in Appendix A. The expression for the component of photocurrent at the modulating frequency f, apart from material and numerical constants, is

$$\iota(f) = -\frac{i}{(\sin\,\theta_o)^{\frac{1}{2}}} P_{\text{opt}}(P_{\text{acoustic}})^{\frac{1}{2}} \exp\,[i2\pi f(t - R/c')] \cdot\frac{\text{Erf}\,(a)}{a^{\frac{1}{2}}}\cdot\frac{\text{Erf}\,(b)}{b^{\frac{1}{2}}}\cdot\exp\left[-\frac{\pi^2}{4\ln 2}\frac{(f - f_o)^2 w_o^2}{v^2\cos^2\,\theta_o}\right], \tag{6}$$

where

$$a \equiv \left(\frac{\ln 2}{2}\right)^{\frac{1}{2}} \frac{h}{w_o}$$

$$b \equiv (\ln 2)^{\frac{1}{2}} \frac{L \sin\,\theta_o}{w_o}$$

$$\sin\,\theta_o \equiv \frac{1}{2}\frac{f_0\lambda}{v}.$$

Again w_o is the full width at half intensity of the Gaussian beam, $2\theta_o$ is the intersection angle inside the modulating medium and f_o is the acoustic frequency for which θ_o is the Bragg angle. The acoustic beam height is h and its width (dimension in the plane formed by the acoustic and optic propagation directions) is L. The distance to the detector is R. Equation (6) is the basic relation which is subject to experimental verification.

The first experimental checks of (6) confirmed that the acoustic and optical power dependences are correct. Experimentally, a 3-dB decrease in optical intensity or a 6-dB decrease in acoustic power as expected decreased the detected current by one-half. The other most interesting predictions made by (6) are contained in the terms $\mathrm{Erf}(b)/b^{\frac{1}{2}}$ and

$$\exp\left[-\frac{\pi^2}{4\ln 2}\frac{(f-f_o)^2 w_o^2}{v^2\cos^2\theta_o}\right].$$

The former is concerned with the maximum detected signal amplitude and the latter with the dynamic modulation bandwidth, both as functions of the diffraction angle of the incident light.

3.1 *Bandwidth vs Beam Waist Diameter*

Consider first the term

$$\exp\left[-\frac{\pi^2}{4\ln 2}\frac{(f-f_o)^2 w_o^2}{v^2\cos^2\theta_o}\right].$$

When the modulating frequency f changes from the value f_o, this term describes the decrease in detected signal. An acoustic half-power bandwidth, Δf_B, for which the detected current is greater than $1/\sqrt{2}$ below its maximum value, may be defined and is given by

$$w_o\,\Delta f_B = \left(\frac{2\sqrt{2}\ln 2}{\pi}\right)v\cos\theta_o\,. \tag{7}$$

The beam waist diameter times the bandwidth is thus a constant for a given material at a given center frequency $f_o(\theta_o)$. By making the beam waist smaller, e.g., by focusing the incident optical beam, the dynamic bandwidth may be increased. As indicated earlier the bandwidth is intimately related to the transit time of the acoustic wave across the optical beam. In fused quartz for longitudinal waves at frequencies low enough that $\cos\theta_o \approx 1$, (7) becomes

$$w_o\,\Delta f_B = 3.70 \times 10^5\ \text{cm/sec}. \tag{8}$$

In Fig. 7 the measured bandwidth Δf_B, for modulation in fused quartz, is plotted against w_o^{-1} and (8) is shown plotted as the solid line. The

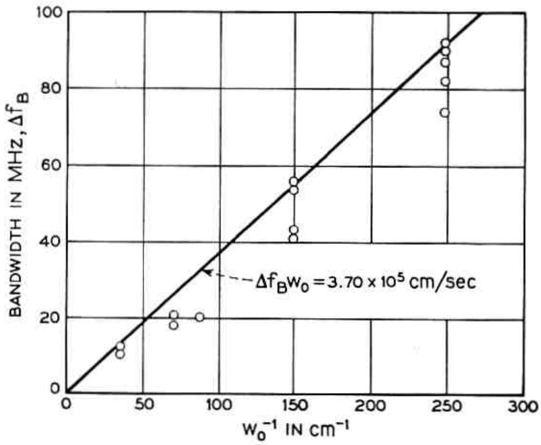


Fig. 7 — Experimental dynamic bandwidth plotted against the reciprocal of the optical beam waist diameter. The solid line is the theoretical expression, (8).

value of w_o was varied by placing lenses of different focal lengths F in the incident optical beam. The relation between w_o and F was assumed to be of the form given by Kogelnik[11] except that w_o is defined as the full width at half intensity of the Gaussian beam. In the limit appropriate for the present experiments

$$w_o \approx \left(\frac{2 \ln 2}{\pi}\right) \frac{F\lambda_o}{W_o}, \tag{9}$$

in which W_o is the beam waist of the incident laser beam and w_o is the beam waist of the beam in the scattering medium following its transformation by a lens of focal length F. Here, λ_o is the free space wavelength. Equation (9) was used to convert from values of F to values of w_o. Using a scanning slit and photomultiplier the value $W_o = (1.00 \pm 0.01)$mm was found. These measurements incidentally verified that the intensity profile of the laser beam was approximately Gaussian.

The experimental points in Fig. 7 were obtained using apparatus which is shown schematically in Fig. 3. The Microdot oscillator produced repetitive pulses of RF energy with each pulse having about 1-μsec duration. Cadmium sulfide thin-film transducers converted these electromagnetic pulses into acoustic energy.[12] The deflected light pulses were detected in the indicated photodiode geometry followed by a standard microwave superheterodyne receiver. In order to avoid IF detector diode nonlinearity the IF output of the system was usually viewed directly on a broadband oscilloscope. Data were taken from the amplitude of the first deflected light pulse (due to the initial outgoing acoustic wave) and were therefore, independent of acoustic resonance effects which would be associated with long acoustic pulses. The RF power incident on the transducer was kept the same at each frequency. A small amount of each RF pulse was fed, using a beyond cutoff attenuator, from in front of the transducer to the input of the receiver. This pulse served as a calibrating signal for the receiver and helped to correct for changes in receiver sensitivity as the frequency was changed. The transducer response in the 100-MHz region around 300 MHz was flat to within 1 dB. The scatter of the experimental points can be attributed to several causes, the most important of which is the unavoidable small changes in optical alignment between the time that each bandwidth curve was taken. It is believed that the largest observed value for any given value of w_o is the most appropriate and the agreement is considered to be good.

3.2 *Signal Amplitude vs Beam Waist Diameter*

Now consider the amplitude terms in (6) which involve the error functions. When $f = f_o$

$$\iota(f_o) \propto \frac{\mathrm{Erf}\,(a)}{a^{\frac{1}{2}}} \; \frac{\mathrm{Erf}\,(b)}{b^{\frac{1}{2}}}\,, \tag{10}$$

where

$$a \equiv \left(\frac{\ln 2}{2}\right)^{\frac{1}{2}} \frac{h}{w_o}$$

$$b \equiv (\ln 2)^{\frac{1}{2}} \frac{L \sin \theta_o}{w_o}.$$

In the experiments, the acoustic beam height h was made sufficiently large (3 mm) compared with the largest beam waist ($\approx$1 mm) that $\mathrm{Erf}(a) \approx 1$ for all values of w_o of interest. Thus, the signal current as a function of w_o is given by

$$A(w_o) = A_o w_o \,\mathrm{Erf}\left[(\ln 2)^{\frac{1}{2}} \frac{L \sin \theta_o}{w_o}\right]. \tag{11}$$

This equation is plotted as the solid curve in Fig. 8; A_o has been considered an adjustable normalizing parameter but the argument of the error function is determined using the experimental acoustic beam width of $L = 7.00$ mm. The experimental points were taken using the configuration shown in Fig. 3 and the beam waist diameter was obtained using (9). Again the agreement is quite good.

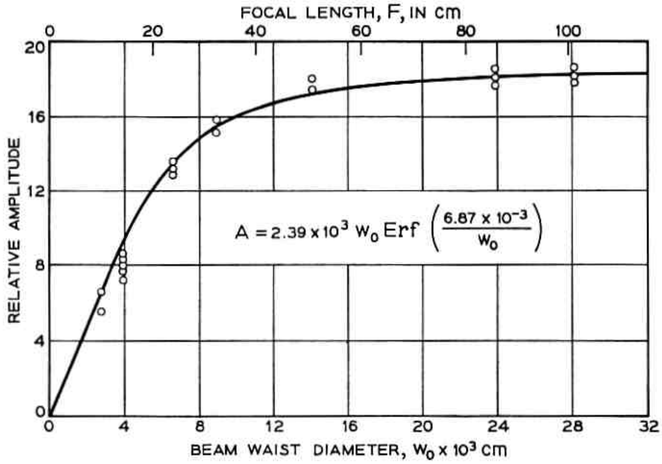


Fig. 8 — Experimental relative modulation frequency photocurrent vs beam waist diameter. The solid curve is the normalized theoretical expression, (11).

3.3 *Signal Amplitude vs Transducer Width*

The main interest thus far has been in determining the result on the signal amplitude of varying the optical diffraction angle and demonstrating that it is solely responsible for the bandwidth. It is instructive to re-emphasize the effect of the acoustic diffraction angle by calculating the detected signal variation as a function of acoustic beam width L. For a given w_o, h, θ_o and acoustic power the detected signal amplitude at $f = f_o$ is given by

$$A(L) = A_o \frac{\mathrm{Erf}\,(b)}{b^{\frac{1}{2}}}$$
$$b = (\ln 2)^{\frac{1}{2}} \frac{L \sin \theta_o}{w_o}. \tag{12}$$

It should be noted that b can also be written as 0.85 × (diffraction angle of light/diffraction angle of the sound).

In the region of small L, ($b \ll 1$)

$$A(L) = A_o(2/\pi^{\frac{1}{2}})b^{\frac{1}{2}} \propto L^{\frac{1}{2}}. \tag{13}$$

Therefore, for a given acoustic power, the signal may be increased by increasing the transducer width. The diffraction angle of the sound is too large for optimum scattering from the given optical beam. In the opposite limit of large L, ($b \gg 1$)

$$A(L) = A_o/b^{\frac{1}{2}} \propto L^{-\frac{1}{2}}. \tag{14}$$

Increasing L decreases the signal. Here, the acoustic diffraction angle is too small to use all of the incident light.

The signal amplitude has a broad maximum at $b = 0.99$ corresponding to approximate equality of the acoustic and optical diffraction angles for optimum modulator efficiency. This conclusion is expected to depend only weakly on the actual beam geometries.

IV. MODULATION DEPTH

Consider now the modulation depth which one would expect from an acoustic modulator. A simple computation of modulation depth is possible only when the waves are considered in a plane wave approximation in the Bragg region. This situation should approximate qualitatively the behavior expected with the present experimental configuration. In the Bragg region, there exists a well-defined angular relationship among the beam directions. Furthermore, all of the light diffracted from the main beam is diffracted into a single Bragg order. Under these circumstances, the solution of the problem of the generation of an optical beam by parametric coupling with an acoustic beam shows that the amplitude of the diffracted beam can be written[13]

$$P_d^{\frac{1}{2}} = \mp i P_{\text{opt}}^{\frac{1}{2}} \sin \eta^{\frac{1}{2}} \exp i2\pi(\nu \pm f)t, \tag{15}$$

while the transmitted beam has the form

$$P_t^{\frac{1}{2}} = P_{\text{opt}}^{\frac{1}{2}} \cos \eta^{\frac{1}{2}} \exp i2\pi\nu t \tag{16}$$

in which η is a scattering parameter defined for a rectangular acoustic beam by[13]

$$\eta \equiv \tfrac{1}{2}\pi^2 \left(\frac{n^6 p^2}{\rho v^3}\right)\left(\frac{LP_a}{\lambda_o^2 h \cos^2 \theta_o}\right), \tag{17}$$

where L is the beam width, P_a the acoustic power, h the acoustic beam height, p the appropriate photoelastic component, v the acoustic velocity, and ρ the mass density, and n the refractive index.

For conventional acoustic modulation, a square law photodetector placed in the transmitted beam will produce a photocurrent proportional to $|P_t|$ which has a maximum value of P_{opt} for no acoustic signal and a minimum value $P_{\text{opt}} \cos^2 \eta^{\frac{1}{2}}$ when the acoustic signal is present. Hence, a modulation depth

$$m_1 = \sin^2 \eta^{\frac{1}{2}} \tag{18}$$

may be defined which is some measure of the ability to detect the influence of the acoustic energy on the light.

Similarly for the superheterodyne case, the photo-current has the form

$$| P_i^{\frac{1}{2}} + P_d^{\frac{1}{2}} |^2 = P_{\text{opt}}(1 \pm \sin 2\eta^{\frac{1}{2}} \sin 2\pi f t)$$

and

$$m_2 = \frac{2 \sin 2\eta^{\frac{1}{2}}}{1 + \sin 2\eta^{\frac{1}{2}}}. \tag{19}$$

Clearly, this modulation depth is superior to that obtained when the carrier is simply intensity modulated. This superiority is most dramatic when small deflected intensities are involved. Fig. 9 shows a comparison of the two modulation depths each plotted as a function of the scattering parameter η (which for $\eta \lessapprox 0.1$ is quite accurately equal to the deflected intensity). It is seen that when 10 percent deflected intensity is obtained, the modulation depth with the optical heterodyne system is 75 percent. For the very small deflected intensity of 0.01 percent, one still has a usable modulation depth of 4 percent in the optical heterodyne detector.

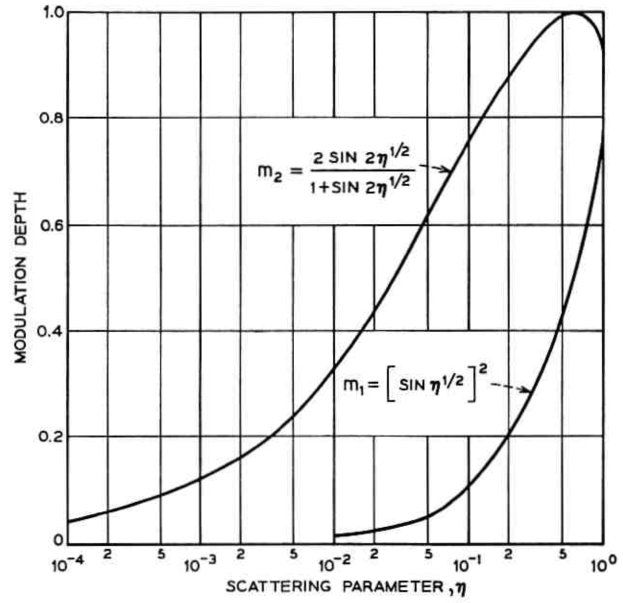


Fig. 9 — Theoretical curves, valid for plane waves, of the ordinary optical modulation depth m_1 and the optical heterodyne modulation depth m_2 both plotted vs acoustic scattering parameter η.

Modulation depth was experimentally investigated using a modulation frequency of 20 MHz which was chosen instead of a frequency near 300 MHz as in the other experiments described in this paper in order to display the modulation depth directly. At 20 MHz one photodiode could be directly connected to an oscilloscope with amplifiers having sufficient baseband bandwidth and gain to display both the dc diode photocurrent and the 20-MHz modulating photocurrent. Fig. 10 shows a typical oscilloscope trace of the diode output for longitudinal acoustic waves propagating in KRS-5 when the relative diffracted light intensity was about 20 percent. The baseline corresponds to no light. Fig. 11 shows a comparison of the modulation depth measured from such photographs compared with the total intensity diffracted from the incident beam under the same experimental conditions. KRS-5 was again the modulating medium. The very large modulation depths obtained for small diffracted intensities are, of course, the most interesting feature of these curves and qualitatively verify the ideas just discussed.

There is some disadvantage in working at a frequency as low as 20 MHz, viz., that the optical-acoustic interaction is not strictly in the Bragg region and a significant amount of light from the main optical beam is diffracted into orders other than that satisfying the Bragg condition. For this reason, the curve of m_2 against acoustic power does not reach 100 percent as it would if the modulating frequency were high enough (greater than about 60 MHz) that Bragg diffraction was dominant. At these increased frequencies a direct display of the modula-

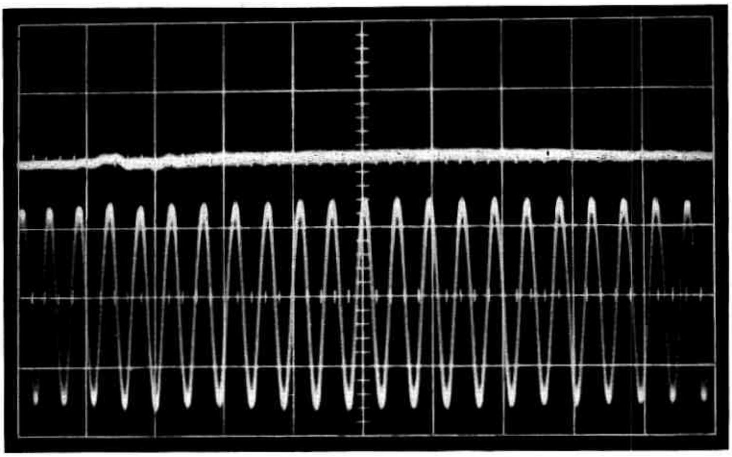


Fig. 10 — Oscilloscope trace showing 80 percent modulation depth at 20 MHz obtained using optical heterodyne detection. The relative diffracted intensity was about 20 percent.

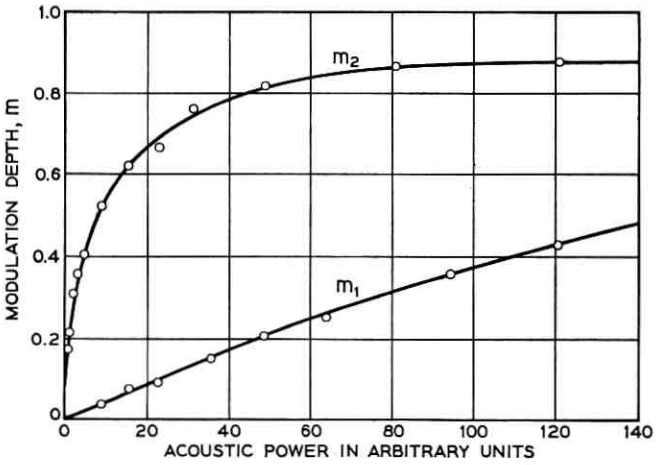


Fig. 11 — Experimental comparison of ordinary modulation depth m_1 and heterodyne modulation depth m_2 for logitudinal acoustic waves in KRS-5.

tion depth, such as that shown in Fig. 10, becomes difficult for the small photocurrents used experimentally.

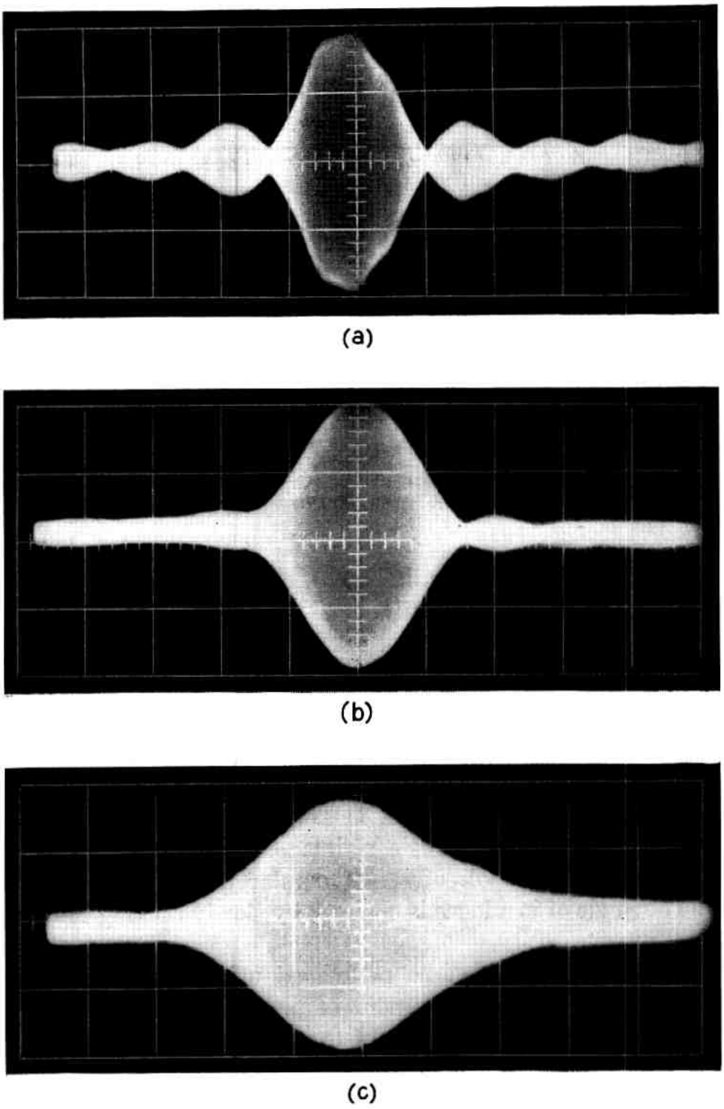
V. MEASUREMENT OF ACOUSTIC DIFFRACTION PATTERN

It is known that Bragg diffraction can be used to determine the angular distribution or diffraction pattern of the acoustic beam.[7] Using a light beam sufficiently collimated that the diffraction angle of the light is much less than that of the sound, measurement of the scattered light power as a function of the angle of incidence of the light relative to the Bragg angle yields directly the angular distribution of acoustic energy. The angle of incidence is changed by slowly rocking the acoustic medium.

A similar experiment can be performed using the arrangement of Fig. 3. In this case, the component of photo-current at the acoustic frequency measures the *amplitude* of the acoustic angular distribution.

Typical results are shown in Fig. 12. Photograph (a) illustrates the case of a well-collimated light beam with diffraction angle much smaller than that of the acoustic beam, (b) and (c) are for progressively larger optical diffraction angles. In (c) the optical diffraction angle is large enough that the curve illustrates the Gaussian character of the light beam. The deviations from the expected sin X/X behavior relate to the lack of antireflection coatings on the acoustic medium.[7]

Homodyne detection of the photocurrent using the acoustic input



(a)

(b)

(c)

Fig. 12 — Oscilloscope display of the modulation frequency photocurrent vs angle of rotation of the modulating medium (cf., Fig. 3). Traces (a), (b), and (c) show results for increasing values of the optical beam diffraction angle.

signal as a reference would allow determination of the phase of the acoustic angular distribution as well as the amplitude. Thus, this technique has some significant advantages over the original experiments described in Ref. 7. The increased sensitivity of this technique should also be noted.

VI. CONCLUSION

A novel type of acoustic modulator arrangement has been described which allows superheterodyne detection of the diffracted light.

Distinct from electro-optic modulators, the optical modulation sidebands of acoustic devices are well separated in angle from the optical carrier and intensity modulation at the subcarrier modulation frequency is not possible. It has been shown that the optical carrier may be reinserted in the appropriate direction in a simple and convenient fashion allowing intensity modulation at the modulation frequency.

Optimum modulator configurations, corresponding to approximate equality of the optical and acoustic diffraction angles have been derived and the modulation bandwidth has been shown to be proportional to this angle or alternately to be equal to the inverse of the acoustic transit time across the optical beam waist. Confirming experiments in the frequency range 250–350 MHz have been described.

VII. ACKNOWLEDGMENT

The authors are indebted to J. S. Wagner for supplying the thin-film transducers and to D. R. Herriott and M. G. Cohen for informative conversations.

APPENDIX A

Calculation of Photodetector Output

The photodetector is usually placed in the focal plane of a collecting lens. For the purpose of calculation, one may assume that the surface of the photodetector is hemispherical, centered on the interaction volume, and sufficiently large that detection occurs in the optical far-field. The photocurrent is proportional to the instantaneous integrated intensity or power falling on the surface. The component of the power or photo-current at the acoustic frequency is proportional to the interference integral of the transmitted and diffracted optical beams.

The calculation is performed by determining the angular dependence of the transmitted and diffracted beams and integrating the product

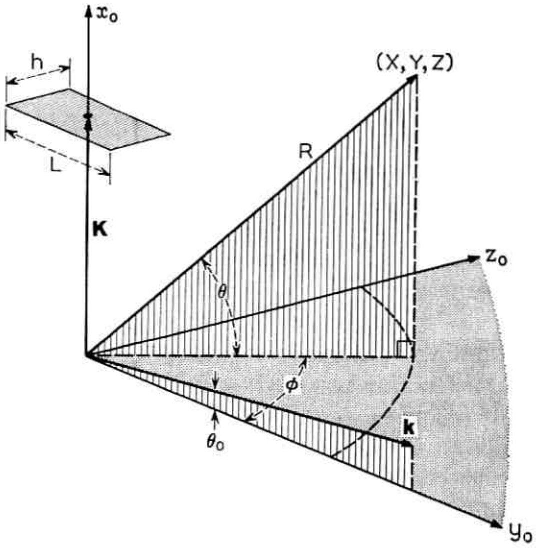


Fig. 13 — Coordinate system used in the calculation of the heterodyne photocurrent.

of the two functions on the surface of a sphere of radius R. The observation point (X, Y, Z) in Fig. 13 has coordinates

$$X = R \sin \theta$$

$$Y = R \cos \theta \cos \varphi$$

$$Z = R \cos \theta \sin \varphi .$$

The near-field amplitude distribution of the transmitted Gaussian beam which is incident in the $x_o - y_o$ plane at angle θ_o is given by

$$\psi(x_o , y_o , z_o , \theta_o)$$
$$= \psi_o \exp - \left[\frac{2 \ln 2}{w_o^2} [(x_o \cos \theta_o - y_o \sin \theta_o)^2 + z_o^2] + ik(y_o \cos \theta_o + x_o \sin \theta_o) \right] , \tag{20}$$

in which w_o is the half-power beam diameter and k is the propagation constant. The amplitude on the sphere is given by

$$\psi_k(X, Y, Z) = \frac{1}{4\pi} \int_{-\infty}^{+\infty} dz_o \int_{-\infty}^{+\infty} dx_o \left[\frac{\exp - ikr}{r} \left(-\frac{\partial \psi}{\partial y_o} \right) + \psi(x_o, y_o, z_o) \frac{\partial}{\partial y_o} \frac{\exp - ikr}{r} \right]_{y_o = 0} \tag{21}$$

$$r = [(X - x_o)^2 + (Y - y_o)^2 + (Z - z_o)^2]^{\frac{1}{2}}$$

corresponding to the usual Green's function solution for an outgoing wave from a source in the x_o, z_o plane. Making the far-field approximations x_o/R, $z_o/R \lll 1$ and $kR \ggg 1$, where applicable, and performing the integration yields for the far-field amplitude

$$\psi_k = \psi_o \frac{ik \exp - ikR}{8R} [\cos \theta_o + \cos \theta \cos \varphi] \left(\frac{w_o^2}{\ln 2 \cos \theta_o} \right) \times \exp - \frac{k^2 w_o^2}{8 \ln 2} [(\cos \theta \sin \varphi)^2 + (\sin \theta_o - \sin \theta)^2 / \cos^2 \theta_a] \tag{22}$$

which can be recognized as the angular dependence of a Gaussian beam including the obliquity factor. The time-dependence is exp $i2\pi\nu t$.

The diffracted beam of frequency $\nu + f$ arises from a perturbation in the optical polarizability of the medium interacting with an optical beam moving at angle $-\theta_o$. The perturbation is proportional to the strain amplitude of the acoustic beam through the appropriate components of the photoelastic tensor. The volume polarization, at frequency $\nu + f$, in the limit of scattering sufficiently weak that the local field amplitude of the incident beam is essentially unchanged by the presence of the acoustic wave, may be written

$$\rho(x_o, y_o, z_o) \exp i2\pi(\nu + f)t = \psi_k(x_o, y_o, z_o, -\theta_o) \exp(i2\pi\nu t) S_o \exp i(2\pi f t - K x_o)$$
$$-\tfrac{1}{2}L \leqq y_o \leqq \tfrac{1}{2}L \qquad -\tfrac{1}{2}h \leqq z_o \leqq \tfrac{1}{2}h \tag{23}$$

and zero otherwise. A rectangular acoustic beam of width L and height h, moving along the x_o-axis, has been assumed. The constant S_o represents the perturbed susceptibility which is proportional to the strain amplitude. The function $\psi_k(x_o, y_o, z_o, -\theta_o)$ is given by (20) with the appropriate change in the sign of θ_o. The diffracted beam amplitude at frequency $\nu + f$ and propagation constant $k' = 2\pi(\nu + f)/c'$ may be written

$$\psi_{k'}(X, Y, Z) = \int_{-\infty}^{+\infty} dx_o \int_{-\frac{1}{2}L}^{+\frac{1}{2}L} dy_o \int_{-\frac{1}{2}h}^{+\frac{1}{2}h} dz_o \rho(x_o, y_o, z_o) \frac{\exp - ik'r}{r} \tag{24}$$

corresponding to the volume Green's function solution for an outgoing wave.

The integral in (24) may be evaluated in a straightforward but lengthy fashion to yield

$$\psi_{k'} = \psi_o S_o \frac{\exp - ik'R}{R} \left(\frac{\pi L}{\beta}\right) \mathrm{ReErf}\left[\tfrac{1}{2}\beta^{\frac{1}{2}}\left(h - \frac{ik' \cos\theta \sin\varphi}{\beta}\right)\right]$$

$$\times \frac{\sin\,(\xi - \eta \tan\,\theta_o)L/2}{(\xi - \eta \tan\,\theta_o)L/2} \exp - \frac{k'^2}{4\beta}\left[\cos^2\,\theta \sin^2\,\varphi + \frac{\eta^2}{k'^2 \cos^2\,\theta_o}\right], \quad (25)$$

in which

$$\beta = \frac{2\ln 2}{w_o^2},$$

$$\eta = k' \sin\theta + k \sin\theta_o - K,$$

$$\xi = k' \cos\theta \cos\varphi - k \cos\theta_o\,,$$

and ReErf(z) is the real part of the error function of complex argument. It may be shown that for the parameters of interest one can make the approximation

$$\mathrm{ReErf}\left[\tfrac{1}{2}\beta^{\frac{1}{2}}\left(h - \frac{ik' \cos\theta \sin\varphi}{\beta}\right)\right] \approx \mathrm{Erf}\,[\tfrac{1}{2}\beta^{\frac{1}{2}}h]$$

with negligible error.

Except for constant factors the photocurrent may be written

$$\iota(f) = R^2 \int_0^{2\pi} d\varphi \int_{-\pi/2}^{\pi/2} d\theta \cos\theta \psi_{k'} \psi_{k'}^* \,. \quad (26)$$

The integral in (26) may be evaluated using (25) and (22). After appropriate manipulations the expression for the photocurrent takes the form

$$\iota(f) = -\frac{i}{8\ln 2}\left(\frac{\pi^3}{\sin\,\theta_o}\right)^{\frac{1}{2}} P_{\mathrm{optical}}(P_{\mathrm{acoustic}})^{\frac{1}{2}} \exp i2\pi f(t - R/c')$$

$$\times \left[\frac{\mathrm{Erf}\,[(\frac{1}{2}\ln 2)^{\frac{1}{2}}h/w_o]}{[(\frac{1}{2}\ln 2)^{\frac{1}{2}}h/w_o]^{\frac{1}{2}}}\right]\left[\frac{\mathrm{Erf}\,[(\ln 2)^{\frac{1}{2}}(L/w_o)\sin\,\theta_o]}{[(\ln 2)^{\frac{1}{2}}(L/w_o)\sin\,\theta_o]^{\frac{1}{2}}}\right] \quad (27)$$

$$\times \exp - \frac{\pi^2(f - f_o)^2 w_o^2}{4\ln 2v^2 \cos^2\,\theta_o},$$

in which

$$f_o = (2v/\lambda)\sin\,\theta_o$$

is the optimum frequency for Bragg diffraction and

$$P_{\text{optical}} = \pi w_o^2 \mid \psi_o \mid^2$$

$$P_{\text{acoustic}} = Lh \mid S_o \mid^2.$$

REFERENCES

1. Forrester, A. T., Gudmundsen, R. A., and Johnson, P. O., Photoelectric Mixing of Incoherent Light, Phys. Rev., *99,* 1955, p. 1691. See also, Forrester, A. T., Mixing as a Spectroscopic Tool, J. Opt. Soc. Amer., *51,* 1961, p. 253.
2. Javan, A., Bennet, Jr., W. R., and Herriott, D. R., Population Inversion and Continuous Optical Maser Oscillation in a Gas Discharge Containing a He-Ne Mixture, Phys. Rev. Letters, *6,* 1961, p. 106. See also, Javan, A., Billik, E. A., and Bond, W. L., Frequency Characteristics of a Continuous-Wave He-Ne Optical Maser, J. Opt. Soc. Amer., *52,* 1962, p. 96.
3. Jennings, D. A. and Takuma, H., Optical Heterodyne Delection of the Forward-Stimulated Brillouin Scattering, Appl. Phys. Letters, *5,* 1964, p. 241; Lastovka, J. B. and Benedek, G. B., *Physics of Quantum Electronics Conference Proceedings,* ed. Kelley, Lax and Tannenwald, McGraw-Hill Book Company, Inc., New York, 1966, p. 231.
4. Cummins, H. Z., Knable, N., and Yeh, Y., Observation of Diffusion Broadening of Rayleigh Scattered Light, Phys. Rev. Letters, *12,* 1964, p. 150.
5. Cummins, H. Z. and Knable, N., Single Sideband Modulation of Coherent Light by Bragg Reflection from Acoustic Waves, Proc. IEEE, *51,* 1963, p. 1246.
6. Goodwin, F. E. and Pedinoff, M. E., Application of CCl_4 and CCl_2:CCl_2 Ultrasonic Modulators to Infrared Optical Heterodyne Experiments, Appl. Phys. Letters, *8,* 1966, p. 60.
7. Cohen, M. G. and Gordon, E. I., Acoustic Beam Probing Using Optical Techniques, B.S.T.J., *44,* May–June, 1965, p. 693.
8. Lucovsky, G., Lasser, M. E., and Emmons, R. B., Coherent Light Deflection in Solid-State Photodiodes, Proc. IEEE, *51,* 1963, p. 166.
9. Kösters, W., Interferenzdoppelprisma Für Messzwecke, Reichspatentamt Patentschrift Nr. 595211, 1931.
10. Saunders, J. B., The Kösters Interferometer, J. Research Nat. Bur. Stand., *58,* 1957, p. 27.
11. Kogelnik, H., Imaging of Optical Modes-Resonators with Internal Lenses, B.S.T.J., *44,* March, 1965, p. 455.
12. Foster, N. F., Cadmium Sulfide Evaporated-Layer Transducers, Proc. IEEE, *53,* 1965, p. 1400.
13. Gordon, E. I. and Cohen, M. G., Electro-Optic Diffraction Grating for Light Beam Modulation and Diffraction, IEEE J. Quantum Electron., *QE-1,* 1965, p. 191. The quantity $\eta^{\frac{1}{2}} \equiv \xi L,$ where ξ is defined in this reference as the incremental phase shift per unit length produced by the acoustic wave.

Water Vapor Permeability of Polyethylene and Other Plastic Materials

By R. L. HAMILTON

(Manuscript received October 6, 1966)

Low-density polyethylene sheathing materials have water vapor permeabilities on the order of 10^{-8} at 22°C. High-density polyethylenes have permeabilities about one-third to one-sixth that of the low-density polyethylene. Copolymers of polyethylene have higher permeabilities than the homopolymer. As an example, 15 percent ethyl acrylate comonomer increased the permeability by a factor of 10 over that of straight low-density polyethylene. The nonolefinic polymers tested have higher permeabilities. For example, polyurethane plug compound has a permeability more than 80 times higher than low-density polyethylene. Finally, it was found that the addition of carbon black decreases thr water vapor permeability roughly in proportion to the amount of carbon black, and that the permeability of these materials increases with increasing porosity.

To make these measurements, two types of laboratory apparatus have been constructed. The first of these makes the permeability measurement on a tubular sample of the material, and the other on films. Both methods used an electrolytic moisture monitor, which is commercially available, to make necessary determinations of water transfer rate through the plastic.

