THE BELL SYSTEM
TECHNICAL JOURNAL

VOLUME XLIV JULY–AUGUST 1965 NUMBER 6

A Survey of Bell System Progress in Electronic Switching*

By W. H. C. HIGGINS

(Revised manuscript received March 31, 1965)

This article is a survey-type discussion of the Bell System's No. 1 Electronic Switching System for central office use and No. 101 Electronic Switching System for business use. For both systems it presents background history and descriptions of the major subsystems and components. It also covers recent manufacture, installation and commercial service of these two systems.

I. INTRODUCTION

Two Electronic Switching Systems have been developed by Bell Telephone Laboratories for general application in the Bell System; both are now in quantity manufacture by the Western Electric Company. One of these, known as No. 1 ESS, is designed for local central offices, and is the commercial successor to the Morris Electronic Central Office. Its system organization is also potentially suitable for tandem and toll applications. The other system, called No. 101 ESS, is designed to provide electronic private branch exchange services in conjunction with existing electromechanical central offices. It brings to the business community modern PBX and Centrex† features which can be provided economically with this new type of system.

* Originally written for the German Bundespost and published in their 1964 Yearbook of Telecommunications, this article has been updated and is published here by permission for readers of the B.S.T.J.

† The principal features of Centrex are direct inward dialing to extensions, identified outward dialing, and certain switchboard attendant features.

The electronic private branch exchange was the first of the two to be placed in commercial service. On November 30, 1963, the Southern Bell Telephone Company initiated Centrex service to about 100 extensions at the Brown Engineering Company, Cape Kennedy, Florida. Two weeks later service began at the Chrysler Corporation's office, also located at Cape Kennedy. During 1964 and early 1965 No. 101 ESS installations were completed for service at such widely separated locations as New York City, Chicago, Cleveland, Los Angeles, and Washington, D. C.

In 1963 installation by the Western Electric Company of the first commercial No. 1 ESS central office began in a new building at Succasunna, New Jersey, for the New Jersey Bell Telephone Company. After undergoing an extensive series of tests, that system was cut over to commercial service on May 30, 1965. It now serves both residence and business telephone users in that community. Additional No. 1 ESS central offices are being installed and tested in Baltimore, New York City, Norfolk, and Washington, D. C., as well as in several locations to serve military customers; the latter provide four-wire switching of lines and trunks as contrasted to two-wire switching for the commercial offices.

This article surveys the work leading to these two developments and describes the production designs which are inaugurating a new era in switching for the Bell System.

II. EARLY WORK

For many years engineers have been intrigued by the idea of applying electronics to switching. As the switching art and digital technology developed, these ideas and speculations became more definitive. With the very high speed operation of electronic components, it was believed that the principal advantages of common control would be enhanced in that very large offices could be controlled with a single common control without the complications of multiple marker usage. At Bell Telephone Laboratories these ideas led to a formalized attack on the problem, beginning shortly after the close of World War II. The aim of the work was to explore various approaches to the electronic switching problem with the ultimate objective of improving service, reducing costs, and providing greater flexibility while maintaining the high reliability of electromechanical switching.

The early work produced many innovations, and several laboratory switching systems were constructed to explore the basic concepts. Among them was a space-division system employing reed-diode switching

matrices with "end-marking" under control of multi-element gas tubes. This system, known as ECASS[1] (Electronically Controlled Automatic Switching System), was brought to a laboratory demonstration level in 1947. In 1948 exploratory work was carried out on a single highway time-division system using vacuum tube gates and quartz delay lines for memory. This was followed in 1949 with DIAD[2] (Drum Information Assembler and Dispatcher). This was a system having a large memory in common control and a space division reed-diode "end-marked" network.

Research on these systems brought valuable insight to both network and common control aspects of electronic switching and provided a firm technical foundation for later work. However, it also pointed up the desirability of new devices for both logic and memory if electronic switching were to become a serious challenge to the highly developed electromechanical systems.

In the early 1950's the transistor (invented at Bell Laboratories in 1948) had reached the stage of development where it could be seriously considered for commercial application. This, together with economical bulk memories based upon the cathode ray and barrier grid tubes, suggested the possibility of developing a commercial electronic switching system. Work on such a system, initiated in 1954, led to a field trial of an electronic central office at Morris, Illinois.

Because of the historical significance of the Morris trial and its impact on subsequent development work, it seems appropriate to present some of the results obtained.

III. THE MORRIS TRIAL

3.1 *The Morris System*

A view of the installation in the Morris central office is shown in Fig. 1. Although the system has been previously described,[3] a brief review of its design will provide useful background.

Fig. 2 is a block diagram of the Morris electronic switching system.[4] Lines and trunks were terminated on a space-division switching matrix having gas tube crosspoints. In this installation the matrix was equipped for 604 customer lines. "End-marking" of the network was under control of a high-speed, stored program, common control system. Because the gas tube crosspoints in the switching network could not carry high-level standard ringing current, each customer was provided with a low-current tone ringer station set. The common control equipment consisted of a central control logic unit associated with barrier grid

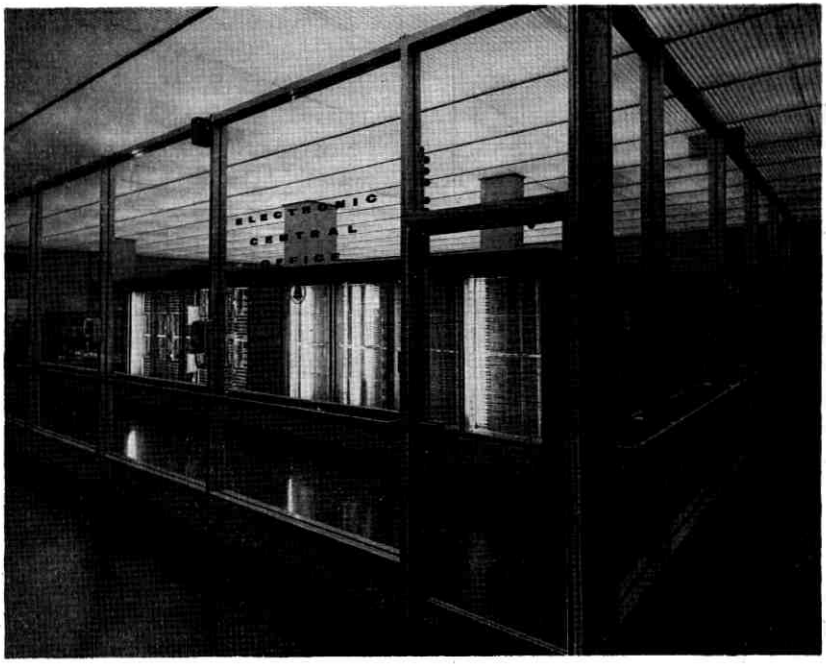


Fig. 1 — Electronic central office trial installation at Morris, Illinois.

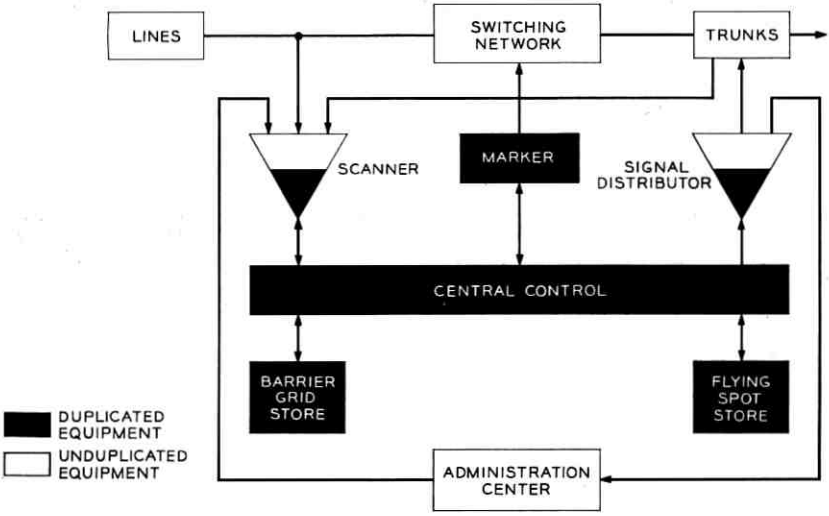


Fig. 2 — Morris ESS block diagram.

stores for temporary memory and a flying spot store for program and translation storage.

The flying spot store was a 2.25 million bit high-speed, random access, semipermanent memory that used a cathode ray tube, a complex optical system, and photographic plates on which program or translation information was placed in the form of a pattern of transparent or opaque spots. Photomultiplier tubes detected the light transmitted through these spots to determine the "1" or "0" condition of the information bit. An ingenious electronic servo system maintained beam position and light intensity with such accuracy that adjacent bits could be placed on 7-mil centers. The 2.25 million bits of program and translation information were stored on four 10 inch $\times$ $12\frac{1}{2}$ inch glass photographic plates. Cycle time of the store was 2.5 microseconds.

The temporary or "scratch pad" memory consisted of two barrier grid tube stores, which provided a memory capacity of 32,768 bits. This memory was also operated on a 2.5 microsecond cycle time.

The semiconductors, transistors, and diodes used in this system were the diffused germanium variety that were available in 1957. In order to insure continuous operation of these devices, the equipment cabinets were air conditioned. In addition, the gas tube switching network required control of ambient temperature within narrow limits for reliable operation. Air conditioning was also required for the two memories because of their high level of heat dissipation.

In spite of the special precautions taken to insure component reliability, it was known at the outset that failures would occur more frequently than could be tolerated for the service continuity required in a switching system. Accordingly, all of the common control equipment and portions of the electronic scanner and signal distributor were provided in duplicate, and arrangements were made to switch automatically from one set of equipment to the other in the event of a malfunction. Also, programs were included in memory to provide automatic fault recognition and diagnosis of the unit in trouble.[5] Since air conditioning was an essential part of the design, this too was provided in duplicate.

The system was installed in the central office in Morris, Illinois, early in 1960, and part-time telephone service was given to a small number of customers beginning in June of that year.[6] Full-time service began in November of 1960 and continued through January of 1962, at which time the trial was terminated. The number of customers and stations served during the trial is shown in Fig. 3.

In addition to the usual telephone service, customers were supplied with one or more special features. One of the most popular of these was

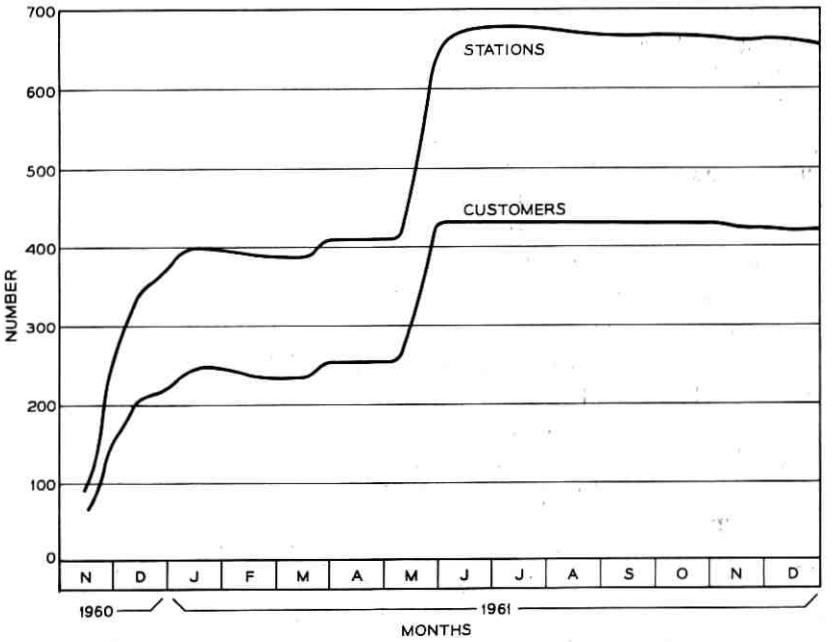


Fig. 3 — Morris ECO — customers and stations.