I. INTRODUCTION

The post war years have seen a phenomenal proliferation of plastic materials throughout industry, and the Bell System has been no exception. At least part of the reason for this widespread and ever increasing use is the attractive ease of fabrication of plastic materials and their relatively low cost. Their inertness to certain environmental factors and their chemical and physical stability also add to their value in a great number of applications, including environmental protection. Not only are these plastics used to enclose and isolate an entire apparatus or structure from its environment, they are also used as protective coatings and as seals such as O-rings in metal containers. One of the undesirable and detrimental factors in the environment is water, and consequently, a need developed for information on the water resistant characteristics of the many plastics in use. This water resistant characteristic is the *water vapor permeability* (WVP) of the material, and has been measured for several materials of interest. This paper discusses the need for the measurements, the method of making and correlating measurements, and the significance of the results.

Several problems have been documented involving moisture transfer rates in complicated, composite plastic materials. Water in small amounts, particularly in pulp insulated cable, has detrimental effects on the electrical characteristics of the core. Splicing an Alpeth cable into a paper insulated cable results in moisture accumulation in the latter cable, presumably as a result of moisture diffusion through the Alpeth sheath and subsequent migration into the paper core. Air fed into pulp cable through polyethylene air tubing also results in moisture accumulation in the pulp because of water transmission through the tubing walls.

Most of the design and investigational calculations made in regard to these problems were based on sparse data for more or less idealized systems. Unfortunately, for the materials of interest, there are no moisture transfer rate data available to the engineer. The studies described in this report were motivated by this paucity of data and it was planned to acquire such data for practically pertinent systems for the design engineer. Although the work is directed toward problems arising in the cable plant, the data and results of this study may be of use in other areas employing plastic materials.

In order to correlate and understand the effects of different parameters on permeation, it was necessary to include in the investigation measurements on systems far removed from practically applicable systems. For example, an investigation of the effect of carbon black loading on permeability must include measurements on natural (unloaded) polymers as well as actual sheathing materials which contain carbon black.

The reader will not be subjected to a long and detailed review and analysis of the previous literature in this field. There are two reasons for this; first, recent reviews[1,2] have been given of the literature on permeation processes in plastics. Secondly, the previous literature is principally concerned with purified materials and is not highly pertinent to the problems and information dealt with in this paper. Previous papers[1,3] have indicated that the permeabilities of polyethylenes to water vapor are quite low, on the order of 2×10^{-8} scc*/sec-cm-cm Hg. Also, the previous work has shown that permeability decreases with increasing polymer density.[3] In addition, it has been found that the permeability is an exponential function of reciprocal absolute temperature[3,4] and that the activation energy for permeation ranges from some 6 kilocalories to almost 10 kilocalories.[3,4]

Water vapor permeabilities of other materials have also been given in the literature. It is generally found that the permeabilities of materials such as nylon, cellophane, and other nonolefinic polymers are much higher than permeabilities of polyethylene.[1,2,4]

There is some disagreement[3,4] among previously published values of permeability of polyethylene, and this has been explained[3] in terms of differences in the samples of materials used, but all the previous work agrees on one point: The transport of water vapor through polyethylene obeys Fick's and Henry's laws. Fick's law relates mass transfer rate M to a concentration gradient and in finite difference form is

$$M = DA \frac{\Delta C}{\Delta X}, \tag{1}$$

where D is the diffusion coefficient, A is an area, and ΔC is a concentration difference across an increment in length, ΔX. Henry's law is

$$C = Sp, \tag{2}$$

where S is solubility and p is the vapor pressure of penetrant. If Henry's law and Fick's law are combined

$$M = DSA \frac{\Delta p}{\Delta X}. \tag{3}$$

Usually,

$$DS \equiv P, \tag{4}$$

the permeability.

These laws are used to correlate moisture diffusion rates by calculating P from the definition and data on transfer rates. These data must include, of course, M, Δp, ΔX. The principal experimental problem is measuring these factors on practically pertinent systems so that P can be calculated.

Essentially, the plastic sample separates two chambers, one of which

* The term scc refers to cubic centimeter of vapor at standard temperature and pressure.

contains water at a known temperature. As this water permeates the sample, it is swept from the other chamber (with a carrier gas) and into a water measuring instrument. From the geometry of the sample, A and ΔX in (3) can be determined and from the water temperature, Δp can be obtained. The problem is in measuring M, the mass flow through the sample. The water measuring instrument must be capable of measuring very small amounts of water: for example, if a plastic tube is 10 cm long, 2.5 cm in diameter with a wall thickness of 2 mm, there will be only 5 μg of water per hour permeating through the wall. Previous workers have indicated the rather large amounts of water that can accumulate in some of the older cable designs but this is in terms of miles of cable and years of time. Obviously, for experimental facility we must use shorter lengths of cable and must be able to work on a much shorter time scale and this, in turn, forces one to work with very minute quantities of water. So there is no inconsistency in the larger amounts of water in the cable plant problems and minute amounts encountered in the laboratory experiments.

II. DESCRIPTION OF EXPERIMENTAL APPARATUS

Moll[2] has described some of the more widely used techniques of measurement of permeability. Most of these methods were not directly applicable to water permeation in the materials of practical engineering interest. One such earlier method uses a thermistor (to detect water) which can become fouled from plasticizers and other volatile additives in sheathing material. The "cup method" is usually used on materials with higher permeabilities. Moreover, both methods are more suited to very thin samples—less than 5 mils and sometimes as low as 1 mil.[4] At these thicknesses, surface imperfections (holes, pits, etc.) can account for a large part of the water transferred across the film. For these reasons, it was felt desirable to construct a new apparatus based on an electrolytic water measuring technique which has proved reliable.[5]

An instrument capable of making the necessary water measurements is available commercially (Consolidated Electrodynamics Corporation). This "moisture monitor" operates as follows: A glass cell is coated with phosphorus pentoxide which is a tenacious absorbent for water. The coat of phosphorus pentoxide is interspersed with platinum electrodes and a carrier gas sweeps the moisture from the test sample and into the cell. As the water is absorbed by the phosphorus pentoxide it is electrolyzed to hydrogen and oxygen and the current necessary for electrolysis is directly related to the amount of water so that the detector is calibrated essentially by the definition of electrical current. The precision of measurement depends on the precision of a microammeter. The detector also has the advantage of being specific for water: other materials (such as antioxidants and light oils) which diffuse through the plastic or out of it do not interfere with the water measurement.

Several factors are involved in deciding on size and form the sample should have. Considerations of accuracy determine the size of the sample; the larger the sample area the greater the rate of water penetration which, in turn, permits more accurate measurement.

The apparatus should be capable of making measurements on tubular samples such as cable sheath and air tubing, for example, but certain test materials are not available in quantities sufficient to extrude tubes and certain others (e.g., polyurethane) cannot be extruded easily so that it was necessary to make measurements on films of these materials. Obviously, two different methods must be used—one for tubes and one for films.

III. TUBULAR APPARATUS

Briefly, the tube apparatus (Fig. 1) consists of one or more tubes of plastic material submerged in a tank of water and connected at one end to the "moisture monitor." Dry carrier gas is forced through the tube to sweep out water which has permeated the tube wall. The rate at which moisture is registered on the moisture monitor gives M in μg/sec and from this and the geometry of the tube P is calculated.

3.1 *Carrier Gas Supply*

The water is swept out of the plastic tube and into the moisture monitor with an inert carrier gas and it is imperative, for accuracy, to have the carrier gas enter the tube dry as possible. The carrier gas used has a water content of approximately three parts per million which is in the order of the water content of the gas in the tube. Thus, it is necessary to remove even this small amount. This operation is accomplished with an electrolytic drier cell as shown in the upper left of Fig. 1. The "wet" carrier gas from the supply cylinder is passed through the cell and the water is absorbed on the phosphorous pentoxide and electrolyzed. This gives a carrier gas *free* of moisture.

3.2 *End Seals*

The ends of the plastic tubes are sealed into the apparatus with "Swagelok" fittings. These fittings were originally designed for metal

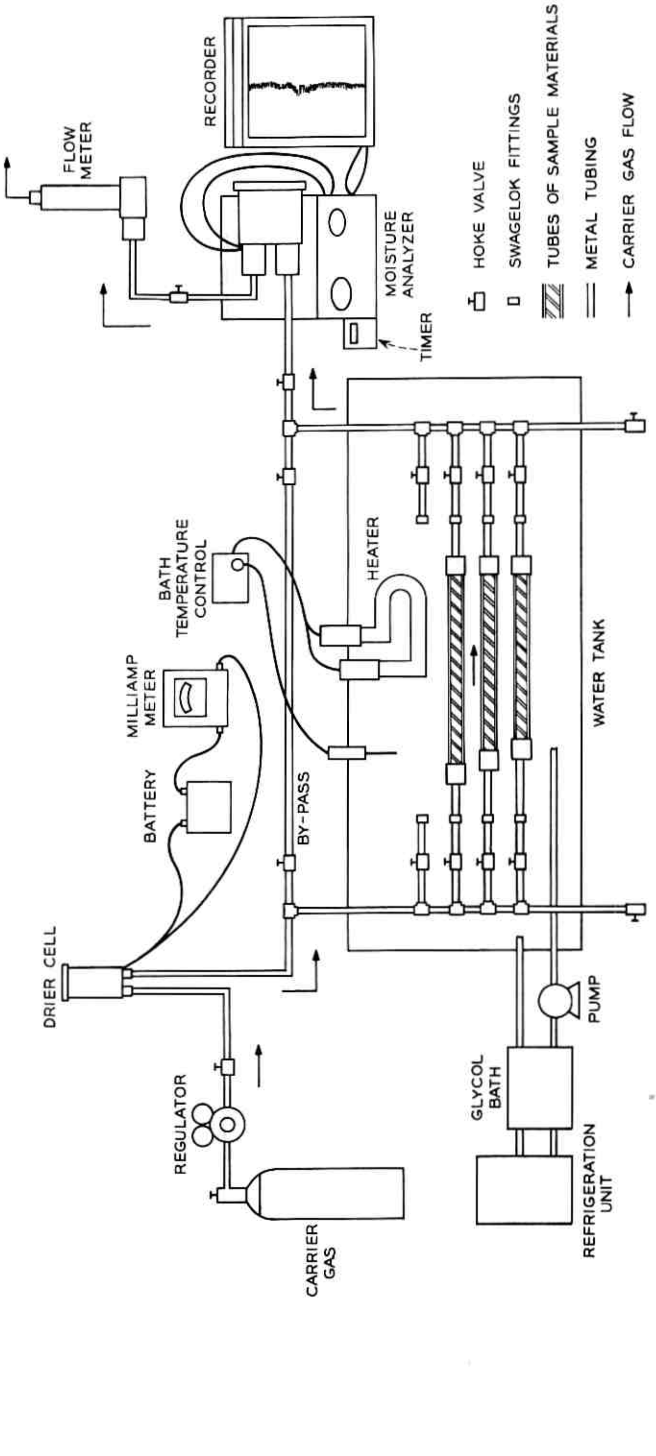


Fig. 1 — Moisture permeability apparatus for tubular samples.

tubing but they were adapted to plastic tubing by inserting a supporting element inside the tubing. These end seals were thoroughly tested and proved very effective as discussed below.

The seal consists of an anvil and chuck and two swaged ferrules supported by an insert inside the plastic tube. When the anvil and chuck are tightened together, the ferrules deform inward against the tubing which is supported by the insert. The ferrules also make a tight metal-to-metal seal against the chuck and anvil, respectively, so there is no path for vapor leakage through the fittings.

3.3 *Fittings and Other Equipment*

Stainless steel tubing is used in parts of the apparatus which could contain water to minimize sorption of water on the tubing walls. In other parts of the apparatus which contain only dry gas, copper tubing is used because it is much easier to fabricate.

The outlet from the moisture monitor is connected to a flow meter to insure that the carrier gas is not flowing so rapidly as to "flood" the electrolytic cell and permit water to escape electrolysis.

3.4 *Temperature Control*

Temperature affects permeability and the temperature of the water must be controlled accurately. This is accomplished with a large, constant temperature heat sink consisting of a glycol bath cooled continuously with a refrigerator. The water is circulated continuously (two gallons per minute) from the sample tank through coils in the cold glycol bath and then back into the tank. This gives a large heat sink and effectively isolates the sample tank temperature from room temperature variations. The temperature of the water is regulated by heating with a temperature controller. Temperature variations can be kept below 0.1°C at 30°C or above. The water in the tank is stirred to keep a uniform temperature throughout. A ten gallon per minute pump, located outside the tank, pumps water from the tank and directly back to the tank through a tube with holes. This effectively stirs and distributes the water and keeps it at a uniform temperature.

3.5 *Measurements*

To calculate P from the measurements on tubes, (4) must be put in cylindrical coordinates. This form of the equation has been given previously[6]

$$P = \frac{M \ln (D_o/D_i)}{2\pi(p_o - p_i)L} ,$$

where L is the sample tube length and D_i and D_o are inside and outside tube diameters. p_i and p_o are water vapor pressures inside and outside the tube. Because the dry carrier gas sweeps water from inside the tube, p_i is small and can be neglected in comparison to p_o.

The inside and outside diameter required care in measurement because the calculated values of P are sensitive to errors in D_o and D_i of the tube samples. These diameters were measured accurately in the case of clear plastic by means of water displacement. A measured length of the sample tube L (usually one meter) was filled with an accurately measured volume V of distilled water. Obviously,

$$V = \frac{\pi D_i^2 L}{4},$$

or

$$D_i = \sqrt{4V/\pi L}.$$

D_i can be calculated from the volume of water V, contained in the length of tube, L. The thickness of the tube wall was measured by encapsulating a short segment of tube in epoxy resin and polishing the mount down until the tube cross section was exposed. The wall thickness was then measured with a stage micrometer. D_o can be obtained from D_i and the wall thickness. The outside diameters of opaque tubes are measured by inserting a sample tube (sealed at one end) into a graduated glass column and measuring the volume ΔV, of water displaced by a given tube length, L. The outside diameter, D_o can then be calculated from

$$D_o = \sqrt{4\Delta V/\pi L}.$$

The inside diameter, $D_i = D_o - 2 \times$ wall thickness.

3.6 *Preliminary Measurements and Tests of Apparatus*

Initial measurements were made to prove the feasibility of the apparatus. It was necessary to check the seals at each end of the tubes to assure that they were watertight. This was done as follows: The permeability was measured for samples of different lengths (67.95 cm, 113.64 cm, and 154.31 cm) of the same material. If the end seals are secure, there should be no difference in permeability for these three samples because P is a property of the plastic and should not depend on the experiment. Fig. 2 gives the results of this test and the good agreement in P for three samples indicates the integrity of the end seals. These tests were made on low density ($\rho = 0.917$) polyethylene tubing supplied by Hydrawlik, Inc., Roselle, New Jersey.

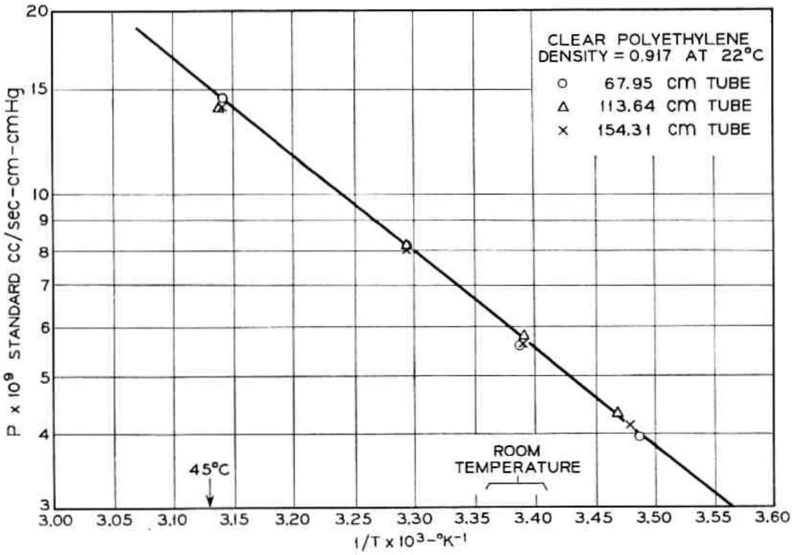


Fig. 2 — Permeability as a function of temperature for various lengths of sample.

IV. FILM APPARATUS

The film apparatus (Fig. 3) consists essentially of two aluminum discs; one on each side of the circular test film. The two discs are identical so that it is necessary to describe only one of them. Two cavities are machined in each disc: one of these is covered with an aluminum plate to form a reservoir for temperature controlled water; the other cavity is adjacent to the test film when the apparatus is assembled. The test film is clamped between the two discs with 6 bolts (see Fig. 3). Water, which is to permeate the test film, is put in the cavity immediately below the film; electrolytically dried N_2 is passed over the film to expel the permeated water. The "wet" gas is then run through the moisture monitor and the rate of water permeation is measured. From this permeation rate, and the thickness and the diameter of the film, the permeability can be calculated.

The diameter of the film exposed to moisture transfer is 5.75 inches. Several considerations influence this dimension of the cell. The first is precision of measurement of water vapor permeation rate. If the diameter of the film is too small, the rate at which water permeates the film will be low and the moisture monitor will be unable to measure accurately the permeation rate. On the other hand, if the diameter of the cell is large (greater than 8 or 10 inches), it becomes difficult to

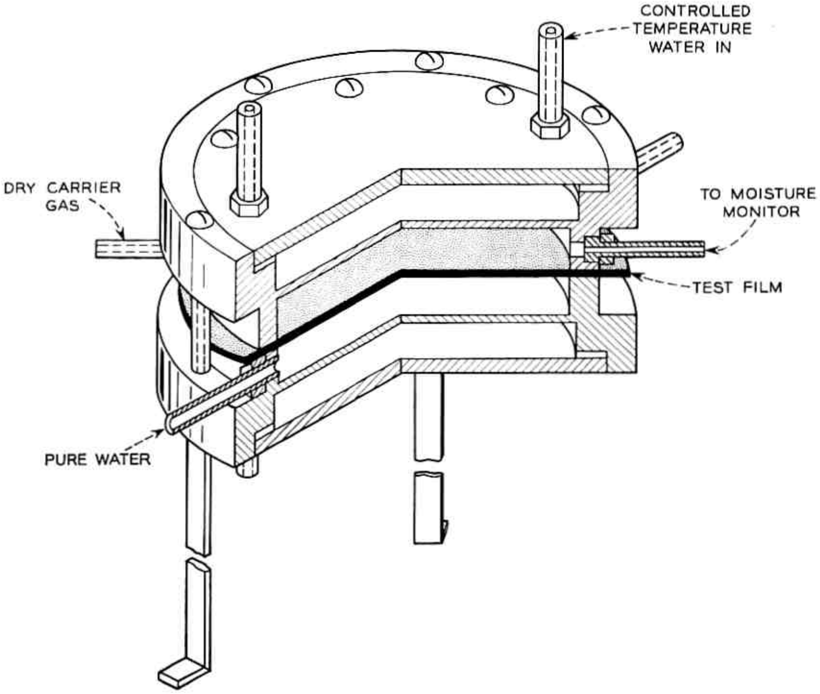


Fig. 3 — Cut-away view of film cell.

make films of uniform thickness. Moreover, large diameter discs are difficult to machine accurately and hard to handle in assembling the cell. The choice of diameter is a compromise between these factors. Consideration of the sensitivity of the moisture monitor and the other factors noted above indicated a 5 to 6-inch diameter film would be optimum. A film of this diameter can be pressed to the necessary uniformity and a cell of this diameter would not be difficult to machine or handle in the laboratory.

The films were made by compression molding and their thickness was measured in two ways. First, the thickness was measured at several points with a micrometer and averaged. In the second method, a circle, $6\frac{1}{2}$ inches in diameter, was cut from the test film and weighed on an analytical balance. The density of this material was then measured with use of density gradient columns. The volume of this circular sample would then be given by the weight, w, of the $6\frac{1}{2}$-inch circle divided by the density, ρ, of the material

$$\text{Vol} = \frac{w}{\rho},$$

and this volume would also be given by the relation

$$\text{Vol} = \frac{\pi}{4} D^2 t$$

(where D is diameter of the circle and t its thickness) so that by combining these two equations and eliminating volume the thickness of the sample can be calculated. In one case, the first method (micrometer) gave an average thickness of 11.9 mils, and the second method a thickness of 11.85 mils. This agreement indicates that the micrometer measurements, which are uncomplicated and much more convenient, would be as reliable and accurate as necessary.

4.1 *Tests of Film Apparatus*

Tests were made to determine the effect of film thickness on measured permeability. If the films are extremely thin, say 1 mil or so, surface imperfections such as pits could contribute substantially to the total moisture transfer rate through the film. These tests were carried out with two materials—a sample of low-density polyethylene (10- and 12-mil films) and a sample of high density polyethylene (7- and 13-mil films). The comparison of the two film thicknesses for each material respectively is shown in Fig. 4 and indicates that the permeability is not dependent on film thickness in the range of film thicknesses (9 to 12 mils) used in these experiments.

4.2 *Additional Tests of Apparatus—Comparison of Films and Tubes*

In Fig. 5 the measured permeability for tubes of the polyethylene materials are compared with the data obtained on films. The permeability of the polyethylene films is about 4 percent lower than for tubes of the same material, while the permeability of the PE-Butyl copolymer film is about 10 to 12 percent lower than that for the tube of this material. These differences are probably due to the differences in processing since the tubes were extruded and essentially quenched, while the films were compression molded and cooled more slowly (about 5°C per minute). The slower cooling anneals the films and they become more highly ordered. Because the permeability of the highly ordered regions is less than the amorphous portion, the permeability of the entire film would be less.

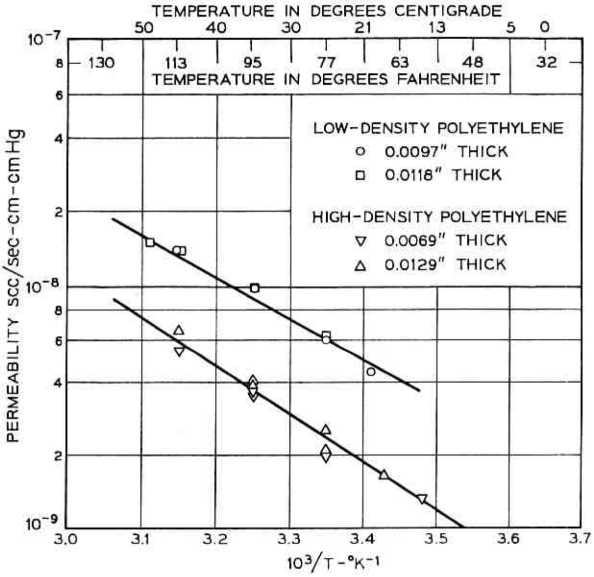


Fig. 4 — Measured permeability is not a function of film thickness.

V. RESULTS AND DISCUSSION

5.1 *A Few Words About Organization*

The discussion will be divided into two categories: the effects of chemical composition and of physical parameters. This is an historical classification rather than a logical one. In this study it was first desirable to measure WVP's of cable sheathing materials, both those in use and proposed. These were all basically polyethylenes but some contained varying contents of comonomers, *viz.* vinyl acetate, ethyl acrylate, and acrylic acid. Polymers of other chemical types (polyester, polyvinyl chloride, *et al.*) are also of interest because they, as sealing and plugging materials, help bear the brunt of nature's attack on the outside plant segment of the Bell System.

As a consequence of this work, it was found that physical parameters such as porosity and other heterogeneities influence WVP and will be discussed last.

5.2 *Effect of Chemical Nature*

Fig. 6 contains the WVP's of representative samples of polyolefins to give an overall picture of these materials and to orient the reader. The point to be noticed in Fig. 6 is the greater WVP's of low-density polyethylene materials compared to those of high density. The ratio of WVP's of high- and low-density polyethylene ranges from $\frac{1}{3}$ to $\frac{1}{6}$.

The water vapor permeability of low-density polyethylene has been reported earlier[4,6] as 2×10^{-8}, whereas the values found here for 0.92 density polyethylene were on the order of 10^{-8}. This difference is most probably due to the differences in the types of polyethylene used and the differences in the thickness of the film samples used. In the previous measurements, the sample films were on the order of 1 to 2 mils so that surface irregularities such as pits could contribute substantially to the overall moisture transfer rate across the film. In the present measurements, the sample films were on the order of 10 to 12 mils thick, and at these thicknesses surface irregularities, if they are present, would not contribute substantially to the moisture transfer rate.

The WVP's of cable sheathing materials (shown in Fig. 7) were of the order of 10^{-8} scc/sec-cm-cm Hg. The homopolymers, in general, have lower WVP's than either copolymers or materials containing low molecular weight additives. This greater WVP is probably due to several factors. Copolymers and additives, even in small amounts, can affect molecular and morphological factors such as branching in the polymer chain, molecular weight, and crystallinity among others. Because all these factors influence permeability, it is not surprising

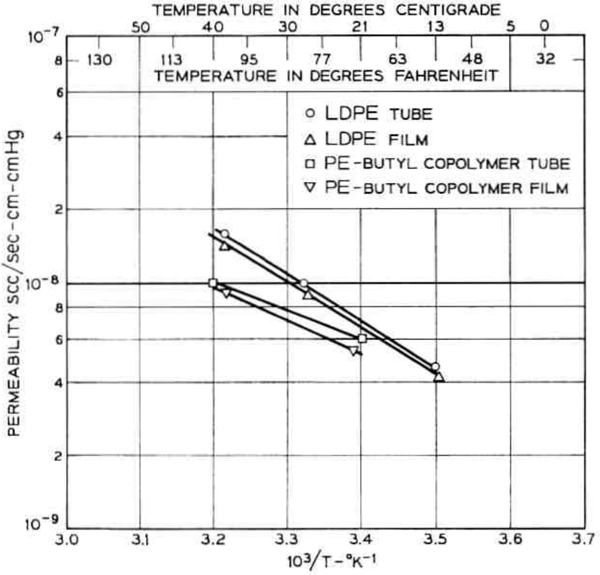


Fig. 5 — Permeabilities of films are slightly lower than tubes.

that copolymers and other additives should cause an increase.

The WVP of polypropylene is intermediate to that of high- and low-density polyethylenes (Fig. 8). This is of interest for two reasons, one of them of practical interest, the other academic. In the first place, polypropylene and low-density polyethylene compete for use in some protective applications. If other factors are equal, advantage should be taken of this lower permeability of polypropylene.

A second point of interest in the findings of polypropylene permeability concerns the effect of density. In the case of polyethylenes, increasing the density decreases the permeability. However, this obviously does not hold true for polyolefins as a class because polypropylene has a density of 0.91 and yet has a lower permeability than low-density (0.92) polyethylene. The lower permeability of polypropylene can be attributed in part to its high crystallinity (about 80 percent). Crystalline regions are much less permeable[3,4] than the noncrystalline regions, hence, the material as a whole has a lower permeability. Moreover, it would be expected that the intermolecular friction would be higher in the case of polypropylene which would, in turn, decrease the diffusion rates through the amorphous regions of the polymer.

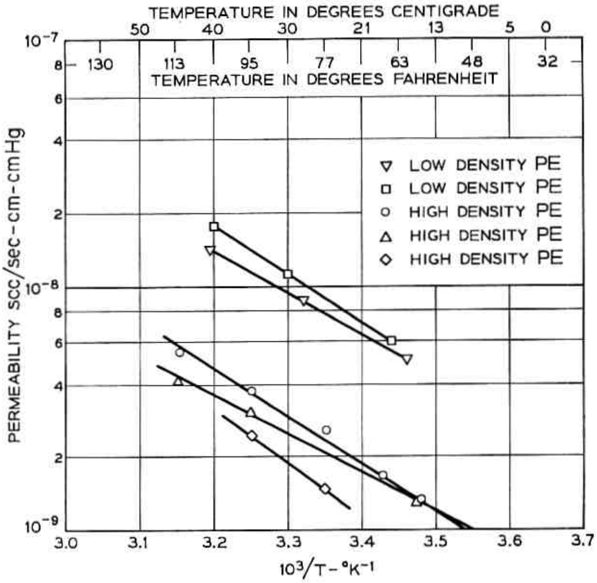


Fig. 6 — Comparison of high-density with low-density polyethylene.

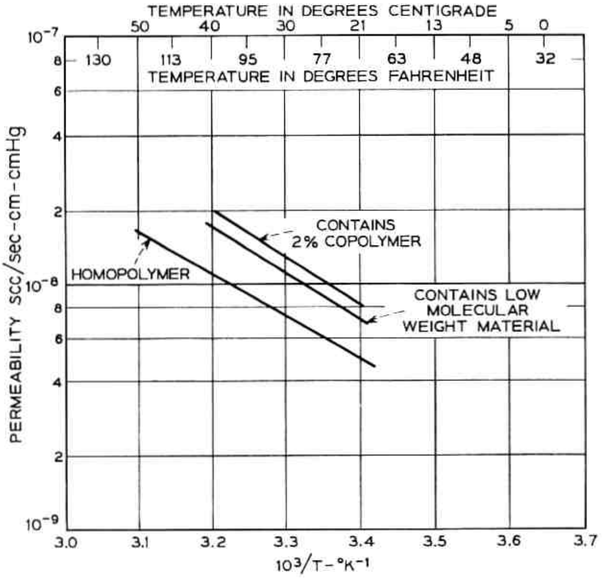


Fig. 7 — Typical cable sheathing materials.

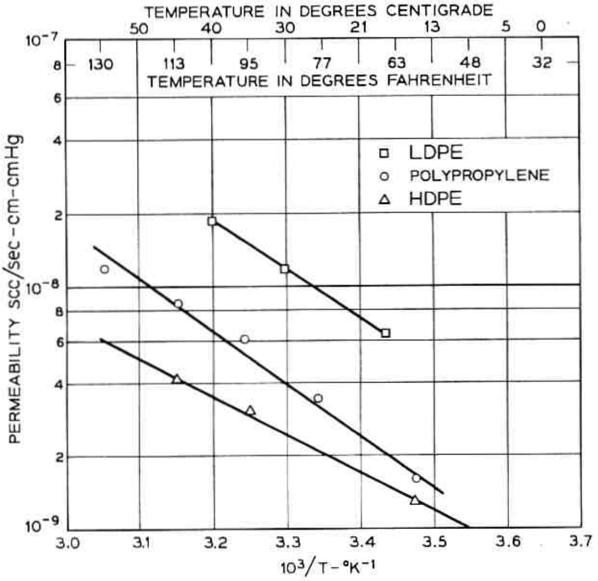


Fig. 8 — Comparison of polypropylene with typical high- and low-density polyethylenes.

Finally, the measurements of permeability of polypropylene bring home another point. In searching for materials for use in environmental protection, we would like to find an inexpensive material with an extremely low water vapor permeability, and if possible, we would like to say that this material is impermeable to water vapor. We must be brought up short in this search, however, in view of the findings in the case of polypropylene. Here is a material that is about 80 percent crystalline so that only 20 percent of the polypropylene contributes to moisture transfer, and even this 20 percent gives the polymer an overall permeability greater than some of the higher density polyethylenes. It is not likely that we could obtain a flexible usable olefinic material with any higher degree of crystallinity, and in this direction polypropylene represents a limit in lowering the water vapor permeability of materials by increasing crystallinity.

5.3 *Copolymers*

Interest in copolymers of olefins stems from two areas. In the first place, with even small percentages of copolymers such as vinyl acetate or ethyl acrylate the polymer is much more flexible and less susceptible to mechanical failure from stress. In the second application, these copolymers are used to compound semiconducting materials by loading them with up to 40 percent carbon black.

These copolymer materials have higher permeabilities than the straight polyethylene homopolymers. The permeabilities of the acrylic acid copolymers are shown in Fig. 9 as a function of acrylic acid copoly-

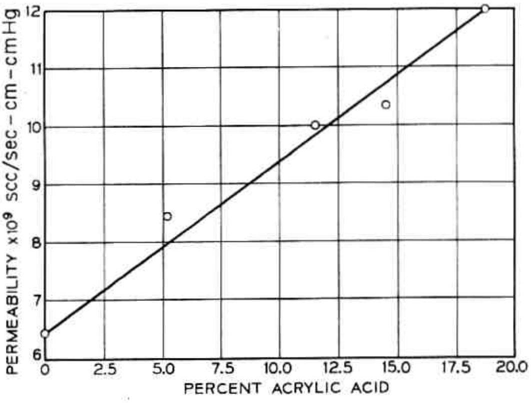


Fig. 9 — Increasing acrylic acid comonomer content increases the permeability of polyethylene.

mer content. These data indicate that the permeability increases with added acrylic acid in the polymer chain. The increase in permeability with the addition of these comonomers is due to morphological factors: the comonomer interrupts the regularity of the polymer chain and thereby reduces the *crystalline content* of the polymer and increases the amorphous content. The amorphous, disordered regions are more permeable; hence, the permeability of the material is increased. In the case of ethyl acetate copolymers, the permeability increased by almost a factor of ten with incorporation of 15 percent copolymer.

5.4 *Oxidized Polymers*

The oxidized materials were made by atmospheric oxidation of unstabilized homopolymer. After oxidation and before the film samples were made the materials were stabilized with Santonox.

The permeability of these materials (Fig. 10) indicates that the introduction of polar groups, such as carbonyl groups in this case, does not *necessarily* cause an *increase* in permeability. Other factors can play a part. Table I compares permeability, density, and carbonyl content of these three materials.

As the carbonyl content goes up the density also increases, and the permeability decreases. Winslow[7] has shown that this increase in density is due to an increase in molecular order in the polymer and because the more highly ordered regions are less permeable, the denser material would have the lower permeability. This reduction in permeability with increased carbonyl content is of interest in cable sheath-

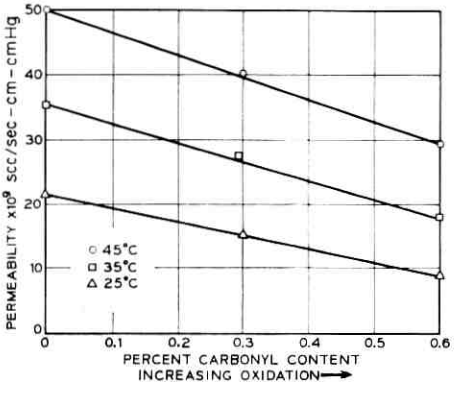


Fig. 10 — Increasing oxidation decreases permeability of polyethylene.

TABLE I—EFFECT OF OXIDATION ON DENSITY AND PERMEABILITY AT 25°C

Percent carbonyl	Density	Permeability
0	0.944	$2.20(10)^{-9}$
0.3	0.956	$1.57(10)^{-9}$
0.6	0.964	$0.82(10)^{-9}$

ing applications for obvious reasons and more materials of this type will be obtained to investigate this effect in more detail.

5.5 *Nonolefinic Polymers*

The results on polyester are compared with low density polyethylene in Fig. 11. The permeability of polyester films was reported[4] several years ago as $1.3(10)^{-8}$ at 25°C. The value measured with the present

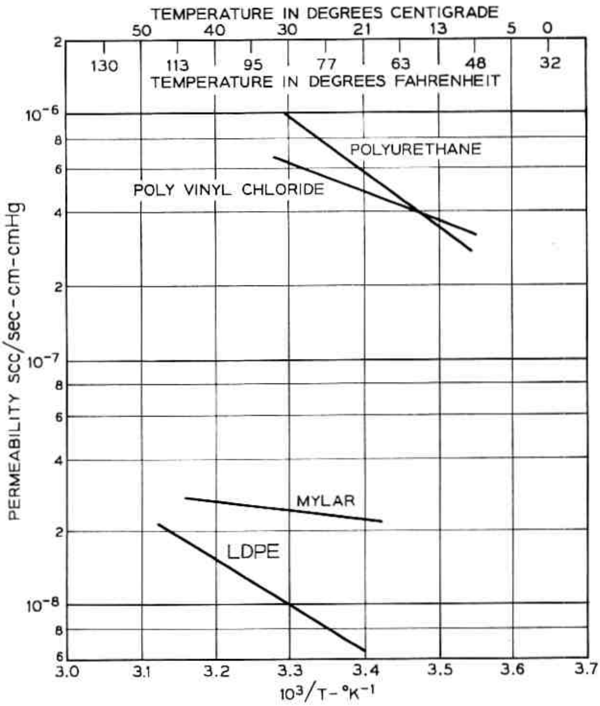


Fig. 11 — Comparison of nonolefinic polymers with a typical low-density polyethylene.

apparatus on polyester commercially available today was $2.3(10)^{-8}$ at 25°C. The older sample of polyester was unplasticized whereas that used in the present study was plasticized slightly and because of these differences a direct comparison is not entirely valid.