"abbreviated dialing" by which a two-digit code could be used to reach a seven-digit telephone number. Only four abbreviated codes were made available in the trial. In spite of this limited repertoire, on some lines as many as 50 per cent of all originations were made using this feature. On the average more than 15 per cent of all originations from lines equipped with this service made use of abbreviated dialing.

Another popular feature was "code calling" which in effect provided an intercom in homes equipped with more than one telephone. Dialing a special code and hanging up initiated a coded ring-back to call a particular member of the household to the nearest extension. Cessation of ring indicated to the calling party that the called party had answered.

Three methods were provided to permit the telephone user to have his incoming calls directed to another telephone. One method permitted the routing of calls to a specific preselected alternate number in the Morris office if the user dialed a special code before leaving his phone. When he returned, he dialed another code to cancel the reroute. Another method required the user to call the telephone company business office to indicate the number to which calls should be routed, the time for service to start, and the time for it to be discontinued. This was a useful feature for people who expected to be out of town for extended

periods and wanted to have their calls answered at another telephone. In the third method the user could initiate the transfer to any number in the central office by dialing a special code and the number to which he wished the calls to be routed. The service could be cancelled by dialing another special code.

3.2 *Trial Results*

Performance of the Morris electronic switching system was measured in several ways. One way was through service observing on selected lines and record keeping on calls in which irregularities occurred. In the early months of the trial, irregularities were much too frequent, but a marked improvement was achieved in February, 1961, as indicated in Fig. 4. A somewhat similar measure of performance, with similar results, was obtained through customer reports, as indicated in Fig. 5.

The marked improvement in February, 1961, was due almost entirely to the introduction of improved programming methods which have had a marked influence on programming philosophy for the commercial design. This improved programming concept was called "guard and defensive" programming. It provided a means of insuring that information being processed within the system did, in fact, agree with reality. For example, information concerning the busy or idle state of a custom-

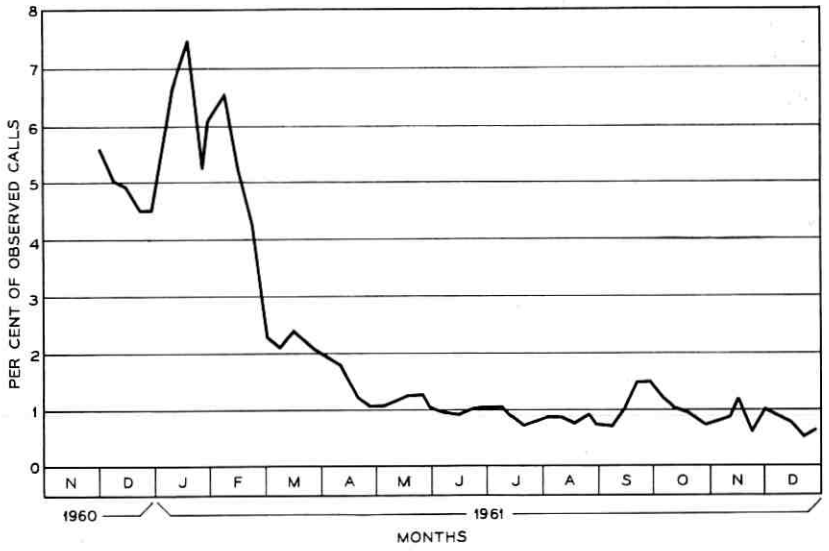


Fig. 4 — Morris ECO — service irregularities.

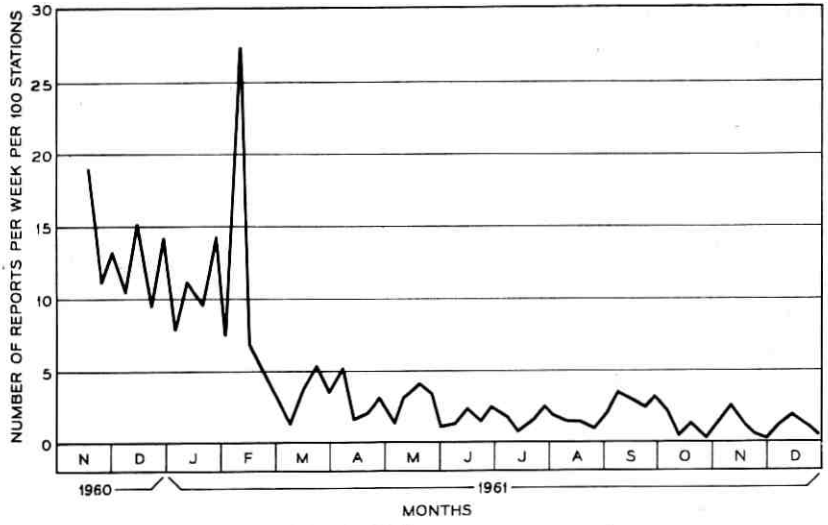


Fig. 5 — Morris ECO — customer reports.

er's line, which is stored in the temporary store, was checked by a guard program every four seconds to determine whether that state agreed with other information concerning that line located elsewhere in the common control equipment. The affect of this important programming change is evident when shown against the background of customer complaints in Fig. 6.

Fig. 6 also shows the period during which improved diagnostic programs for automatic maintenance were installed in the system. The small vertical lines on the Figure indicate dates on which major changes in the program contained in the flying spot store were made. Because of the duplication of the common control equipment, such changes could be readily installed without interrupting customer service. With the facilities provided for processing new photographic plates, the entire program could be changed in about 45 minutes.

In order to aid the maintenance personnel in locating equipment faults, a maintenance dictionary was prepared and became available during the last half of the trial, as indicated in Fig. 6. Whenever a fault occurred in the system, diagnostic routines in the program analyzed the situation and provided a print-out on a teletypewriter associated with the central office equipment. The print-out could then be used as an entry point in the maintenance dictionary to determine which plug-in electronic package required replacement. In most cases in the highly complex central control equipment, the diagnostic print-out and dic-

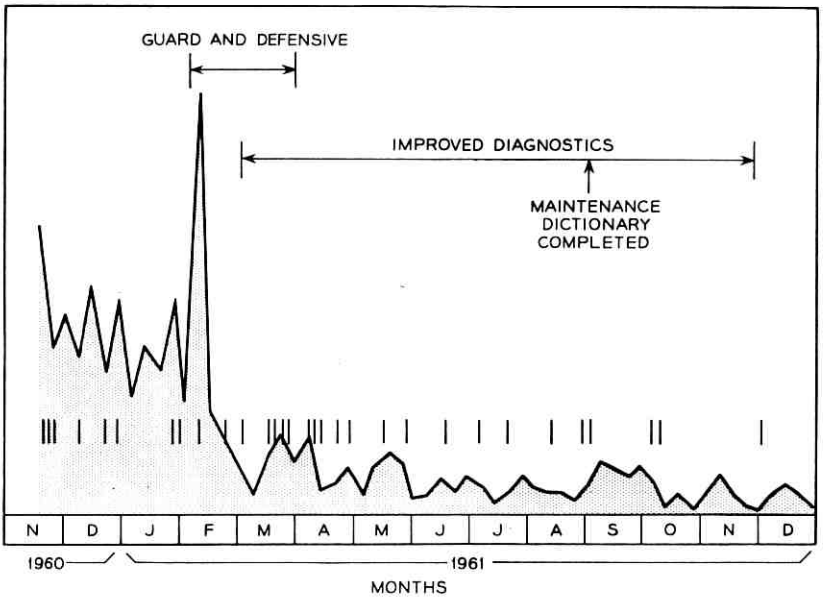


Fig. 6 — Morris ECO — program changes.

tionary could isolate the fault to a single plug-in package. In a relatively small number of fault conditions, a group of several packages might be indicated as the possible source of trouble.

Another measure of system performance is contained in the record of electronic package failures, shown in Fig. 7. It will be noted that the largest number of electronic packages in the system were semiconductor logic packages and that the failure rates for these were very low. As might be anticipated, packages containing relatively high power semiconductors failed at a somewhat greater rate, while electron tube and gas tube failures were highest of all.

Although the failure rates dropped off during the course of the trial, the shape of this trend as seen in Fig. 7 is markedly different from that of the service irregularities and customer reports discussed earlier. It is believed that this difference is due to the guard and defensive programming. The difference gives rise to the concept of "dependability" as a service measure while reserving the term "reliability" as a measure of component performance.

The component failures for the semiconductor logic packages as a function of time are shown in Fig. 8. The marked difference between the failures in the first half and last half of the trial can probably be attrib-

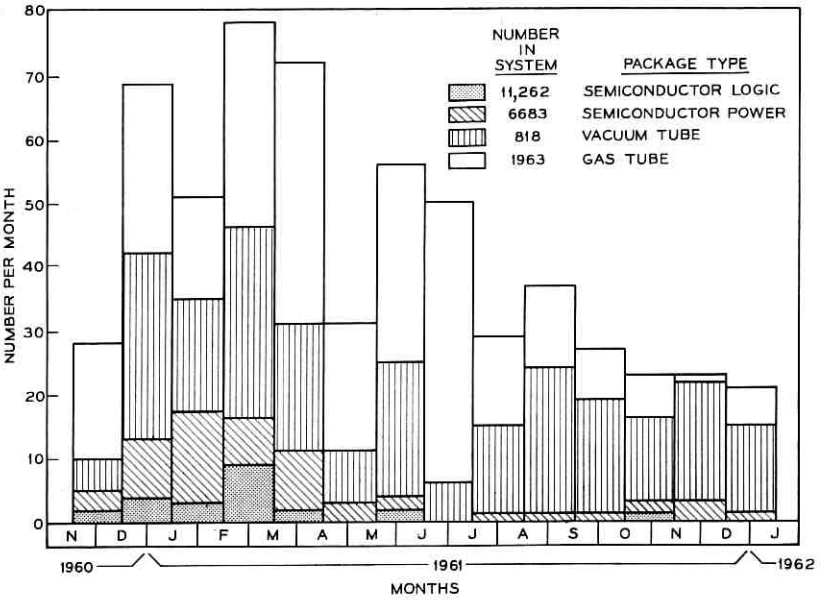


Fig. 7 — Morris ECO — package failures.

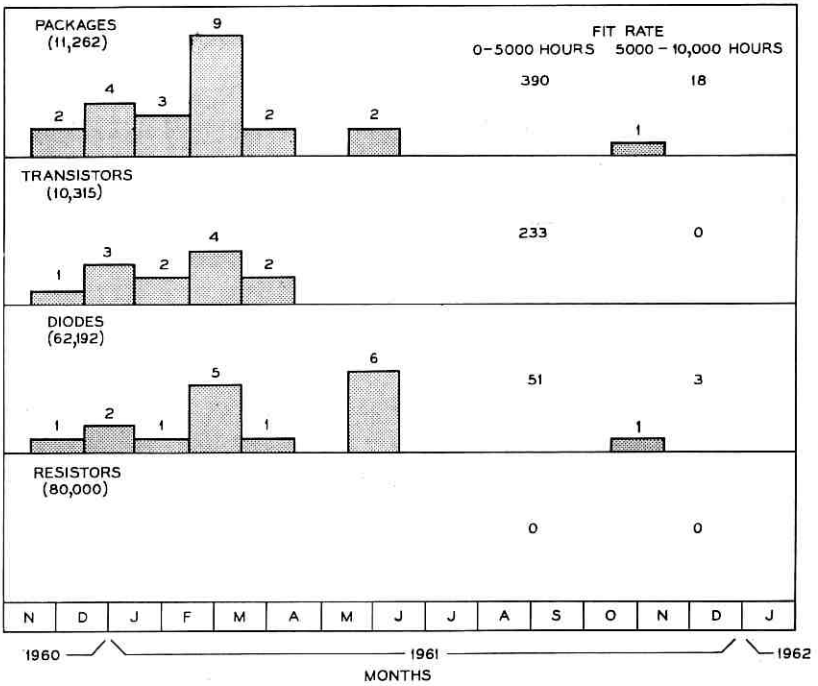


Fig. 8 — Morris ECO — logic package device failures.

uted to less human intervention in system operation as well as to the weeding out of marginal units. The failure rate for the first and last 5,000 hours of system operation is compared in the so-called "fit rate" shown on the right-hand side of the Figure. One fit corresponds to one failure on 10^9 hours of operation.