The permeabilities of plasticized poly(vinyl chloride) and polyurethane are also shown in Fig. 11; both are over an order of magnitude greater than typical low density polyethylene.

VI. EFFECT OF PHYSICAL PARAMETERS

One is not likely to find pure, homogeneous plastic materials used in the telephone plant. In addition to the deliberately added and necessary heterogeneities such as carbon black, these materials have inadvertent imperfections such as pores or solid particles. To give an example a cross-section of polyurethane cable plugging compound is shown in Fig. 12. The black areas represent pores formed in polymerization and indicate a porosity of up to 20 percent by volume although, superficially, the material usually appears to be homogeneous. Another example of porous material is the foamed polyethylene dielectric used in some coaxial cables.

6.1 *Porosity*

The effect of pores on permeability might not be straightforward: permeability involves both diffusion and solubility and although diffusion would be expected to increase with porosity, solubility decreases,

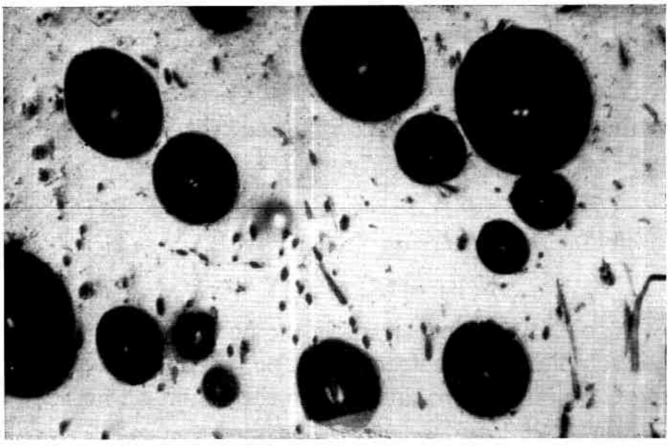


Fig. 12 — Photomicrograph (50×) showing pores in polyurethane.

because water is more soluble in the polymers of interest than in air or gas in the pore. First, the effect of pores on P will be calculated from a model for transport properties of heterogeneous, two-phase materials. These will then be compared with measurements of the effect with foamed silicone rubber.

If we can calculate the effect of pores on diffusion coefficient D and on solubility S, we can calculate the effect on P because

$$P(V) \equiv D(V)S(V) \tag{5}$$

where V is 'porosity' or volume fraction pores.

In calculating $D(V)$, use is made of models developed originally by Maxwell[8] in connection with the electrical transport properties of heterogeneous materials. With the analogies between Ohm's law and Fick's law, Maxwell's model can be used to calculate the diffusion coefficient of plastics containing small amounts of gas-filled cavities or pores. From Maxwell's work, the diffusion coefficient of a porous polymer is given by

$$D = \frac{D_p[D_a + 2D_p - 2V(D_p - D_a)]}{D_a + 2D_p + V(D_p - D_a)}, \tag{6}$$

where D_p and D_a are, respectively, the diffusion coefficients of the penetrant in polymer and in air (or the gas in the pore). It should be understood that the model from which (6) is derived ignores the interactions between adjacent pores so that for the case of porous plastics the value of D would be somewhat low. For polyethylene

$$D_p \cong 10^{-8} \text{ cm}^2/\text{sec} \tag{7}$$

and for air

$$D_a \cong 10^{-1} \text{ cm}^2/\text{sec}. \tag{8}$$

Neglecting D_p in comparison to D_a, Maxwell's relation reduces to

$$D(V) = \frac{D_p(1 + 2V)}{1 - V}. \tag{9}$$

The solubility S(V) decreases linearly with V:

$$S(V) = (S_a - S_p)V + S_p, \tag{10}$$

where S_a is solubility of penetrant in the pore or in the gas within the pore; S_p is solubility of penetrant in the polymer.

Putting these last two equations into the definition for P

$$P(V) = [(S_a - S_p)V + S_p]\left[D_p \frac{1 + 2V}{1 - V}\right]. \tag{11}$$

For the case of silicone rubber foam $S_p \gg S_a$ and (11) becomes

$$P(V) = [(1 - V)S_p]\left[D_p \frac{(1 + 2V)}{(1 - V)}\right]. \tag{12}$$

The factors $(1-V)$ cancel so that P is proportional to V and a plot of $P(V)/D_pS_p$ *vs* V will have a slope of 2 and intercept ($V = 0$) of unity. Such a plot is shown in Fig. 13 for foamed silicone rubber and as anticipated the measured WVP's are somewhat greater than those calculated from the model.

As Maxwell noted:[8]

> "When the distance between the [pores] is not great compared with their radii . . . other terms enter into the result, which we shall not now consider."

For the time being, we will follow his 70 year old cue and use the result in a qualitative manner only. Pore interaction in moderately and in highly foamed materials and its effect on diffusion is the subject of continuing research and will be reported on in a subsequent paper.

6.2 *Effect of Carbon Black*

Two magnitudes of carbon black content were investigated. Those used in cable sheath materials (about 2.5 percent by weight) and those higher contents (up to 40 percent) proposed for semiconducting sheathing materials. Fig. 14 shows some representative data for materials

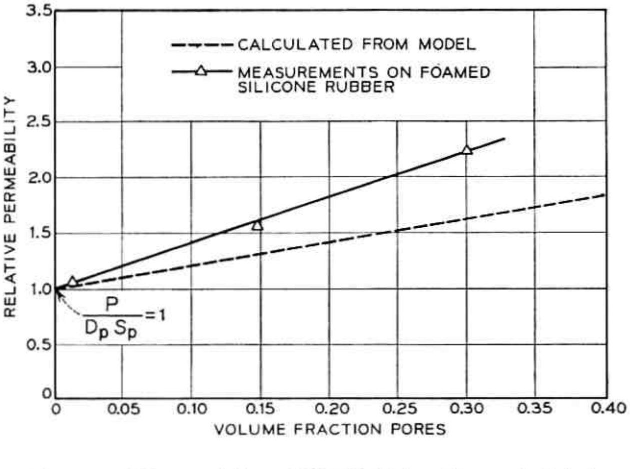


Fig. 13 — Measured foam WVP is higher than calculated.

containing about 2.5 percent carbon black compared with the natural materials. The carbon black at these contents lowered the permeability only slightly. Note also, that at these low percentages, the temperature dependence of the filled and unfilled material are essentially identical.

Materials highly loaded with carbon black also showed a linear decrease in permeability with increasing carbon black content. Fig. 15 gives some representative data. Qualitatively, this decrease in permeability with increasing carbon black content is not difficult to explain. The carbon black particles are most probably impermeable to moisture; hence, their presence in the polymer decreases the volume available to moisture diffusion. Again, we can use Maxwell's model discussed earlier to describe this decrease in permeability with increasing carbon black content. For the case of a polymer interspersed with *impermeable* particles, Maxwell's equation for the diffusion coefficient is given by

$$D = \frac{2D_p(1 - V_{cb})}{2 + V_{cb}}, \tag{13}$$

where D_p is the diffusion coefficient of water vapor in the polymer and V_{cb} is the volume fraction of carbon black.

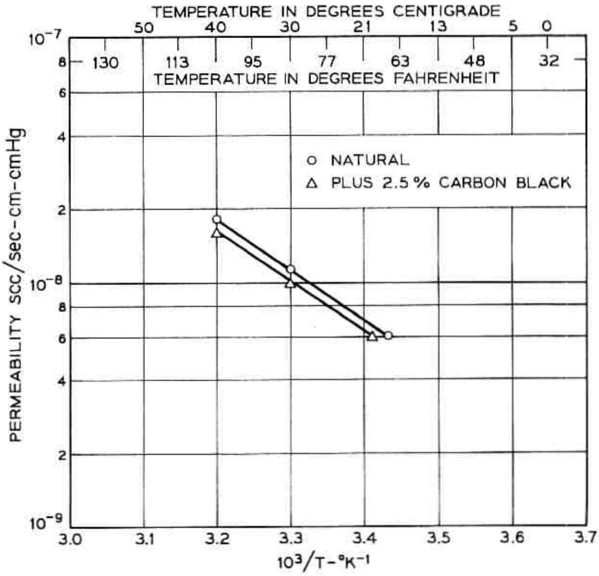


Fig. 14 — Effect of 2.5 percent carbon black.

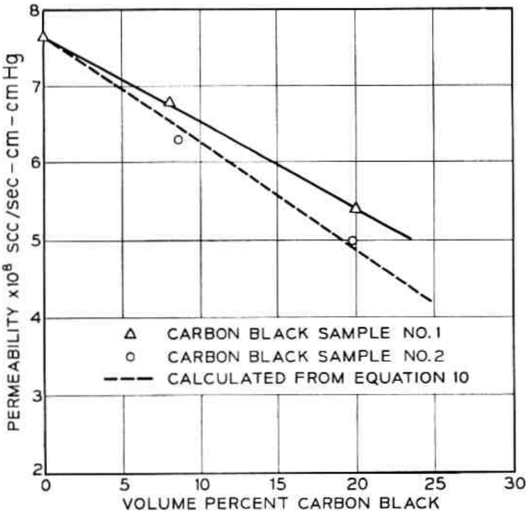


Fig. 15 — Effect of carbon black on permeability of poly (ethylene ethyl acrylate) copolymer.

The solubility of the loaded polymers decreases linearly with carbon black content

$$S(V_{cb}) = (1 - V_{cb})S_p. \tag{14}$$

Combining the equation for D and S with the definition of P gives

$$P(V_{cb}) = \frac{2D_pS_p(1 - V_{cb})^2}{2 + V_{cb}}. \tag{15}$$

This equation is shown as a dotted line in Fig. 15 and is slightly lower than the experimental values for WVP. In the equation for $S(V_{cb})$ above, it was assumed that the solubility of water on the carbon black was negligible, but as shown below there is some interaction and to be precise, we would be justified in assigning some contribution to $S(V_{cb})$ due to the carbon black. This would increase the calculated values of P to agree more closely with the data.

At higher carbon black contents (10 percent and above) the temperature dependence of filled and unfilled materials is markedly different. Fig. 16 gives an example of this behavior and shows that the activation energy for permeation, E_p, is decreased by carbon black. The decrease in E_p for the case of water sorption in polyethylene

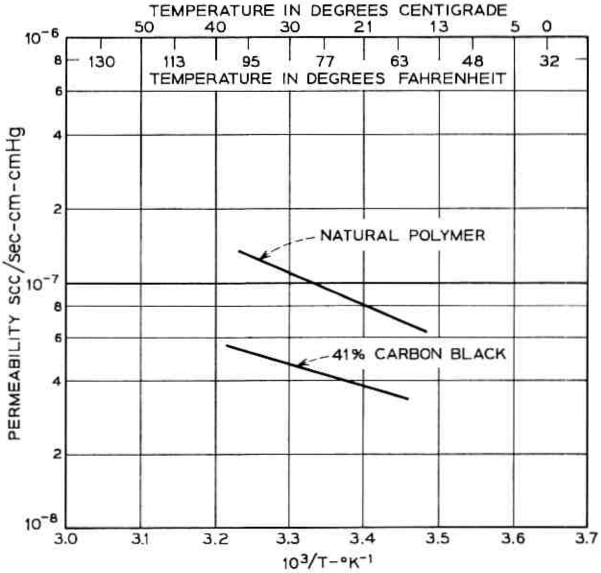


Fig. 16 — Effect of adding 41 percent carbon black to poly (ethylene ethyl acrylate) copolymer.

highly loaded with carbon black is not unexpected. For the water-polyethylene system,

$$D = D_o \exp(-E_D/RT), \tag{16}$$

$$S = S_o \exp(-\Delta H/RT), \tag{17}$$

and

$$P = P_o \exp(-E_p/RT), \tag{18}$$

where D_o, S_o, and P_o are constants; E_D and E_p are activation energies for diffusion and permeation, respectively, and ΔH is the heat of sorption. Substituting these equations into the definition of P (4) gives

$$P_o \exp(-E_p/RT) = D_oS_o \exp[-(E_D + \Delta H)/RT]. \tag{19}$$

From this

$$E_p = E_D + \Delta H. \tag{20}$$

For the case of water sorption on carbon black, ΔH is negative and increasing the carbon black to 41 percent gives a greater negative contribution and E_p is therefore decreased.

VII. SUMMARY

In general, the low-density polyethylene sheathing materials have permeabilities on the order of 10^{-8} scc/sec-cm-cm Hg at 22°C. High-density polyethylenes have permeabilities from 1/3 to 1/6 that of low-density polyethylene. The permeability of polypropylene is intermediate to high- and low-density polyethylenes. Copolymers of polyethylene (for example, vinyl acetate, ethyl acrylate, and acrylic acid) have higher permeabilities than the homopolymer; in one case 15 percent ethyl acrylate increased the permeability by a factor 10.

Nonolefinic polymers, in general, have higher permeabilities; for example, polyurethanes have permeabilities more than 80 times higher than low-density polyethylene.

Heterogeneities in the plastic such as carbon black and pores influence permeability. Addition of carbon black decreases the WVP roughly in proportion to the amount of carbon black. The permeability of these materials increases with increasing porosity.

VIII. ACKNOWLEDGMENT

Finally, there should come a long list of people to whom the author is indebted for many helpful suggestions, comments, and criticism. In particular, J. P. McCann and J. J. Lamont helped make the measurements. Discussions with Messrs. Windeler, Biskeborn, and Howard were helpful and Mr. Connolly and Mr. Snoke listened patiently to the author's appeals for more time and equipment with which to do the work. From outside the Bell Laboratories, I must acknowledge my debt to Mr. George Brown of Hydrawlic, Incorporated of Roselle, New Jersey, and Mr. Bluestein of Cooke Color and Chemical Company of Hackettstown, New Jersey, both of whom helped obtain some of the sample materials.

REFERENCES

1. Li, N. N., *et al.,* Ind. Engr. Chem., *57,* 1965, pp. 18–29.
2. Moll, W. L. H., Kolloid Z., *195,* 1964, pp. 43–52.
3. Klinte, C. H. and Franklin, P. J., J. Polymer Science, *32,* 1958, pp. 161–176.
4. Stannett, V., *et al.,* Permeability of Plastic Films and Coated Paper to Gases and Vapors, Tappi Monograph #23; Technical Association of the Pulp and Paper Industry, New York, 1962.
5. Toren, P. E., Anal. Chem., *37,* 1965, pp. 922–923.
6. Ferrari, A. G., Modern Plastics, *41,* 1964, p. 153.
7. Winslow, F. H., *et al.,* Trans. New York Academy of Sciences, Ser. II, *28,* December, 1965, pp. 304–315.
8. Maxwell, James C., *Treatise on Electricity and Magnetism,* Vol. 1, Third Edition, Oxford, 1892, pp. 440–441.

Statistical Analysis of the Level Crossings and Duration of Fades of the Signal from an Energy Density Mobile Radio Antenna

By W. C.-Y. LEE

(Manuscript received September 30, 1966)

A theoretical analysis of signal fading using an energy density antenna is developed and compared with that from an isotropic antenna. The energy density antenna provides a signal proportional to the energy density of the mobile radio field. The number of crossings that the signal makes of a given signal level and the average duration of fades below a given signal level have been derived theoretically for these two cases using a simple statistical model. Comparing the number of level crossings of the electric field with that of the energy density, it is shown that the energy density fades less frequently than the electric field by at least a factor of two. The average duration of fades of the electric field is greater than that of the energy density only for lower signal levels. These results are in reasonable agreement with experimental measurements.

I. INTRODUCTION

The study of signal fading appears to be very important to mobile radio systems. When a steady sine wave is sent out from a fixed station, the signal received by a mobile receiver in motion fluctuates, or, in radio jargon, fades. The received signal fluctuates more rapidly as both the frequency of the transmitted wave and the speed of the mobile radio increase. For a field received by a moving isotropic antenna, the maximum fading frequency f_d, as Ossanna[1] has pointed out, is $f_d = 2V/\lambda$, where V is the speed of the mobile radio and λ is the wavelength of the steady sine wave. For instance, at 836 MHz and a speed of 15 miles/hr, the signal fades at a rate of about 40 times every second and is a serious disturbance to the mobile radio communication.

There have been many investigations of the fading problem. Aikens and Lacy[2] made a test using 450-MHz transmission to a mobile receiver in an urban area. Bullington[3] investigated radio propagation variation at VHF and UHF. Young[4] pointed out that for the test samples of signal strength taken over a small area, the amplitude follows a Rayleigh distribution to a fair approximation. Also S. O. Rice[5] pointed out that the fluctuations of a received radio signal have the same behavior as the envelope of a narrow-band Gaussian noise. Recently, Ossanna[1] measured the power spectra of a mobile radio fading signal. They all treated the signal as obtained from an isotropic antenna.

In this paper, a theoretical analysis of fading using an energy density antenna[6] is developed and compared with that from an isotropic antenna. The concept of using the energy density antenna to reduce the effect of signal fading was suggested by J. R. Pierce.[7] It will be discussed in detail later. The number of crossings $n(\Psi)$ that the signal makes of a given signal level Ψ, and the average duration of fades $\mathbf{t}(\Psi)$ below a given signal level Ψ, have been derived theoretically from a statistical model using Gaussian random amplitudes and equal angles of arrival of an infinite number of incoming waves. The two statistical factors, n and $\mathbf{t}$, first expressed by Rice,[8] can describe the property of individual signal fading very well. In this paper, n and $\mathbf{t}$ for the isotropic antenna will be compared with the values for the energy density antenna. These theoretical results also will be compared with the experimental data.

II. THE METHOD OF OBTAINING THE EXPECTED NUMBER OF LEVEL CROSSINGS AND AVERAGE DURATION OF FADES

From Kac's[9] and Rice's[8] paper, a brief derivation of the expected number of level crossings $n(\Psi)$ of a given signal level Ψ and average duration of fades below a given signal level Ψ is as follows. We assume a random function ψ which is statistically stationary in time, and for which the joint probability density function of ψ and its slope $\dot{\psi}$ is $p(\psi, \dot{\psi})$. Any given slope $\dot{\psi}$ can be obtained by

$$\dot{\psi} = \frac{d\psi}{\tau}, \tag{1}$$

where τ is the time required for a change of ordinate $d\psi$, as shown in Fig. 1. The expected number of crossings of a random function ψ in the interval $(\Psi, \Psi - d\psi)$ for a given slope $\dot{\psi}$ in time dt is

$$\left.\frac{\begin{array}{l}\text{the expected amount of time spent in the interval}\\ \qquad d\psi \text{ for a given } \dot{\psi} \text{ in time } dt\end{array}}{\begin{array}{l}\text{the time required to cross once for a given } \dot{\psi} \text{ in}\\ \qquad \text{the interval } d\psi\end{array}}\right|_{\text{at } \psi=\Psi}$$

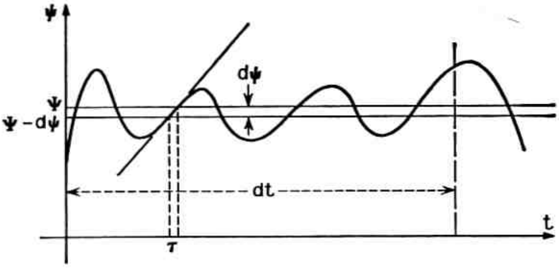


Fig. 1 — The notation used in obtaining the expected number of level crossings $n(\Psi)$ and the average duration of fades $\mathsf{t}(\Psi)$.

$$= \frac{E(t)}{\tau} = \left.\frac{p(\psi,\ \dot{\psi})\,d\psi\,d\dot{\psi}\,dt}{\dfrac{d\psi}{\dot{\psi}}}\right|_{\text{at } \psi=\Psi} = \dot{\psi}p(\Psi,\ \dot{\psi})\,d\dot{\psi}\,dt. \tag{2}$$

The expected number of crossings for a given $\dot{\psi}$ in time T is

$$\int_0^T \dot{\psi}p(\Psi,\ \dot{\psi})\,d\dot{\psi}\,dt = \dot{\psi}p(\Psi,\ \dot{\psi})\,d\dot{\psi}\,T. \tag{3}$$

The total expected number of upward crossings in time T is

$$N(\Psi) = T\int_0^\infty \dot{\psi}p(\Psi,\ \dot{\psi})\,d\dot{\psi}. \tag{4}$$

The total expected number of crossings per second is

$$n(\Psi) = \frac{N(\Psi)}{T} = \int_0^\infty \dot{\psi}p(\Psi,\ \dot{\psi})\,d\dot{\psi}. \tag{5}$$

Since the expected number of crossings at a particular level Ψ per second can also be stated as

$$n(\Psi) = \frac{\begin{array}{c}\text{the expected amount of time where the}\\ \text{function } \psi \text{ is below level } \Psi \text{ in one second}\end{array}}{\text{the average duration of fades below level } \Psi}$$

$$= \frac{P(\psi < \Psi)}{\mathsf{t}(\Psi)}, \tag{6}$$

hence, the average duration of fades below level Ψ is

$$\mathsf{t}(\Psi) = \frac{P(\psi < \Psi)}{n(\Psi)}. \tag{7}$$

Hence, the results will be derived from the joint probability density function $p(\psi, \dot{\psi})$, and the problem is to derive this probability density function for the various signals.

III. THE EXPECTED NUMBER OF LEVEL CROSSINGS AND THE AVERAGE DURATION OF FADES FOR A VERTICALLY POLARIZED WAVE

In order to obtain the expected number of level crossings of a given signal level R and the average duration of fades below a given signal level R for the three field components of a vertically polarized wave, first we need to specify the forms of the three field components. Then a statistical model of the field components is assumed. From such a model, we find the joint probability density functions of amplitude R and its slope $\dot{R}$ for the three field components. Finally, we use (5) and (7) to obtain the result for each field component.

Following Gilbert[10] a vertically polarized plane wave E_z traveling in a direction $\mathbf{u}$ in the (x, y) plane is assumed. The three field components referenced to a receiver moving with velocity vector $\mathbf{V}$ can be written

$$E_z = e_z = A_u \exp(-j\beta\mathbf{u}\cdot\mathbf{V}t) \exp(j\omega t) \text{ volt}/m$$

$$H_x = \eta(h_x \text{ amp}/m) = A_u \sin\theta_u \exp(-j\beta\mathbf{u}\cdot\mathbf{V}t) \exp(j\omega t) \text{ volt}/m$$

$$H_y = \eta(h_y \text{ amp}/m) = -A_u \cos\theta_u \exp(-j\beta\mathbf{u}\cdot\mathbf{V}t) \exp(j\omega t) \text{ volt}/m,$$

where β is a wave number and A_u is a complex amplitude of an electric wave propagating at a direction $\mathbf{u}$. $\mathbf{u}$ is a unit vector related to an angle θ_u between the positive x-axis and the unit vector itself. η is free-space wave impedance. The time variation exp $j\omega t$ can be dropped out of three field components for simplifying the derivation. Moreover, from now on, we will treat the units of all three components E_z, H_x, and H_y in volt/m which will also simplify the calculation.

When N vertical polarized waves coming from N directions are received by an isotropic antenna of the mobile radio, the three components become

$$E_z = \sum_{u=1}^{N} A_u \exp(-j\beta\mathbf{u}\cdot\mathbf{V}t) = \sum_{u=1}^{N} A_u \exp[-j\beta Vt \cos(\theta_u - \alpha)] \quad (8)$$

$$\begin{aligned} H_x &= \sum_{u=1}^{N} A_u \sin\theta_u \exp(-j\beta\mathbf{u}\cdot\mathbf{V}t) \\ &= \sum_{u=1}^{N} A_u \sin\theta_u \exp[-j\beta Vt \cos(\theta_u - \alpha)] \quad (9) \end{aligned}$$

$$H_y = \sum_{u=1}^{N} - A_u \cos \theta_u \exp(-j\beta \mathbf{u} \cdot \mathbf{V}t)$$

$$= -\sum_{u=1}^{N} A_u \cos \theta_u \exp[-j\beta Vt \cos(\theta_u - \alpha)], \quad (10)$$

where θ_u is the angle between the positive x-axis and the direction of uth wave $\mathbf{u}$, and $0 \leqq \theta_u \leqq 2\pi$. α is the angle between the x-axis and the velocity $\mathbf{V}$, and $0 \leqq \alpha \leqq 2\pi$. Both θ_u and α are shown in Fig. 2.

In this paper, a statistical model is used as follows: The complex amplitude A_u can be separated into a real and an imaginary part $A_u = R_u + jS_u$, hence N incoming waves have N real values of R_u and S_u. We suppose all those $2N$ real values are Gaussian independent variables with mean zero and variance one. Also, we assume the N waves have uniform angular distribution, i.e., the kth wave u_k has an angle of arrival $\theta_u = 2\pi k/N$. Moreover, in this paper an infinite number of multiply reflected waves ($N \rightarrow \infty$) are assumed for finding the expected number of level crossings $n(R)$ of a given signal level R, and the average duration $\mathbf{t}(R)$ of fades below a given signal amplitude R.

3.1 *Finding the Values of $n(R)$ and $\mathbf{t}(R)$ from the E_z Field*

First of all, we need to obtain the joint probability density function of signal amplitude R and its slope $\dot{R}$ for the electric field component E_z using the statistical model we mentioned previously. We start from (8). The alternate form of (8) can be written as

$$E_z = \sum_{u=1}^{N} (R_u + jS_u)[\cos\{\beta Vt \cos(\theta_u - \alpha)\} - j \sin\{\beta Vt \cos(\theta_u - \alpha)\}]. \quad (11)$$

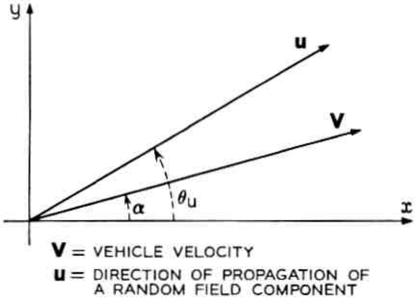


Fig. 2 — The coordinate system.

Equation (11) can be separated into real and imaginary parts

$$E_z = X_1 + jY_1 \,. \tag{12}$$

The real part of E_z is

$$X_1 = \sum_{\mu=1}^{N} (R_u \cos \varphi_u + S_u \sin \varphi_u) \tag{13}$$

and the imaginary part of E_z is

$$Y_1 = \sum_{u=1}^{N} (S_u \cos \varphi_u - R_u \sin \varphi_u), \tag{14}$$

where

$$\varphi_u = \beta V t \cos (\theta_u - \alpha). \tag{15}$$

We assume that all N values of R and S in (13) and (14) are time independent. Then the derivatives with respect to time of (13) and (14) are

$$\dot{X}_1 = \beta V \sum_{u=1}^{N} (-R_u \sin \varphi_u + S_u \cos \varphi_u) \cos (\theta_u - \alpha) \tag{16}$$

$$\dot{Y}_1 = \beta V \sum_{u=1}^{N} (-S_u \sin \varphi_u - R_u \cos \varphi_u) \cos (\theta_u - \alpha). \tag{17}$$

The mean values, variances, and covariances of X_1 , Y_1 , $\dot{X}_1$, and $\dot{Y}_1$ are

$$m_1 = \langle X_1 \rangle = \langle Y_1 \rangle = \langle \dot{X}_1 \rangle = \langle \dot{Y}_1 \rangle = 0$$

$$\mu_{11} = \langle X_1^2 \rangle = \langle Y_1^2 \rangle = N \quad \text{for any} \quad N \tag{18}$$

$$\mu_{11}' = \langle \dot{X}_1^2 \rangle = \langle \dot{Y}_1^2 \rangle = (\beta V)^2 \frac{N}{2} \quad \text{for} \quad N \geqq 3 \tag{19}$$

and

$$\langle X_1 Y_1 \rangle = \langle X_1 \dot{X}_1 \rangle = \langle Y_1 \dot{X}_1 \rangle = \langle Y_1 \dot{Y}_1 \rangle = \langle \dot{X}_1 \dot{Y}_1 \rangle = \langle X_1 \dot{Y}_1 \rangle = 0.$$

The above results are shown in the Appendix.

From the central limit theorem, it follows that X_1 , Y_1 , $\dot{X}_1$, and $\dot{Y}_1$ are four independent random variables which are distributed normally as the value N approaches infinity. The probability density function of four independent real random variables X_1 , Y_1 , $\dot{X}_1$, and $\dot{Y}_1$ is[11]

$$p(X_1 , Y_1 , \dot{X}_1 , \dot{Y}_1) = \frac{1}{(2\pi)^2 \, |\mu|^{\frac{1}{2}}} \exp \left\{ -\frac{1}{2} \left(\frac{X_1^2 + Y_1^2}{\mu_{11}} + \frac{\dot{X}_1^2 + \dot{Y}_1^2}{\mu_{11}'} \right) \right\}, \tag{20}$$

where the determinant $|\mu|$ of the covariance matrix is

$$|\mu| = (\mu_{11}\mu_{11}')^2 = \frac{N^4}{4}(\beta V)^4.$$

We may introduce the concept of the envelope

$$E_s = X_1 + jY_1 = r_e \, \underline{/\eta_e} \, .$$

The quantity r_e is the envelope and η_e is the phase, both of which are slowly varying functions of the time. Then,

$$X_1 = r_e \cos \eta_e \qquad ; \qquad Y_1 = r_e \sin \eta_e \tag{21a}$$

$$\dot{X}_1 = \dot{r}_e \cos \eta_e - r_e \dot{\eta}_e \sin \eta_e \; ; \qquad \dot{Y}_1 = \dot{r}_e \sin \eta_e + r_e \dot{\eta}_e \cos \eta_e \, . \tag{21b}$$

The Jacobian of the transformation from $(X_1, Y_1, \dot{X}_1, \dot{Y}_1)$-space to $(r_e, \eta_e, \dot{r}_e, \dot{\eta}_e)$-space is[12] $|J| = r_e^2$.

Therefore, the change of variables gives the probability density the form

$$\begin{aligned} p(X_1, Y_1, \dot{X}_1, \dot{Y}_1) &= r_e^2 q(r_e, \eta_e, \dot{r}_e, \dot{\eta}_e) \\ &= p(r_e, \eta_e, \dot{r}_e, \dot{\eta}_e) \\ &= \frac{r_e^2}{(2\pi)^2 |\mu|^{\frac{1}{2}}} \exp\left\{-\frac{1}{2}\left(\frac{r_e^2}{\mu_{11}} + \frac{r_e^2\dot{\eta}_e^2 + \dot{r}_e^2}{\mu_{11}'}\right)\right\}, \end{aligned} \tag{22}$$

where $q(r_e, \eta_e, \dot{r}_e, \dot{\eta}_e)$ is the density obtained on substituting for X_1, Y_1, etc., their values in r_e, η_e, etc., obtained from (21a) and its time derivative (21b). To obtain $p(r_e, \dot{r}_e)$, the probability density of the envelope and its rate of change, we must integrate over η_e and $\dot{\eta}_e$, the range of which are, respectively, (0 to 2π) and ($-\infty$ to $+\infty$). From (22) we obtain

$$p(r_e, \dot{r}_e) = \frac{r_e}{\sqrt{2\pi\mu_{11}'\mu_{11}}} \exp\left\{\frac{1}{2}\left(\frac{r_e^2}{\mu_{11}} + \frac{\dot{r}_e^2}{\mu_{11}'}\right)\right\}. \tag{23}$$

It is observed that the expression on the right of (23) is independent of t. Hence, the expected number of level crossings $n(R_e)$ at a given signal amplitude ($r_e = R_e$) can be obtained from (5) by using $p(R_e, \dot{r}_e)$ in (23).

$$n(R_e) = \int_0^\infty \dot{r}_e p(R_e, \dot{r}_e) d\dot{r}_e = \sqrt{\frac{\mu_{11}'}{\mu_{11}}} \frac{R_e}{\sqrt{2\pi\mu_{11}}} \exp\left(-\frac{R_e^2}{2\mu_{11}}\right). \tag{24}$$

Now the variance of r_e is

$$\langle r_e^2 \rangle = \langle X_1^2 \rangle + \langle Y_1^2 \rangle = 2\mu_{11} = 2N.$$

Let

$$\tilde{R}_e = R_e/\sqrt{\langle r_e^2\rangle} = R_e/r_{e(rms)} = R_e/\sqrt{2N}. \qquad (25)$$

Substituting the values of variances μ_{11} and μ_{11}' from (18) and (19) into (24), also applying the relations in (25), we obtain

$$n(\tilde{R}_e) = \frac{\beta V}{\sqrt{2\pi}} \tilde{R}_e \exp(-\tilde{R}_e^2). \qquad (26)$$

Equation (26) is plotted in Fig. 3 where the abscissa is $\tilde{R}_e$ in dB (20 log $\tilde{R}_e$) and the ordinate is $(\sqrt{2\pi}/\beta V)n(\tilde{R}_e)$.

The average duration of fades $\mathbf{t}(R_e)$ of E_z can be obtained as follows: The probability that the envelope r_e is less than a given amplitude level R_e is

$$P(r_e(t) < R_e) = \int_0^{R_e} p(r_e(t))dr_e(t), \qquad (27,$$

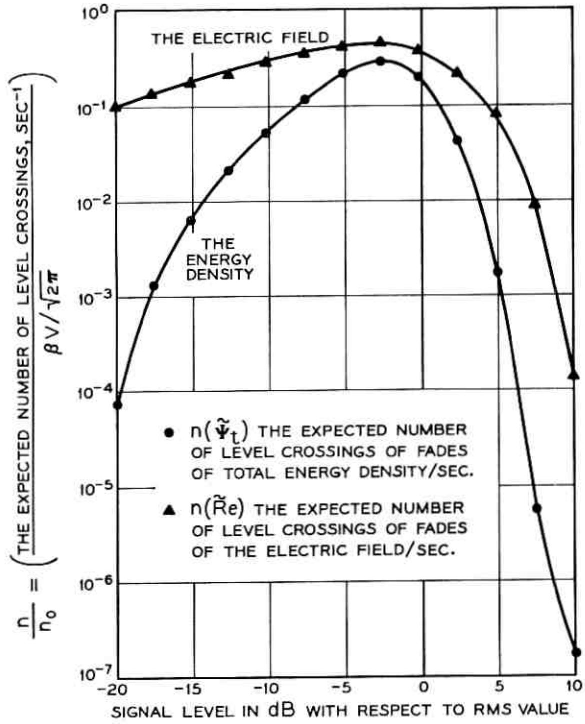


Fig. 3 — Comparison of the level crossing rate of the electric field with that of the energy density.

where $p(r_e(t))$ is the probability density obtained from $p(X_1, Y_1)$. By changing the variables, we obtain $p(r_e(t), \eta_e(t))$ from $p(X_1, Y_1)$. Then we integrate $p(r_e(t), \eta_e(t))$ over from 0 to 2π to obtain[13]

$$p(r_e(t)) = \frac{r_e(t)}{\mu_{11}} \exp\left[-\frac{r_e^2(t)}{2\mu_{11}}\right], \tag{28}$$

where μ_{11} is obtained from (18). Substituting (28) into (27) we get

$$P(r_e(t) < R_e) = 1 - \exp(-R_e^2/2\mu_{11}) = 1 - \exp(-\tilde{R}_e^2). \tag{29}$$

The expected number of times per second that $r_e(t)$ passes upward (or downward) across the level R_e is obtained from (24). The average duration of fades during which $r_e(t) < R_e$ may then be obtained by substituting (24) and (29) into (7)

$$\mathbf{t}(R_e) = \frac{P(r_e(t) < R_e)}{n(R_e)} = \frac{P(\tilde{r}_e(t) < \tilde{R}_e)}{n(\tilde{R}_e)} = \frac{\sqrt{2\pi}}{\beta V}\frac{1}{\tilde{R}_e}\left[\exp(\tilde{R}_e^2) - 1\right] \tag{30}$$

which is shown in Fig. 4.

3.2 *Finding the Values of $n(R_{hx})$ and $\mathbf{t}(R_{hx})$ from the H_x Field*

Following the same steps as above, we are going to find the joint probability density function $p(r_{hx}, \dot{r}_{hx})$ of the envelope r_{hx} and its slope $\dot{r}_{hx}$ of the H_x field component first. From (9) we find the real and imaginary parts of H_x which are expressed in the Appendix. The means, variances, and covariances of four real Gaussian random variables shown in the Appendix are

$$m_2 = \langle X_2 \rangle = \langle Y_2 \rangle = \langle \dot{X}_2 \rangle = \langle \dot{Y}_2 \rangle = 0$$

$$\mu_{22} = X_2^2 = Y_2^2 = \frac{N}{2} \qquad \text{for any } N$$

$$\mu_{22}' = \dot{X}_2^2 = \dot{Y}_2^2 = \frac{N}{8}(\beta V)^2 \cdot [\cos^2\alpha + 3\sin^2\alpha] \quad \text{for} \quad N = 3 \quad \text{and} \quad N \geqq 5$$

$$\langle X_2 Y_2 \rangle = \langle X_2 \dot{X}_2 \rangle = \langle X_2 \dot{Y}_2 \rangle = \langle Y_2 \dot{X}_2 \rangle = \langle Y_2 \dot{Y}_2 \rangle = \langle \dot{X}_2 \dot{Y}_2 \rangle = 0.$$

The probability density of the envelope of H_x field and its rate of change $p(r_{hx}, \dot{r}_{hx})$ is then obtained by following the same procedure used in deriving $p(r_e, \dot{r}_e)$.

$$p(r_{hx}, \dot{r}_{hx}) = \frac{1}{\sqrt{2\pi\mu_{22}'}}\frac{r_{hx}}{\mu_{22}} \exp\left\{-\frac{1}{2}\left(\frac{r_{hx}^2}{\mu_{22}} + \frac{\dot{r}_{hx}^2}{\mu_{22}'}\right)\right\}. \tag{31}$$

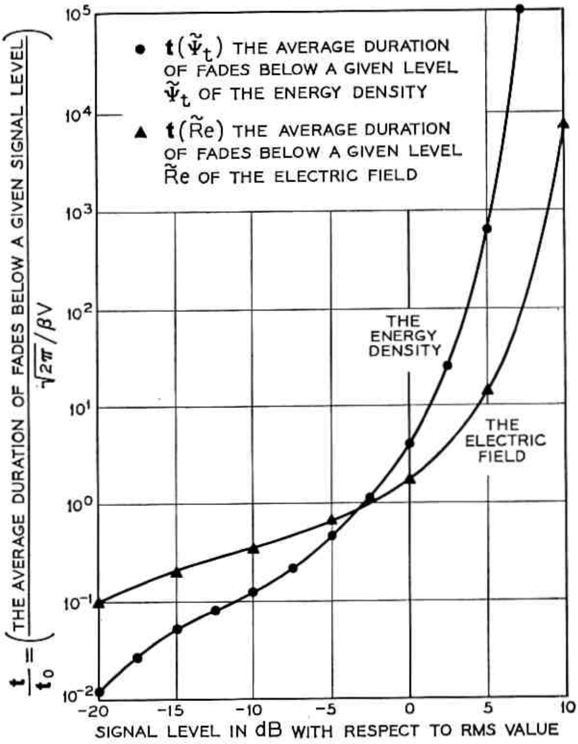


Fig. 4 — Comparison of the duration of fades of the electric field with that of the energy density.

Hence, the expected number of level crossings $n(R_{hz})$ at a given signal amplitude ($r_{hz} = R_{hz}$) can be obtained from (5) by using $p(R_{hz}, \dot{r}_{hz})$ in (31)

$$n(R_{hz}) = \int_0^\infty \dot{r}_{hz} p(R_{hz}, \dot{r}_{hz}) d\dot{r}_{hz} = \sqrt{\frac{\mu_{22}'}{\mu_{22}}} \left(\frac{R_{hz}}{\sqrt{2\pi\mu_{22}}}\right) \exp\left(-\frac{R_{hz}^2}{2\mu_{22}}\right). \quad (32)$$

The variance of r_{hz} is

$$\langle r_{hz}^2 \rangle = \langle X_2^2 \rangle + \langle Y_2^2 \rangle = 2\mu_{22} = N.$$

Let

$$\tilde{R}_{hz} = \frac{R_{hz}}{\sqrt{\langle r_{hz}^2 \rangle}} = \frac{R_{hz}}{r_{hz(rms)}} = \frac{R_{hz}}{\sqrt{N}}. \quad (33)$$

Substituting the values of the variances μ_{22} and μ'_{22} into (32) and replacing N by $\langle r_{hx}^2 \rangle$ we get

$$\begin{aligned} n(\tilde{R}_{hx}) &= \frac{\beta V}{\sqrt{2\pi}} \sqrt{\frac{\cos^2 \alpha + 3 \sin^2 \alpha}{2}}\, \tilde{R}_{hx} \exp -\tilde{R}_{hx}^2 \\ &= \frac{\beta V}{\sqrt{2\pi}} \sqrt{1 - \tfrac{1}{2} \cos 2\alpha}\, \tilde{R}_{hx} \exp -\tilde{R}_{hx}^2 . \end{aligned} \tag{34}$$

Equation (34) is the same form as (26) except for a multiplying factor which is a function of α shown in Fig. 5. Hence, $n(\tilde{R}_{hx})$ is also a function of angle α. Thus, when the mobile is moving along the x-axis $\alpha = 0$ or π, and

$$n(\tilde{R}_{hx}) = \frac{\beta V}{\sqrt{2\pi}} \frac{1}{\sqrt{2}} \tilde{R}_{hx} \exp -\tilde{R}_{hx}^2 \tag{34a}$$

which is the minimum value of $n(\tilde{R}_{hx})$. When the mobile is moving on $\pm y$-axis $\alpha = \pm\pi/2$, and

$$n(\tilde{R}_{hx}) = \frac{\beta V}{\sqrt{2\pi}} \sqrt{\frac{3}{2}}\, \tilde{R}_{hx} \exp -\tilde{R}_{hx}^2 \tag{34b}$$

which is the maximum value of $n(\tilde{R}_{hx})$. The ratio of level crossings for these two cases is

$$\frac{n(\tilde{R}_{hx})(\alpha = 0°,\ 180°)}{n(\tilde{R}_{hx})(\alpha = \pm 90°)} = \frac{1}{\sqrt{3}}. \tag{34c}$$

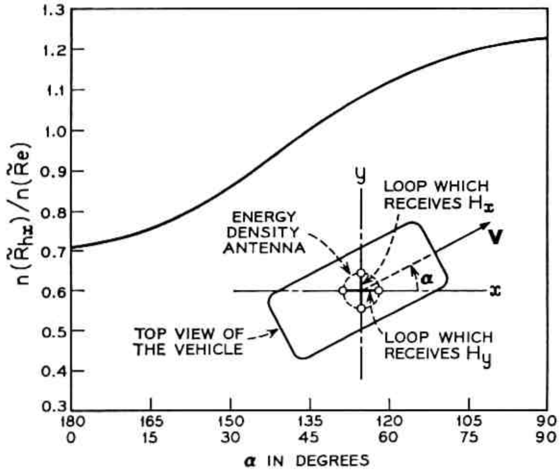


Fig. 5 — The effect of the angle α on the ratio of level crossing rates of the electric field to the x-component of the magnetic field.

For $\alpha = \pm 45°$ and $\pm 135°$,

$$n(\tilde{R}_{hz}) = \frac{\beta V}{\sqrt{2\pi}} \tilde{R}_{hz} \exp -\tilde{R}_{hz}^2 \tag{34d}$$

which is the same expression as (26) for $\tilde{R}_e$. However, although (34d) and (26) are the same form, the magnitudes $\tilde{R}_e$ and $\tilde{R}_{hz}$ are different, since $\langle r_e^2 \rangle$ in (25) is equal to two times $\langle r_{hz}^2 \rangle$ in (33).

The average duration of fades $\mathsf{t}(R_{hz})$ of H_z can be obtained without difficulty. It is easy to prove that the expression for the average length of the intervals during which $r_{hz} < R_{hz}$ will have the same form as $\mathsf{t}(R_e)$ in (30), except for a multiplying factor that depends on α, as follows:

$$\mathsf{t}(\tilde{R}_{hz}) = \frac{\sqrt{2\pi}}{\beta V} \frac{1}{\sqrt{1 - \frac{1}{2}\cos 2\alpha}} \frac{1}{\tilde{R}_{hz}} [\exp(\tilde{R}_{hz}^2) - 1]. \tag{35}$$

When $\alpha = 0°$ and $180°$:

$$\mathsf{t}(\tilde{R}_{hz}) = \frac{\sqrt{2\pi}}{\beta V} \sqrt{2} \frac{1}{\tilde{R}_{hz}} [\exp(\tilde{R}_{hz}^2) - 1] \tag{35a}$$

which is the maximum value of $\mathsf{t}(R_{hz})$, and when $\alpha = \pm 90°$:

$$\mathsf{t}(\tilde{R}_{hz}) = \frac{\sqrt{2\pi}}{\beta V} \sqrt{\frac{2}{3}} \frac{1}{\tilde{R}_{hz}} [\exp(\tilde{R}_{hz}^2) - 1] \tag{35b}$$

which is the minimum value of $\mathsf{t}(R_{hz})$. The ratio of level crossings for these two cases is

$$\frac{\mathsf{t}(\tilde{R}_{hz})(\alpha = 0°, 180°)}{\mathsf{t}(\tilde{R}_{hz})(\alpha = \pm 90°)} = \sqrt{3} \tag{35c}$$

which is the inverse of (34c). This tells us that when $n(\tilde{R}_{hz})$ reaches a maximum value, the average duration of fades reaches a minimum value and vice versa.

For $\alpha = \pm 45°, \pm 135°$

$$\mathsf{t}(\tilde{R}_{hz}) = \frac{2\pi}{\beta V} \frac{1}{\tilde{R}_{hz}} [\exp(\tilde{R}_{hz}^2) - 1] \tag{35d}$$

which is the same form as the expression for $\mathsf{t}(\tilde{R}_e)$ in (30). We may say at these angles $\alpha = \pm 45°$ and $\pm 135°$, the E field and the H_z field have the same average duration of fades below the level $\tilde{R}_e = \tilde{R}_{hz}$.

3.3 *Finding the Values of* $n(R_{hy})$ *and* $\mathsf{t}(R_{hy})$ *from the* H_y *Field*

Similarly, we obtain $n(R_{hy})$, the expected number of level crossings of a given signal level R_{hy} from the H_y field. It is very easy to see that $n(R_{hz})$ and $n(R_{hy})$ are the same forms of distribution as expressed in (34), except for the multiplying factor that depends on α. The average duration of fades $\mathsf{t}(R_{hy})$ during which $r_{hy}(t) < R_{hy}$ is also of the same form as $\mathsf{t}(R_{hz})$ in (35) except for the multiplying factor depending on α. The variances μ_{33} and μ_{33}' are given in the Appendix. We may thus write directly

$$n(\tilde{R}_{hy}) = \frac{\beta V}{\sqrt{2\pi}} \sqrt{1 + \tfrac{1}{2} \cos 2\alpha}\, \tilde{R}_{hy} \exp -\tilde{R}_{hy}^2 \tag{36}$$

$$\mathsf{t}(\tilde{R}_{hy}) = \frac{\sqrt{2\pi}}{\beta V} \frac{1}{\sqrt{1 + \tfrac{1}{2} \cos 2\alpha}} \frac{1}{\tilde{R}_{hy}} [\exp (\tilde{R}_{hy}^2) - 1], \tag{37}$$

where

$$\tilde{R}_{hy} = \frac{R_{hy}}{\sqrt{\langle r_{hy}^2 \rangle}} = \frac{R_{hy}}{r_{hy(rms)}}.$$

It is obvious that

$$n(\tilde{R}_{hy})_{\alpha=0^\circ,180^\circ} = n(\tilde{R}_{hz})_{\alpha=\pm 90^\circ}$$

$$n(\tilde{R}_{hy})_{\alpha=\pm 90^\circ} = n(\tilde{R}_{hz})_{\alpha=0^\circ,180^\circ}$$

when $\tilde{R}_{hy} = \tilde{R}_{hz}$.

IV. THE EXPECTED NUMBER OF LEVEL CROSSINGS AND THE AVERAGE DURATION OF FADES OF THE SIGNAL FROM AN ENERGY DENSITY ANTENNA

J. R. Pierce[7] has suggested utilization of the energy density concept as a possible means for reducing the signal fading in mobile radio. If we pick up the electric field e and the magnetic field h in free space and amplify the two fields by their appropriate relative gains, square and add these two fields, we obtain a signal proportional to electromagnetic energy density

$$W = \tfrac{1}{2}(\epsilon e^2 + \mu h^2), \tag{38}$$

where ϵ is dielectric constant, and μ is permeability. This idea can be realized by using a special antenna[6] which receives three field components e_z, h_x, and h_y simultaneously. The three signals enter separate square-law detectors, and the three detector outputs are added to obtain the energy density,

$$W = \tfrac{1}{2}(\epsilon \, |e_z|^2 + \mu \, |h_x|^2 + \mu \, |h_y|^2).$$

We may express W in a different form

$$\begin{aligned} W &= \frac{\epsilon}{2}\left[\left(|e_z|^2 + \frac{\mu}{\epsilon}\,|h_x|^2 + \frac{\mu}{\epsilon}\,|h_y|^2\right)\text{volt}^2/m^2\right] \\ &= \frac{\epsilon}{2}\,[|E_z|^2\,(\text{volt}^2/m^2) + |H_x|^2\,(\text{volt}^2/m^2) + |H_y|^2\,(\text{volt}^2/m^2)] \\ &= \frac{\epsilon}{2}\,[\psi_t(\text{volt}^2/m^2)] = \frac{\epsilon\psi_t}{2}\ \text{Joules}/m^3. \qquad (39) \end{aligned}$$

We define ψ_t as a normalized energy density

$$\begin{aligned} \psi_t &= |E_z|^2 + |H_x|^2 + |H_y|^2\,\text{volt}^2/m^2 \\ &= \psi_e + \psi_{hx} + \psi_{hy} \qquad (40) \\ &= (X_1^2 + Y_1^2) + (X_2^2 + Y_2^2) + (X_3^2 + Y_3^2). \qquad (41) \end{aligned}$$

Gilbert[10] has done some work on finding power spectra in energy reception for mobile radio. His work provides very useful background for this paper.

In this section, we are attempting to derive the number of crossings $n(\Psi_t)$ at a given level of signal magnitude Ψ_t using (5) in Section II. First of all, we need to find the joint probability density function $p(\psi_t, \dot\psi_t)$ of signal ψ_t and its slope $\dot\psi_t$. Since ψ_t is a function of $(X_1, Y_1, X_2, Y_2, X_3, Y_3)$, and $\dot\psi_t$ is assumed to be a function of $(\dot\psi_e, \dot\psi_{hx}, \dot\psi_{hy})$, we will find out that the variables $(X_1, Y_1, X_2, Y_2, X_3, Y_3)$ and $(\dot\psi_e, \dot\psi_{hx}, \dot\psi_{hy})$ are two independent Gaussian variable groups. Then,

$$\begin{aligned} &p[\psi_t(X_1, Y_1, X_2, Y_2, X_3, Y_3), \dot\psi_t(\dot\psi_e, \dot\psi_{hx}, \dot\psi_{hy})] \\ &\quad = p[\psi_t(X_1, Y_1, X_2, Y_2, X_3, Y_3)] \times p[\dot\psi_t(\dot\psi_e, \dot\psi_{hx}, \dot\psi_{hy})] \\ &\quad = p(\psi_t)p(\dot\psi_t). \qquad (42) \end{aligned}$$

A brief sketch of the method of finding $p(\psi_t)$ and $p(\dot\psi_t)$ is discussed below.

4.1 *To get $p(\psi_t)$*

Since we know from (41) that

$$\psi_t = (X_1^2 + Y_1^2) + (X_2^2 + Y_2^2) + (X_3^2 + Y_3^2)$$

and since X_1, Y_1, $\cdots$ are independent Gaussian variables, it is easy to get $p(\psi_t)$ from the Fourier transform of the characteristic function $M_{\psi_t}(v)$, where $M_{\psi_t}(v) = M_{X_1^2} M_{Y_1^2} M_{X_2^2} M_{Y_2^2} \cdots$ etc., and we can get these M's very easily.

4.2 *To get* $p(\dot{\psi}_t)$

All the terms in the summations of equations (79), (80), and (81) which represent $\dot{\psi}_e$, $\dot{\psi}_{hz}$ and $\dot{\psi}_{hy}$, respectively, in the Appendix are statistically independent. Then by the central limit theorem these three variables $\dot{\psi}_e$, $\dot{\psi}_{hz}$, and $\dot{\psi}_{hy}$ are Gaussian distributed. Hence, the joint probability density function $p(\dot{\psi}_e, \dot{\psi}_{hz}, \dot{\psi}_{hy})$ can be established. Since $\dot{\psi}_t = \dot{\psi}_e + \dot{\psi}_{hz} + \dot{\psi}_{hy}$ we can get $p(\dot{\psi}_t)$ from the Fourier transform of the characteristic function $M_{\dot{\psi}_t}$, but $M_{\dot{\psi}_t}$ must now be obtained from the general definition

$$M_{\dot{\psi}_t}(v) = E[e^{jv\dot{\psi}_t}] = \iiint e^{jv\dot{\psi}_t} p(\dot{\psi}_e, \dot{\psi}_{hz}, \dot{\psi}_{hy}) d\dot{\psi}_e d\dot{\psi}_{hz} d\dot{\psi}_{hy} \quad (43)$$

since we have no simple way of getting $M_{\dot{\psi}_e}$, $M_{\dot{\psi}_{hz}}$ and $M_{\dot{\psi}_{hy}}$ separately.

Let us introduce a new variable ϵ which can be any one of the above Gaussian random variables. It has a zero mean and variance μ. Then the probability density function of the square ϵ^2 is[14]

$$p(\gamma = \epsilon^2) = \frac{1}{\sqrt{2\pi\mu\gamma}} \exp\left(-\frac{1}{2\mu}\gamma\right) \quad (44)$$

for $\gamma > 0$. The characteristic function corresponding to this probability density is

$$M_\gamma(jv) = \int_0^\infty e^{jv\gamma} p(\gamma) d\gamma = (1 - j2\mu v)^{-\frac{1}{2}}. \quad (45)$$

From the Appendix we know all six variables X_1, Y_1, X_2, Y_2, X_3 and Y_3 are independent Gaussian variables. It is not hard to see that the X_1^2, Y_1^2, X_2^2, Y_2^2, X_3^2, Y_3^2 are independent variables by obtaining $p(X_1^2, Y_1^2, X_2^2, Y_2^2, X_3^2, Y_3^2)$ from the Jacobian of the transformation[12] of $p(X_1, Y_1, X_2, Y_2, X_3, Y_3)$. Then X_1^2 and Y_1^2 have the same characteristic function $(1 - 2j\mu_{11}v)^{-\frac{1}{2}}$. Also X_2^2, Y_2^2, X_3^2, and Y_3^2 have the same characteristic function $(1 - 2\,\mu_{22}v)^{-\frac{1}{2}}$. Thus, by the addition theorem[15] the sum ψ_t which is defined in (41) has the characteristic function

$$M_{\psi_t}(jv) = \frac{1}{(1 - 2j\mu_{11}v)(1 - 2j\mu_{22}v)^2}. \quad (46)$$

Then the probability density function $p(\psi_t)$ can be obtained from the Fourier transformation of the characteristic function

$$p(\psi_t) = \frac{1}{2\pi}\int_{-\infty}^{\infty} e^{-jv\psi_t} M_{\psi_t}(jv)dv$$

$$= \frac{\mu_{11}}{2(\mu_{22}-\mu_{11})^2}\left[\exp\left(-\frac{\psi_t}{2\mu_{11}}\right) - \exp\left(-\frac{\psi_t}{2\mu_{22}}\right)\right] + \frac{\psi_t \exp\left(-\dfrac{\psi_t}{2\mu_{22}}\right)}{4\mu_{22}(\mu_{22}-\mu_{11})}. \tag{47}$$

The joint probability density $p(\dot\psi_e, \dot\psi_{hx}, \dot\psi_{hy})$ has been derived in the Appendix

$$p(\dot\psi_e , \dot\psi_{hx} , \dot\psi_{hy})$$

$$= \frac{1}{(2\pi)^{\frac{3}{2}} \mid \Lambda \mid^{\frac{1}{2}}} \exp \{-\tfrac{1}{2}(\mid \Lambda_{11} \mid \dot\psi_e^2 + \mid \Lambda_{22} \mid \dot\psi_{hx}^2 + \mid \Lambda_{33} \mid \dot\psi_{hy}$$

$$+ 2 \mid \Lambda_{12} \mid \dot\psi_e\dot\psi_{hx} + 2 \mid \Lambda_{13} \mid \dot\psi_e\dot\psi_{hy} + 2 \mid \Lambda_{23} \mid \dot\psi_{hx}\dot\psi_{hy})\}, \tag{48}$$

where $[\Lambda]$ is the covariance matrix of $\dot\psi_e$, $\dot\psi_{hx}$, $\dot\psi_{hy}$ and the $|\Lambda_{nm}|$ are the cofactors of $|\Lambda|$, given in the Appendix. From (40), $\dot\psi_t$ is the sum of the three random variables $\dot\psi_e$, $\dot\psi_{hx}$, and $\dot\psi_{hy}$. Then the characteristic function for $\dot\psi_t$ is

$$M_{\dot\psi_t}(jv) = E[\exp\{jv(\dot\psi_e + \dot\psi_{hx} + \dot\psi_{hy}\}]$$

$$= \iiint_{-\infty}^{\infty} \exp\{jv(\dot\psi_e + \dot\psi_{hx} + \dot\psi_{hy})\}p(\dot\psi_e , \dot\psi_{hx} , \dot\psi_{hy})d\dot\psi_e d\dot\psi_{hx} d\dot\psi_{hy} . \tag{49}$$

The details of this computation are given in the Appendix with the result (92)

$$M_{\dot\psi_t}(jv) = \exp\{-\tfrac{1}{2}\rho_t' v^2\}, \tag{50}$$

where

$$\rho_t' = \rho_{11}' + \rho_{22}' + \rho_{33}' + 2\rho_{12}' + 2\rho_{13}' . \tag{51}$$

The probability density of total $\dot\psi_t$ is then

$$p(\dot\psi_t) = \frac{1}{2\pi}\int_{-\infty}^{\infty} M_{\dot\psi_t}(jv)e^{-jv\dot\psi_t}dv = \frac{1}{\sqrt{2\pi\rho_t'}}\exp\left(-\frac{1}{2\rho_t'}\dot\psi_t^2\right). \tag{52}$$

The joint probability density $p(\psi_t , \dot\psi_t)$ of ψ_t and $\dot\psi_t$ can now be obtained by substituting (47) and (52) into (42).

The expected number of crossings $n(\Psi_t)$ at a given signal level $\psi_t = \Psi_t$ made by the total energy density signal in one second can be obtained from (5)

$$n(\Psi_t) = \int_0^\infty \dot{\psi}_t p(\Psi_t\,,\,\dot{\psi}_t) d\dot{\psi}_t = p(\Psi_t) \int_0^\infty \dot{\psi}_t p(\dot{\psi}_t) d\dot{\psi}_t$$

$$= \sqrt{\frac{\rho_t'}{2\pi}} \left\{ \frac{\mu_{11}}{2(\mu_{22} - \mu_{11})^2} \left[\exp\left(-\frac{\Psi_t}{2\mu_{11}}\right) - \exp\left(-\frac{\Psi_t}{2\mu_{22}}\right) \right] + \frac{\Psi_t \exp -\Psi_t/2\mu_{22}}{4\mu_{22}(\mu_{22} - \mu_{11})} \right\}, \qquad (53)$$

where ρ_t' is given by (51) and $\mu_{11} = N$, $\mu_{22} = N/2$ as shown in the Appendix.

Also we know from Gilbert[10]

$$\langle \psi_t^2 \rangle = 22N^2.$$

In addition, $\rho_t' = \rho_{11}' + \rho_{22}' + \rho_{33}' + 2\rho_{12}' + 2\rho_{13}' = 2N^2(\beta V)^2$ and letting

$$\tilde{\Psi}_t = \frac{\Psi_t}{\sqrt{\langle \psi_t^2 \rangle}} = \frac{\Psi_t}{\psi_{t\,(\mathrm{rms})}}$$

we can simplify (53) as follows:

$$n(\tilde{\Psi}_t) = \frac{\beta V}{\sqrt{2\pi}} \left\{ 2\sqrt{2} \left[\exp\left(-\frac{\sqrt{22}}{2}\,\tilde{\Psi}_t\right) - \exp\left(-\sqrt{22}\tilde{\Psi}_t\right) \right] - 2\sqrt{11}\tilde{\Psi}_t \exp\left(-\sqrt{22}\tilde{\Psi}_t\right) \right\}. \qquad (54)$$

Equation (54) is a distribution which is independent of the angle α. When $\tilde{\Psi}_t = 1$, it means that Ψ_t is equal to its rms value. Equation (54) then becomes

$$n(\tilde{\Psi})_{\tilde{\Psi}=1} = \frac{\beta V}{\sqrt{2\pi}} \times 0.1839.$$

Also let $\tilde{R}_e = 1$ in (26):

$$n(\tilde{R}_e)_{\tilde{R}_e=1} = \frac{\beta V}{\sqrt{2\pi}} \times 0.3678.$$

It is shown that the expected number of crossings of the total energy density is one half the expected number of crossings of the envelope of

E field at their rms levels ($\tilde{\Psi}_t = \tilde{R}_e^2$). In other words, the energy density fades less frequently than the E field by at least a factor of 2.

$$n(\tilde{\Psi}_t) \leqq \tfrac{1}{2} n(\tilde{R}_e)$$

for $\tilde{\Psi}_t = \tilde{R}_e^2$ with respect to their rms values. The theoretical values of $n(\Psi)$ and $n(R_e)$ are in Fig. 3.

The average duration of fades below a given level Ψ_t is given by (7)

$$\mathsf{t}(\Psi_t) = \frac{P(\psi_t(t) < \Psi_t)}{n(\Psi_t)}, \tag{55}$$

where $P(\psi_t(t) < \Psi_t)$ is the probability function obtained from $p(\psi_t(t))$.

$$\begin{aligned} P(\psi_t(t) < \Psi_t) &= \int_0^{\Psi_t} p(\psi_t(t)) d\psi_t(t) \\ &= 1-4 \exp\left(-\frac{\sqrt{22}}{2}\tilde{\Psi}_t\right) + (3 + \sqrt{22}\tilde{\Psi}_t) \exp(-\sqrt{22}\tilde{\Psi}_t), \end{aligned} \tag{56}$$

where

$$\tilde{\Psi}_t = \frac{\Psi_t}{\psi_{t\,(\mathrm{rms})}}.$$

Substituting (54) and (56) into (55), we obtain the average duration of a fade below a given level Ψ_t :

$$\begin{aligned} \mathsf{t}(\tilde{\Psi}_t) &= \frac{\sqrt{2\pi}}{\beta V} \frac{1}{\sqrt{2}} \\ &\cdot \left\{ \frac{1 - 2 \exp\left(-\frac{\sqrt{22}}{2}\tilde{\Psi}_t +\right) \exp(-\sqrt{22}\tilde{\Psi}_t)}{2[\exp\left(-\frac{\sqrt{22}}{2}\tilde{\Psi}_t\right) - \exp(-\sqrt{22}\tilde{\Psi}_t)] - \sqrt{22}\tilde{\Psi}_t \exp(-\sqrt{22}\tilde{\Psi}_t)} - 1 \right\}. \end{aligned} \tag{57}$$

When $\tilde{\Psi}_t = 1$, (57) becomes

$$\mathsf{t}(\tilde{\Psi}_t) = \frac{\sqrt{2\pi}}{\beta V} \times 3.74.$$

When $\tilde{R}_e = 1$, (30) becomes

$$\mathsf{t}(\tilde{R}_e) = \frac{\sqrt{2\pi}}{\beta V} \times 1.7183.$$

It is shown that the average duration of fades of the energy density below a level $\tilde{\Psi}_t$ is larger than the average duration of fades of the E

field below a level $\bar{R}_e$ where $\bar{\Psi}_e = \bar{R}_e^2 = 1$. The curves of $\mathbf{t}(\bar{\Psi}_t)$ and $\mathbf{t}(\bar{R}_e)$ have been plotted in Fig. 4 for comparison.

V. DISCUSSION OF THE THEORETICAL RESULTS

From the above derivation we know that ψ_t and $\dot{\psi}_t$ are two independent variables as shown by (42):

$$p(\psi_t\ ,\ \dot{\psi}_t) = p(\psi_t)p(\dot{\psi}_t),$$

and therefore (5) can be written as follows:

$$\begin{aligned} n(\Psi_t) &= p(\Psi_t) \int_0^\infty \dot{\psi}_t p(\dot{\psi}_t) d\dot{\psi}_t \\ &= p(\Psi_t)\ \{\dot{\psi}_t\} \end{aligned} \tag{58}$$

where $\{\dot{\psi}_t\}$ represents an integral. Equation (58) simply shows that the expected number of crossings $n(\Psi_t)$ at given level Ψ can be obtained from the probability density of level Ψ_t times the integral $\{\dot{\psi}_t\}$. The average duration of fades, then, turns out to be

$$\begin{aligned} \mathbf{t}(\Psi) &= \frac{P(\psi_t < \Psi_t)}{n(\Psi_t)} \\ &= \frac{1}{\{\dot{\psi}_t\}} \frac{P(\psi_t < \Psi_t)}{p(\Psi_t)}. \end{aligned} \tag{59}$$

We emphasize that (58) and (59) are valid only when ψ_t and $\dot{\psi}_t$ are two independent variables.

The two curves, $n(\bar{R}_e)$ and $n(\bar{\Psi}_t)$, are plotted in Fig. 3, normalized by the common factor $\sqrt{2\pi}/\beta V$. Both curves are plotted as functions of the signal level normalized to their own rms values. The value of $n(\bar{R}_e)$ is, as shown, always higher than the value of $n(\bar{\Psi}_t)$ for any signal level. From these two curves, it may be said that the fading of the energy density is less frequent than the fading of the envelope of the electric field. The maximum expected numbers of crossings of both $n(\bar{R}_e)$ and $n(\bar{\Psi}_t)$ are at the -3 dB level, which means for signal level at $1/\sqrt{2}$ and $\frac{1}{2}$ of their rms values, respectively, we will count the most fades. The curve of $n(\bar{\Psi}_t)$ has dropped faster on both sides of 0 dB than the curve of $n(\bar{R}_e)$, which means that the range of the signal amplitude ψ_t is less than the range of the signal amplitude r_e.

The average duration of the signal below a given amplitude level is another way of looking at the fading problem. Fig. 4 shows that the average duration of fades of the energy density $\mathbf{t}(\bar{\Psi}_t)$ is always larger than the average duration of fades of the electric field $\mathsf{t}(R_e)$ when the given signal levels are above −3 dB with respect to their rms values ($\tilde{\Psi}_e = \tilde{R}_e^2 > -3$ dB). The value of $\mathsf{t}(\tilde{\Psi}_t)$ is less than $\mathsf{t}(\tilde{R}_e)$ when the given signal levels are more than 3 dB below the rms values $\tilde{\Psi}_t = \tilde{R}_e^2 <$ −3 dB. When the given signal level is at −3 dB ($\tilde{\Psi}_t = \tilde{R}_e^2 = -3$ dB), the average duration of fades $\mathsf{t}(-3$ dB) of both the energy density and the electric field are the same.

VI. COMPARISON OF THE THEORETICAL PREDICTION WITH THE EXPERIMENTS

The three field components E, H_x, and H_y have been received by a special antenna[6,7] mounted on a mobile van moving at a speed of 15 mile/hr. All the figures shown in this section were taken on Commonwealth Avenue, New Providence, New Jersey, from a transmitting

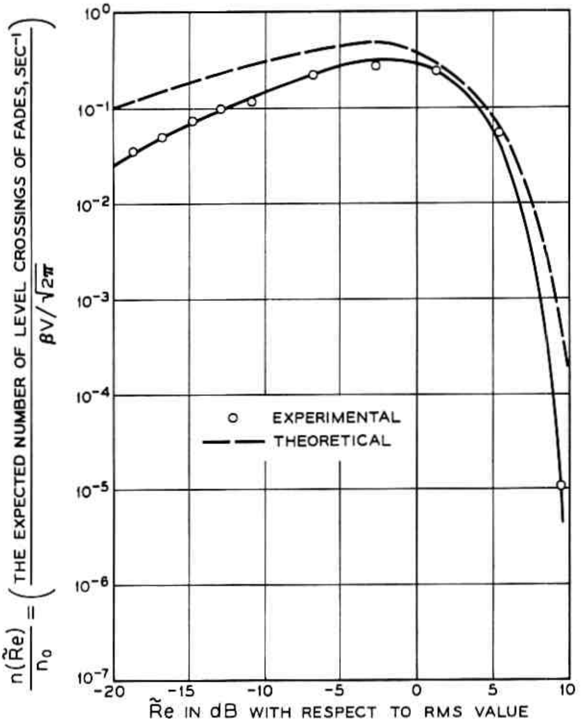


Fig. 6 — Comparison of the predicted level crossing rates to the observed rates for the electric field on Commonwealth Avenue, New Providence, New Jersey.

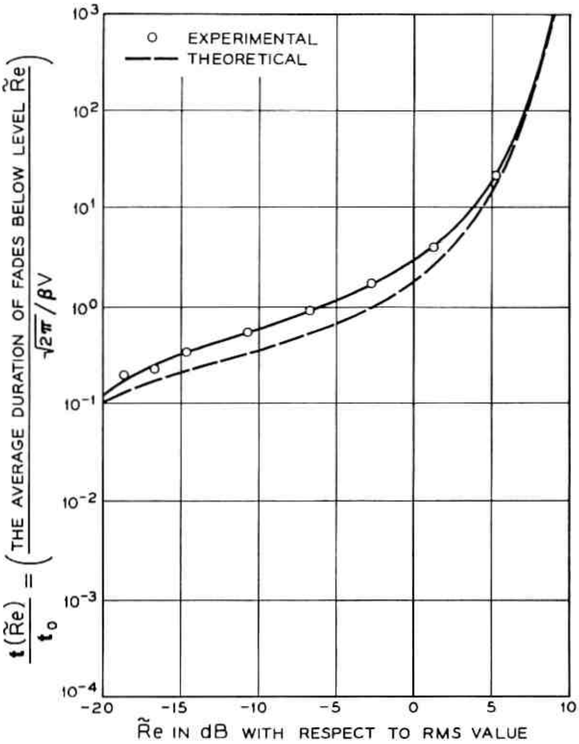


Fig. 7 — Comparison of the predicted average duration of fades to the observed average duration of fades for the electric field on Commonwealth Avenue, New Providence, New Jersey.

antenna at 836 MHz at Bell Laboratories, Murray Hill. After adjusting the appropriate relative gains of the three fields, the energy density can be obtained by squaring and summing these three fields by computer

$$\psi_t = | E |^2 + | H_x |^2 + | H_y |^2, \qquad \text{volt}^2/m^2.$$

Since the distance between the transmitting antenna and the mobile unit is relatively short, the angle swept out by the radius vector from the base station to the mobile unit varies considerably over a typical length of run. To reduce the variation of this angle the data for the entire run were cut into sections 8 seconds long, corresponding to 175 feet of travel, for computer processing. Each section, either the envelope r_e of the E field or the energy density ψ_t, was used to obtain the number of level crossings n and the average duration of fades **t** by com-

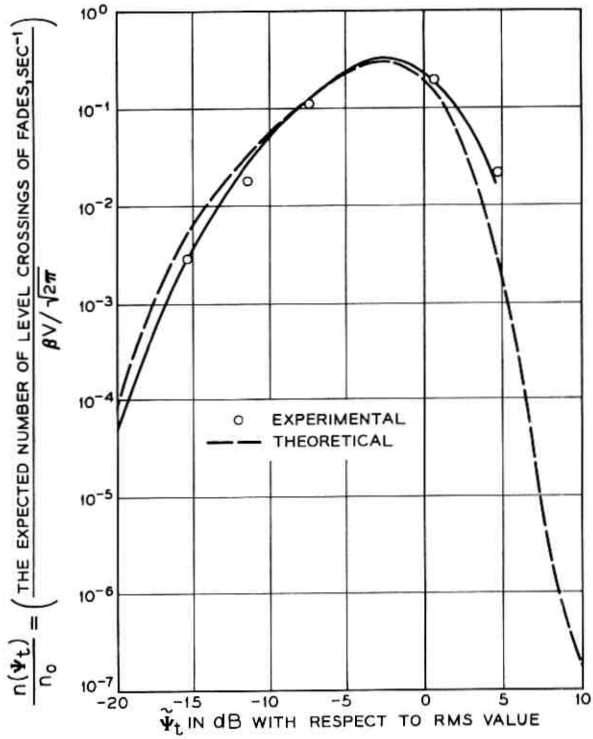


Fig. 8 — Comparison of the predicted level crossing rates to the observed rates for the energy density on Commonwealth Avenue, New Providence, New Jersey.

puter program. However, since the experimental curves of n and $\mathbf{t}$ were almost all alike for all sections, we used only one for comparison with the theoretical curve.