3.3 *Impact on System Design*

The Morris trial provided important background for the development of No. 1 ESS. In particular, it demonstrated the feasibility of providing dependable service with stored program control which has major advantages in manufacture, maintenance, and flexibility of office administration. The practicability of providing automatic diagnostic programs to assist the maintenance personnel was confirmed. Special programming strategies of guard and defensive programming were evolved to greatly increase system dependability.

The trial also indicated that a major effort should be made to eliminate electron tubes and to remove the requirement for expensive air conditioning equipment. Furthermore, the experience suggested an improvement in the method used for switching between duplicate system equipment so that calls which were in the process of being set up would not be mutilated during the switching interval.

As a result of experience with Morris, the hardware design of No. 1 ESS differs markedly from that used in the trial, although the basic philosophy of stored program control remains the same. A description of this commercial successor to Morris is contained in the next section.

IV. NO. 1 ESS

4.1 *Design Considerations*

Any switching system intended for general Bell System application throughout the United States must cover a broad range in office size and traffic capability and must provide for orderly office growth. An analysis of the Bell System lines in service as of 1960 indicated that 75 per cent of the lines terminated in central office buildings containing over 7500 lines. Fifty per cent were in buildings serving over 19,000 lines and 25 per cent terminated in buildings serving over 32,000 lines. On the other hand, 75 per cent of the central office buildings served less than 3000 lines if one includes community dial offices. An effort was made in the design of No. 1 ESS to provide a configuration with suffi-

cient growth potential to cover a wide range of needs throughout the Bell System.

The stored program control concept demonstrated in the Morris trial was selected for implementation. In fact, it was concluded that the flexibility of stored program control, made possible by high-speed electronics, is more important for switching systems of the future than is the use of electronics per se. The method is adaptable to the wide range in size and growth and simplifies the introduction of changes in operating methods or service features after installation by changing program rather than office wiring. From the factory point of view, the stored program concept permits uniform production with a minimum of wired options; it also should result in less installation effort both initially and for office growth.

To replace the Morris gas-tube switching matrix, a search was undertaken for a suitable metallic crosspoint having control compatibility with high-speed electronics. This was considered desirable for two principal reasons. First, it would avoid the need for special telephone instruments required at Morris and second, it would simplify testing of lines and trunks. To meet this need for an electronically controlled metallic crosspoint, the ferreed was invented, about which more will be said later.

Economic considerations made it desirable to avoid air conditioning. This was made possible by (1) the advent of silicon epitaxial semiconductor devices, which will withstand higher ambient temperatures than the germanium devices used in Morris, (2) the development of new types of random access memory, to be described later, and (3) the ferreed development already mentioned.

The various considerations of Bell System requirements and the experience gained from Morris led to the design of a system which will serve the needs of offices varying in size from a few thousand lines to a maximum of 65,000 lines. The lower limit is determined strictly by economics because of the relatively fixed and rather substantial cost of the common control portion of the system. The upper limit in number of lines to be handled is determined largely by traffic considerations, the speed of the common control, and configuration of the switching network. In high-traffic offices, such as might be found in metropolitan New York, the maximum number of lines to be served by a single switching system will be substantially lower than the 65,000 maximum mentioned above, largely because of common control speed limitations.

4.2 *System Organization*

The organization of No. 1 ESS is very similar to that of the Morris system, as indicated by the block diagram shown in Fig. 9. It consists of an eight-stage space-division switching network utilizing ferreed crosspoints. A central control logic unit interprets instructions contained in the semipermanent memory and carries out the various operations required in handling the telephone traffic. Temporary memory used in conjunction with central control provides call processing registers and

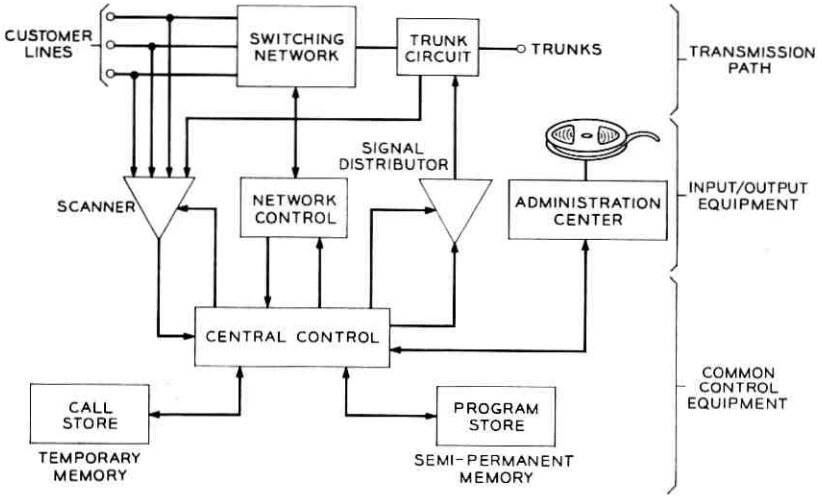


Fig. 9 — No. 1 Electronic Switching System organization.

other "scratch pad" type memory needed in central control operations. Input to this high-speed information processing complex is obtained via scanners which examine the state of lines and trunks on a time-shared basis. When it is addressed by central control, the scanner will examine the state of a particular group of lines and place into the temporary memory information concerning the "on-hook" or "off-hook" state of these lines. Normally lines are scanned at 200-millisecond intervals for detection of originations. Upon detection of an origination, the rate is increased to give a scanning interval of 10 milliseconds. This shorter interval is required to count dial pulses or to detect the outputs of receivers used to convert TOUCH-TONE* calling signals to dc signals.

The signal distributor provides a means for converting the short

* Reg. U.S. Pat. Off.

electronic pulses from central control to appropriate signals on trunks to distant offices. Thus for low-speed outpulsing, central control may request the distributor to close a relay contact, and then the central control will continue to perform many additional logic operations on other calls. Several tens of milliseconds later, at the appropriate time, another order to the distributor would call for opening the relay contact. By this means, the stored program control can perform many complex functions in trunk circuits, thereby minimizing the types and complexity of trunk circuits now found in electromechanical switching offices.

A second output from central control provides for closing the appropriate crosspoints in the switching network, while a third provides information to an administration center. The latter contains the teletypewriter for machine maintenance and a magnetic tape recorder for automatic message accounting information.

4.3 *Design for Dependability*

A more detailed block diagram of No. 1 ESS is shown in Fig. 10. Incoming lines and trunks enter the system at the protector blocks shown at the top of the Figure and thence are connected through a main distributing frame to appropriate portions of the switching network. Electronic control for these network frames is provided over a peripheral unit bus from duplicated central control units. Similar bus arrangements are used for interconnecting program stores and call stores to central control.

This bussing arrangement is another innovation in No. 1 ESS. It permits switching among duplicated common control equipment at electronic speeds. The improved method protects calls that are being processed at the time a switch is made and provides a convenient means for electronically organizing a working system from random combinations of duplicate units; either central control may associate itself with any program store or call store while other units may be in a trouble condition.

In carrying out call processing operations, both of the central controls and their normally associated program stores and call stores simultaneously process the information for the complete call. Interconnections between duplicated portions of the system provide for cross-checking of information. If a mismatch occurs at any point, a fault recognition program is called into play to determine whether the mismatch is due to an error (which does not repeat) or to a true fault. If a fault has occurred in the on-line system, the duplicate equipment immediately takes over

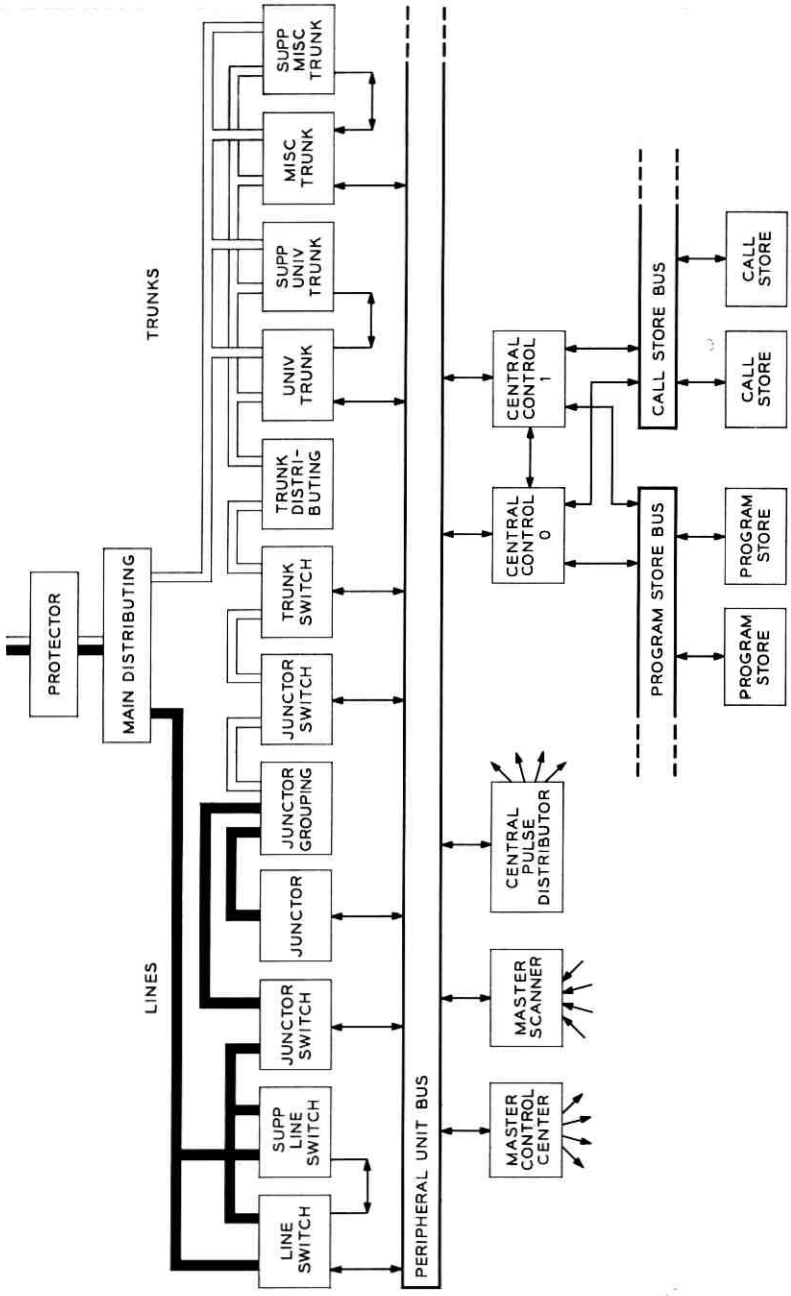


Fig. 10 — No. 1 Electronic Switching System block diagram.

the call processing operations. In its spare time central control carries out a diagnostic routine on the faulty unit. The results of this diagnosis are printed out on a teletypewriter for use by the maintenance man.

In the following sections the design and functions of the major portions described above will be covered in somewhat more detail.