Fig. 6 shows a comparison of the curves of the expected number of crossings $n(\tilde{R}_e)$ at any level $\tilde{R}_e$ for both experiment and theory. The shape of the experimental curve is in fairly good agreement with the theoretical curve. Since the receiving antenna on Commonwealth Avenue is in line of sight with the transmitting antenna at Bell Laboratories, a small direct wave component may be introduced. This small direct wave component is not considered in our theoretical analysis, hence the values $n(\tilde{R}_e)$ from the experiments should be less than the theoretical results as we would predict.

Fig. 7 shows a comparison of the curves of the average duration of fades $\mathbf{t}(\tilde{R}_e)$ for both experiment and theory. They are quite alike. Since a small direct wave component does exist, the average duration of fades for the experimental data should be higher than the theoretical results.

Fig. 8 shows a comparison of the curves of the expected number of crossings $n(\Psi_t)$ at any level $\tilde{\Psi}_t$ for both experiment and theory. The shape of the experimental curve is very much like the theoretical curve. It shows that the theoretical model used in this paper is quite acceptable.

Fig. 9 shows a comparison of the curves of the average duration of fades $\mathbf{t}(\tilde{\Psi}_t)$ for both experiment and theory. The difference between the experimental curve and the theoretical curve may be caused by the small direct wave component. A small direct wave component introduced into our theoretical model may cause a little higher average duration of fades than it might expect, but does not affect the number of level crossings.

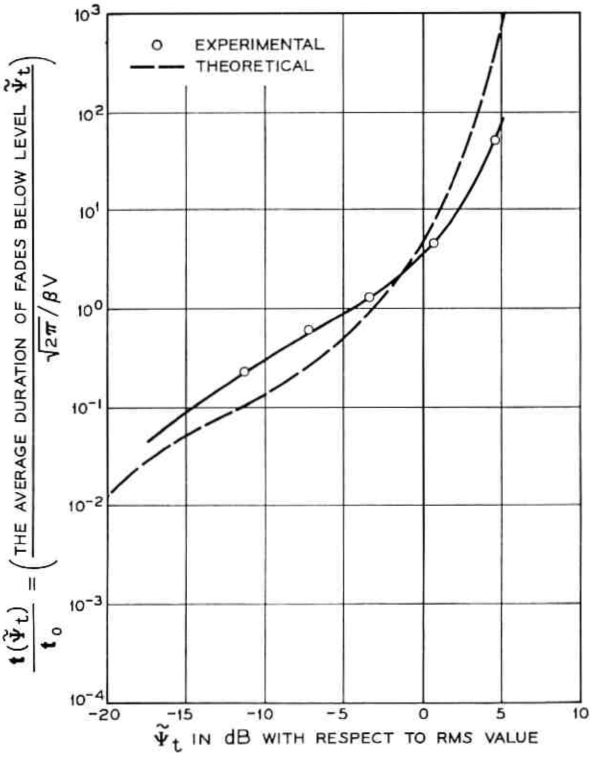


Fig. 9 — Comparison of the predicted average duration of fades to the observed average duration of fades for the energy density on Commonwealth Avenue, New Providence, New Jersey.

VII. CONCLUSIONS

Comparing the expected number of level crossings and the average durations of fades of the energy density with that of the E field, we see that the fading of the energy density is much less severe than the fading of the envelope of E field.

Referring to Fig. 5, which shows the fading rate related to the orientation of the energy density antenna and the direction of vehicle motion, we see that when the two orthogonal loops are at 45° to the direction of motion, the fades of all three field components are the same. When one loop is lined up with the direction of motion and the other normal to it, the H field component received from the loop normal to the motion has less fading than either of the other two field components.

The expected number of crossings/second of fades at a given signal level, n, for both R_e and Ψ_t is proportional to the carrier frequency f_c and the mobile speed V, as shown in (26) and (54). They have the common factor, $(\beta V/\sqrt{2\pi} = \sqrt{2\pi}(Vf_c/c)$, where c is the velocity of light. Hence, if either V or f_e goes higher, n becomes greater.

The average duration of fades, $\mathbf{t}$, is inversely proportional to the carrier frequency f_c and mobile speed V, as shown in (30) and (57). Hence, if either V or f_c goes higher, $\mathbf{t}$ becomes smaller.

The foregoing theoretical analysis is based on a Gaussian model and does not include a direct wave component. Even so, this analysis is compared with the experimental results in Section VI with fairly good agreement.

VIII. ACKNOWLEDGMENT

Discussions with W. C. Jakes, Jr. and R. H. Clarke have been very helpful. H. H. Hoffman helped with the experimental data and P. Liu did the computer programming to obtain the statistical results.

APPENDIX

A.1 *Finding the mean values, variances and covariances of nine variables* $(X_1, Y_1, X_2, Y_2, X_3, Y_3, \dot{\psi}_e, \dot{\psi}_{hx}, \dot{\psi}_{hy})$.

From (8), (9), and (10) we may express in the following forms

$$E_z = X_1 + jY_1$$

$$H_x = X_2 + jY_2$$

$$H_y = X_3 + jY_3 ,$$

where

$$X_1 = \sum_{u=1}^{N} R_u \cos \Phi_u + S_u \sin \Phi_u \tag{60}$$

$$Y_1 = \sum_{u=1}^{N} S_u \cos \Phi_u - R_u \sin \Phi_u \tag{61}$$

$$X_2 = \sum_{u=1}^{N} (R_u \cos \Phi_u + S_u \sin \Phi_u) \sin \theta_u \tag{62}$$

$$Y_2 = \sum_{u=1}^{N} (S_u \cos \Phi_u - R_u \sin \Phi_u) \sin \theta_u \tag{63}$$

$$X_3 = -\sum_{u=1}^{N} (R_u \cos \Phi_u + S_u \sin \Phi_u) \cos \theta_u \tag{64}$$

$$Y_3 = -\sum_{u=1}^{N} (S_u \cos \Phi_u - R_u \sin \Phi_u) \cos \theta_u \tag{65}$$

also

$$\Phi_u = \beta V t \cos (\theta_u - \alpha) \tag{66}$$

and the angles θ_u and α are shown in Fig. 2. The time derivatives of X_1, Y_1, X_2, Y_2, X_3, and Y_3 are

$$\dot{X}_1 = \beta V \sum_u (-R_u \sin \Phi_u + S_u \cos \Phi_u) \cos (\theta_u - \alpha) \tag{67}$$

$$\dot{Y}_1 = \beta V \sum_u (-S_u \sin \Phi_u - R_u \cos \Phi_u) \cos (\theta_u - \alpha) \tag{68}$$

$$\dot{X}_2 = \beta V \sum_u (-R_u \sin \Phi_u + S_u \cos \Phi_u) \sin \theta_u \cos (\theta_u - \alpha) \tag{69}$$

$$\dot{Y}_2 = \beta V \sum_u (-S_u \sin \Phi_u - R_u \cos \Phi_u) \sin \theta_u \cos (\theta_u - \alpha) \tag{70}$$

$$\dot{X}_3 = -\beta V \sum_u (-R_u \sin \Phi_u + S_u \cos \Phi_u) \cos \theta_u \cos (\theta_u - \alpha) \tag{71}$$

$$\dot{Y}_3 = -\beta V \sum_u (-S_u \sin \Phi_u - R_u \cos \Phi_u) \cos \theta_u \cos (\theta_u - \alpha). \tag{72}$$

The mean values of all above random variables are zero $\langle X_1 \rangle = \langle Y_1 \rangle = \langle X_2 \rangle = \langle Y_2 \rangle = \langle X_3 \rangle = \langle Y_3 \rangle = \langle \dot{X}_1 \rangle = \langle \dot{Y}_1 \rangle = \langle \dot{X}_2 \rangle = \langle \dot{Y}_2 \rangle = \langle \dot{X}_3 \rangle = \langle \dot{Y}_3 \rangle = 0$. The variances of all above random variables are

$$\left.\begin{aligned}\mu_{11} &= \langle X_1^2\rangle = \langle Y_1^2\rangle = N \\ \mu_{22} &= \langle X_2^2\rangle = \langle Y_2^2\rangle = \frac{N}{2} \\ \mu_{33} &= \langle X_3^2\rangle = \langle Y_3^2\rangle = \frac{N}{2}\end{aligned}\right\} \qquad \text{for any } N \qquad \begin{matrix}(73)\\(74)\\(75)\end{matrix}$$

$$\mu'_{11} = \langle \dot{X}_1^2\rangle = \langle \dot{Y}_1^2\rangle = \frac{N}{2}(\beta V)^2 \qquad \text{for } N \geqq 3 \quad (76)$$

$$\left.\begin{aligned}\mu'_{22} &= \langle \dot{X}_2^2\rangle = \langle \dot{Y}_2^2\rangle = \frac{N}{8}(\beta V)^2[\cos^2\alpha + 3\sin^2\alpha] \\ \mu'_{33} &= \langle \dot{X}_3^2\rangle = \langle \dot{Y}_3^2\rangle = \frac{N}{8}(\beta V)^2[3\cos^2\alpha + \sin^2\alpha]\end{aligned}\right\} \text{for} \begin{cases} N = 3 & (77) \\ N \geqq 5. & (78)\end{cases}$$

Remark: The values of μ'_{11} for $N \geqq 3$ is derived as follows: The summations of sine and cosine functions can be expressed

$$\sum_{k=1}^{N} \sin kx = \frac{\cos\frac{x}{2} - \cos(2N+1)\frac{x}{2}}{2\sin\frac{x}{2}}$$

$$\sum_{k=1}^{N} \cos kx = \frac{\sin(2N+1)\frac{x}{2} - \sin\frac{x}{2}}{2\sin\frac{x}{2}}.$$

Then in (67)

$$\theta_u = \frac{2\pi u}{N}$$

and

$$\sum_{u=1}^{N} \cos^2(\theta_u - \alpha) = \frac{1}{2}\sum_{u=1}^{N}\left[1 + \cos\left(u\frac{4\pi}{N} - 2\alpha\right)\right]$$

$$= \frac{N}{2} + \frac{\cos 2\alpha}{2}\sum_{u=1}^{N}\cos u\frac{4\pi}{N} + \frac{\sin 2\alpha}{2}\sum_{u=1}^{N}\sin u\frac{4\pi}{N}.$$

But,

$$\left.\begin{aligned}
&\sum_{u=1}^{N} \cos u\,\frac{4\pi}{N} = \frac{\sin\left(4\pi + \dfrac{2\pi}{N}\right) - \sin\dfrac{2\pi}{N}}{2\sin\dfrac{2\pi}{N}} = 0\\
&\text{and}\\
&\sum_{u=1}^{N} \sin u\,\frac{4\pi}{N} = \frac{\cos\dfrac{2\pi}{N} - \cos\left(4\pi + \dfrac{2\pi}{N}\right)}{2\sin\dfrac{2\pi}{N}} = 0
\end{aligned}\right\}\text{for}\quad N \geqq 3.$$

Thus, the average value of X_1^2 or Y_1^2 is

$$\left\langle \sum_{u=1}^{N} \cos^2\left(\frac{2\pi u}{N} - \alpha\right)\right\rangle_{\text{av}} = \frac{N}{2} \quad \text{for} \quad N \geqq 3.$$

Following the same derivation, we obtain the valid range of N for μ_{22}' and μ_{33}'. Later on we will obtain the range of N for ρ_{11}', ρ_{22}', ρ_{33}', ρ_{12}', ρ_{13}' and ρ_{23}' in the same way.

QED.

Now we are going to find the relations between all the six variables X_1, X_2, X_3, Y_1, Y_2, Y_3 and their time derivatives. Since we know if two variables a and b are Gaussian, and also uncorrelated, $\langle ab\rangle = 0$, then a and b are independent.[16] Therefore, the covariance of X_1, Y_1, $\dot{X}_1$, and $\dot{Y}_1$ are

$$\langle X_1Y_1\rangle = \langle X_1\dot{X}_1\rangle = \langle X_1\dot{Y}_1\rangle = \langle Y_1\dot{X}_1\rangle = \langle Y_1\dot{Y}_1\rangle = \langle \dot{X}_1\dot{Y}_1\rangle = 0$$

hence, the four variables X_1, Y_1, $\dot{X}_1$, and $\dot{Y}_1$ are statistically independent. The covariances of X_2, Y_2, $\dot{X}_2$, and $\dot{Y}_2$ are

$$\langle X_2Y_2\rangle = \langle X_2\dot{X}_2\rangle = \langle X_2\dot{Y}_2\rangle = \langle Y_2\dot{X}_2\rangle = \langle Y_2\dot{Y}_2\rangle = \langle \dot{X}_2\dot{Y}_2\rangle = 0$$

hence, the four variables X_2, Y_2, $\dot{X}_2$, and $\dot{Y}_2$ are statistically independent. The covariances of X_3, Y_3, $\dot{X}_3$, and $\dot{Y}_3$ are

$$\langle X_3Y_3\rangle = \langle X_3\dot{X}_3\rangle = \langle X_3\dot{Y}_3\rangle = \langle Y_3\dot{X}_3\rangle = \langle Y_3\dot{Y}_3\rangle = \langle \dot{X}_3\dot{Y}_3\rangle = 0$$

hence, the four variables X_3, Y_3, $\dot{X}_3$, and $\dot{Y}_3$ are statistically independent.

Also we may show that

$$\begin{aligned}
\langle X_mY_n\rangle &= 0 && \text{for all } m \text{ and } n\\
\langle X_mX_n\rangle = \langle Y_mY_n\rangle &= 0 && \text{for } m \neq n,
\end{aligned}$$

where

$$\left.\begin{matrix} m \\ n \end{matrix}\right\} = 1, 2, 3,$$

hence the six variables X_1, X_2, X_3, Y_1, Y_2, Y_3 are independent. The rate of change of energy densities of three fields E_z, H_x, and H_y are

$$\begin{aligned} \dot{\psi}_e &= \frac{d}{dt}\psi_e = \frac{d}{dt}(X_1^2 + Y_1^2) \\ &= \beta V \sum_{u,v} [-(R_uR_v + S_uS_v)\sin(\Phi_u - \Phi_v) \\ &\quad + (S_uR_v - R_uS_v)\cos(\Phi_u - \Phi_v)] \times [\cos(\theta_u - \alpha) - \cos(\theta_v - \alpha)], \end{aligned} \tag{79}$$

$$\begin{aligned} \dot{\psi}_{hx} &= \frac{d}{dt}\psi_{hx} = \frac{d}{dt}(X_2^2 + Y_2^2) \\ &= \beta V \sum_{u,v} [-(R_uR_v + S_uS_v)\sin(\Phi_u - \Phi_v) \\ &\qquad + (S_uR_v - R_uS_v)\cos(\Phi_u - \Phi_v)] \\ &\qquad \cdot \sin\theta_u \sin\theta_v [\cos(\theta_u - \alpha) - \cos(\theta_v - \alpha)], \end{aligned} \tag{80}$$

$$\begin{aligned} \dot{\psi}_{hy} &= \frac{d}{dt}\psi_{hy} = \frac{d}{dt}(X_3^2 + Y_3^2) \\ &= \beta V \sum_{u,v} [-(R_uR_v + S_uS_v)\sin(\Phi_u - \Phi_v) \\ &\qquad + (S_uR_v - R_uS_v)\cos(\Phi_u - \Phi_v)] \\ &\qquad \cdot \cos\theta_u \cos\theta_v [\cos(\theta_u - \alpha) - \cos(\theta_v - \alpha)]. \end{aligned} \tag{81}$$

The only terms that exist in (79), (80), (81) are those for which $u \neq v$. There are $N(N - 1)/2$ different terms which are all statistically independent in (79), (80), and (81). Hence, by the central limit theorem, $\dot{\psi}_e$, $\dot{\psi}_{hy}$, and $\dot{\psi}_{hx}$ are Gaussian random variables.

The variance of $\dot{\psi}_e$, $\dot{\psi}_{hx}$ and $\dot{\psi}_{hy}$ are

$$\rho_{11}' = \langle \dot{\psi}_e^2 \rangle = (\beta V)^2\, 4N(N - 1) \qquad \text{for} \quad N \geqq 3 \tag{82}$$

$$\left.\begin{aligned} \rho_{22}' &= \langle \dot{\psi}_{hx}^2 \rangle = (\beta V)^2 \frac{N(N - 1)}{2} [\cos^2\alpha + 3\sin^2\alpha] \quad (83) \\ \rho_{33}' &= \langle \dot{\psi}_{hy}^2 \rangle = (\beta V)^2 \frac{N(N - 1)}{2} [3\cos^2\alpha + \sin^2\alpha] \quad (84) \end{aligned}\right\} \text{for} \begin{cases} N = 3 \\ N \geqq 5. \end{cases}$$

The covariances of $\dot{\psi}_e$, $\dot{\psi}_{hx}$, and $\dot{\psi}_{hy}$ are

$$\left.\begin{aligned} \rho'_{12} &= \rho'_{21} = \langle \dot{\psi}_e \dot{\psi}_{hx} \rangle = -(\beta V)^2 2N(N-1) \sin^2 \alpha \\ \rho'_{13} &= \rho'_{31} = \langle \dot{\psi}_e \dot{\psi}_{hy} \rangle = -(\beta V)^2 2N(N-1) \cos^2 \alpha \end{aligned}\right\} \text{for } N \geqq 3 \qquad \begin{matrix} (85) \\ (86) \end{matrix}$$

$$\rho'_{23} = \rho'_{32} = \langle \dot{\psi}_{hx} \dot{\psi}_{hy} \rangle = 0. \tag{87}$$

It is very easy to show from (60) to (66) and (79) to (81) that the covariances of the variables between two groups ($\dot{\psi}_e$, $\dot{\psi}_{hx}$, $\dot{\psi}_{hy}$) and (X_1, Y_1, X_2, Y_2, X_3, Y_3) are zero. We may write

$$\langle \dot{\psi}_m X_n \rangle = \langle \dot{\psi}_m Y_n \rangle = 0 \qquad \text{for all } n \text{ and } m \tag{88}$$

$$\left.\begin{matrix} m \\ n \end{matrix}\right\} = 1, 2, 3$$

hence ($\dot{\psi}_e$, $\dot{\psi}_{hx}$, $\dot{\psi}_{hy}$) and (X_1, Y_1, X_2, Y_2, X_3, Y_3) are two independent variable groups.[16]

A.2 *Derivation of* $M_{\dot{\psi}_t}(jv)$ *in* (50)

The mean values of all three Gaussian random variables $\dot{\psi}_e$, $\dot{\psi}_{hx}$, and $\dot{\psi}_{hy}$ we observed from (79) to (81) are zero. Also (87) gives the covariance $\langle \dot{\psi}_{hx} \dot{\psi}_{hy} \rangle = 0$. The joint probability density function of three variables $\dot{\psi}_e$, $\dot{\psi}_{hx}$, $\dot{\psi}_{hy}$ can be obtained[11]

$$\begin{aligned} p(\dot{\psi}_e, \dot{\psi}_{hx}, \dot{\psi}_{hy}) = \frac{1}{(2\pi)^{\frac{3}{2}} \mid \Lambda \mid^{\frac{1}{2}}} \exp \Bigg\{ -\frac{1}{2 \mid \Lambda \mid} (\mid \Lambda_{11} \mid \dot{\psi}_e^2 + \mid \Lambda_{22} \mid \dot{\psi}_{hx} \\ + \mid \Lambda_{33} \mid \dot{\psi}_{hy}^2 + 2 \mid \Lambda_{12} \mid \dot{\psi}_e \dot{\psi}_{hx} + 2 \mid \Lambda_{13} \mid \dot{\psi}_e \dot{\psi}_{hy} \\ + 2 \mid \Lambda_{23} \mid \dot{\psi}_{hx} \dot{\psi}_{hy}) \Bigg\}, \end{aligned}$$

where

$[\Lambda]$ is a covariance matrix, and
$|\Lambda_{mn}|$ is a cofactor of ρ_{mn} in the covariance matrix $[\Lambda]$

$$[\Lambda] = \begin{bmatrix} \rho'_{11} & \rho'_{12} & \rho'_{13} \\ \rho'_{12} & \rho'_{22} & 0 \\ \rho'_{13} & 0 & \rho'_{33} \end{bmatrix} \tag{89}$$

$$|\Lambda| = \text{determinant of } [\Lambda] = \rho'_{33}(\rho'_{11}\rho'_{22} - \rho'^{2}_{12}) - \rho'_{22}{\rho'_{13}}^{2}$$

$$\begin{aligned}
\mid \Lambda_{11} \mid &= \rho'_{22}\rho'_{33} \\
\mid \Lambda_{22} \mid &= \rho'_{11}\rho'_{33} - {\rho'_{13}}^{2} \\
\mid \Lambda_{33} \mid &= \rho'_{11}\rho'_{22} - {\rho'_{12}}^{2} \\
\mid \Lambda_{12} \mid &= \mid \Lambda_{21} \mid = -\rho'_{12}\rho'_{33} \\
\mid \Lambda_{13} \mid &= \mid \Lambda_{31} \mid = -\rho'_{22}\rho'_{13} \\
\mid \Lambda_{23} \mid &= \mid \Lambda_{32} \mid = \rho'_{13}\rho'_{12} \, .
\end{aligned}$$

From (38), $\dot{\psi}_t$ is the sum of three Gaussian random variables $\dot{\psi}_e$, $\dot{\psi}_{hx}$, $\dot{\psi}_{hy}$:

$$\dot{\psi}_t = \dot{\psi}_e + \dot{\psi}_{hx} + \dot{\psi}_{hy} \, .$$

The characteristic function for $\dot{\psi}_t$ is then

$$\begin{aligned}
M_{\dot{\psi}_t}(jv) &= E[\exp \{jv(\dot{\psi}_e + \dot{\psi}_{hx} + \dot{\psi}_{hy}\}] \\
&= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \exp \{jv(\dot{\psi}_e + \dot{\psi}_{hx} + \dot{\psi}_{hy})\} p(\dot{\psi}_e \, , \dot{\psi}_{hx} \, , \dot{\psi}_{hy}) d\dot{\psi}_e d\dot{\psi}_{hx} d\dot{\psi}_{hy} \\
&= \frac{1}{(2\pi)^{\frac{3}{2}} \mid \Lambda \mid^{\frac{1}{2}}} \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} e^{jv(\dot{\psi}_e + \dot{\psi}_{hx})} \\
&\quad \cdot \exp \left\{ -\frac{1}{2 \mid \Lambda \mid} (\mid \Lambda_{11} \mid \dot{\psi}_e^2 + \mid \Lambda_{22} \mid \dot{\psi}_{hx}^2 + 2 \mid \Lambda_{12} \mid \dot{\psi}_e \dot{\psi}_{hx}) \right\} \\
&\quad \times \int_{-\infty}^{\infty} e^{jv\dot{\psi}_{hy}} \exp \left\{ -\frac{\mid \Lambda_{33} \mid}{2 \mid \Lambda \mid} \left(\dot{\psi}_{hy}^2 + \frac{2(\mid \Lambda_{13} \mid \dot{\psi}_e + \mid \Lambda_{23} \mid \dot{\psi}_{hx})}{\mid \Lambda_{33} \mid} \dot{\psi}_{hy} \right) \right\} \\
&\quad \cdot d\dot{\psi}_{hy} d\dot{\psi}_{hx} d\dot{\psi}_e \, .
\end{aligned} \tag{90}$$

The last integrand of $\dot{\psi}_y$ is

$$\begin{aligned}
&\exp \left\{ -j \frac{\mid \Lambda_{13} \mid \dot{\psi}_e + \mid \Lambda_{23} \mid \dot{\psi}_{hx}}{\mid \Lambda_{33} \mid} v + \frac{\mid \Lambda_{33} \mid}{2 \mid \Lambda \mid} \left(\frac{\mid \Lambda_{13} \mid \dot{\psi}_e + \mid \Lambda_{23} \mid \dot{\psi}_{hx}}{\mid \Lambda_{33} \mid} \right)^2 \right\} \\
&\qquad \times \int_{-\infty}^{\infty} \exp \left[-jv \left(\dot{\psi}_{hy} + \frac{\mid \Lambda_{13} \mid \dot{\psi}_e + \mid \Lambda_{23} \mid \dot{\psi}_{hx}}{\mid \Lambda_{33} \mid} \right) \right] \\
&\exp \left\{ -\frac{\mid \Lambda_{33} \mid}{2 \mid \Lambda \mid} \left(\dot{\psi}_y + \frac{\mid \Lambda_{13} \mid \dot{\psi}_e + \mid \Lambda_{23} \mid \dot{\psi}_{hx}}{\mid \Lambda_{33} \mid} \right)^2 \right\} d\dot{\psi}_{hy} \\
&\qquad = \exp \left(-jv \left(\frac{\mid \Lambda_{13} \mid \dot{\psi}_e + \mid \Lambda_{23} \mid \dot{\psi}_{hx}}{\mid \Lambda_{33} \mid} \right) \right. \\
&\qquad\qquad \left. + \frac{\mid \Lambda_{33} \mid}{2 \mid \Lambda \mid} \left(\frac{\mid \Lambda_{13} \mid \dot{\psi}_e + \mid \Lambda_{23} \mid \dot{\psi}_{hx}}{\mid \Lambda_{33} \mid} \right)^2 \right\} \\
&\qquad\qquad \times \int_{-\infty}^{\infty} \exp \left(jv\xi - \frac{h_1}{2} \xi^2 \right) d\xi.
\end{aligned}$$

From Cramer,[15] p. 99 we obtain

$$\int_{-\infty}^{\infty} \exp\left(jv\xi - \frac{h_1}{2}\,\xi^2\right)d\xi = \sqrt{\frac{2\pi}{h_1}}\exp -\frac{v^2}{2h_1},$$

where

$$h_1 = \frac{|\,\Lambda_{33}\,|}{|\,\Lambda\,|}.$$

Then following the same techniques we find

$$\begin{aligned} M_{\dot{\psi}_t}(jv) &= \frac{1}{(2\pi)^{\frac{3}{2}}\,|\,\Lambda\,|^{\frac{1}{2}}}\int_{-\infty}^{\infty}\exp\left(jv_3\xi_3 - \frac{h_3}{2}\,\xi_3^2\right)d\xi_3 \\ &\quad\cdot\int_{-\infty}^{\infty}\exp\left(jv_2\xi_2 - \frac{h_2}{2}\,\xi_2^2\right)d\xi_2\int_{-\infty}^{\infty}\exp\left(jv_1\xi_1 - \frac{h_1}{2}\,\xi_1^2\right)d\xi_1 \\ &= \frac{1}{(2\pi)^{\frac{3}{2}}\,|\,\Lambda\,|^{\frac{1}{2}}}\sqrt{\frac{(2\pi)^3}{h_1h_2h_3}}\left\{\exp\left[-\frac{1}{2}\left(\frac{v_1^2}{h_1}+\frac{v_2^2}{h_2}+\frac{v_3^2}{h_3}\right)\right]\right\}, \end{aligned} \tag{91}$$

where

$$\begin{aligned} h_1 &= \frac{|\,\Lambda_{33}\,|}{|\,\Lambda\,|} \\ h_2 &= \frac{B}{|\,\Lambda\,|\,|\,\Lambda_{33}\,|} \\ h_3 &= \frac{1}{|\,\Lambda\,|\,|\,\Lambda_{33}\,|\,B}\,[BC - A^2] \\ v_1 &= v \\ v_2 &= \left(1 - \frac{|\,\Lambda_{23}\,|}{|\,\Lambda_{33}\,|}\right)v \\ v_3 &= \left[1 - \frac{|\,\Lambda_{13}\,|}{|\,\Lambda_{33}\,|} - \left(1 - \frac{|\,\Lambda_{23}\,|}{|\,\Lambda_{33}\,|}\right)\frac{A}{B}\right]v \\ A &= |\,\Lambda_{12}\,|\,|\,\Lambda_{33}\,| - |\,\Lambda_{13}\,|\,|\,\Lambda_{23}\,| = -\rho_{12}'\,|\,\Lambda\,| \\ B &= |\,\Lambda_{22}\,|\,|\,\Lambda_{33}\,| - |\,\Lambda_{23}\,|^2 = \rho_{11}'\,|\,\Lambda\,| \\ C &= |\,\Lambda_{11}\,|\,|\,\Lambda_{33}\,| - |\,\Lambda_{13}\,|^2 = \rho_{22}'\,|\,\Lambda\,| \\ BC - A^2 &= |\,\Lambda_{33}\,|\,|\,\Lambda\,|^2. \end{aligned}$$

The constant value outside the bracket of (91) is

$$\frac{1}{(2\pi)^{\frac{3}{2}}\,|\,\Lambda\,|^{\frac{1}{2}}}\sqrt{\frac{(2\pi)^3}{h_1h_2h_3}} = \frac{1}{|\,\Lambda\,|^{\frac{1}{2}}}\sqrt{\frac{|\,\Lambda\,|^3\,|\,\Lambda_{33}\,|}{BC - A^2}} = 1$$

and the expression inside the bracket of (91) is

$$\begin{aligned}\exp -\frac{1}{2}\left[\frac{v_1^2}{h_1}+\frac{v_2^2}{h_2}+\frac{v_3^2}{h_3}\right] &= \exp\left\{-\frac{1}{2}\frac{\mid \Lambda \mid}{\mid \Lambda_{33} \mid}\left[1+\frac{(\mid \Lambda_{33}\mid - \mid \Lambda_{23}\mid)^2}{B}\right.\right.\\ &\qquad \left.\left.+\frac{\left[\mid \Lambda_{33}\mid - \mid \Lambda_{13}\mid - (\mid \Lambda_{33}\mid - \mid \Lambda_{23}\mid)\dfrac{A}{B}\right]^2 B}{BC-A^2}\right]v^2\right\}\\ &= \exp\left\{-\tfrac{1}{2}(\rho_{11}'+\rho_{22}'+\rho_{33}'+2\rho_{12}'+2\rho_{13}')v^2\right\} = \exp -\tfrac{1}{2}\rho_t'v^2.\end{aligned}$$

Thus,

$$M_{\dot\psi_t}(jv) = \exp -\tfrac{1}{2}\rho_t'v^2. \tag{92}$$

REFERENCES

1. Ossanna, J. F., Jr., A Model for Mobile Radio Fading Due to Building Reflections: Theoretical and Experimental Fading Waveform Power Spectra, B.S.T.J., *43,* November, 1964, pp. 2935–2971.
2. Aikens, A. J. and Lacy, L. Y., A Test of 450-Megacycle Urban Area Transmission to a Mobile Receiver, IRE Proc., *38,* November, 1950, p. 1317.
3. Bullington, K., Radio Propagation Variations at VHF and UHF, IRE Proc., *38,* January, 1950, p. 27.
4. Young, W. R., Jr., Mobile Radio Transmission Compared at 150 to 3700 Mc, B.S.T.J., *31,* November, 1952, pp. 1068–1085.
5. Rice, S. O., Distribution of the Duration of Fades in Radio Transmission, B.S.T.J., *37,* May, 1958, pp. 581–636.
6. Lee, W. C.-Y., Theoretical and Experimental Study of the Properties of the Signal from an Energy Density Mobile Radio Antenna, presented at IEEE 1966 Vehicular Communications Conference, Montreal, Quebec, December, 1966.
7. Pierce, J. R., private communication.
8. Rice, S. O., Statistical Properties of a Sine Wave Plus Random Noise, B.S.T.J., *37,* January, 1948, p. 109.
9. Kac, M., On the Average Number of Real Roots of a Random Algebraic Equation, Bull. Am. Math. Soc., *49,* 1943, p. 314.
10. Gilbert, E. N., Energy Reception for Mobile Radio, B.S.T.J., *44,* October, 1965, pp. 1779–1803.
11. Davenport, W. B., Jr. and Root, W. L., *Random Signals and Noise,* McGraw-Hill Book Co., Inc., New York, 1958, p. 151.
12. *Ibid.,* p. 38.
13. *Ibid.,* p. 160.
14. *Ibid.,* p. 35.
15. Cramer, H., *Mathematical Methods of Statistics,* Princeton University Press, 1946, p. 212.
16. Davenport and Root, *op. cit.* p. 149.

Some Simple Self-Synchronizing Digital Data Scramblers

By J. E. SAVAGE

(Manuscript received July 28, 1966)

Two types of self-synchronizing digital data scramblers and descramblers are introduced and examined. The descramblers recover synchronization quickly after the insertion or deletion of channel bits, and they are relatively insensitive to channel errors. The scramblers act to increase the period of periodic data sequences, and the periodic channel sequences produced have approximately half as many transitions in one period as there are bits in a period. These circuits find application in common carrier systems where short-period data sequences produce high-level tones in the transmission band and, as a consequence, interchannel interference. And they have application when receiver clocks derive synchronization from transitions in the channel signal. A number of variations and modifications of the scramblers which affect their cost and size are considered.

The scramblers and descramblers are similar in construction and consist of linear sequential filters with either feed-forward or feedback paths, counters, storage elements and peripheral logic. The counters, storage elements and peripheral logic monitor the channel sequence but react infrequently so that the scramblers and descramblers behave principally as linear sequential filters.

I. INTRODUCTION AND SUMMARY

In this paper, we present two basic types of self-synchronizing digital data scramblers and descramblers. A scrambler is a digital machine which maps a data sequence into a channel sequence and, when the data sequence is periodic, into a periodic channel sequence with period which is many times the data period. When the source is periodic, the channel sequence produced by the scrambler also has many transitions.

A simple scrambler and one which is often used is a machine which adds a maximal-length shift-register sequence[1,2] to the data signal.

The scrambled data signal is then descrambled by the subtraction of the same maximal-length sequence. While this procedure is simple and easily implemented, it suffers from the serious disadvantage that the insertion or deletion of bits in the channel signal results in a descrambled sequence which is a garbled version of the data signal. The scramblers presented in this paper have the self-synchronizing property* and recover quickly from the insertion or deletion of channel bits.

There are two important applications for our scramblers. In common carrier systems small nonlinearities are present in modulators and demodulators which are used to frequency multiplex a bank of channels. Consequently, high-level tones in one channel may produce interference in other channels as a result of the nonlinearities in the mixing process. For this reason, systems engineers place limits on the levels of isolated tones in a customer's transmission band. Tones, in turn, are produced in digital data transmission systems by periodic data sequences and the limit on tone levels is then translated into a lower bound on the period of periodic channel sequences. Thus, our first application is to insure that any periodic source sequence is mapped by a scrambler onto a periodic channel sequence with sufficiently large period.

The second application for our scramblers concerns the need for transitions in the amplitude of the channel signal so that receiver clocks can derive bit or frame synchronization from the channel sequence. Receiver clocks often derive synchronization by passing the received signal through a filter tuned to the spectral component corresponding to the basic baud length and then observing the zero crossings of the filter output. Since the amplitude of the filter output will decrease to the background noise level if no transitions occur in the amplitude of the received signal or if the density of transitions is small, it is clear that in this application it is desirable to guarantee many closely spaced transitions in the amplitude of channel sequence when the source is periodic.

We introduce two basic types of self-synchronizing, digital data scramblers called multi-counter scramblers and single-counter scramblers and they are discussed in Sections IV and VI, respectively. Each scrambler consists of a "basic scrambler" and a "monitoring logic" which consists of additional storage elements, counters and incidental logic. We show in Section II that the "basic scrambler," which is a linear sequential filter with feedback paths and tap polynomial $h(x)$, responds to a periodic data sequence of period s by producing a periodic

* R. D. Fracassi and T. Tammaru introduced the self-synchronizing descrambler in a special scrambler for which they have a patent pending.[3]

channel sequence whose period is either s or the least common multiple of s and $p^m - 1$, where m is the number of stages in the basic scrambler and p is a prime greater than or equal to the number of elements in the source alphabet. The basic scrambler responds in this way to periodic inputs when its tap polynomial $h(x)$ is primitive over the modular field of p elements, $GF(p)$. The counters, logic, and storage elements of the monitoring logic monitor the channel sequence and respond whenever this sequence has as periods, one of the known data periods. The monitoring logic then reacts and changes the state of the basic scrambler, forcing it to have the long-period output.

We show in Sections IV and V that a multi-counter scrambler exists for binary as well as non-binary sources and we find the smallest thresholds required on counters in the monitoring logic of this scrambler. The single-counter scrambler is considered in Sections VI and VII and because of analytical difficulties we are only able to show the existence of this scrambler when the source is binary ($p = 2$) and the source periods are all prime to $2^m - 1$, where m is the size of the basic scrambler. Mixtures of the scramblers for binary sources are examined in Section VIII.

In Section IX we show that the scrambler output, when the input is periodic, contains many closely spaced transitions and that there are half as many transitions in one period as there are digits in that period. In Section XI we perform representative calculations to determine the spectrum of the scrambler output and find when the source is periodic that the output spectrum has P times as many tones as the unscrambled spectrum and each tone has $1/P$th the energy, where P is the factor by which the source period is increased.

The descramblers for each of the scramblers are discussed in the sections in which the scramblers are introduced and they are also discussed separately in Section X. In that section, we show that the descramblers recover synchronization rapidly after the insertion or deletion of channel digits and we observe that the principal effect of infrequent channel errors on the descramblers is to multiply the number of channel errors by $w(h)$, where $w(h)$ is the number of nonzero terms in the tap polynomial $h(x)$. In Section X we also note that the monitoring logics at the scrambler and descrambler reach threshold infrequently when the source is random and at most once when the source is periodic so that the descrambler monitoring logic may be removed and the descrambler considerably simplified as long as thresholding in the monitoring logic occurs at a tolerably low rate.

An example is given in Section XII of the application of the scramblers and descramblers and representative calculations are performed to determine which scrambler configuration is least expensive. Section XIII closes with conclusions.

II. BASIC SCRAMBLER AND DESCRAMBLER

The shift register circuit shown in Fig. 1(a) is a linear sequential filter with feedback paths[4] and is an example of the scrambling circuit which is basic to the multi-counter scrambler and to the single-counter scrambler discussed in later sections. The linear sequential filter with feed-forward paths[4] shown in Fig. 1(b) is the complementary circuit to that shown in Fig. 1(a) and regenerates the data sequence from the channel sequence. We assume in these two examples that data is presented as a binary sequence, that addition is taken modulo 2 and that the storage elements provide one bit of delay.

Examination of the circuits of Fig. 1 show that they have the required synchronization property since the effect of a bit lost or added in the line sequence is felt only as long as the values stored in the descrambler disagree with those stored at the scrambler, which is five bit intervals in our example.

A more general form for the basic scrambler when the data is assumed to be a sequence of digits from the modular field of p elements, $GF(p) = \{0, 1, \cdots, p - 1\}$, where p is prime, is shown in Fig. 2. Here, addition

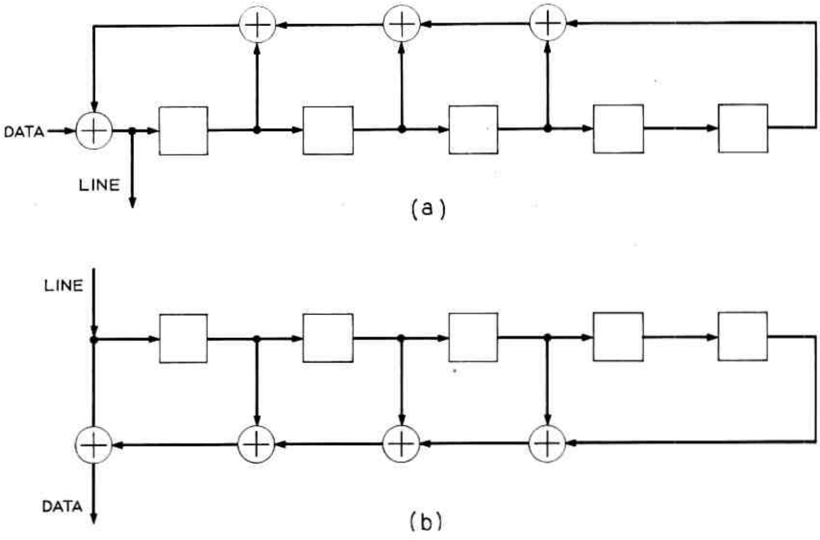


Fig. 1 — (a) A basic scrambler; (b) a basic descrambler.

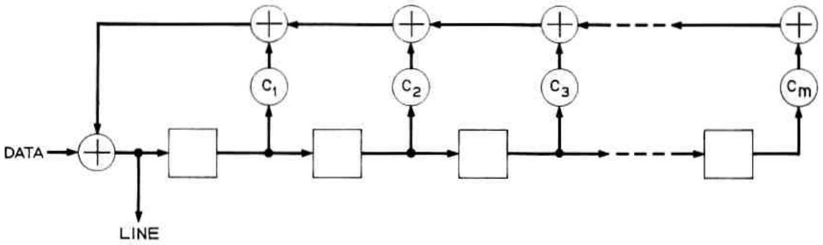


Fig. 2 — General basic scrambler.

is taken modulo p and the outputs of the storage elements are multiplied by the tap constants $\{c_1, c_2, \cdots, c_m\}$ drawn from $GF(p)$. Here, multiplication is also taken modulo p. The tap constants must be constrained in a particular way if our scramblers are to extend the period of periodic sequences in the desired manner. Namely, the tap polynomial $h(x)$ in the indeterminate x given below

$$h(x) = x^m - c_1x^{m-1} - \cdots - c_m \tag{1}$$

must be a primitive polynomial* over the field $GF(p)$. This condition will guarantee that the sequence generated by the basic scrambler in the absence of input will be either all zero or a maximal length sequence, that is a sequence which repeats but once every $p^m - 1$ digits. In the example given in Fig. 1, the tap polynomial is primitive over the binary field and it will generate a maximal length sequence of period $2^5 - 1 = 31$. ($h(x)$ is a primitive polynomial of degree m over the field $GF(p)$ if it is irreducible, that is, has no factors except 1 and itself, and if it divides $x^n - 1$ for $n = p^m - 1$ but does not divide it for any smaller n.)

Theorem 1: The basic scrambler described above when excited by a periodic sequence of period† s will respond with a periodic line sequence which has either period s or a period which is the least common multiple (LCM) of s and $p^m - 1$ ($LCM(s, p^m - 1)$). The period with which the scrambler responds is a function of the initial values stored in the scrambler storage elements, that is, its initial state, and there is but one such state (for each phase of the input sequence) for which the line sequence has period s. For all other such initial states the line sequence has the larger period.

This theorem is basic to all later results. It states that for only one starting state will the basic scrambler respond with period s to a data

* A nonprimitive polynomial may produce more than two output periods for an input period (see Theorem 1).

† A sequence will be said to be of period s if it has no smaller period.

sequence of that period. Thus, our objective, which is to extend the period of periodic sequences, is equivalent to detecting whether data preceding a periodic sequence has left the scrambler in the critical state for that sequence. Two basic methods of detecting the presence of the critical starting state when sequences of different periods are expected in the data are given in later sections.

III. PROOF OF THE BASIC SCRAMBLER THEOREM

Model a periodic input to the basic scrambler with a circulating register, as shown in Fig. 3 for an input of period 3. The initial state of the circulating register will be the first period of the periodic sequence. We let the vector **y** represent the state of the new circuit. Thus, if the input has period s and the basic scrambler has m stages, then **y** has $s + m$ components where the first s represent (in reverse order) the first period of the periodic input and the last m components represent the values stored in the corresponding storage elements when the periodic sequence begins. For example, $\mathbf{y} = (101101001)$ if the basic scrambler has the stored values 01001 when the sequence 1101, 1101, 1101, $\cdots$ arrives.

The circuit of Fig. 3 is linear since the next set of stored values is a linear combination of the preceding set. Thus, the state $\mathbf{y}'$ following **y** can be found by a matrix operation on **y** by the matrix T given below,* that is, $\mathbf{y}' = T\mathbf{y}$ where **y** and $\mathbf{y}'$ are taken to be column vectors.

$$T = \left[\begin{array}{ccc|ccccc} 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ \hline 0 & 0 & 1 & 1 & 1 & 1 & 0 & 1 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \end{array}\right]. \tag{2}$$

For the general basic scrambler and an input of arbitrary period,

* For an excellent discussion of the matrix approach to linear sequential switching circuits see B. Elspas, The Theory of Autonomous Linear Sequential Networks, IRE Trans. Circuit Theory, *6*, pp. 45–60, 1959, which is reprinted in Ref. 13.

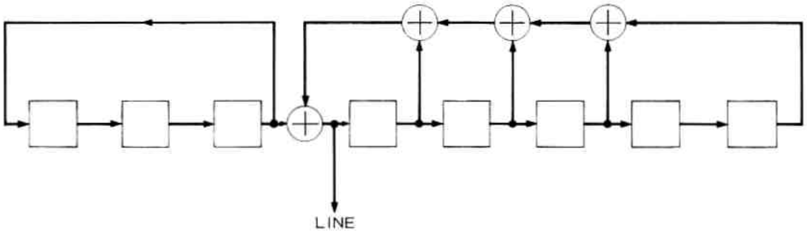


Fig. 3—Model of basic scrambler with periodic input.

say s, the matrix T has the following form

$$T = \left[\begin{array}{c|c} R & \underline{0} \\ \hline \underline{0}^{t} & T_h \end{array}\right], \tag{3}$$

where R is $s \times s$ and is shown below

$$R = \begin{bmatrix} 0 & 0 & 0 & \cdots & 1 \\ 1 & 0 & 0 & \cdots & 0 \\ 0 & 1 & 0 & \cdots & 0 \\ \vdots & & & & \vdots \\ 0 & 0 & & \cdots & 0 \end{bmatrix}. \tag{4}$$

T_h is $m \times m$ and is given as

$$T_h = \begin{bmatrix} c_1 & c_2 & \cdots & c_m \\ 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & 0 \end{bmatrix}. \tag{5}$$

Since the state $\mathbf{y}'$ is found from $\mathbf{y}' = T\mathbf{y}$, all succeeding states are found by taking powers of T, that is, the ith state succeeding $\mathbf{y}$ is

$$\mathbf{y}_i = T^i\mathbf{y}. \tag{6}$$

The line sequence generated by a periodic input to the basic scrambler is periodic of the same period as the state $\mathbf{y}_i$ of the circuit which models the basic scrambler and periodic input. Thus, we prove Theorem 1 by studying the cycles of (6).

There is an indirect approach[5] that one can take to study the cycles of (6). It amounts to a proof that these cycles are isomorphic to cycles of a matrix $T*$ obtained from T by eliminating the solitary 1 shown in (3). This amounts to disconnecting the circulating register from the basic scrambler and observing that the basic scrambler, which is a maximal length sequence generator,[6] has period 1 or $p^m - 1$, $m = \deg h(x)$, so that cycles of $T*$ have period s or $LCM(s, p^m - 1)$. The proof that the register can be disconnected amounts to showing that the minimal and characteristic polynomials of T are the same and equal those of $T*$. Then, the elementary divisors of T and $T*$ are the same and their cycles are isomorphic.

Since there is a direct proof of Theorem 1 which contains many results important to the remainder of the paper, we present it here. If the basic scrambler with periodic input starts with state $\mathbf{y}$, then it has a cycle of length g if $T^g\mathbf{y} = \mathbf{y}$. The basic scrambler output will then be periodic with period g. We now ask for those values of g for which $T^g\mathbf{y} = \mathbf{y}$ has a solution. We begin by writing

$$\mathbf{y} = \mathbf{y}_s + \mathbf{y}_m \,, \tag{7}$$

where $\mathbf{y}_s$ is such that its first s components equal those of $\mathbf{y}$ and its last m components are zero. The vector $\mathbf{y}_m$ is zero in its first s components and is equal to $\mathbf{y}$ in its last m components. We can interpret $\mathbf{y}_m$ as the "starting state" of the basic scrambler and $\mathbf{y}_s$ as the state of the model for the periodically driven basic scrambler when the starting state is zero.

If

$$T^g\mathbf{y} = \mathbf{y} \tag{8}$$

then

$$-T^g\mathbf{y}_s + \mathbf{y}_s = T^g\mathbf{y}_m - \mathbf{y}_m \tag{9}$$

since T is a linear operator. We assume that the periodic input is fixed and has period which is strictly s. Then, the left-hand side of (9) is fixed and we ask whether a solution $\mathbf{y}_m$ for it exists for a given value of g. We note that

$$T^g = \left[\begin{array}{c|c} R^g & 0 \\ \hline * & T_h^g \end{array}\right], \tag{10}$$

where the asterisk indicates some submatrix. Therefore, $T^g\mathbf{y}_m - \mathbf{y}_m$ is a vector whose first s components are zero. The left-hand side of (9) has its first s components zero only when g is a multiple of s because in that case $R^g = I_s$, the $s \times s$ identity matrix, and otherwise $R^g -$

$I_s \neq 0$ which means that a cyclic shift of the first s components of $\mathbf{y}_s$ when added to $\mathbf{y}_s$ is nonzero unless $g = ks$ for some integer $k \geqq 1$.

If we use the notation $(\mathbf{y})'$ to indicate the last m components of $\mathbf{y}$ we have for (9) the following when $g = ks$

$$(-T^{ks}\mathbf{y}_s + \mathbf{y}_s)' = [T_h^{ks} - I](\mathbf{y}_m)' \tag{11}$$

where I is now $m \times m$. We use the following theorem on (11).

Theorem 2: The matrix T_h has characteristic polynomial $h(x)$ which is assumed primitive over $GF(p)$. Therefore, $T_h^i - I$ is nonsingular for $i = 1, 2, \cdots, p^m - 2$ and $T^n = I$ for $n = p^m - 1$.

Proof: See appendix.

Since $T_h^n = I$ for $n = p^m - 1$, we can reduce ks modulo n so that T_h^{ks} can be written as a power of T_h less than n. In particular for $k < k_0$ where k_0s is the least common multiple of s and n, which we call e, that is,

$$e = k_0s = LCM(s, p^m - 1), \tag{12}$$

the matrices T_h^{ks} can be written as $T_h^{i_k}$ where $0 < i_k < n$. We have

$$T_h^{k_0s} = T_h^e = (T_h^n)^{e/n} = (I)^{e/n} = I. \tag{13}$$

Returning to (11) we see that $T^{ks} - I$ is nonsingular for $1 \leqq k < k_0$. Therefore, when $k = 1$, (11) possesses a unique solution $\mathbf{y}_m$. That is, there exists a unique starting state $\mathbf{y}_m$ for each periodic sequence (modeled by $\mathbf{y}_s$) having period strictly s such that $T^s(\mathbf{y}_m + \mathbf{y}_s) = \mathbf{y}_m + \mathbf{y}_s$. Similarly, there exists a unique solution to (11) for each $2 \leqq k < k_0$. However, if $T^s\mathbf{y} = \mathbf{y}$, $\mathbf{y} = \mathbf{y}_m + \mathbf{y}_s$, then $T^{ks}\mathbf{y} = \mathbf{y}$ so that the cycles having period ks are really repetitions of the single cycle having period s. Also, when $k = k_0$, $T^{k_0s} = I$ and $T^{k_0s}\mathbf{y} = \mathbf{y}$ for all $\mathbf{y}$. We conclude that *for a prescribed input having period strictly s, the basic scrambler will respond with period s for only one starting state* $\mathbf{y}_m$ *and for all other starting states will respond with period e given by* (12). This proves Theorem 1.

We have finished our discussion of the basic scrambler. We now consider the techniques used to detect the presence of a periodic sequence of low period on the line and present the first of two methods for altering the starting state of the basic scrambler. This first method is more general than the second and allows for the simultaneous detection of sequences of several periods. The second method applies only when the sequences expected on the line have periods which divide one of two numbers.

IV. THE MULTI-COUNTER SCRAMBLER

The general form of the multi-counter scrambler (MCS) is shown in Fig. 4. (The descrambler is shown in Fig. 5.) There are N counters, one for each period s_i , $1 \leqq i \leqq N$, and the ith counter will generate $+1$ if it reaches its threshold t_{s_i} . A counter is reset whenever the reset lead is nonzero so that t_{s_1} consecutive zeros on the reset lead of the ith counter will cause it to reach its threshold. All counter outputs are fed to the OR circuit shown so that a 1 is generated at the exclusive OR and added to the "tap sum"* whenever a counter reaches threshold. At the same time, all counters are automatically reset.

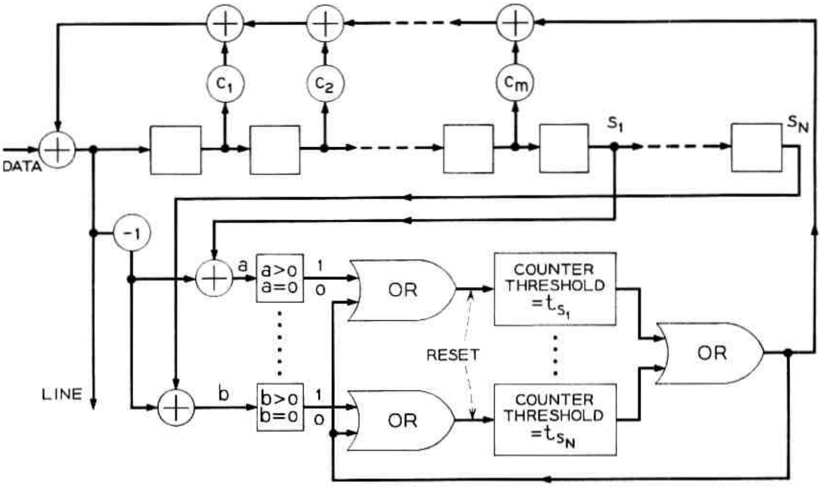


Fig. 4 — Multi-counter scrambler.

The input to the ith counter is the difference between the present line digit and the digit transmitted s_i clock intervals earlier. If the line sequence has period s_i , then these two digits agree and the difference is zero. Then, the ith counter will reach threshold, the tap sum will be altered and the state of the basic scrambler changed.† The line sequence will then be changed from period s_i to period $LCM(s_i , p^m - 1)$ where p is the size of the modular field $GF(p)$ and m is the number of stages in the basic scrambler.

* We define the "tap sum" as the quantity added to the next data bit at the input to the basic scrambler.

† If the starting state of the basic scrambler is critical for a sequence of period s_i , then the state after j clock intervals is critical for the jth cyclic shift of the input sequence. Hence, a change in the tap sum will force the next state to be noncritical.

We observe, then, that the multi-counter scrambler for any choice of thresholds $\{t_{s_i}, 1 \leqq i \leqq N\}$ will force the basic scrambler to switch from a critical state to a noncritical state whenever the input has period $s_1, s_2, \cdots$, or s_N or some period which divides an s_i. It should be clear, however, that it is not necessary and perhaps not desirable to change the tap sum and the next state of the basic scrambler when the output does not have period s_i, $1 \leqq i \leqq N$, or some period which divides an s_i. The next theorem specifies the minimum values of the thresholds t_{s_i}, $1 \leqq i \leqq N$, so that the tap sum is changed only when "necessary." (Note that random data may generate line sequences which resemble periodic sequences and in such cases it will be "necessary" to change the tap sum.)

Theorem 3 (MCS Theorem): The multi-counter scrambler shown in Fig. 4 will scramble a periodic sequence of period s if s divides s_i for some i, $1 \leqq i \leqq N$, and will produce a periodic line sequence of period LCM $(s, p^m - 1)$ if the following two conditions are met:

(i) The tap polynomial $h(x)$ of degree m is primitive over $GF(p)$ where data sequences have components from $GF(p)$.

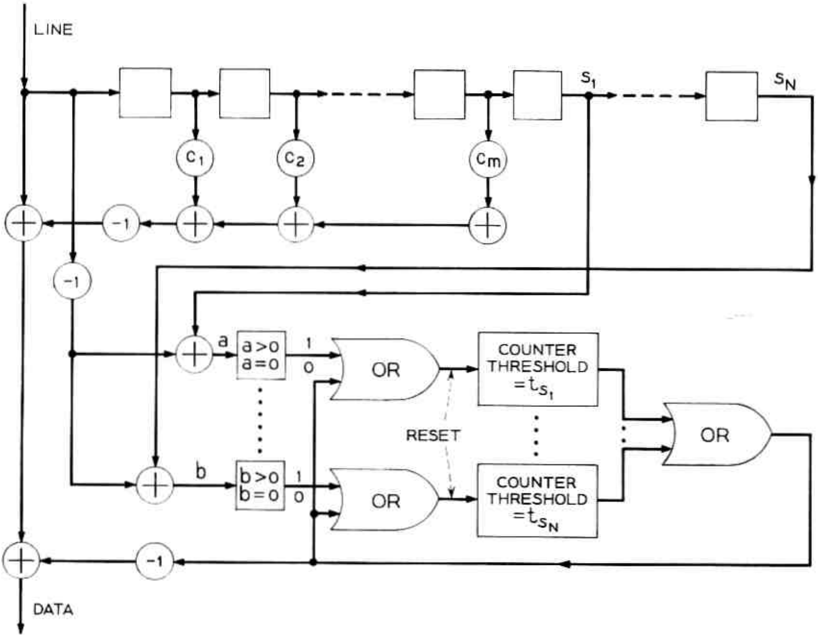


Fig. 5—Multi-counter descrambler.

(ii) The thresholds t_{s_i}, $1 \leqq i \leqq N$, are chosen as

$$t_{s_i} \geqq (m - 1) + \max_{\substack{1 \leq i \leq N \\ j \neq i}} s_j .$$

If all input periods divide s_0, then the theorem holds when condition (i) is met and a threshold of $t_{s_0} \geqq m$ is used.

The descrambler for the MCS, shown in Fig. 5, has the self-synchronization property as long as line errors do not occur. When line errors occur, the counters in the descrambler may not read the same levels as the corresponding counters in the MCS. However, as seen from the MCS theorem, the counters must reset at least once every $(m - 1) + \max_i s_i$ clock intervals when the input is periodic so that counter synchronization is easily established in this case. With random data the situation is not quite so clear. The question of descrambler synchronization, including the effect of channel errors, is considered in detail in Section X.

V. PROOF OF THE MCS THEOREM

We use the notation developed in Section III for the proof of Theorem 3. Fig. 6 shows the basic scrambler with periodic input of period s and one counter. The only input to the basic scrambler other than the data input is the lead from the counters which is used to change the tap sum. The counter shown is assigned to the detection of line sequences whose periods divide s_i.

To prove Theorem 3 we must show that t_{s_i} can be chosen such that

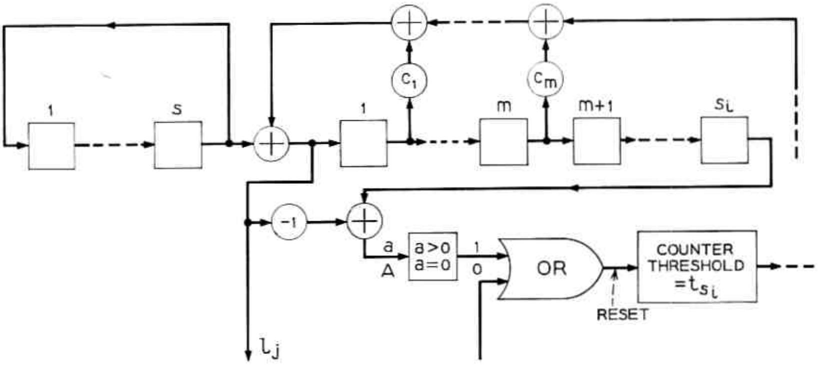


Fig. 6—One counter of the MCS with periodic input.

the ith counter does not reach threshold when the data sequence is periodic unless the line sequence has period s_i or some period which divides s_i. Before we begin our proof we introduce some notation. In Fig. 6, we use l_j to indicate the jth line digit calculated with data from a periodic input of period s. The basic scrambler is shown and we indicate with the vector $\mathbf{y}$ the state of the linear sequential filter composed of the circulating register of s stages and the m stages of the basic scrambler. Call this filter of $s + m$ stages the driven basic scrambler. Then, from (6) the next state of the driven basic scrambler, $\mathbf{y}'$, is

$$\mathbf{y}' = T\mathbf{y} \tag{8}$$

provided that the monitoring logic is not active. If it is active, that is, if one or more counters reach threshold, then

$$\mathbf{y}' = T\mathbf{y} + \mathbf{y}_t \tag{14}$$

where $\mathbf{y}_t$ contains a single one in its $(s + 1)$th position.

The first line digit calculated with the periodic input, l_1 , is

$$l_1 = [T\mathbf{y} + u_1\mathbf{y}_t]_s \tag{15}$$

where

$$[\mathbf{z}]_s = z_{s+1} , \tag{16}$$

the $(s + 1)$th component of the vector $\mathbf{z}$, and

$$u_1 = \begin{cases} 1 & \text{monitoring logic active at first calculation,} \\ 0 & \text{otherwise.} \end{cases} \tag{17}$$

In general, the jth line digit is

$$l_j = \left[T^j\mathbf{y} + \sum_{k=1}^{j} u_k T^{j-k} \mathbf{y}_t \right]_s \tag{18}$$

where

$$u_k = \begin{cases} 1 & \text{monitoring logic active at } k\text{th calculation,} \\ 0 & \text{otherwise.} \end{cases} \tag{19}$$

Now consider the sequence $\{a_j\}$ calculated at point A of Fig. 6. If a run of consecutive zeros in this sequence is large enough, the ith counter will reach threshold unless some other counter reaches threshold before it. When the periodic input begins, the counters in the MCS will be at unknown levels and the $(\max_i s_i)$ stored values will be, in general, unrelated to the input; thus one or more counters may reach threshold before l_1 reaches the s_ith storage element of the MCS.* We now wish to show that *the sequence* $\{a_j\ ,\ j \geqq s_i + 1\}$ *will contain a run of no more than* $\max_{i \neq i} (m - 1 + s_i)$ *zeros if the line sequence is periodic with a period which does not divide* s_i .

We have for $j \geqq s_i + 1$

$$a_j = -l_j + l_{j-s_i} \tag{20}$$

and

$$a_j = -\left[T^{j-s_i}(T^{s_i}\mathbf{y} - \mathbf{y}) + \sum_{k=1}^{j} u_k T^{j-k}\mathbf{y}_t - \sum_{k=1}^{j-s_i} u_k T^{j-s_i-k}\mathbf{y}_t\right]_s . \tag{21}$$

If $u_{s_i+1} = 0$ then let j_0 be such that $u_j = 0$, $s_i + 1 \leqq j < j_0$ and $u_{j_0} = 1$, that is, the most recent thresholding occurs at $j = j_0$. (The case $u_j = 0$ for all $j \geqq s_i + 1$ is trivial, so we assume that $u_{j_0} = 1$ for some j_0 .) Then, if we write $\mathbf{z}_i$ as

$$\mathbf{z}_i = \sum_{k=1}^{j_0} u_k T^{j_0-k}\mathbf{y}_t - \sum_{k=1}^{j_0-s_i} u_k T^{j_0-s_i-k}\mathbf{y}_t \tag{22}$$

and if we ignore all counters except the ith, we have

$$a_j = -[T^{j-j_0}(T^{j_0-s_i}\{T^{s_i}\mathbf{y} - \mathbf{y}\} + z_i)]_s \tag{23}$$

for $j_0 \leqq j \leqq j_0 + t_{s_i} - 1$. For this range of j the a_j can be viewed as the values appearing in the $(s + 1)$th storage element of the driven basic scrambler with starting state $\mathbf{y}_i^*$

$$\mathbf{y}_i^* = T^{j_0-s_i}\{T^{s_i}\mathbf{y} - \mathbf{y}\} + z_i\ . \tag{24}$$

Now, assume that input period s does not divide s_i . Then, the first s components of $T^{j_0-s_i}\{T^{s_i}\mathbf{y} - \mathbf{y}\}$ are not all zero. Since z_i is zero in its first s components, the starting state $\mathbf{y}_i^*$ is nonzero in some of its first s components. Consequently, the state of the driven basic scrambler (of $s + m$ stages) can never be completely zero† so that the sequence $\{a_j\ ,\ j \geqq j_0\}$ cannot contain more than $s + m - 1$ consecutive zeros if s does not divide s_i .‡ We shall now show that, in fact, the sequence $\{a_j\ ,\ j \geqq j_0\}$ cannot contain more $s + m - 2$ consecutive zeros if $s \nmid s_i$. We shall also show that there exists an input of period s if $s_i = ks \pm 1$

* Note that the lead from the counters will be active at most once during the first s_i calculations if $s_i \leqq \min_j t_{s_j}$.

† Note that the matrix operator T just circulates the first s components of $\mathbf{y}^*$.

‡ We will use the notation $s \nmid s_i$ to mean s does not divide s_i .

for some integer $k \geqq 1$ such that the sequence $\{a_j, j \geqq j_0\}$ will have this many consecutive zeros.

In (24) the first s components of $\mathbf{y}_i^*$ are a cyclic shift of the first s components of $T^{s_i}\mathbf{y} - \mathbf{y}$. Recalling the definition of T from (3) we see that these s components are the components of the vector $(R^{s_i} - I)(\mathbf{y}_s)''$ where R is $s \times s$ and is given by (4) and $(\mathbf{y}_s)''$ is the vector containing the first s components of $\mathbf{y}_s$. The vector $(R^{s_i} - I)(\mathbf{y}_s)''$ cannot contain a single nonzero element if $s \nmid s_i$ as seen below by example: Let $s = 4$, $s_i = 5$ and $(\mathbf{y}_s)$ have components y_1, y_2, y_3, y_4. Then, we have

$$(R^{s_i} - I)(\mathbf{y}_s)'' = \begin{bmatrix} -1 & 0 & 1 & 0 \\ 0 & -1 & 0 & 1 \\ 1 & 0 & -1 & 0 \\ 0 & 1 & 0 & -1 \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \end{bmatrix} = \begin{bmatrix} b \\ 0 \\ 0 \\ 0 \end{bmatrix}, \tag{25}$$

where b is the single nonzero element. Thus,

$$\begin{aligned} -y_1 + y_3 &= b \neq 0 \\ -y_2 + y_4 &= 0 \\ +y_1 - y_3 &= 0 \\ +y_2 - y_4 &= 0. \end{aligned} \tag{26}$$

It is clear that two equations, the first and third, cannot both be satisfied. This will be true regardless of the location of the single nonzero element. Hence, *there must be at least two nonzero elements in the first s components of* $\mathbf{y}_i^*$. Consequently, $\{a_j, j \geqq j_0\}$ cannot contain a run of more than $s + m - 2$ consecutive zeros. If $\mathbf{y}_s$ contains a single nonzero element and if $s_i = ks \pm 1$ then $(R^{s_i} - I)(\mathbf{y}_s)''$ will contain two consecutive nonzero elements. Also, since $\mathbf{y}_m$ is in general arbitrary, it can be chosen so that the first $s + m - 2$ digits generated with $\mathbf{y}_i^*$ as the starting state will be zero.

At this point, we have shown that a periodic input of period s, where $s \mid s_j$, some $j \neq i$ but $s \nmid s_i$, will not cause the ith counter to reach threshold more than once after the s_ith line digit is transmitted if we choose t_{s_i} to be

$$t_{s_i} = (m - 1) + \max_{j \neq i} s_j . \tag{27}$$

This is true since the sequence generated at point A of Fig. 6 will not show more than t_{s_i} consecutive zeros after the first time the monitoring logic is active following the transmission of the s_ith line digit. We also note that a threshold of the size given above may be necessary if there exists an s such that $s_i = ks \pm 1$, some $k \geqq 1$, where $s \mid s_j$, some j.

Consider next the case where $s \mid s_i$. If the line sequence is periodic of period s (see Theorem 1) at any time after the periodic sequence begins, the sequence at point A will contain an indefinite number of zeros so that the ith counter will definitely reach threshold (unless $s \mid s_j$, some $j \neq i$, and $t_{s_j} < t_{s_i}$, in which case the jth counter may reach threshold first). Since there is only one critical state for each periodic sequence, the change in the tap sum resulting from the detection of the period s line sequence will cause the output to have period $LCM(s, p^m - 1)$. In this case the vector $\mathbf{y}_i^*$ in (24) cannot be entirely zero (its first s components are zero, however) because it would then result in an all zero sequence at point A. Thus, the last m components of $\mathbf{y}_i^*$ must contain at least a single nonzero component. But $[T^{i-i_0}\mathbf{y}_i^*]_s$ (which generates $\{a_j, j \geqq j_0\}$) then is just the output of a maximal length sequence generator (see appendix) so that no more than $m-1$ consecutive zeros will be seen at point A if $s \mid s_i$ and the output has period $LCM(s, p^m - 1)$.

In conclusion, *if $s \mid s_i$ but the output does not have period s or if $s \nmid s_i$ but $s \mid s_j$ some $j \neq i$, then the ith counter will reach threshold at most once after the transmission of the s_ith line digit if the ith threshold t_{s_i} is chosen as*

$$t_{s_i} = (m - 1) + \max_{j \neq i} s_j \, . \tag{28}$$

Of course, the same is true for any threshold larger than t_{s_i}.

VI. THE SINGLE-COUNTER SCRAMBLER

The single-counter scrambler (SCS) is shown in Fig. 7 (and the descrambler is shown in Fig. 8). This scrambler is designed to scramble periodic *binary* sequences whose periods divide either s_1 or s_2 or both. It has a single counter and for some applications may be less costly to build than the multi-counter scrambler. And while we consider the SCS when the input periods divide either s_1 or s_2 or both, one may be able to design for the case of many more input periods.

The SCS has two circuits for detecting periodic sequences. If either or both of the two detecting circuits produces 0 at any one time, one cannot with a single measurement determine whether the line sequence has period s_1 or s_2. On the contrary, if both circuits produce a nonzero

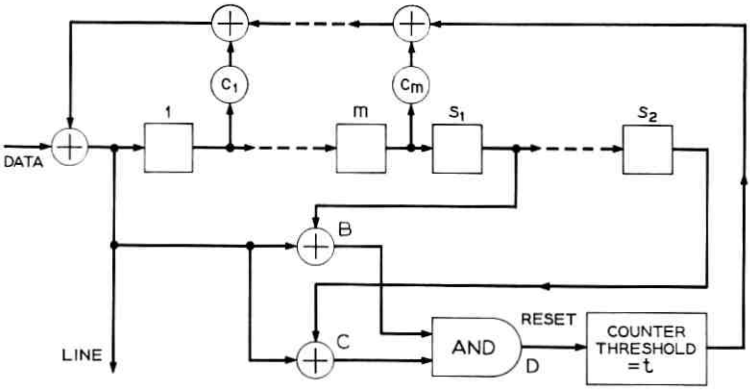


Fig. 7 — Single counter scrambler.

output, it is clear that the line sequence does not have period s_1 or s_2 and that the counter should be reset. A 2-input AND gate has a nonzero output only when both inputs are nonzero, consequently, we use it as input to a counter, as shown in Fig. 7. This counter will reach threshold after t line transmissions if each of t consecutive pairs of outputs of the detecting circuits contains one or more 0's.

The major design problem of the SCS is the choice of the counter threshold. This is not an easy problem, unfortunately, and all that we

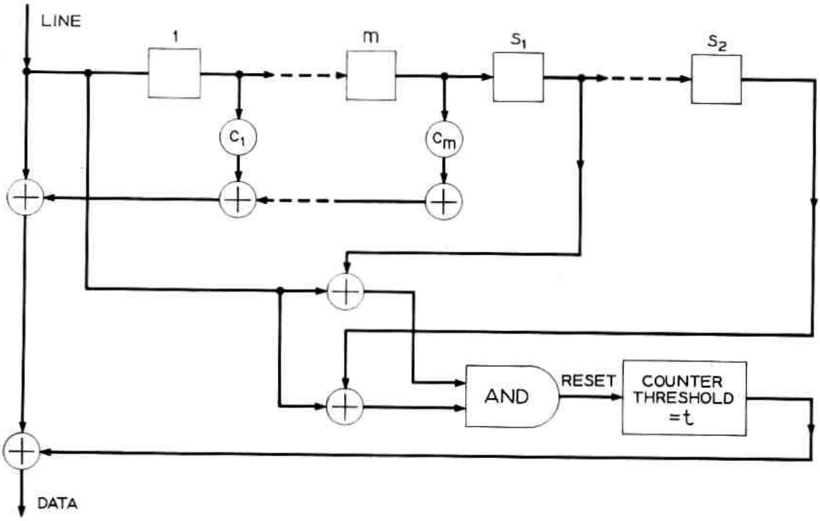


Fig. 8 — Single counter descrambler.

have been able to say about it is that counter thresholds do exist when $p = 2$ (the source is binary) and the input periods are relatively prime to $2^m - 1$ and then to only give a gross upper bound on the smallest permissible threshold. The following theorem states what is known about the threshold for the SCS.