4.4 *Switching Network*

A schematic representation of the eight-stage switching network is shown in Fig. 11. Line link networks, consisting of line switch frames (LSF's) and junctor switch frames (JSF's), contain four switching stages; the remaining four stages are contained in trunk link networks consisting of junctor switch frames and trunk switch frames (TSF's). Wire junctors are used between line link and trunk link networks for line-to-trunk interconnections, and between appearances on the trunk

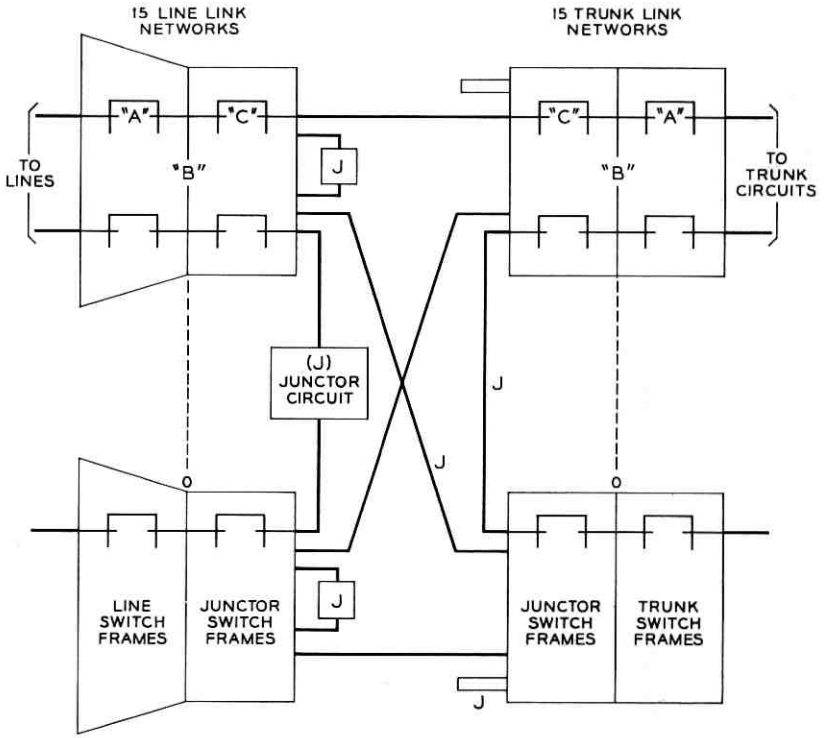


Fig. 11 — Over-all network plan showing line link networks, trunk link networks and typical connections.

link networks for tandem switching. Line-to-line switching is accomplished through a junctor circuit which includes the necessary transmission apparatus and facilities for supervising the individual lines. Because of the wide variety of offices which this electronic switching system is intended to serve, the line link networks are arranged to cover various concentration ratios from 2:1 up to 8:1.

A line switch frame for 4:1 concentration is shown in Fig. 12. The double bay of equipment at the left contains the switching, supervisory and electronic control equipment for interconnecting 512 lines to 128 junctors. A supplementary line switch frame on the right increases this switching capacity to 1024 lines. In this configuration the electronic

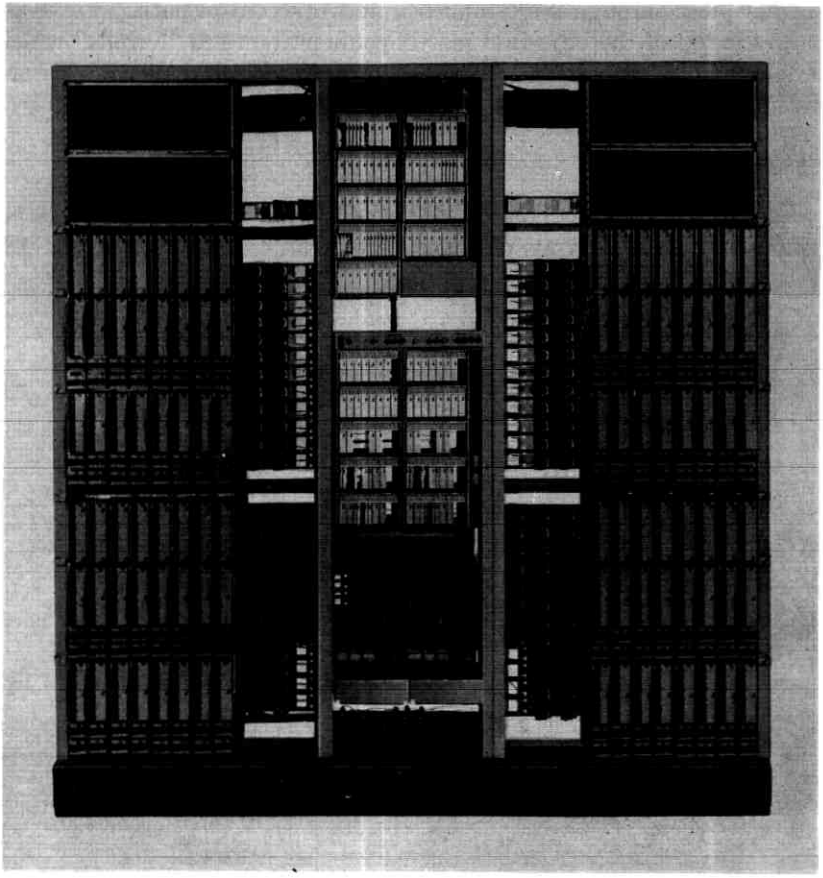


Fig. 12 — No. 1 ESS line switch frames.

network control serves both the basic and supplementary line switch frames.

A number of wire spring relays can be seen on each of these frames. These are used for setting up steering circuits for crosspoint control in the ferreed switches contained in the rectangular cases seen in this Figure. Driving current to operate the ferreed crosspoints is obtained from a solid-state pulser employing a high-power silicon triode. This pulser is located near the bottom of the bay containing the control electronics.

To set up a connection, central control, through the switching frame electronics, orders the establishment of a pulsing path to the appropriate crosspoints. The crosspoints are then closed by applying a high-current pulse through the established network control path.

At the top of the bays in a matrix configuration are "ferrods" which provide line supervision. Both the ferreeds and ferrods were invented especially for No. 1 ESS and are described further below.

4.5 *The Ferreed*[7]

Each of the rectangular ferreed switches shown in the photograph contains an 8 × 8 array of ferreed crosspoints, as shown in Fig. 13.

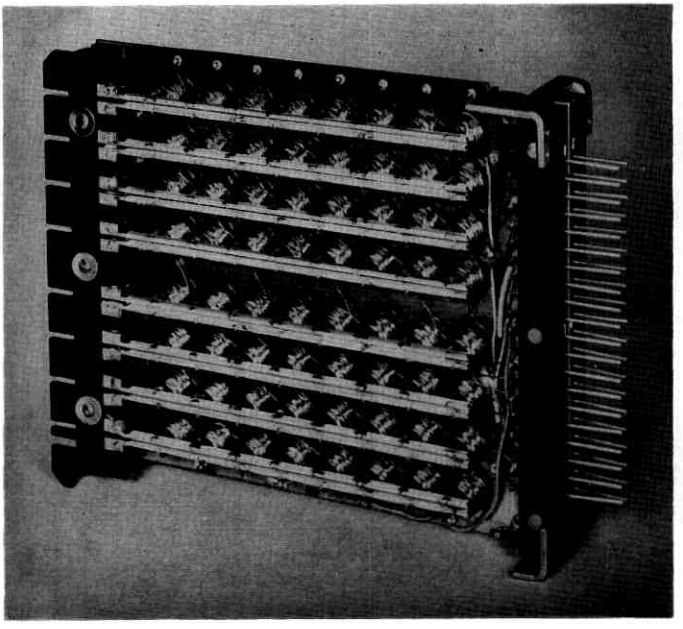


Fig. 13 — No. 1 ESS ferreed switch assembly.

Each crosspoint consists of a pair of dry reed glass-encapsulated switches molded into a small subassembly and inserted with two remendur plates into a solenoid consisting of two control windings as shown by the exploded view in Fig. 14. When a high-current pulse is transmitted simultaneously in the appropriate direction through the two solenoid windings, the remendur plates are poled to produce a north-south magnetic field from top to bottom. Remendur, being a square loop material, remains magnetized after removal of the pulse, causes closure of the reed contacts, and holds them closed without further expenditure of power. Operate current from the pulser flows through appropriate interconnections on the wire spring relays to a given column and row of

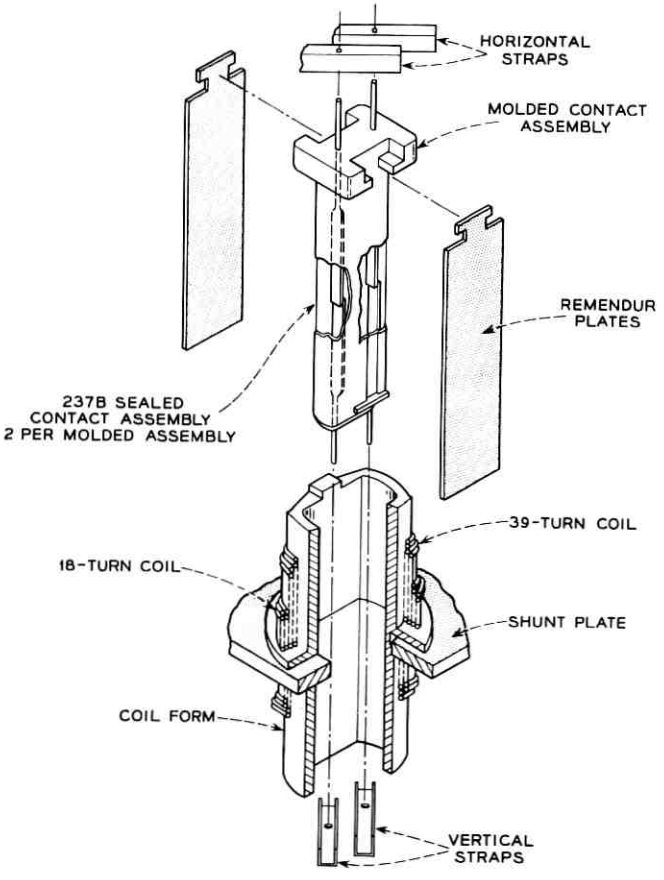


Fig. 14 — Two-wire ferreed crosspoint.

the 8 × 8 ferreed switch unit, thus operating the crosspoint at the intersection of that column and row.

The two windings of the control solenoid are arranged so that a pulse of current through only one of the windings will produce magnetization of the remendur plates in a north-south/south-north distribution about the magnetic shunt plate shown in the diagram. The opposing magnetic fields of the two halves of the remendur plates thus permit the contact to open. This arrangement, together with the matrix interconnection in the 8 × 8 array, produces a crosspoint configuration of the "destructive mark" type. There is no need to release a connection upon completion of a call since the half select current on a subsequent network connection will cause the release of the crosspoint if it is no longer required in the new connection. A network map indicating the closed or open state of the crosspoints is recorded in the temporary memory described later. Thus there is no need for a sleeve lead to be provided in the network as in electromechanical switching systems.

An early model of a machine developed by Western Electric Company for automatically winding the solenoids for the ferreed crosspoints is shown in Fig. 15. Here the entire shunt plate containing the molded assemblies for the crosspoints is oscillated in such a way as to wind four solenoids simultaneously. In the foreground are a two-wire and four-wire ferreed switch assembly before the crosspoints and remendur plates have been inserted.

A second type of ferreed is also required to act as a cut-off relay for ferrod sensors used for line supervision. This design, shown schematically in Fig. 16, may be operated or released by reversing the direction of current through the control winding. When energized in one direction, the remendur rod in the control winding is poled in a direction to aid the magnetic field from a permanent magnet. This causes closure of the reed contact. A pulse of current in the opposite direction switches the remendur field to oppose the permanent magnet to release the contact. The use of this device in connection with line supervision is described below.

4.6 *The Ferrod*

Line supervision is obtained by means of ferrods mentioned earlier. Several varieties of this device are illustrated in Fig. 17. Each of these assemblies contains two ferrods, one at either end of the assembly in a molded wire arrangement that is well adapted to mechanized manufacture using wire spring relay manufacturing techniques. The devices

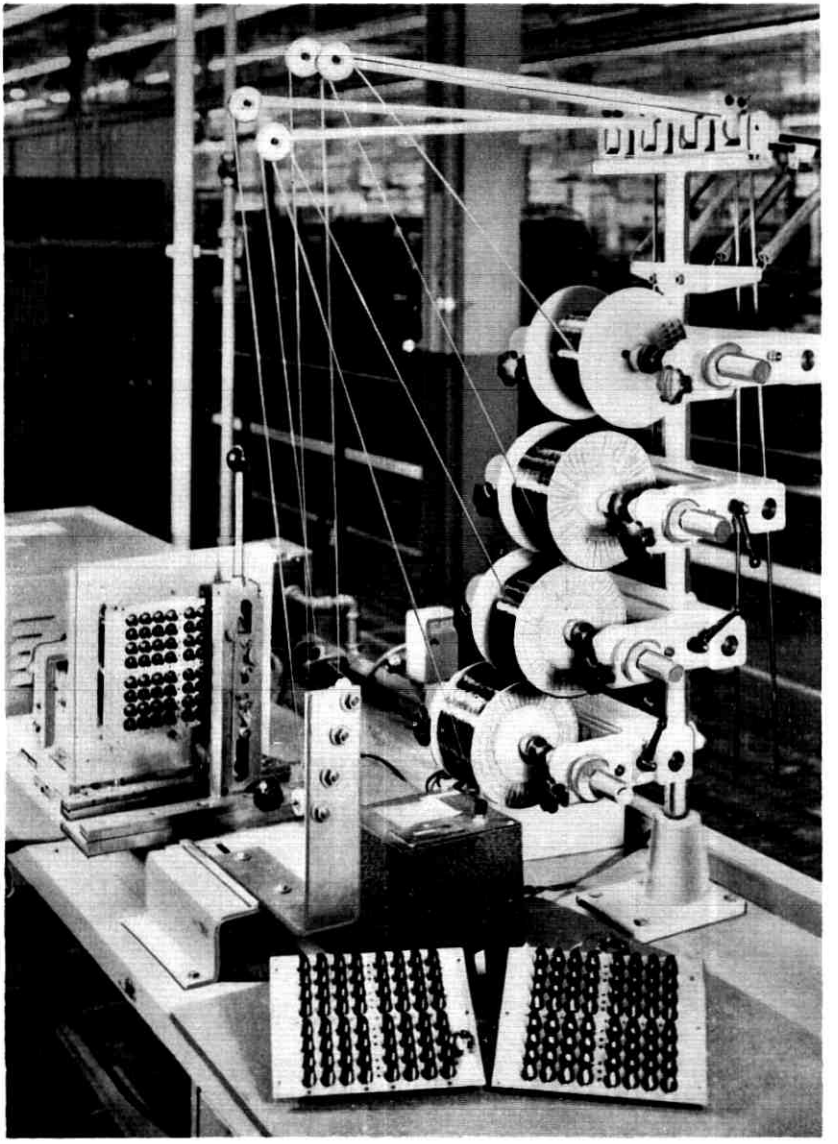


Fig. 15 — Ferreed switch automatic winding machine.

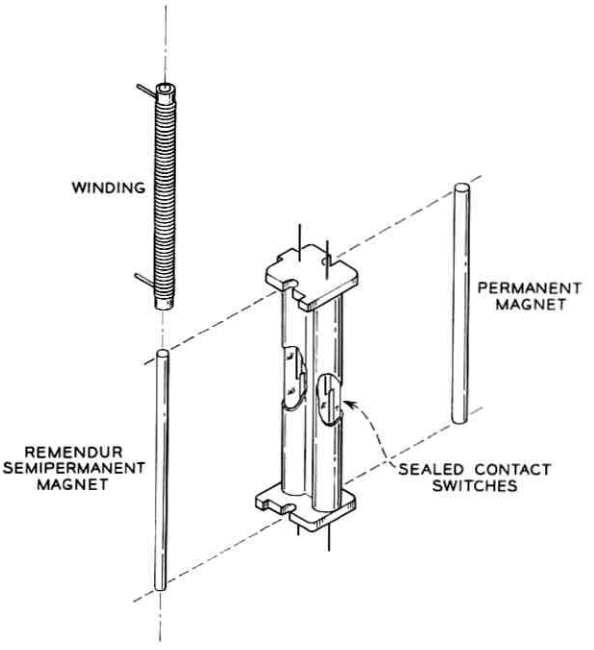


Fig. 16 — Bipolar ferreed assembly.

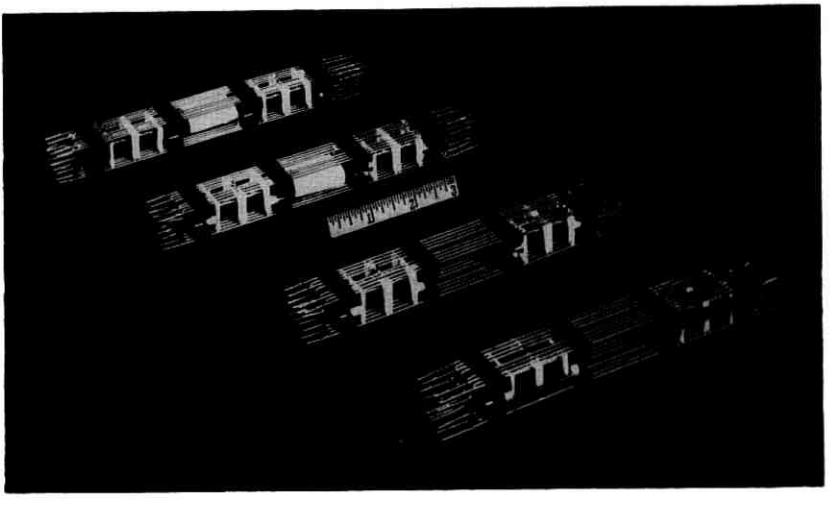
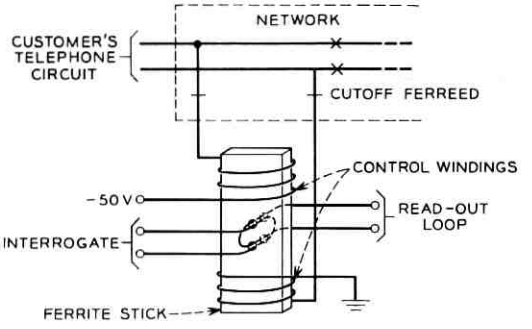


Fig. 17 — Four dual ferrod assemblies.

previously seen in the line switch frames mounted in a matrix configuration were the ends of a number of these dual ferrod assemblies.

A schematic diagram of this simple and reliable device is shown in Fig. 18(top). It consists of a rectangular ferrite stick surrounded by solenoid control windings connected in series with the customer's telephone line and talking battery. In the center of the ferrite stick are two holes through which two small coupling loops are inserted. In the absence of line current, i.e., when the customer is "on-hook," the ferrite stick is unsaturated and good coupling exists between the two single loop windings. Thus a 4-microsecond interrogating pulse transmitted from the scanner to the loop will produce a corresponding pulse in the read-out loop. When the customer goes "off-hook," the resulting cur-



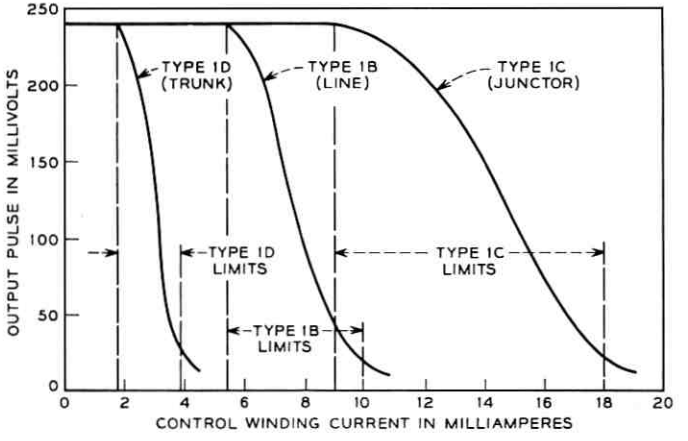


Fig. 18 — Top: ferrod in customer's line circuit; bottom: typical response of ferrods.

rent through the control windings saturates the ferrite stick, with the result that very little coupling exists between the interrogating and read-out loops. Thus it can be seen how this device provides a means for sensing the state of a customer's line at speeds compatible with electronic data processing.

The ferrod control windings are connected to the customer's line via a cut-off ferreed described above. When a service request is detected, an appropriate dial tone connection is set up through the switching network and supervision is transferred to a junctor or trunk circuit. The cut-off ferreed disconnects the ferrod associated with that customer's line to remove any transmission impairment which might otherwise be incurred. The cut-off ferreeds are mounted in a 1 × 8 ferreed switch assembly and may be seen adjacent to the 8 × 8 ferreed switches in Fig. 12.

Ferrods are used not only for customer line supervision but also at various other places throughout the system where high-speed sensing of direct current states is required. The sensitivities needed in these various applications call for three ferrod types, as indicated by the response curves shown in Fig. 18(bottom).

4.7 *Scanner*

Interrogate pulses for the ferrods are obtained from an electronic scanner of 1024 points. The ferrods are arranged in 64 rows of 16 ferrods per row, and the scanner selects one row of 16 ferrods simultaneously when requested to do so by central control. This is accomplished by the arrangement shown schematically in Fig. 19. Half microsecond pulses from the central control address bus are stretched to 4 microseconds and through a ferrite core matrix drive the appropriate row of 16 ferrods. Separate output amplifiers from the ferrod read-out loops supply central control with the "0" or "1" state of the corresponding ferrods through separate output amplifiers.

Fig. 19 also shows some of the features included to sense any malfunction in the scanning processes. One of these shown to the right and labeled ASW check is an "all seems well" pulse. This pulse indicates to central control that a particular row, and only that row, of ferrods was indeed interrogated. Another check feature is shown at the bottom left of the diagram in which a pulse is returned to central control to verify the fact that an "enable" pulse for the scanner was in fact received. Morris experience played a strong role in suggesting these provisions.

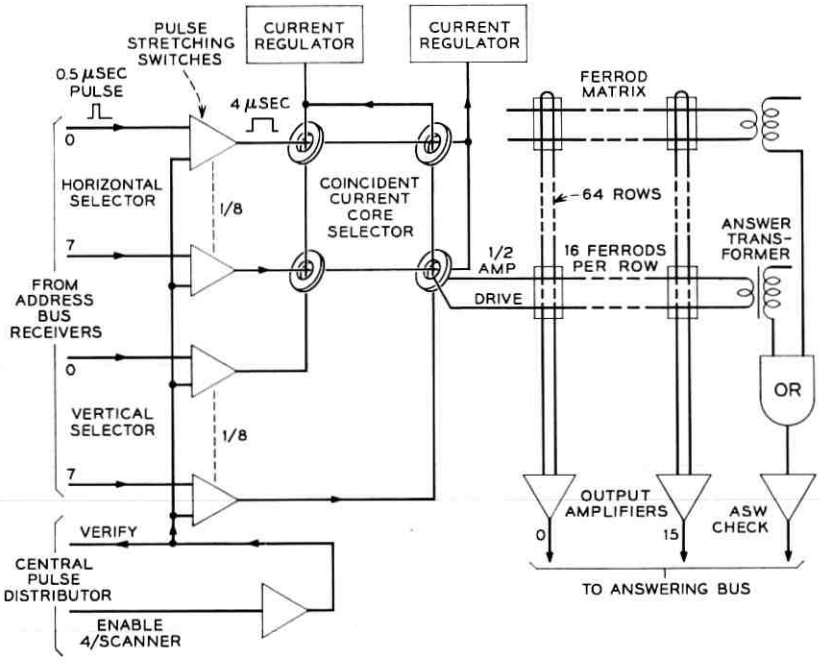


Fig. 19 — Functional diagram of a 1024 point scanner.

4.8 *Trunk Circuits*

Earlier it was mentioned that stored program control permits a major simplification in trunk circuits. Through this type of operation, it has been possible to reduce drastically the number of different types of trunk circuits required and to provide many of them on a plug-in basis with standardized factory wired frames for the receptacles. Compartments for plugging in the trunk circuits are shown in the universal trunk frame illustrated in Fig. 20. Each compartment accepts a trunk package containing two trunk circuits of a type indicated schematically in Fig. 21. From the notes on the diagram, the reader will observe the wide variety of circuit configurations made possible with program control. Other trunk circuits of this same general type are provided to meet special needs for interconnection with existing electromechanical offices. Some of these incorporate special networks to improve return loss.