Theorem 4 (SCS Theorem): A single-counter scrambler which will scramble all periodic binary sequences with periods which divide s_1 *or* s_2 $(s_1 < s_2 , s \nmid s_2)$ *exists if*

(*i*) *the tap polynomial* $h(x)$ *of degree* m *is primitive over* $GF(2)$,
(*ii*) s_1 *and* s_2 *are relatively prime to* $2^m - 1$, *and*
(*iii*) *a counter threshold,* t, $t \leqq s_2(2^m - 1) - 2^{m-1} + 2$ *is chosen.*

This theorem does not rule out the possibility that an SCS exists when s_1 and s_2 are not both relatively prime to $2^m - 1$ nor does it rule out an SCS for nonbinary data. It simply states that when conditions (i) and (ii) are met, one can show that a counter with threshold t, $t \leqq s_2(2^m - 1) - 2^{m-1} + 2$, will not reach threshold when the output of the basic scrambler has period $s \times (2^m - 1)$ where s divides s_1 or s_2 or both. In fact, the bound on the threshold required to prevent the counter from reaching threshold prematurely is many times larger than necessary. In the example given in Section XII the bound is more than 35 times too large.

VII. PROOF OF THE SCS THEOREM

For the proof of Theorem 4, we recall the proof of Theorem 3. In particular, it is instructive to review the discussion surrounding equations (20) through (24). We recall that a_j of (20) is the jth digit calculated (after the arrival of the periodic data sequence) at the input to the ith counter of Fig. 6. We argued that if the ith counter reaches threshold on the j_0th calculation, $j_0 \geqq s_i + 1$, then a_j could be calculated from

$$a_j = -[T^{j-j_0}\mathbf{y}_i^*]_s \tag{29}$$

for $j \geqq j_0$ and until the next time the ith counter reaches threshold. Here $[\mathbf{y}_s]$ indicates the $(s + 1)$st component of $\mathbf{y}$ and $\mathbf{y}_i^*$ is given by (24). Thus, the sequence generated at point A of Fig. 6, namely $a_{j_0}, a_{j_0+1}, \cdots$ can be viewed as generated by the basic scrambler with periodic input and $\mathbf{y}_i^*$ as starting state.

In Theorem 4 we assume that the data sequence is binary so that the above equations apply if we interpret subtraction as addition since they are equivalent on the binary field. In Fig. 7 the sequences $\{b_j\}$ and $\{c_j\}$ are generated at points B and C, respectively. We wish to show that the largest run of consecutive zeros in the logical AND of $\{b_j\}$ and $\{c_j\}$ after a certain transient period cannot exceed $s_2(2^m - 1) - 2^{m-1} + 1$ when $s_1 < s_2$ and s_1 and s_2 are both relatively prime to $2^m - 1$.

Consider the sequences $\{b_j, j \geqq s_2 + 1\}$ and $\{c_j, j \geqq s_2 + 1\}$. Then, if the counter reaches threshold at* j_1, $j_1 \geqq s_2 + 1$, these two sequences will cause the counter to reach threshold only once more if the line sequence has period s, $s \mid s_1$ or $s \mid s_2$ or both. If the line sequence has period† $s \times (2^m - 1)$, then neither $\{b_j, j \geqq j_1\}$ nor $\{c_j, j \geqq j_1\}$ can be all zero since this would imply that the line sequence has period s_1 or s_2.

We now consider two cases, case I when s divides both s_1 and s_2 and case II when s divides s_1 but not s_2 or vice versa. From (24) it is clear that in case I both $\{b_j, j \geqq j_1\}$ and $\{c_j, j \geqq j_1\}$ are the outputs of basic scramblers with no input and with nonzero starting states so that they repeat with period $2^m - 1$. In case II when $s \mid s_1$, say, but not s_2, $\{b_j, j \geqq j_1\}$ is the output of a basic scrambler with nonzero starting state and has period $2^m - 1$ while $\{c_j, j \geqq j_1\}$ is the output of a driven basic scrambler with input period s. (The input may not be strictly of period s, however, as we shall see later.)

The logical AND of the sequences generated at points B and C of Fig. 7 can be interpreted as the sequence generated by the normal arithmetic multiplication of b_j and c_j. Thus, the sequence at point D of Fig. 7 has period $2^m - 1$ in case I and period $s \times (2^m - 1)$ in case II. Let $\mathbf{B}_n$ and $\mathbf{C}_n$ be n component vectors with $B_l = b_{j_1+l}$ and $C_l = c_{j_1+l}$. Then, at point D, the vectors $\mathbf{B}_n$ and $\mathbf{C}_n$ generate the n-vector $\mathbf{D}_n = \mathbf{B}_n \cdot \mathbf{C}_n$ where multiplication of $\mathbf{B}_n$ and $\mathbf{C}_n$ is term-by-term, i.e.,

$$\mathbf{D}_n = (B_1C_1, B_2C_2, \cdots, B_nC_n). \tag{30}$$

Let $w(\mathbf{y}_n)$ be the Hamming weight[7] of the n-vector $\mathbf{y}_n$, that is, the number of 1's in $\mathbf{y}_n$. Then we have[8]

$$w(\mathbf{D}_n) = w(\mathbf{B}_n \cdot \mathbf{C}_n) = \frac{w(\mathbf{B}_n) + w(\mathbf{C}_n) - w(\mathbf{B}_n + \mathbf{C}_n)}{2}, \tag{31}$$

where addition is modulo 2. We now wish to use this last equation to find a lower bound on the number of 1's $\mathbf{D}_n$. From this we can obtain an upper bound on the number of consecutive 0's in D_n and an upper

* It may indeed reach threshold for $1 \leqq j_1 \leqq s_2$ but this does not affect our analysis.

† s_1 and s_2 are relatively prime to $2^m - 1$ so that the line sequence has period $s \times (2^m - 1)$ if $s \mid s_1$ or $s \mid s_2$ as seen from Theorem 1.

bound to the threshold required to prevent the counter of Fig. 7 from making unnecessary changes in the tap sum.

In case I we let $n = 2^m - 1$, which is the period of the sequence generated at point D. Thus, $\mathbf{D}_n$ is one period of this sequence. $\mathbf{B}_n$ and $\mathbf{C}_n$ are each one period of the output of the basic scrambler (which is a maximal length sequence generator). Thus, $\mathbf{B}_n$ can be obtained from $\mathbf{C}_n$ by a cyclic shift and they both have the same Hamming weight. Then, we have the following result.

Lemma 1: If $n = 2^m - 1$ *and* $\mathbf{B}_n$ *and* $\mathbf{C}_n$ *are periods of a maximal length sequence with* $\mathbf{B}_n = \mathbf{C}_n$, *then* $w(\mathbf{D}_n) = w(\mathbf{C}_n) = 2^{m-1}$. *If* $\mathbf{B}_n \neq \mathbf{C}_n$, *then* $w(\mathbf{D}_n) = w(\mathbf{C}_n)/2 = 2^{m-2}$.

Proof: We need only show that $w(\mathbf{C}_n) = 2^{m-1}$. From the comments at the end of the appendix we have that the state of the autonomous basic scrambler, as a binary m-tuple, ranges through all $2^m - 1$ nonzero binary m-tuples. Since the first digit of each m-tuple is a line digit, there will be exactly 2^{m-1} 1's in one period of the line sequence generated by the autonomous basic scrambler. (Note that the scrambler does not start with the zero state.) Q.E.D.

In case I, then, the number of consecutive zeros in the sequence at D cannot exceed $2^m - 1 - 2^{m-2}$ and a threshold of $2^m - 2^{m-2}$ will guarantee unnecessary tap sum changes in this case.

Consider now case II where $\{b_j, j \geqq j_1\}$ has period $2^m - 1$ and $\{c_j, j \geqq j_1\}$ is the output of a driven basic scrambler characterized by

$$c_j = [T^{j-j_1}\mathbf{y}_0^*]_s, \tag{32}$$

where $\mathbf{y}_0^*$ is an $(s + m)$-vector which from (24) has the form

$$\mathbf{y}_0^* = T^{j_1-s_2}\{T^{s_2}\mathbf{y} + \mathbf{y}\} + \mathbf{z} \tag{33}$$

where $\mathbf{y}$ is an arbitrary $(s + m)$-vector, except that its first s components model a periodic sequence of strictly period s, and $\mathbf{z}$ is zero in its first s components and arbitrary in its last m components. The first s components of $T^{s_2}\mathbf{y} + \mathbf{y}$ cannot be all zero if $s \nmid s_2$. It may model a periodic sequence of period s_0, however, where $s_0 < s$ and $s_0 \mid s$. In particular, we may have $s_0 = 1$ in which case the first s components of $\mathbf{y}_0^*$ may be 1's and $\{c_j, j \geqq j_1\}$ may have an output of period 1 consisting of the all 1 sequence. If this is true $\mathbf{D}_n = \mathbf{B}_n \cdot \mathbf{C}_n = \mathbf{B}_n$ and $\mathbf{D}_n$ will have no more than $m - 1$ consecutive zeros. If $\{c_j, j \geqq j_1\}$ has period $s_0 > 1$ it will contain no less than a single nonzero component in each period nor no more than $s_0 - 1$ nonzero components in each period.

Thus, if $n = s_0$, $s_0 > 1$, and $\mathbf{C}_n$ represents one period in the output of period s_0, we have

$$1 \leqq w(\mathbf{C}_n) \leqq (s_0 - 1). \tag{34}$$

When $\{c_j, j \geqq j_1\}$ has period $s_0 \times (2^m - 1)$ let $n = s_0 \times (2^m - 1)$ in (31). We now show that for this case

$$s_0(2^{m-1} - 1) \leqq w(C_n) \leqq s_0 \times 2^{m-1}. \tag{35}$$

The state of the driven basic scrambler with input of period s_0 can be represented with an $(s_0 + m)$-vector. There are $s_0 \times 2^m$ admissible state vectors since the last m components are arbitrary and the first s_0 components must be a cyclic shift of the first s_0 components of some other state vector. The last m components of these $s_0 \times 2^m$ vectors range through each of the 2^m m-tuples s_0 times. Since the $(s + 1)$st component of each state vector is a line digit $w(\mathbf{C}_n) \leqq s_0 \times 2^{m-1}$ which is the number of 1's shown in these positions. We also have $w(\mathbf{C}_n) \geqq s_0 \times 2^{m-1} - s_0$ since the components of $\mathbf{C}_n$ are generated by only $s_0 \times (2^m - 1)$ of the $s_0 \times 2^m$ admissible state vectors and the missing state vectors may all contain 1 in the $(s + 1)$st components.

Returning to (31) we see that the vector $\mathbf{B}_n + \mathbf{C}_n$ appears. It represents the first n components of $\{b_j + c_j, j \geqq j_1\}$. This is the output sequence of a driven basic scrambler driven with period s_0 and which has as a starting state the state which produces $\{c_j, j \geqq j_1\}$ and which is modified by the addition in its last m components of the starting state of the autonomous basic scrambler which produces $\{b_j, j \geqq j_1\}$. Since this last state is arbitrary, $\mathbf{B}_n + \mathbf{C}_n$ can be expected to have period s_0 or $s_0 \times (2^m - 1)$ and the bounds on the weight of $\mathbf{C}_n$ for these two periods apply to $\mathbf{B}_n + \mathbf{C}_n$.

We now combine our bounds with (31) to obtain a lower bound to $w(\mathbf{D}_n)$ for case II. Remember that $n = s_0(2^m - 1)$.

(*i*) Let $\mathbf{C}_n$ have period s_0, then $\mathbf{B}_n + \mathbf{C}_n$ has period $s_0 \times (2^m - 1)$ and

$$w(D_n) \geqq \frac{s_0 \times 2^{m-1} + (2^m - 1) - s_0 \times 2^{m-1}}{2} = \frac{2^m - 1}{2} \tag{36}$$

where $w(B_n) = s_0 \times 2^{m-1}$ from Lemma 1.

(*ii*) Let $\mathbf{C}_n$ have period $s_0 \times (2^m - 1)$ and $\mathbf{B}_n + \mathbf{C}_n$ have period s_0. Then

$$w(D_n) \geqq \frac{s_0 \times 2^{m-1} + s_0(2^{m-1} - 1) - (s_0 - 1)(2^m - 1)}{2} = \frac{2^m - 1}{2}\cdot \tag{37}$$

(*iii*) Let $\mathbf{C}_n$ and $\mathbf{B}_n + \mathbf{C}_n$ have period $s_0(2^m - 1)$. Then

$$w(\mathbf{D}_n) \geqq \frac{s_0 \times 2^{m-1} + s_0(2^{m-1} - 1) - s_0 \times 2^{m-1}}{2} \geqq \frac{s_0}{2}(2^{m-1} - 1). \quad (38)$$

Therefore, the number of 1's in the sequence D_n of $n = s_0 \times (2^m - 1)$ components for case II must exceed $(2^m - 1)/2 - \frac{1}{2}$ and the number of consecutive zeros cannot exceed $s_0 \times (2^m - 1) - (2^m - 1)/2 + \frac{1}{2}$.

Combining the results for cases I and II we find that the number of consecutive zeros at point D of Fig. 7 when $s_1 < s_2$, $s_1 \nmid s_2$ and the input has period s, $s \mid s_1$ or $s \mid s_2$ or both, will not exceed $s_2(2^m - 1) - (2^m - 1)/2 + \frac{1}{2}$ unless the line sequence has period s. The threshold then need not be any larger than $s_2(2^m - 1) - 2^{m-1} + 2$ to prevent unnecessary changes in the tap sum. Q.E.D.

VIII. MIXTURES OF THE SCRAMBLERS

The two types of scramblers given above are distinguished by the structure of their monitoring logics. The MCS has one counter for each of the input periods $s_1, s_2, \cdots s_N$ and the SCS has a single counter to detect the presence of one of two periods, s_1 or s_2. We have found the smallest threshold required on each counter of the MCS so that they change the tap sum only when necessary. Also, we have shown the existence of a finite threshold on the single counter of the SCS when the source is binary and input periods are relatively prime to $2^m - 1$, where m is the number of stages in the basic scrambler.

Since the monitoring logic for both counters acts to detect the presence of periodic sequences of known periods in the line sequence, it should be clear that a monitoring logic containing a mixture of the MCS logic and the SCS logic may be used. We know of an SCS monitoring logic only when the source is binary, however, so that the mixture must be restricted to the binary source case. Thus, we may now consider a scrambler with a monitoring logic, a portion of which has counters detecting the presence of one of a pair of periods and another portion consisting of individual counters for single periods. The outputs of all counters are fed to an OR gate which in turn is added modulo 2 to the tap sum. The output of the OR gate is also used to reset all counters.

IX. TRANSITIONS IN A SCRAMBLED SEQUENCE

The basic scramblers described above may have applications in situations where bit framing at the receiver is derived from transitions in the line signal. In this section we show that transitions occur frequently in a scrambled periodic sequence and that in one period of a scrambled sequence there are approximately half as many transitions as there are digits. These results are shown when the source is binary and the scrambler input periods are relatively prime to $2^m - 1$, where m is the size of the basic scrambler.

Let $\mathbf{l}$ represent one period of the line sequence generated by the basic scrambler when the input has period s. If the source is binary, if the basic scrambler has m stages and if s is relatively prime to $2^m - 1$, then $\mathbf{l}$ is an $s(2^m - 1)$ component vector. If we assume that the binary line sequence is converted into a line signal by the mapping $1 \rightarrow 1$ $0 \rightarrow -1$, and if it is linearly modulated, then transitions in the channel signal occur whenever transitions in the line sequence appear. Thus, we should like to know the number of transitions in $\mathbf{l}$ and the maximum separation between transitions.

Theorem 5: The binary vector $\mathbf{l}$ *of length* $s(2^m - 1)$ *representing the response of a binary scrambler to an input of period* s, *when* s *and* $2^m - 1$ *are relatively prime, has at least one transition every* $s + m$ *digits and has a total of* $Tr(\mathbf{l})$ *transitions where*

$$\frac{1}{2}\left(\frac{2^m - 2}{2^m - 1}\right) \leqq \frac{\mathrm{Tr}\,(\mathbf{l})}{s(2^m - 1)} \leqq \frac{1}{2}\left(\frac{2^m}{2^m - 1}\right). \tag{39}$$

We begin by showing that every set of $s + m$ consecutive line digits must contain at least one transition. The scrambled sequence is the response of the basic scrambler of Fig. 2 to an input of period s. We note that if the basic scrambler is in the all zero state then the tap sum (which is added to the data bit) is zero. Similarly, if it is in the all 1 state the tap sum is zero because if not, $h(1) = 0$ and $h(x)$ is divisible by $x - 1$ which is impossible since $h(x)$ is irreducible. Then, if $s + m$ consecutive outputs of the scrambler are identical, the last s of the $(s + m)$ corresponding tap sums are zero so that s consecutive data bits must be identical. This cannot happen if the source is periodic with period greater than 1. When $s = 1$, the line sequence must have period 1 if $s + m$ consecutive line digits are identical, which also cannot happen since the line sequence has period $2^m - 1$ in this case.

We now bound Tr $(\mathbf{l})$, the number of transitions in one period, $\mathbf{l}$, of the line sequence. We use the notation of Section V so that the jth digit of $\mathbf{l}$, namely l_j is written

$$l_j = [T^j\mathbf{y}]_s\,, \tag{40}$$

where T is given by (3) through (5) and $\mathbf{y}$ is the state of the driven basic scrambler at the beginning of a period of the data sequence. Let us now observe that a *transition occurs in between two digits in* $\boldsymbol{l}$ *if they sum to* 1 *modulo* 2. Thus, the number of 1's in $\boldsymbol{l} + \boldsymbol{l}'$ (where $\boldsymbol{l}'$ is one cyclic shift of $\boldsymbol{l}$ and addition is term-by-term) is the number of transitions in $\boldsymbol{l}$. For example, if $\boldsymbol{l} = 10110$, $\boldsymbol{l}' = 01011$ and $\boldsymbol{l} + \boldsymbol{l}' = 11101$ then the number of transitions in $\boldsymbol{l}$, including the implicit transition at the first digit is the Hamming weight of $\boldsymbol{l} + \boldsymbol{l}'$.

In the process of proving Theorem 4 we have shown (see (35)) that the Hamming weight of one period of the output of the basic scrambler when the input is binary of period s_0 and s_0 and $2^m - 1$ are relatively prime lies between $s_0(2^{m-1} - 1)$ and $s_0 2^{m-1}$. Hence, if we can show that $\boldsymbol{l} + \boldsymbol{l}'$ is one period of the output of the scrambler with input period s_0, we will have established Theorem 5.

We note that

$$l_i + l_i' = [T^i\mathbf{y} + T^{i-1}\mathbf{y}]_s \tag{41}$$

so that we now examine $T^{i+1}\mathbf{y} + T^i\mathbf{y}$. We have

$$T^i\mathbf{y} + T^{i-1}\mathbf{y} = T^{i-1}(T + I)\mathbf{y}. \tag{42}$$

As in (7), let $\mathbf{y} = \mathbf{y}_s + \mathbf{y}_m$ where $\mathbf{y}_s$ is zero in its last m components, $\mathbf{y}_m$ is zero in its first s components and they represent the periodic input and starting state of the basic scrambler, respectively. Then,

$$(T + I)\mathbf{y} = \mathbf{y}_s + \mathbf{y}_s' + (\underset{s}{\underline{0}}, y_s, \underset{m-1}{\underline{0}}) + \mathbf{y}_m + T\mathbf{y}_m, \tag{43}$$

where $\mathbf{y}_s'$ is a single cyclic shift of $\mathbf{y}_s$ in its first s places and $(\underline{0}, \mathbf{y}_s, \underline{0})$ is a vector with a single component y_s in the $(s + 1)$st position. If we use $(\mathbf{z})'$ to represent the last m components of $\mathbf{z}$, then

$$(\mathbf{y}_m + T\mathbf{y}_m)' = (T_h + I_m)(\mathbf{y}_m)'. \tag{44}$$

In the appendix it has been established that $T_h + I_m$ is a nonsingular matrix. From this we deduce that the last m components of $(T + I)\mathbf{y}$ range over all 2^m m-tuples as $\mathbf{y}_m$ ranges over all m-tuples.

Now consider $\mathbf{y}_s + \mathbf{y}_s'$, which represents the first s components of $(T + I)\mathbf{y}$. While $\mathbf{y}_s$ models one period of a data sequence with period exactly s, $\mathbf{y}_s + \mathbf{y}_s'$ may model a sequence with period s_0, $s_0 \mid s$. For example, let $\mathbf{y}_4 = (1001000)$, then $\mathbf{y}_4 + \mathbf{y}_4' = (0101000)$ and its first 4 components represent two periods of a period 2 sequence. Thus, we must view $(T + I)\mathbf{y}$ as the starting state of a driven basic scrambler with input period s_0 where $s_0 \mid s$. We then ask if the sequence generated by this state has period s_0 or $s_0(2^m - 1)$. Since $\mathbf{y}$ is noncritical, the sequence generated by $(T + I)\mathbf{y}$ must have the larger period because if $\mathbf{y}$ were critical $(T^i + I)\mathbf{y} = 0$ for some i, $1 \leqq i \leqq s_0(2^m - 1) - 1$ and $(T^i + I)(T + I)\mathbf{y} = 0$ as well for some i in this range so that $(T + I)\mathbf{y}$ is a critical state. But we have shown in Theorem 1 that there is only one critical state for each periodic input. In the last paragraph we have seen that there is a one-to-one mapping between the last m components of $\mathbf{y}$ and the last m components of $(T + I)\mathbf{y}$, hence if $\mathbf{y}$ is noncritical, $(T + I)\mathbf{y}$ is noncritical and the line sequence generated by $(T + I)\mathbf{y}$ has period $s_0(2^m - 1)$ where $s_0 \mid s$.

The vector $\boldsymbol{l} + \boldsymbol{l}'$ contains s/s_0 periods of a sequence of period s_0. Let $\mathbf{C}_n$ represent one such period. Then from (35), the number of 1's in $\mathbf{C}_n$, $w(\mathbf{C}_n)$, is bounded by

$$s_0(2^{m-1} - 1) \leqq w(\mathbf{C}_n) \leqq s_0 2^{m-1}. \tag{35}$$

Then,

$$\mathrm{Tr}\,(\boldsymbol{l}) = w(\boldsymbol{l} + \boldsymbol{l}')$$

and

$$s(2^{m-1} - 1) \leqq \mathrm{Tr}\,(\boldsymbol{l}) \leqq s2^{m-1} \tag{45}$$

which gives the desired result after division by $s(2^m - 1)$.

X. THE SELF-SYNCHRONIZING DESCRAMBLERS

In this section, we show that the descrambler for each of the scramblers given above has the self-synchronizing property, that it is relatively insensitive to channel errors and that in some applications it can be considerably simplified by removal of the monitoring logic.

Each scrambler is of the form shown in Fig. 9. Each descrambler can be represented as shown in Fig. 10. The output marked "data" in Fig. 10 is indeed data if the scrambler and descrambler are both started in the same state and no channel errors occur since (*i*) the line sequence will then pass through both basic scramblers and (*ii*) the modulo p sum of a data bit, tap sum, line bit, and monitoring logic output is zero at both the scrambler and descrambler.

If there are no channel errors we would like to show that the descrambler will synchronize itself should it ever lose synchronism. The descrambler will be said to be out of synchronism with the scrambler if either the values stored in the basic scrambler and the delay elements differ from those stored in corresponding sections of the scrambler or if the counters in the monitoring logic are not at the same levels as those at the scrambler or both. It is clear that the s_N stages (if the largest

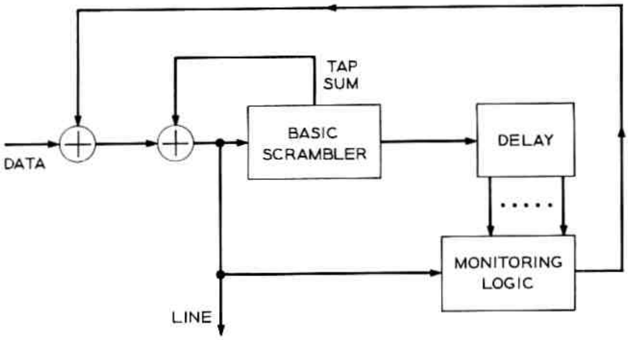


Fig. 9 — Block diagram of the scrambler.

expected period is s_N) of the basic scrambler in the descrambler and delay section will be purged after s_N clock intervals and replaced with accurate information if there are no channel errors. Then, after s_N clock intervals the monitoring logic at the scrambler and descrambler both are fed the same information. The monitoring logics will then reach synchronism when either (*i*) counters at the scrambler and descrambler reach threshold together in which case all counters are reset simultaneously or (*ii*) the last $s_N + 1$ digits of the line sequence is found to be inconsistent with a periodic sequence of period s_1, s_2, $\cdots$ or s_N and the counters at the descrambler are reset individually but in synchronism with those at the scrambler. When the data sequence is *periodic* of period s_1, s_2, $\cdots$ or s_N the *i*th counter of the MCS is reset (following the transient interval associated with the arrival of

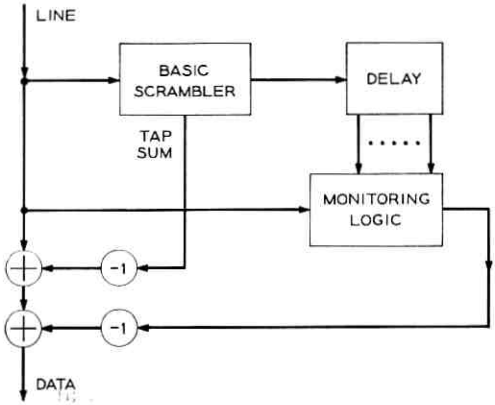


Fig. 10 — Block diagram of the descrambler.

the periodic sequence) at least once every $t_{s_i} = m - 1 + \max_{j \neq i} s_j$ clock intervals. With the SCS the single counter is reset at least once every $s_2(2^m - 1) - 2^{m-1} + 2$ clock intervals when the input has period s_1 or s_2, $s_1 < s_2$. Should the input sequence be *random*, the monitoring logics may be brought into synchronism because one of the counters reaches threshold and all counters are reset, which is unlikely, or because the counters are reset individually in synchronism with the scrambler counters, which is very probable and increases in probability very rapidly to one. (If the source is binary with independent, equiprobable outputs, the ith counter of the MCS descrambler is resynchronized in the second manner after n clock intervals with probability $1 - 2^{-n}$; similarly, the counter of the SCS descrambler is resynchronized with probability $1 - (\frac{3}{4})^n$.)

Channel errors can affect the process of resynchronization. However, if we assume that they are relatively few in number, say, occurring once in every 10^5 transmissions, there will be long intervals during which resynchronization can take place. Since the descrambler requires at most $s_N + \max t_{s_i}$ (which equals $2s_N + m - 1$ in the MCS case when $s_i \leqq s_N$ and is at most $s_2 2^m - 2^{m-1} + 2$ in the SCS case) clock intervals to resynchronize when the source is periodic, resynchronization will not be a problem with periodic inputs if m and s_N (or s_2) are reasonable in size. When the source output is random and is a sequence of independent, equiprobable binary digits, the average number of clock cycles required by the ith counter of the MCS descrambler to resynchronize (in the second way described in the preceding paragraph) is two so that the MCS descrambler will resynchronize on the average in $s_N + 2$ clock intervals. The counter of the SCS descrambler will require four clock intervals on the average to resynchronize so that the SCS descrambler will be resynchronized on the average in $s_N + 4$ clock intervals. Hence, we may conclude that resynchronization in the presence of channel errors which are relatively few in number will not be a problem when the source is random. In fact, it may be easier to resynchronize when the data is random than it is when the data is periodic.

Now assume that the scrambler and descrambler are operating in synchronism and consider the effect of channel errors on the descrambler output. If we neglect the monitoring logic for a moment, it will be seen that an isolated channel error, as it passes through the basic scrambler, will cause $w(h)$ output errors, where $w(h)$ is the number of nonzero terms in the tap polynomial $h(x)$. The monitoring logic, however, may fail to act when it should or act when it should not and thereby introduce additional errors. If we consider the effect of a single channel error on the monitoring logic, we see that this error has a direct effect on the ith counter of the MCS at two occasions, when it enters the basic scrambler and when it reaches the s_ith storage element. A single channel error has a direct effect on the counter of the SCS three times, once when it enters the basic scrambler and again when it enters the s_1th and s_2th storage elements. When the channel error effects a counter of the descrambler, it may cause it to reset when it should not, which will not cause any harm if the counter is about to be reset before reaching threshold, as is the case for the known periodic inputs or as frequently happens with a random source. A channel error which causes a counter to continue to count when it should reset may indeed be harmful since it may result in its reaching threshold and introduce an unnecessary change in the descrambler output. This event is unlikely to happen for the known periodic source sequences since the counters reset frequently, and the number of clock intervals between a set of three normal counter resets is often less than a given counter threshold. It is also unlikely that a channel error will eliminate a reset and cause a counter to reach threshold when the source is random. For example, when the source is a binary, equiprobable, independent letter source the average separation between three resets on the MCS counters is four clock intervals and is eight clock intervals on the SCS. We may conclude then *that channel errors have a small effect on the monitoring logic and thus affect the descrambler primarily by producing approximately $w(h)$ as many output errors as channel errors.*

The descrambler can be considerably simplified, the problem of synchronization loss in the descrambler monitoring logic eliminated, and the problem of output errors due to the monitoring logic solved, all by the removal of the monitoring logic at the descrambler. This is not the drastic solution that it might seem for the monitoring logic reacts infrequently on random data and at most twice on known periodic inputs (if counter thresholds all are larger than the largest expected input period). With a binary, independent, equiprobable letter source, one or more of the N counters of the MCS reaches threshold in n transmissions with a probability, $P_M(n)$, which is less than or equal to

$$P_M(n) \leqq \sum_{i=1}^{N} (n - t_i + 1)2^{-t_i} , \tag{46}$$

where t_i is the threshold on the ith counter and $t_i \geqq t_{s_i}$. The single counter of the SCS reaches threshold t in n transmissions with probability $P_S(n)$ where

$$P_S(n) \leqq (n - t + 1)(\tfrac{3}{4})^{-t}. \tag{47}$$

Hence, if the thresholds are large enough so that $P_M(n)$ or $P_s(n)$ is less than 0.1, say, when n equals the average number of transmissions between channel errors, then we may safely say that the monitoring logic at the descrambler is not necessary on random data inputs.

When the source is periodic of period s however, one of the p^m starting states* of the basic scrambler will result in a line sequence of period s which subsequently will require at least one and at most two outputs from the monitoring logic. Thus, if the data preceding a periodic input is random, the monitoring logic at the descrambler will with probability $1/p^m$ change at least 1 digit in the descrambler output. Hence, if a customer can tolerate such an error rate and if the thresholds are large enough, the monitoring logic at the descrambler can be removed and the descrambler will then simply consist of a basic scrambler.

XI. THE SPECTRUM OF THE SCRAMBLER OUTPUT

In this section, we perform representative calculations to show the effect of scrambling on the spectrum of a linearly modulated carrier. Assume that the source is binary and that a binary sequence is converted into a waveform by the mapping $0 \rightarrow -1$, $1 \rightarrow +1$. Let T_0 be the time interval alloted to each binary digit and let $\hat{l}(t)$ be the waveform generated by the binary sequence $\boldsymbol{l}$. Then, we have

$$\hat{l}_1(t) \cdot \hat{l}_2(t) = -\widehat{(\boldsymbol{l}_1 + \boldsymbol{l}_2)}(t) \tag{48}$$

where addition is taken modulo 2 and multiplication is on the reals.

The autocorrelation function of a waveform $\hat{l}(t)$ is defined as

$$R_l(\tau) = \lim_{T \to \infty} \frac{1}{2T} \int_{-T}^{T} \hat{l}(t) l(t + \tau)\, dt. \tag{49}$$

If $\boldsymbol{l}$ is the output of the scrambler when the input is an equiprobable, independent letter source, then $\boldsymbol{l}$ is a sequence of independent, equiprobable, binary digits. Then, we have

$$R_l(\tau) = \begin{cases} \left(1 - \dfrac{|\tau|}{T_0}\right) & |\tau| \leqq T_0\,, \\ 0 & |\tau| > T_0\,. \end{cases} \tag{50}$$

[*] p is the input alphabet size and m is the number of stages in the basic scrambler.

The power density spectrum of $\hat{l}(t)$, which is the Fourier Transform of $R_l(\tau)$ is for the random binary source

$$S_l(f) = T_0\left(\frac{\sin \pi f T_0}{\pi f T_0}\right)^2. \tag{51}$$

Now let the source be periodic and assume, as an example, that it has period 8 and that the following sequence is one period of the source output: 10110010. Then, if $\boldsymbol{l}$ represents this sequence and if it is transmitted without scrambling, we find using (48) that it has the autocorrelation function, $R_l(\tau)$, of Fig. 11. The power density spectrum of $\hat{l}(t)$, $S_l(f)$, is given below and shown in Fig. 12.

$$S_l(f) = 2T_0\left[\left(\frac{\sin \pi f T_0}{\pi f T_0}\right)^2 - \left(\frac{\sin 2\pi f T_0}{2\pi f T_0}\right)^2\right] \sum_{j=-\infty}^{\infty} \frac{1}{8T_0}\, \delta\left(f - \frac{j}{8T_0}\right). \tag{52}$$

Here $\delta(\cdot)$ is the Dirac delta function. Thus, $S_l(f)$ contains isolated tones spaced by $1/T_1$, $T_1 = 8T_0$, the period of the data sequence.

If the periodic data source of period s is now scrambled, the line sequence has period $T_0(LCM\ (s, 2^m - 1))$. Assume now, as an example, that s and $2^m - 1$ are relatively prime so that the line sequence has period PT_1, $P = 2^m - 1$, the scale-up factor, and $T_1 = sT_0$, the source period. Now let $\boldsymbol{l}$ represent *one period* of the binary line sequence. Then, if $\boldsymbol{l}_k$ represents k cyclic shifts of $\boldsymbol{l}$ we have

$$R_l(kT_0) = \frac{1}{PT_1}\int_{-PT_1/2}^{PT_1/2} \hat{l}(t)\hat{l}_k(t)\, dt. \tag{53}$$

When $k = \pm 1, \pm 2, \cdots, \pm(P - 1)$, we have

$$R_l(kT_0) = -\frac{T_0}{PT_1}\,(\text{No. 1's in } (\boldsymbol{l} + \boldsymbol{l}_k) - \text{No. 0's in } \boldsymbol{l} + \boldsymbol{l}_k). \tag{54}$$

Since $R_l(\tau)$ is linear in τ for $(k - 1)T_0 \leqq \tau \leqq kT_0$ we need only have $R_l(\tau)$ at $\tau = kT_0$, $k = 0, \pm 1, \pm 2, \cdots$. We note that $R_l(kPT_1) = 1$,

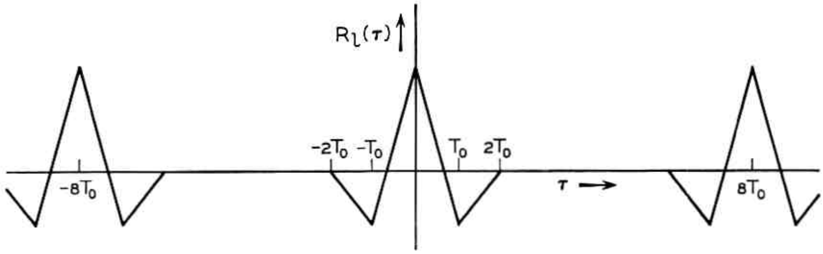


Fig. 11 — Autocorrelation function of period 8 sequence.

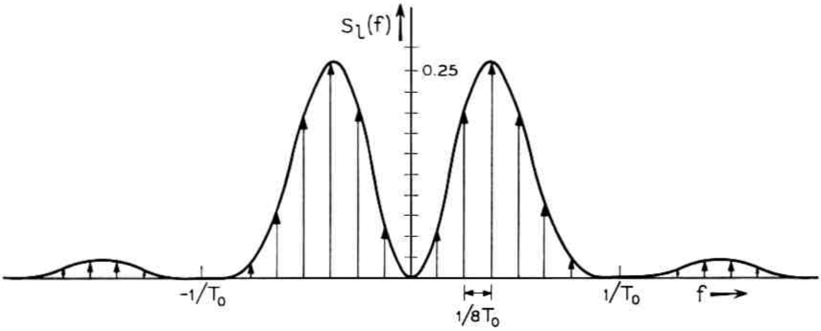


Fig. 12—Spectrum of period 8 sequence.