A miscellaneous variety of trunk or service circuits are also required

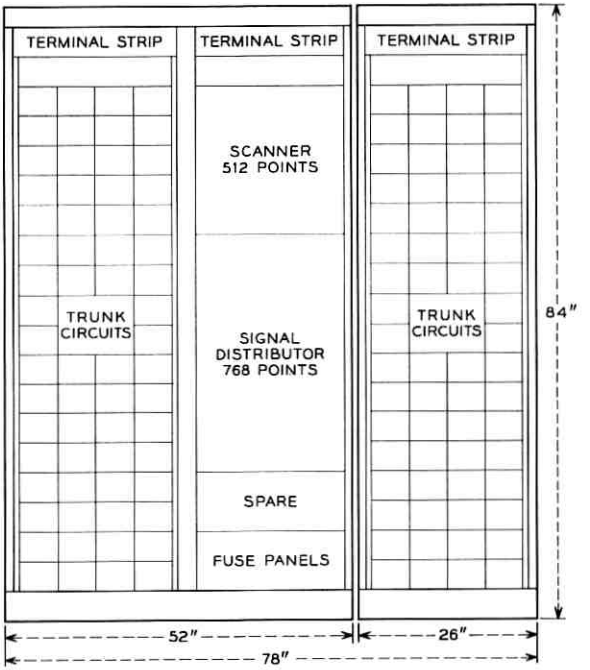


Fig. 20 — Universal trunk frame.

to provide such functions as dial tone, ring, audible ring, TOUCH-TONE receivers, and the like. These are mounted in a miscellaneous trunk frame as required and are permanently wired at the factory.

4.9 *Central Control*

Central control is the heart of the high-speed information processing common control equipment. It is a high-speed semiconductor logic machine designed to interpret instructions contained in the program store and to carry out the appropriate logical operations contained in each instruction. A photograph of one of the two central controls used in the system is shown in Fig. 22. Each central control is made up of approximately 2300 circuit packages containing approximately 14,000 transistors and 45,000 diodes. Typical plug-in packages and the nest into which they are plugged are shown in Fig. 23. Considerable development effort was devoted to the design of a highly reliable connector which could be manufactured at low cost. This was essential in view of the large number of plug-in packages used in the system.

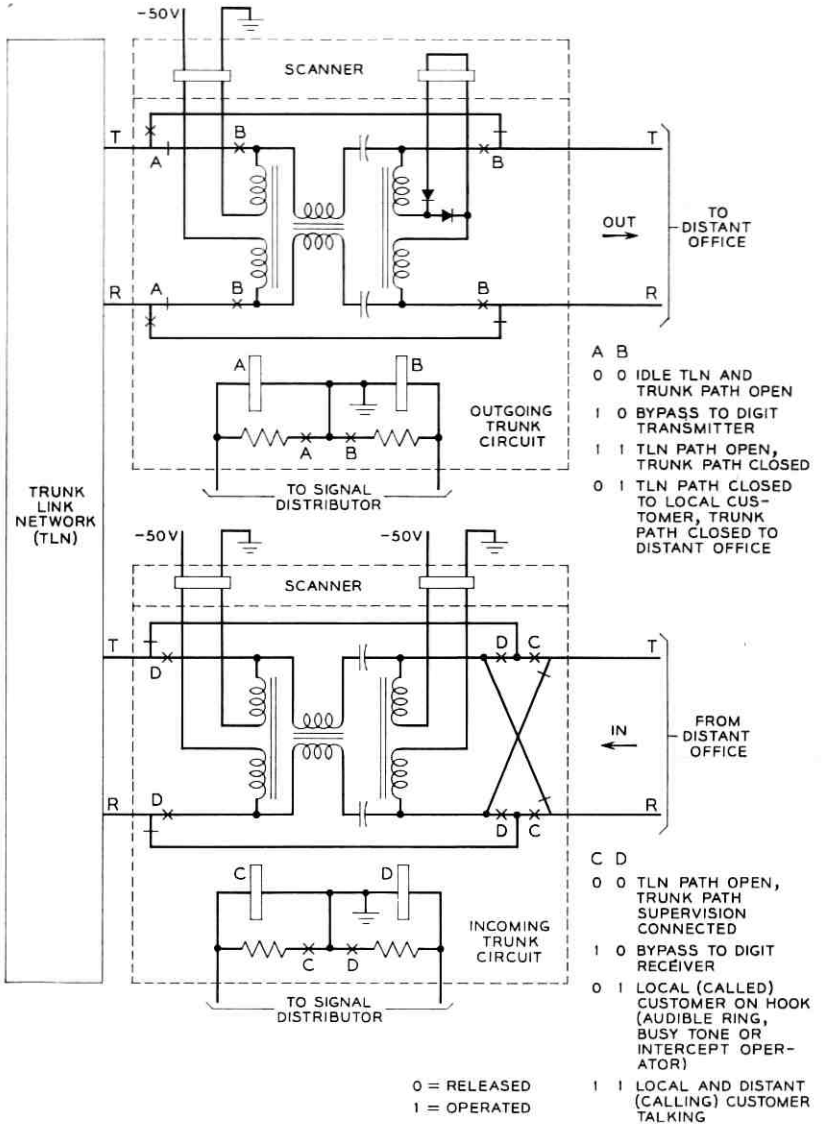


Fig. 21 — Simplified No. 1 ESS trunk circuits.

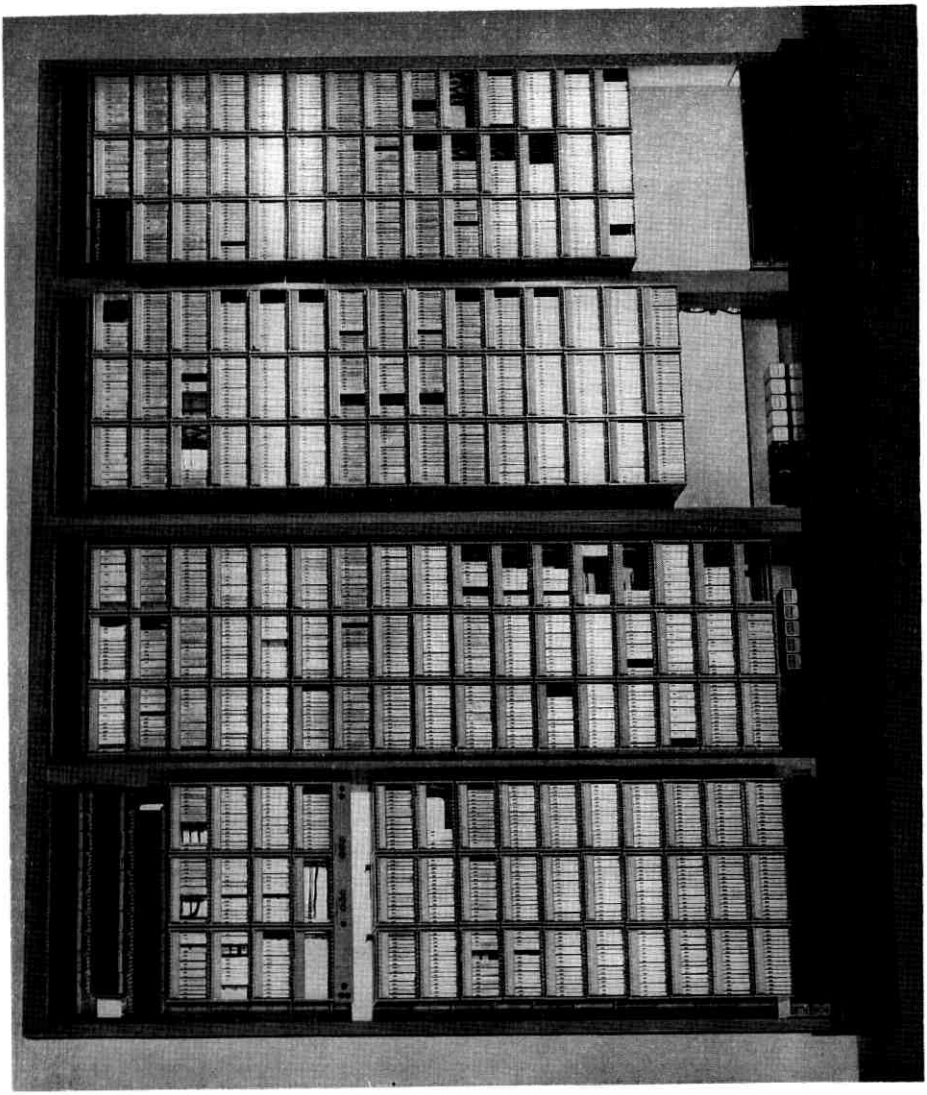


Fig. 22 — No. 1 ESS central control

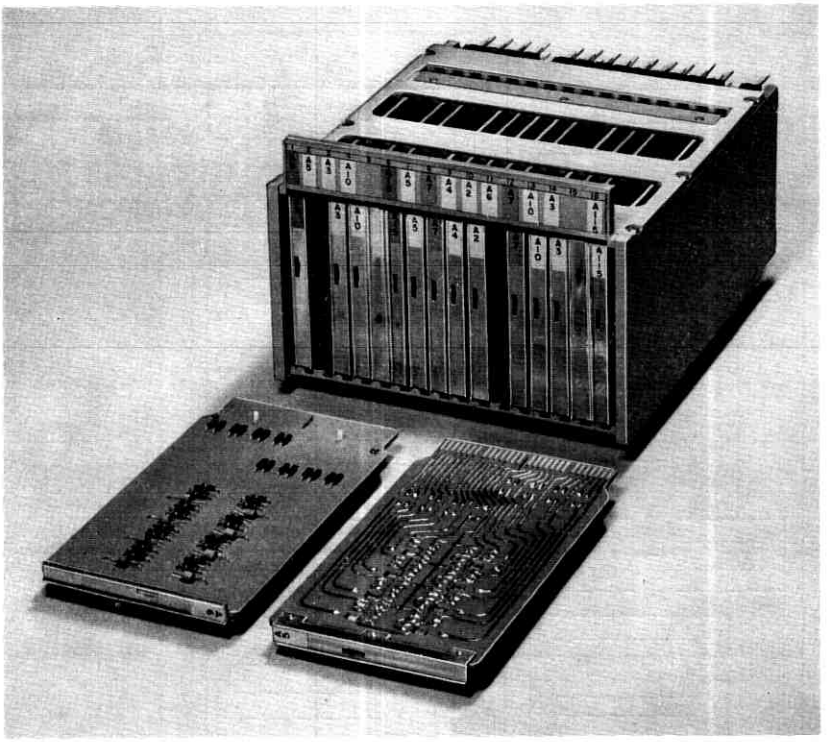


Fig. 23 — No. 1 ESS logic packages and nest.

Central control is word-organized to receive 44-bit instructions from the program store and to process information with the aid of the call store on a 24-bit word basis. Its cycle time is 5.5 microseconds.

A considerably simplified block diagram of central control is shown in Fig. 24. No attempt will be made here to describe this diagram. Instead, the types of actions the central control is designed to perform will be outlined briefly. The organization differs from that of a general purpose computer since the functions required in a telephone switching office are primarily logical rather than arithmetic operations.

A special program language with appropriate symbolic codes was evolved to optimize the performance of the system in processing information for telephone switching type operations. Each word in the program store shown at the top of the diagram is identified by a specific address designated in binary code. The program word located at that address contains an operation order, information as to where the data

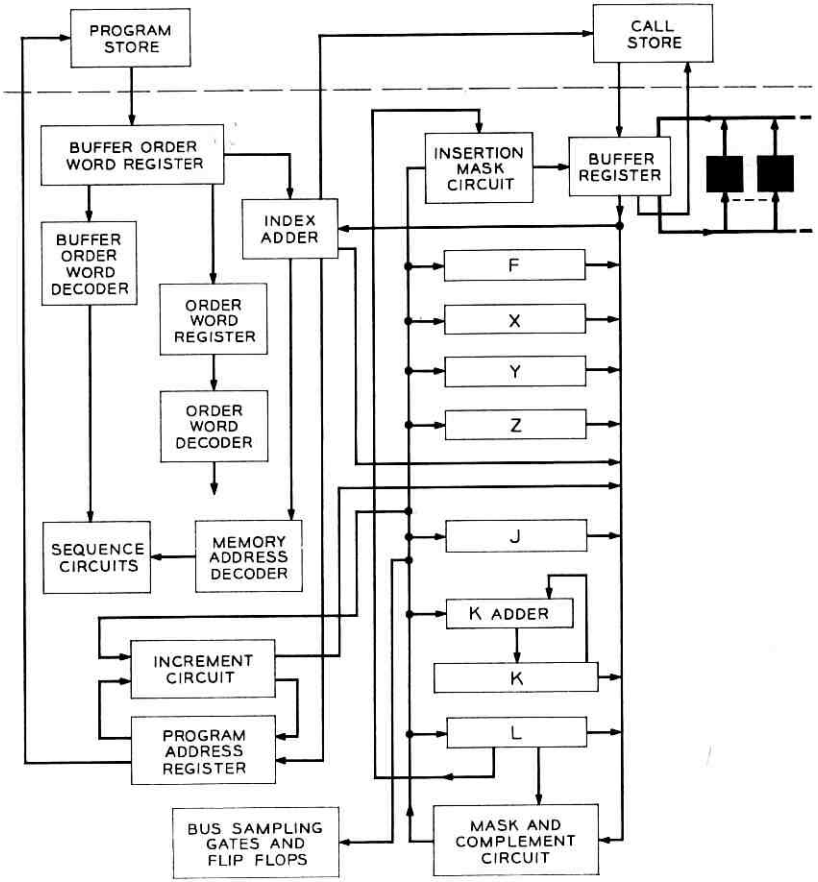


Fig. 24 — No. 1 ESS central control — simplified block diagram.

to be processed may be found in memory, and various other symbols which may request such operations as masking a portion of a word, complementing the word to be processed, or performing the logical union or product with another word. Various types of operation codes are included in the programming language. The class of operation code, an example of the symbology used, and the meaning of that symbology for several types of operation are shown in Fig. 25. The logic design of central control provides for interpreting these types of instructions and for carrying out the specified operation.

A frequently used order related to the "shift" order shown in the

CLASS	EXAMPLE	MEANING	NUMBER OF TYPES
MOVE	MK	MEMORY TO ACCUMULATOR (K)	28
ADD	AWK	ADD WORD TO K	11
SUBTRACT	SBR	SUBTRACT BUFFER FROM REGISTER	10
COMPARE	CMK	COMPARE MEMORY WITH K	5
LOGICAL	PMK	PRODUCT OF MEMORY WITH K ("AND")	24
	UWX	UNION OF WORD WITH REG ("OR")	
	H	SHIFT	
TRANSFER	TKAZ	TRANSFER IF K IS ZERO	26
COMBINED	TZRFZ	TRANSFER IF K IS ZERO, IF NOT FIND FIRST ONE AND ZERO IT AND SAVE BIT POSITION IN F REGISTER	12
	QMX	ROTATE K, MOVE MEMORY TO X REGISTER	66

Fig. 25 — Example of operation codes.

figure is called "rotate." This is similar to a shift order except that the bits of a word which might be shifted off the right-hand end of the register are saved by bringing them back to the left end of the register. For example, this instruction may be used to determine the "right-most one" in a binary word. Suppose that the bits of this word represent the busy ("0") or idle ("1") state of a group of trunks. The single order to determine the "right-most '1' " would immediately locate the first idle trunk.

In addition to logic circuitry to carry out operations of the types described above, central control includes a number of features provided for automatic maintenance purposes. These are listed in Table I.

4.10 *Program Store*

In Morris the program store was the flying spot store as already noted. For the commercial system, it was deemed desirable to eliminate electron tubes wherever possible, both from a reliability point of view and to simplify power supply arrangements. Fortunately, the twistor memory,[8] invented at Bell Laboratories, appeared on the scene early enough for this purpose.

An over-all view of the program store incorporating sixteen twistor modules together with access and read-out electronics is shown in Fig. 26. This store, of which at least two are provided in each office, provides a memory capacity of about 5.8 million bits organized into 131,000 words of 44 bits each. Any word in the store may be randomly accessed

TABLE I — CENTRAL CONTROL MAINTENANCE FACILITIES

1. Internally and externally generated maintenance interrupts.
2. Information in call store encoded with a parity bit checking both data and address.
3. Information in program store stored in Hamming code, the parity bit checking both address and data; can correct single errors, detect double errors in data, can detect single and double errors in address.
4. Program and call store reread facilities.
5. Round trip check of central pulse distributor enables output to peripheral units.
6. An internal check signal ("all seems well") is generated in program stores, call stores, and scanners; absence is detected by central control.
7. Synchronizing signal on all store communications.
8. Word matching between central controls of selected and selectable internal central control points:
   a. all call store communications normally matched,
   b. program store replies matched after a transfer,
   c. selected matching of busses, program store address register, and key decoder and sequence circuit outputs.
9. Error counters.
10. Emergency action circuit.
11. Off-line operation possible for selected system configurations.

by request from central control. The 44 bits in each word consist of 37 information bits, a 6-bit Hamming code for single error correction-double error detection, and a final over-all parity bit. The error detection and correction code is computed across not only the word of information to be read out of memory but also its address.

The vertical slots which may be seen in the twistor modules are designed to receive aluminum cards such as that shown in Fig. 27. Each of the cards has 64 columns of vicalloy spots arranged in 45 rows. Forty-four rows are used to store the 44 bits of a program word. The 45th bit, together with a row of elongated magnets shown on the upper edge of the card, serve to condition the magnetic properties of the twistor wire as the card is inserted. One hundred twenty-eight cards are used in each of the twistor modules, thus providing a storage capacity of 8196 44-bit words per module.

Of the 16 modules in the program store, 13 are allotted to program and the remaining 3 to translation information. Approximately half of the program is devoted to telephone call processing and administrative operations, and half is devoted to fault recognition and diagnostic programs designed to ensure dependability and simplify office maintenance.

The modules devoted to translation contain such information as line-to-directory number translation, class of service marks, trunk translation, abbreviated dialing lists, and the like. Approximately 16 types of translation information are used, and class of service designations are

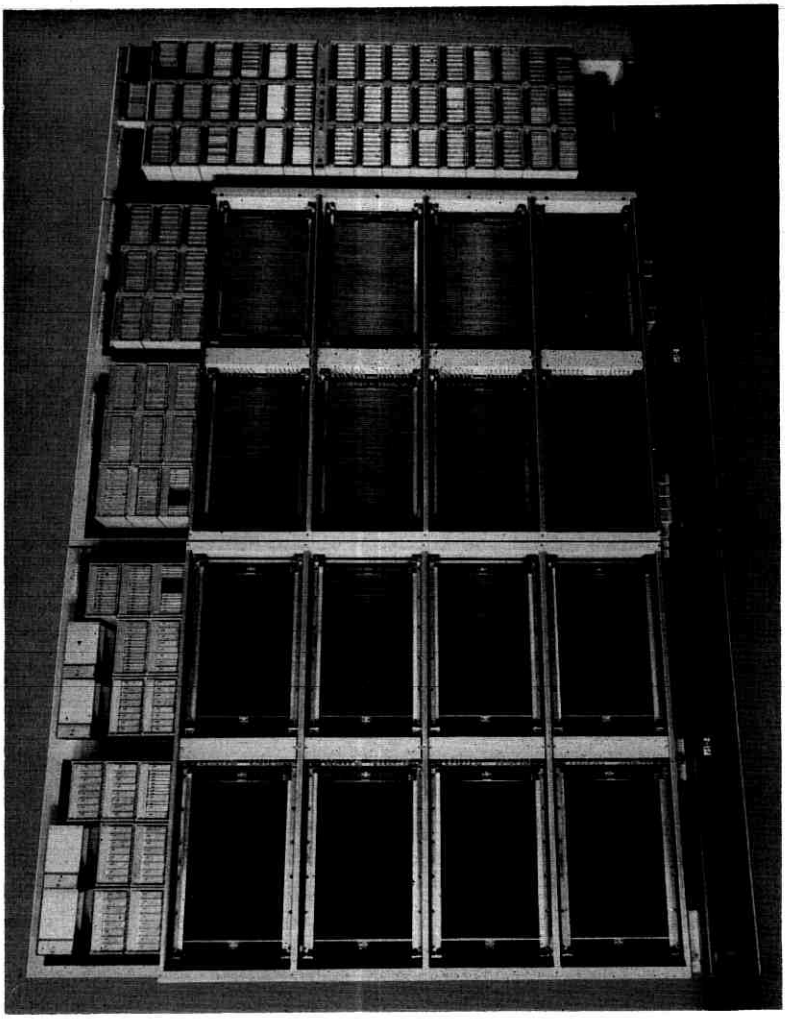


Fig. 26 — No. 1 ESS program store.

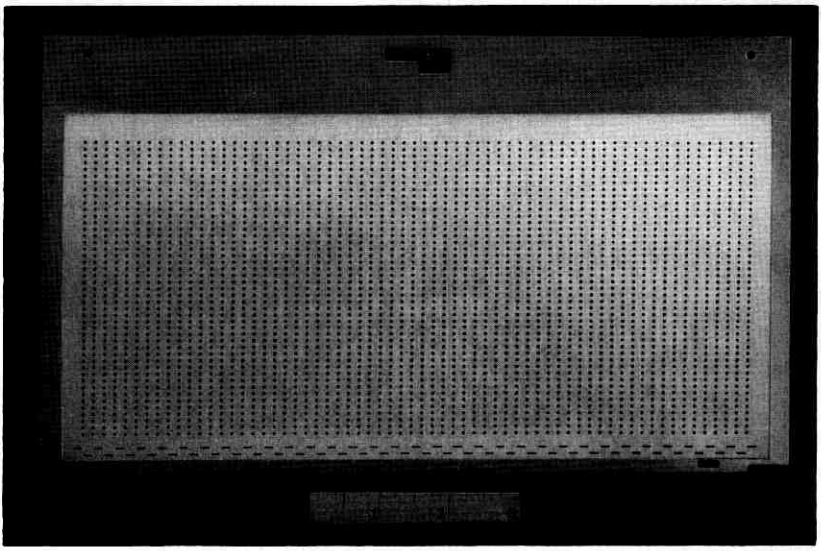


Fig. 27 — Magnet card for twistor store.

practically limitless within the capacity of the store. On the average, approximately three words of program store are required per line. Thus, the three modules provide translation for an office of about 8,000 lines. For larger offices, additional program stores would be required to provide additional translation capacity.