$k = 0, \pm 1, \pm 2, \cdots$. To further evaluate (54), however, we must return to Section IX.

We have seen in Section IX that $\mathbf{l} + \mathbf{l}_1$ represents several periods in the output of a basic scrambler driven by an input of period s_0 , $s_0 \mid s$, and started with a noncritical state $\mathbf{y}$. The proof of this result amounted to showing that the operation $(T + I)$ on $\mathbf{y}$ mapped the last m components of $\mathbf{y}$ one-to-one onto the last m components of $(T + I)\mathbf{y}$. Thus, if $\mathbf{y}$ is critical so is $(T + I)\mathbf{y}$ and since there is only one critical state for each periodic input $(T + I)\mathbf{y}$ is noncritical if $\mathbf{y}$ is noncritical. We can show in a similar manner that $(T^k + I)\mathbf{y}$ is noncritical when $\mathbf{y}$ is noncritical as long as k is not a multiple of $2^m - 1$. Thus, $\mathbf{l} + \mathbf{l}_k$, which is produced by the starting state $(T^k + I)\mathbf{y}$ when $\mathbf{y}$ generates $\mathbf{l}$, is the output of a basic scrambler with input period s_0 and output period $s_0(2^m - 1)$ when k is not a multiple of $2^m - 1$ and $s_0 \mid s$. Then, invoking (35), we have

$$-1/P \leqq R_l(kT_0) \leqq 1/P, \; k \text{ not a multiple of } P = 2^m - 1. \tag{55}$$

We note, however, that $R_l(kT_0)$ for such k may not all be equal.

Next consider $\mathbf{l} + \mathbf{l}_k$ when k is a multiple of $2^m - 1$. If $\mathbf{l} + \mathbf{l}_k$ represents an output which has period which divides s, then $(T^s + I)(T^k + I)\mathbf{y} = 0$. We now show that $(T^s + I)(T^k + I) = 0$ for all $\mathbf{y}$ when k is a multiple of $2^m - 1$. We observe that

$$T^s + I = \left[\begin{array}{c|c} 0 & 0 \\ \hline Q_s & T_h^s + I_m \end{array}\right] \tag{56}$$

and

$$T^k + I = \left[\begin{array}{c|c} R^k + I_s & 0 \\ \hline Q_k & 0 \end{array}\right] \tag{57}$$

since $T_h^k = I$. Because $T^s + I$ and $T^k + I$ commute, we have

$$(T^s + I)(T^k + I) = (T^k + I)(T^s + I) = \underline{0}. \tag{58}$$

Thus, $\boldsymbol{l} + \boldsymbol{l}_k$ represents a scrambler output of period s_0, where $s_0 \mid s$. It is clear then that the number of 1's in one period of $\boldsymbol{l} + \boldsymbol{l}_k$ is greater than or equal to 1 and less than or equal to $s_0 - 1$. Also, $\boldsymbol{l} + \boldsymbol{l}_k$ is the same sequence for all multiples of $P = 2^m - 1$ which are not multiples of sP. Thus, for k a multiple of P which is not a multiple of sP, we have from (52) that

$$-\frac{(s-2)}{s} \leqq -\left(\frac{s_0 - 2}{s_0}\right) \leqq R_l(kT_0) \leqq \left(\frac{s_0 - 2}{s_0}\right) \leqq \frac{s-2}{s} \tag{59}$$

when $s_0 \geqq 2$.

To calculate a representative spectrum of the scrambled data sequence, we assume that $R_l(\tau)$ has the following form, where u, $2 \leqq u \leqq 2s$, is a function of the scrambler input (the number of 1's in $\boldsymbol{l} + \boldsymbol{l}_k$, k a multiple of P, depends on the input):

$$R_l(kT_0) = \begin{cases} 1 & k = nsP, \quad n = 0, \quad \pm 1, \pm 2, \cdots, \\ \dfrac{s-u}{s} & k = nP, \quad n \neq 0, \quad \pm s, \pm 2s, \cdots, \\ \dfrac{1}{P} & \text{all other } k. \end{cases} \tag{60}$$

The power density spectrum $S_l(f)$ then is

$$S_l(f) = \frac{1}{P}\,\delta(f) + T_0\left(\frac{\sin \pi f T_0}{\pi f T_0}\right)^2 \left\{\frac{u}{SPT_1} \sum_{j=-\infty}^{\infty} \delta\left(f - \frac{j}{PT_1}\right)\right. \\ \left. + \left(1 - \frac{u}{s} - \frac{1}{P}\right) \frac{1}{PT_0} \sum_{j=-\infty}^{\infty} \delta\left(f - \frac{j}{PT_0}\right)\right\}. \tag{60}$$

When u is of the order of s we see that the second term in curly brackets has amplitudes which are proportional to $1/P^2$ and are thus much smaller than terms in the first sum. We show $R_l(\tau)$ with $u = s$, $R_l(T_0) = \epsilon$ in Fig. 13 and $S_l(f)$ in Fig. 14. The assumption that $u = s$ is equivalent to the assumption that $\boldsymbol{l} + \boldsymbol{l}_k$ contains an equal number of 1's and 0's when k is a multiple of P.

We deduce from this discussion of spectra that *the principal effect of scrambling* when the scrambled sequence is converted to a signal waveform in the manner given above *is to increase the number of tones in a given bandwidth by a factor which is approximately P and to decrease the level of each tone by approximately the same factor.*

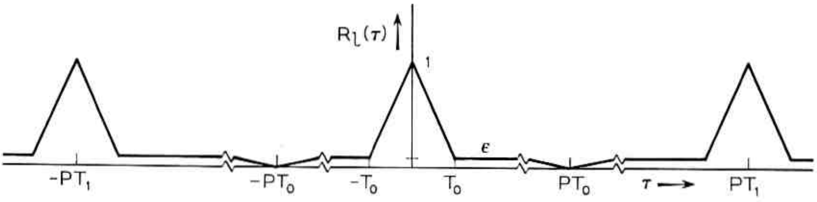


Fig. 13 — Autocorrelation function of a scrambled periodic sequence.

XII. AN EXAMPLE

We shall now consider an application for the scramblers and we shall compare the relative cost and effectiveness of the MCS and the SCS. We will report on a computer simulation directed at the determination of the smallest SCS counter threshold for our example.

Assume that the source is binary and that it may occasionally contain sequences of period 1 (there are two—the all 0 sequence and the all 1 sequence), period 2 (there is only one—the 1010 $\cdots$ sequence, known as dotting), period 7 or period 8. Assume also that a line sequence of period less than 100 is undesirable from spectra considerations. Since the least common multiple of 1 and $2^m - 1$ is $2^m - 1$, we will require that $2^m - 1 > 100$. The smallest value of m for which this is true is $m = 7$ for which $2^m - 1 = 127$, a prime. We next require a primitive, degree 7 binary polynomial for the tap polynomial. The polynomial $h(x) = 1 + x^4 + x^7$ is one such. Given $h(x)$ our basic scrambler is fixed. We next observe that 1 divides 7 and 8 and that 2 divides 8 so that we may build a scrambler which detects two periods $s_1 = 7$ and $s_2 = 8$.

We next consider whether the MCS or the SCS should be used for our problem. We see immediately from Theorem 3 that the threshold

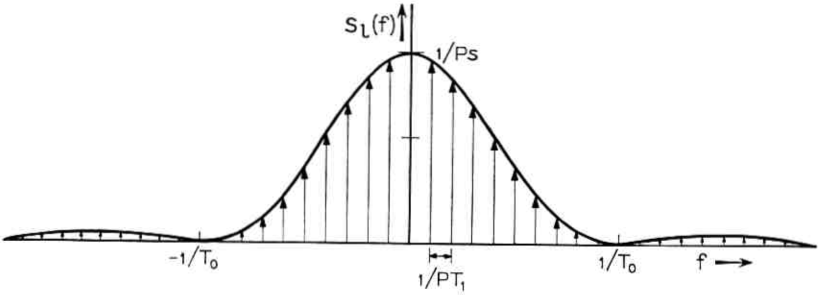


Fig. 14 — Spectrum of a scrambled periodic sequence.

on the first counter of the MCS, t_{s_1}, must be at least 14. Similarly, t_{s_2} of the MCS must be at least 13. Since a 4-stage binary counter will count to 16, we see directly that 8 counter stages, 8 shift register stages, 3 OR gates and some peripheral logic will suffice to build an MCS for our problem.

From Theorem 4 we see that the threshold on the SCS need not be any larger than 954 or require more than 10 stages of a binary counter since $2^{10} = 1024 > 954$. A computer simulation of the SCS, however, shows that the bound of 954 is more than 34 times larger than the smallest required threshold, which was found to be 28. The results of this simulation are tabulated in Table I. The largest run of consecutive zeros at counter input was found for all period 7 and period 8 sequences when the line sequences had periods $7 \cdot 127$ and $8 \cdot 127$, respectively. In Table I we list the fraction of the 384 periodic sequences which have the gap lengths (maximum run of zeros) shown. (We note that it is only necessary to simulate the SCS with one starting state of the basic scrambler when $2^m - 1$ is prime since all $2^m - 1$ noncritical starting states appear as states of the basic scrambler. Note also that we can neglect the first eight inputs to the counter following the argument of the third paragraph of Section VII.)

The SCS will scramble our periodic inputs if we choose a counter threshold of 32 which can be realized with a 5-stage binary counter. It will also require eight shift register stages, an AND gate and peripheral logic.

As far as random data is concerned, we see from (46) that the MCS

Table I—Gap Lengths for Periodic Inputs

	Period 7		Period 8	
Gap length	No.	%	No.	%
13	14	10.92	0	0
14	28	21.84	16	6.25
15	14	10.92	16	6.25
16	0	0	16	6.25
17	0	0	16	6.25
18	28	21.84	60	23.41
19	14	10.92	18	7.04
21	0	0	56	21.85
22	2	1.56	2	0.78
24	0	0	24	9.38
25	0	0	16	6.25
26	14	10.92	0	0
27	14	10.92	16	6.25
	128		256	

for our application reaches threshold at least once in n transmissions with probability

$$P_M(n) \leqq (n - 15)(3.06)10^{-5} \tag{61}$$

and we see from (47) that the SCS reaches threshold with probability

$$P_S(n) \leqq (n - 31)10^{-4}. \tag{62}$$

Thus, the MCS has a slight edge on the SCS when it comes to scrambling random data since it is desirable to keep the frequency of threshold crossings low.

In sum, it is safe to say that the SCS has the edge for our problem primarily because it is simpler and less expensive. Also, we note that the addition of a single-counter stage will reduce $P_S(n)$ to $(n - 31)10^{-8}$. The autocorrelation function of the scrambled data sequence will be like that of Fig. 13 with $| \epsilon | \leqq 0.008$.

XIII. CONCLUSIONS

We have introduced two major classes of self-synchronizing, digital data scramblers called multi-counter scramblers and single-counter scramblers. We have shown that these scramblers and combinations of the two will map a periodic sequence of period s into a periodic sequence of period $LCM(s, p^m - 1)$, where p is the size of the source alphabet (the SCS results require that $p = 2$ and that s and $2^m - 1$ be relatively prime), if the basic scrambler tap polynomial $h(x)$ of degree m is a primitive polynomial over $GF(p)$. We have found the smallest values for the counter thresholds in the MCS and have shown the existence of finite thresholds for the successful operation of the SCS.

We have shown that there are many transitions in the scrambled sequence and that they are well distributed. We have shown that the descramblers possess the self-synchronizing property and we have considered the effect of channel errors on the descrambling process. We have seen that the principal effect of infrequent channel errors (occurring at a rate of one in 10^5 transmissions, say) is to cause approximately $w(h)$ as many output errors, where $w(h)$ is the number of nonzero terms in $h(x)$. Channel errors were shown to have a relatively small effect on the output of the descrambler monitoring logic.

We have found the power density spectrum of the waveform generated by the scrambler output for a representative case, namely, when the source is binary and the scrambled sequence is mapped onto a ± 1 sequence. We have seen that scrambling does not affect the spectrum of the line signal when the source is random and that its principal effect when the source is periodic is to introduce P times as many tones each having $1/P$th as much energy where P is the factor by which the source period is increased.

It has been shown that the counters in the scrambler and descrambler reach threshold infrequently when the source is random and at most once each time the source becomes periodic. Thus, it has been argued that the counters at the descrambler might be removed if the rate at which the counters at the scrambler reach threshold is less than the rate of occurrence of channel errors, and if the customer can tolerate occasional output errors when his data is periodic.

XIV. ACKNOWLEDGMENTS

The author acknowledges helpful early conversations with R. D. Fracassi in this work and many discussions with his colleagues in the Data Communications Laboratory.

APPENDIX

Proof of Theorem 2

Let T_h be the matrix shown below where the coefficients c_1, c_2, $\cdots$, c_m are elements of the modular field $GF(p)$ of p elements, p a prime

$$T_h = \begin{bmatrix} c_1 & c_2 & \cdots & c_{m-1} & c_m \\ 1 & 0 & \cdots & 0 & 0 \\ 0 & 1 & \cdots & 0 & 0 \\ \vdots & \vdots & & \vdots & \vdots \\ 0 & 0 & \cdots & 1 & 0 \end{bmatrix} \Bigg\updownarrow m. \tag{63}$$

Let $h(x)$ be the polynomial shown below in the indeterminate x where coefficients are those appearing in (63).

$$h(x) = x^m - c_1 x^{m-1} - \cdots - c_m . \tag{64}$$

Then, one can show by direct calculation that the characteristic polynomial of T_h, $\varphi(x)$, defined by

$$\varphi(x) = \det (T_h - xI), \tag{65}$$

is related to $h(x)$[9,10] by

$$\varphi(x) = (-1)^m h(x). \tag{66}$$

The matrix T_h is called the "companion matrix" for the polynomial $h(x)$.

We assume that $h(x)$ is a primitive polynomial over the field $GF(p)$. A polynomial $h(x)$ is primitive if

(*i*) $h(x)$ is irreducible over $GF(p)$, that is, if there is no polynomial with coefficients in $GF(p)$ which divides $h(x)$ except 1 and $h(x)$, itself, and

(*ii*) $h(x)$ of degree m divides $x^n - 1$ for $n = p^m - 1$ but for no smaller integer n.

If we replace the term c_m in $h(x)$ given in (64) by the matrix $c_m I$, where I is the $m \times m$ identity matrix and replace x by T_h, where powers of T_h are defined as successive matrix products, then we have the well-known Cayley-Hamilton theorem[11]

$$\varphi(T_h) = 0, \tag{67}$$

where $\varphi(x)$ is the characteristic polynomial of T_h. Thus, a matrix T_h satisfies its own characteristic polynomial. There is a smallest degree monic polynomial (coefficient of the highest degree term is 1), called the minimal polynomial, $m(x)$, such that

$$m(T_h) = 0. \tag{68}$$

Since $h(x)$ is irreducible, we have

$$m(x) = h(x). \tag{69}$$

We now wish to prove the following theorem.

Theorem 2: The matrices $T_h^k - I$ *are nonsingular for* $1 \leqq k \leqq p^m - 2$.

We first prove the following two lemmas.

Lemma 2: If $0 \leqq i, j \leqq p^m - 2, i \neq j$, *then* $T_h^i \neq T_h^i$.

Proof: If $T^i = T^j$ for the i, j given above and $i < j$ then

$$T^i(T^{j-i} - I) = 0$$

implies

$$T^{j-i} - I = 0$$

since det $T_h = \varphi(0) \neq 0$. (If $\varphi(0) = 0$ then $\varphi(x)$ is divisible by x and $h(x)$ is not primitive.) Consider now the polynomial $x^{j-i} - 1$. Using the Euclidean division algorithm we have

$$x^{j-1} - 1 = h(x)q(x) + s(x)$$

for unique $q(x)$ and $s(x)$ and degree $s(x) <$ degree $h(x)$. Therefore,

$$T_h^{i-1} - I = 0 = h(T_h)q(T_h) + s(T_h)$$

which implies that

$$s(T_h) = 0.$$

But $m(x) = h(x)$ is the minimal polynomial of T_h so that $s(x) = 0$. Therefore, $h(x)$ divides $x^n - 1$, $n = j - i < p^m - 1$. Contradiction. Hence, $T^i \neq T^j$, $i \neq j$, $0 \leqq i, j \leqq p^m - 2$. QED

Lemma 3: All nonzero polynomials in T_h with coefficients in $GF(p)$ and of degree $m - 1$ or less are nonsingular.

Proof: Let $p(x)$ be a polynomial of degree $m - 1$ or less with coefficients in $GF(p)$. Then, using the Euclidean division algorithm, we have that the greatest common divisor, $d(x)$, of $p(x)$ and $h(x)$ is given by

$$d(x) = a(x)p(x) + b(x)h(x),$$

where $a(x)$ and $b(x)$ are unique polynomials. Since $h(x)$ has degree m and is irreducible $d(x) = 1$ and

$$1 = a(x)p(x) + b(x)h(x).$$

Taking these polynomials in T_h, we have

$$I = a(T_h)p(T_h) + b(T_h)h(T_h)$$

or since $h(T_h) = 0$ we have

$$I = a(T_h)p(T_h) = p(T_h)a(T_h),$$

where the latter equality follows since the polynomials $a(x)$ and $p(x)$ commute. Thus, the polynomial $p(T_h)$ of degree $m - 1$ or less with coefficients over $GF(p)$ in the matrix T_h has both a left inverse and a right inverse and is nonsingular. QED

Proof of Theorem 2:

Since $h(T_h) = 0$ we have

$$T_h^m = c_1 T_h^{m-1} + c_2 T_h^{m-2} + \cdots + c_m I.$$

Thus, every power of T_h, such as T_h^i can be written as a polynomial in T_h of degree $m - 1$ or less. Hence, $T_h^i - T_h^j$ can be written as a polynomial of degree $m - 1$ or less in T_h. From Lemma A1, $T_h^i - T_h^j \neq 0$, $i \neq j$, $0 \leqq i, j \leqq p^m - 2$ so that $T_h^i - T_h^j$ as a polynomial in T_h of degree $m - 1$ or less is nonzero. From Lemma A2, $T_h^i - T_h^j$ is nonsingular and it follows by choosing $j = 0$, $i = k$ with $1 \leqq k \leqq p^m - 2$ that $T_h^k - I$ is nonsingular. QED

Theorem 2 in effect says that if $\mathbf{y}$ is some arbitrary, nonzero column vector of m components chosen from $GF(p)$ then $T^k\mathbf{y}$ runs through all $p^m - 1$ nonzero vectors $\mathbf{y}$ as k ranges between 0 and $p^m - 2$. Thus, *the linear sequential filter* with feedback paths *described by* T_h *is* a *maximal-length sequence generator*. Elspas[5] comments that these results were noted by Zierler[2] and Golomb.[12]

REFERENCES

1. Golomb et al., *Digital Communications with Space Applications,* Prentice-Hall, New Jersey, 1964.
2. Zierler, N., Several Binary Sequence Generators, Lincoln Lab., MIT, Lexington, Mass., Tech. Rep. No. 95; September, 1956, also reprinted in (13).
3. Fracassi, et al., Patent Application, Case 8-1, Serial 482498, Filed August 25, 1965.
4. Huffman, D. A., The Synthesis of Linear Sequential Coding Networks, Proc. Third London Symp. on Information, Theory, September, 1955, pp. 77–95, also reprinted in *Linear Sequential Switching Circuits* (see 13).
5. Elspas, B., The Theory of Autonomous Linear Sequential Networks, IRE Trans. Circuit Theory, *CT-6,* No. 1, March, 1959, pp. 45–60, also reprinted in *Linear Sequential Switching Circuits* (see 13).
6. Zierler, N., Linear Recurring Sequences, SIAM Journal, *7,* March, 1959, pp. 31–48, also reprinted in *Linear Sequential Switching Circuits* (see 13).
7. Peterson, W. W., *Error-Correcting Codes,* MIT Press and John Wiley & Sons, 1961, p. 7.
8. Bose, R. C. and Kuebler, Jr., R. R., A Geometry of Binary Sequences Associated with Group Alphabets in Information Theory, Annals Math. Stat., *31,* March, 1960, pp. 113–139.
9. Albert, A. A., *Fundamental Concepts of Higher Algebra,* The University of Chicago Press, 1956, p. 86.
10. Birkhoff, G. and MacLane, S., *A Survey of Modern Algebra,* Macmillan, 1941, pp. 316–318.
11. *Ibid.,* pp. 319–321.
12. Golomb, S. W., Sequences with Randomness Properties, Glenn, L. Martin Company, Baltimore, Md., Final Rep. on Contract No. SC54–33611, dated June 14, 1955 and cited as Ref. 7 in Ref. 5 above.
13. Kautz, W. H. (ed.), *Linear Sequential Switching Circuits,* Holden-Day, San Francisco, 1965.

Contributors to This Issue

Martin B. Brilliant, B.A., 1955, Washington and Jefferson College; S.B., S.M., 1955, ScD., 1958, Massachusetts Institute of Technology; Bell Telephone Laboratories, 1955 and 1966—. Mr. Brilliant has also held positions with the Air Force Cambridge Research Center; National Company, Inc.; Hazeltine Research Corporation; University of Kansas; and Booz·Allen Applied Research, Inc. At Bell Telephone Laboratories, he worked in 1955 on a transistor pulse generator for the Electronic Central Office. He is now concerned with systems engineering problems in the switching of time-multiplexed signals. Member, Sigma Xi, AAAS.

Herbert Y. Chang, B.S., 1960, M.S., 1962 and Ph.D. in EE, 1964, University of Illinois; Bell Telephone Laboratories, 1964—. Mr. Chang has been associated with the maintenance dictionary project for the No. 1 Electronic Switching System. At present he is engaged in studies of techniques for the design of self-diagnosable digital machines and the reliability and maintainability studies for digital systems. Member, IEEE, Sigma Xi, Tau Beta Pi, Eta Kappa Nu, Pi Mu Epsilon, ACM.

Richard W. Dixon, A.B., 1958, Harvard College; M.A., 1960 and Ph.D., 1964, Harvard University; Bell Telephone Laboratories, 1965—. Since joining the Bell Telephone Laboratories, Mr. Dixon has been concerned with problems connected with the interaction of light and elastic waves in solids and liquids. Member, American Physical Society.

Detlef Gloge, Dipl. Ing., 1961, D.E.E., 1964, Braunschweig Technische Hochschule (Germany); research staff, Braunschweig Technische Hochschule 1961–1965; Bell Telephone Laboratories, 1965—. In Braunschweig, Mr. Gloge was engaged in research on lasers and optical components. At Bell Telephone Laboratories, he has concentrated in the study of optical transmission techniques. Member, VDE, IEEE.

EUGENE I. GORDON, B.S. in Physics, 1952, City College of New York; Ph.D. in Physics, 1957, Massachusetts Institute of Technology; Bell Telephone Laboratories, 1957—. Mr. Gordon is presently Head of the Optical Device Department concerned with gas lasers and their application as well as devices for *Picturephone*®. He is an associate editor of the IEEE Journal of Quantum Electronics. Member, APS, Phi Beta Kappa, Sigma Xi.

R. L. HAMILTON, B.S. Chem. Engr., 1957, Ph.D., 1960, Oklahoma University; Bell Telephone Laboratories, 1963—. Mr. Hamilton has been engaged in developing techniques for measuring water vapor permeabilities of plastic materials as well as water vapor ingress rates into cable sheaths, apparatus cases, and other enclosures used in outside plant. At present, he is developing composite materials for use in underground conduits. Member, Tau Beta Pi, Sigma Xi.

WILLIAM C.-Y. LEE, B.Sc. in Engineering, 1954, Chinese Naval Academy; M.Sc., E.E., 1960, and Ph.D. in E.E., in 1963, The Ohio State University; Bell Telephone Laboratories, 1964—. Mr. Lee has been concerned with the study of wave propagation in anisotropic medium and antenna theory. His present work has included studies of mobile radio antennas and signal fading problems. Member, Sigma Xi, IEEE.

J. E. SAVAGE, S.B., S.M., 1962, and Ph.D., 1965, Massachusetts Institute of Technology; Bell Telephone Laboratories, 1965—. Mr. Savage has been engaged in research on problems in coding and information theory. He holds the position of Lecturer in Electrical Engineering at Columbia University. Member, Tau Beta Pi, Sigma Xi, Eta Kappa Nu, IEEE.

WENDL THOMIS, B.S., 1956, Illinois Institute of Technology; M.S., 1959, Purdue University; Bell Telephone Laboratories, 1959–1966. Mr. Thomis has done systems work in the maintenance planning area of the No. 1 Electronic Switching System since joining the Laboratories. He headed the maintenance dictionary project for No. 1 ESS and is currently working on improved maintenance dictionary techniques.

B.S.T.J. BRIEFS

A Camera Tube with a Silicon Diode Array Target

By M. H. CROWELL, T. M. BUCK, E. F. LABUDA, J. V. DALTON, and E. J. WALSH

A variety of electronic cameras have been developed for television systems.[1] Among these the vidicon[2] and the Plumbicon[3] have the inherent advantages of high sensitivity, small size, and simple mechanical construction. The operating principles of the vidicon and the Plumbicon are quite similar since they both utilize a thin photoconductive layer to convert the optical image to a stored charge pattern which is periodically scanned and erased by an electron beam. Erasing the charge pattern creates the video signal. However, there is a distinct difference in overall device performance since the photoconducting target in the Plumbicon (PbO) is deposited in a manner to form a single, large area, graded p-n junction, each layer having high resistivity. In the vidicon, the evaporated layers of Sb_2S_3 forming the target behave like a semi-insulating photoconductor.

A new type of target consisting of an array of electrically isolated reverse-biased diodes, as first suggested by Reynolds,[4] later discussed by Heijne[5] and more recently by Wendland[6], has several valuable attributes.

(*i*) The dark current and the light-induced current can be essentially independent of target (reverse bias) voltage and the response characteristic can have a gamma of unity as in the Plumbicon.

(*ii*) The time constant associated with the charge leakage of an array of reverse-biased diodes can be very much larger than the intrinsic (dielectric relaxation) time constant of the bulk material. This implies that an infrared responsive camera operating at room temperature can be realized.

(*iii*) The spectral response can cover a wide range including the visible and consequently much greater and more uniform sensitivity can be achieved than in the vidicon or Plumbicon.

(*iv*) The target performance is insensitive to electron beam bombardment and is unaffected by intense light sources so that deleterious burn-in does not occur.

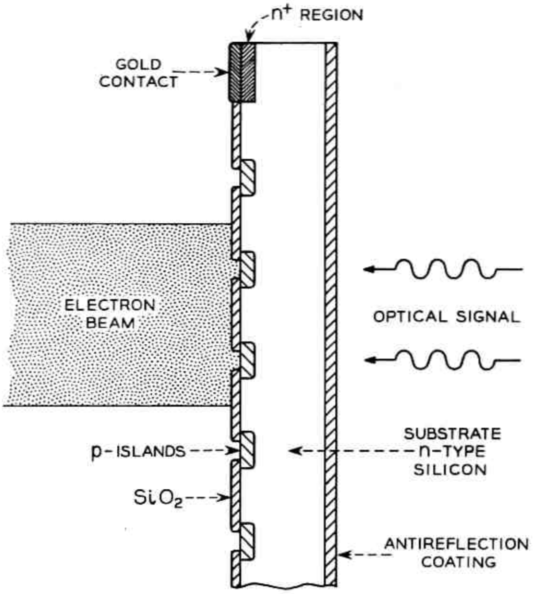


Fig. 1 — Schematic drawing of the diode array. In practice the perimeter thickness was made ≈4 mils to obtain a self-supporting structure.

(*v*) There is no image persistence due to photoconductive lag.

(*vi*) The assembled tube may be processed using standard vacuum techniques including a high temperature bake.

(*vii*) The operating lifetime can be expected to exceed that of the vidicon and Plumbicon by a considerable margin.

In this brief, experimental results obtained from targets consisting of a 540 × 540 array of reverse-biased Si diodes are reported. The substrate is 10 Ω-cm, n-type Si, is self-supporting and can be anti-reflection coated. The p-type islands are formed by diffusing boron through 8μ diameter holes in the SiO_2 film, the center-to-center spacing between holes being 20μ. This arrangement provides sufficient diode capacitance $\approx 1000\ \mu\mu fd/\text{cm}^2$) to integrate the diode photoresponse over the time interval of 1/30 sec (one frame period in commercial television). Ohmic contact to the array is obtained via the gold ring evaporated onto the n^+ region near the perimeter of the Si wafer chip.

In normal operation, the electron beam, the diameter of which is larger than that of a single diode, periodically charges the p-type islands down to cathode (ground) potential while the potential of the n-type material is held at ≈5 to 10 volts. This potential difference can be sustained for a normal television frame time so long as the dark current is $<5 \times 10^{-13}$ amps/diode. The SiO_2 film, also charged down to cathode potential by the beam, remains there and isolates the substrate from the beam. The incident light associated with the image is absorbed in the Si, creating hole-electron pairs. Since the thickness of a self-supporting wafer is $\geqq 10^{-3}$cm and the absorption coefficient of Si for visible light is greater than 3000 cm^{-1}, most of the hole-electron pairs will be generated near the incident surface; the minority carriers (holes) then diffuse to the depletion region of the diodes, discharging the diodes by an amount proportional to the light intensity. The recharging of the diodes by the scanning beam creates the video signal.

An exact analytical evaluation of the performance of the diode array shown in Fig. 1 is quite complicated. However, with a simpler model in which the p-regions of the array are replaced by one large homogenious p-region with no lateral conductivity, it is possible to estimate the loss in light sensitivity and resolution due to minority carrier recombination and diffusion. An analysis of this simpler model indicates that for a minority carrier lifetime of ≈ 10 μsec, a surface recombination

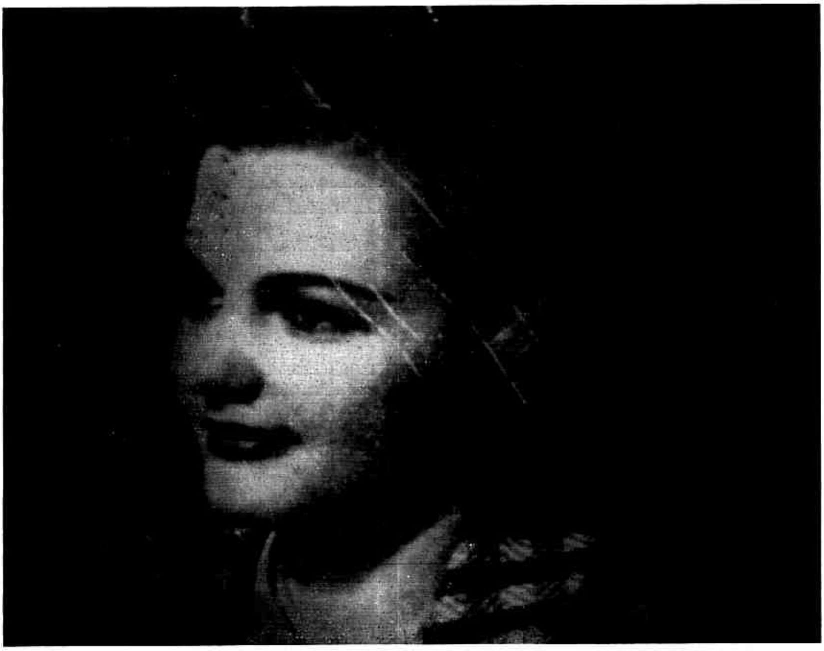


Fig. 2 — Photograph obtained with the 540 × 540 diode array target. The subject was a black and white transparency illuminated with a tungsten lamp.

velocity of $\approx 10^4$ cm/sec, and a wafer thickness of $\approx 10^{-3}$ cm, the collection efficiency (ratio of collected holes to generated holes) for uniform illumination with visible light is $\approx$80 percent. Lateral diffusion will degrade the spatial resolution. For example, if the spatial variation in the visible light were sinusoidal with a period corresponding to 4×10^{-3} cm or twice the center-to-center spacing of the diodes, the ac signal would be reduced to $\frac{5}{8}$ of the dc signal.

The performance of a Si diode array is illustrated by Fig. 2. This photograph was obtained from a Kintel[7] closed circuit system with commercial television standards. The usual vidicon camera tube was replaced by a tube using a 540 × 540 diode array target. The defects in the picture reflect a localized high dark current and can be partly attributed to defects in the bulk crystal from which the array was fabricated and to defects in the SiO_2 film.

The measured spectral response of a camera tube with a diode array target is given in Fig. 3 for two wafer thicknesses. In these measure-

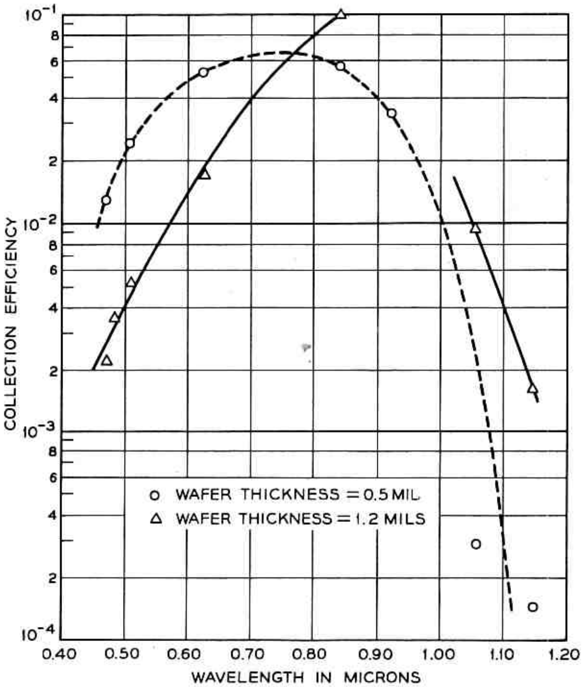


Fig. 3 — Spectral response without reflection loss corrections of experimental diode array targets for two wafer thicknesses. The photograph shown in Fig. 2 was obtained with the 1.2-mil target.

ments, the whole diode array was illuminated with a uniform light intensity and the light induced dc current (average video signal) in the target lead was measured. Continuous laser transitions were used to obtain the absolute response at several wavelengths. The actual target collection efficiency was better than that indicated in the figure since no anti-reflection coating was used and no reflection loss corrections were made. For Si, the reflection coefficient varies from 30 percent in the near infrared to $\approx$65 percent in the blue portion of the spectrum.[8] With a single layer anti-reflection coating, the reflection can be reduced to a few percent. This implies that if such a coating had been used on the experimental targets a maximum collection efficiency of $\approx$20 percent would have been obtained. At this maximum, the sensitivity would have been 0.16 μamps/μwatt. Because of its wider spectral response, the camera tube with a diode array target was $\approx$25 times more sensitive than an 8134 RCA vidicon for illumination with an incandescent lamp at normal operating temperature. The measured gamma was unity.

The observed dark current for the entire array was $\approx 5 \times 10^{-8}$ amps for a reverse bias of 5 to 10 volts. This implies that the leakage current per diode was $\approx 2 \times 10^{-13}$ amps. The resolution was not limited by leakage between diodes.

ACKNOWLEDGMENTS

The authors gratefully acknowledge the advice of E. I. Gordon who initiated the work reported here.

REFERENCES

1. Zworykin, V. K. and Morton, G. A., *Television*, 2nd Edition, John Wiley and Sons, Inc., New York, 1954.
2. Weimer, P. K., Forgue, J. V., and Goodrich, R. R., Electronics, *23*, May, 1950, p. 70.
3. deHaan, E. F., van der Drift, A., and Schampers, P. P. M., Philips Tech. Rev., *25*, 1964, p. 133.
4. Reynolds, F. W., Unpublished Work, 1951 and Solid State Light Sensitive Storage Device, U. S. Patent No. 3,011,089, issued Nov. 28, 1961.
5. Heijne, L., Philips Res. Rept. Suppl., No. 4, 1961.
6. Wendland, P. H., paper presented at Eighth Conference on Tube Techniques in New York on September 22, 1966.
7. Cohu Electronics, Inc., Kintel Division.
8. Runyan, W. R., *Silicon Semiconductor Technology*, McGraw-Hill Book Co., New York, 1965, p. 198.