An understanding of the operation of the twistor may be obtained with the aid of Fig. 28. Forty-four pairs of copper read-out wires (of which four pairs are shown in the diagram) run adjacent to the vicalloy spots on the magnet card; each pair forms a balanced transmission line feeding a sensing amplifier. At each word position a single-turn coupling loop is disposed at right angles to the 44 pairs of twistor wires. A pulse can be driven through this loop by switching a ferrite core accessed by appropriate X and Y currents. One wire of each of the 44 pairs is surrounded by a spiral of permalloy tape. The vicalloy spots on the magnet card are located at the intersections of the twistor wire and the single-turn interrogating loop. In the absence of a magnet at that intersection, an access pulse in the interrogating loop switches the permalloy twistor wire and produces an output at the end of that wire. However, a permanent magnet at that intersection will prevent the permalloy tape from switching, with the result that substantially no output is obtained. Thus, the vicalloy spots on the magnet cards can be used to define the "0's"

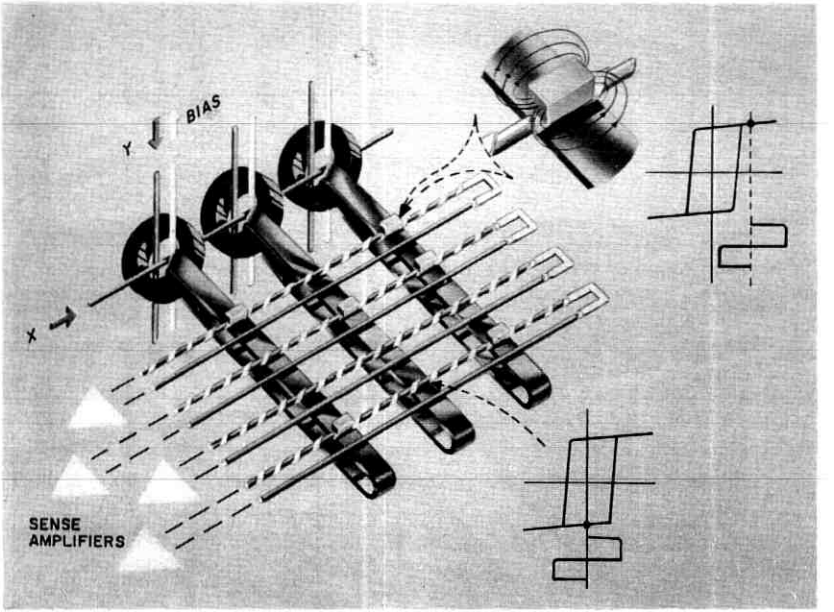


Fig. 28 — Principles of permanent magnet twistor.

and "1's" of a 44-bit word by either magnetizing or demagnetizing the tiny vicalloy magnetic material.

The 44 pairs of twistor wires are encapsulated in a plastic tape which is cemented in a continuous run to the vertical supporting members of the twistor module. This may be understood more clearly from Fig. 29, which shows an early version of a machine designed by Western Electric for fabricating twistor modules. A rear view of a completed module showing the access core matrix is shown in Fig. 30.

4.11 *Memory Card Writer*

It should be evident that either the program or translation information can be modified by simply changing the pattern of magnetic spots on the removable magnet cards. This permits a great deal of flexibility in office administration, not only in modifying translation but also in providing new service features. For these types of changes, no hardware or wiring modifications are required and service changes can be made in a minimum of time.

Changes in the magnetic bit pattern can be made with the aid of a

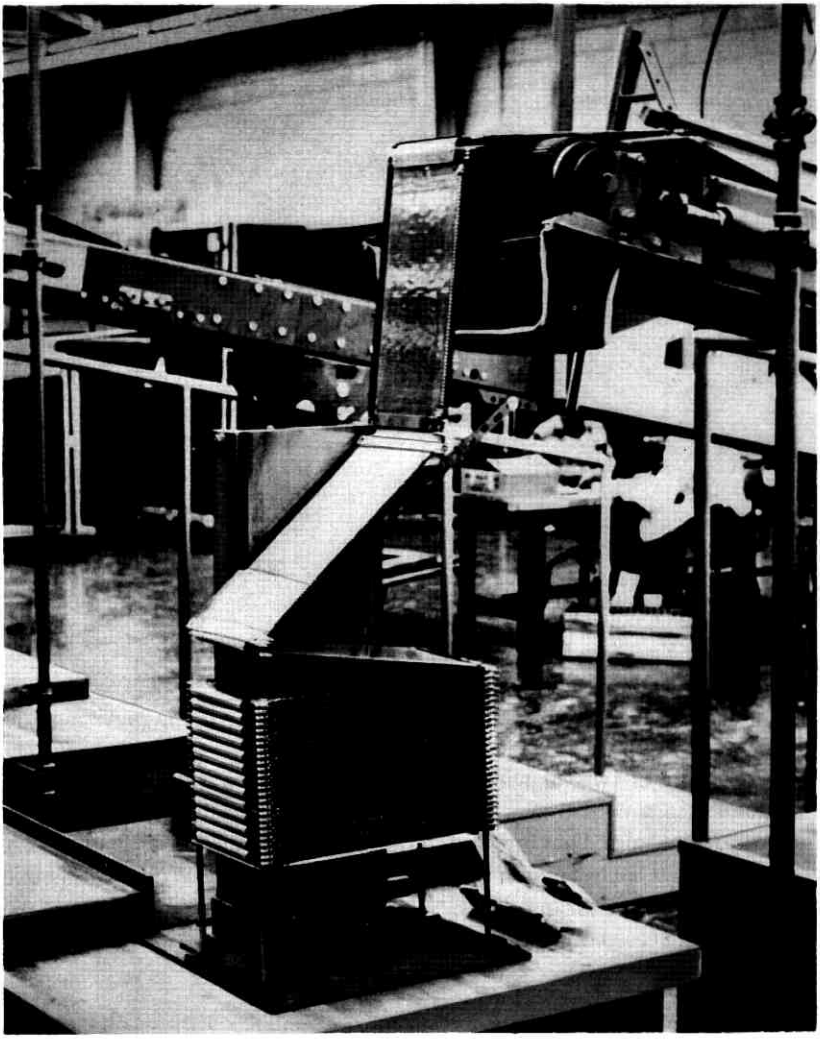


Fig. 29 — Twistor module assembly operation in Western Electric Company plant.

memory card writer, shown in Fig. 31. All of the cards in one twistor module are removed by a motorized program store card loader shown mounted vertically on the card writer. The card writer is arranged to withdraw one card at a time from this card loader and pass a 44-bit writing head across its surface. The card is then automatically replaced

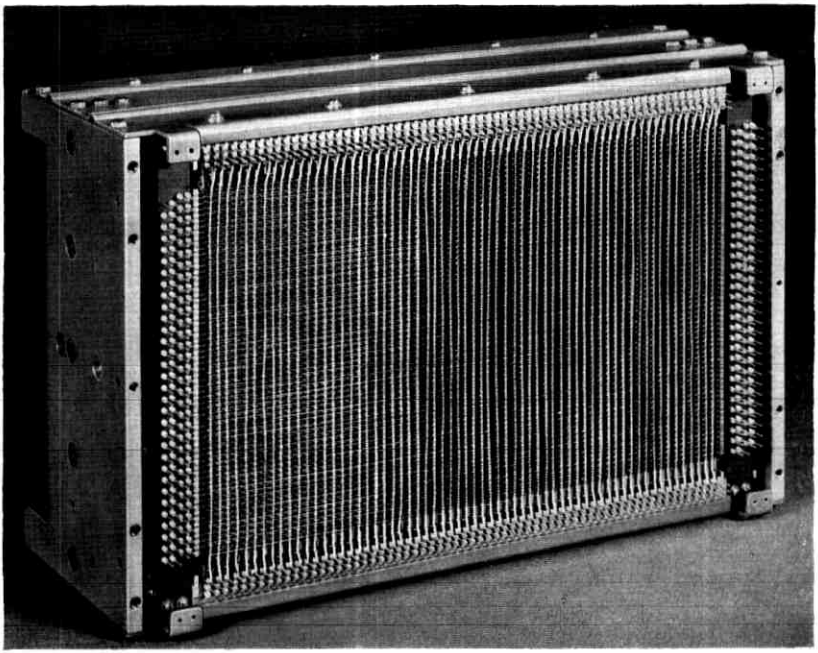


Fig. 30 — Twistor module access matrix.

in the loader, and the unit is indexed upward to place the next card into position for withdrawal.

The appropriate information for the 44 bits on the writing head can be obtained from a tape reader for initial magnetization of the cards at the factory, or from the call store via central control in an operating office. In the latter case, information is inserted into the switching system via the maintenance teletypewriter into a "recent change" space in the call store. Before this space is completely full, the temporary translation information can be automatically transferred to the twistor memory in the program store via the card writer as outlined above.

By arranging to use the call store as a temporary repository for change information, it is possible to respond very rapidly to customer requests for a change of service. For example, a remote teletypewriter can be provided to a service order clerk who can type information (such as abbreviated dialing lists) directly into the system. Service can be activated as soon as the service order clerk has finished typing the information. In processing telephone calls, central control examines the "recent change" space of the call store before referring to the program store

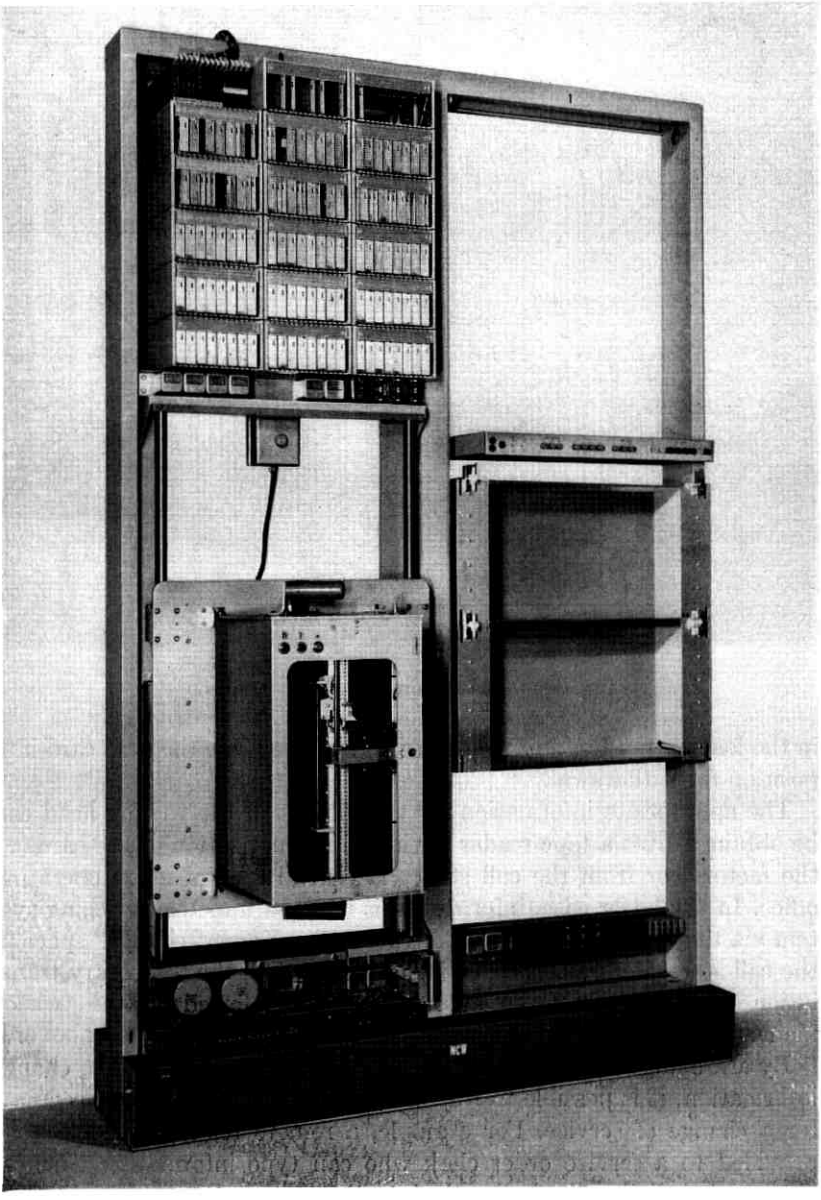


Fig. 31 — No. 1 ESS twistor memory card writer.

translation tables. It is expected that clearing of the "recent change" space and transferring the information to the program store will be required no more than once a week, even in a fairly active office.

4.12 *Call Store*

Reasons similar to those given for replacing the Morris flying spot store with the twistor store led also to replacement of the Morris barrier grid store by a solid-state temporary memory. For this purpose the ferrite sheet was chosen as the memory element.

A single ferrite sheet is shown in Fig. 32. This sheet contains an array of 256 holes on a 16 × 16 grid, each of which acts as an individual ferrite core. The difficult threading operation common to a ferrite core matrix is largely overcome by the technique of plating one of the leads in a continuous path through the holes in the ferrite sheet. A number of these sheets can then be stacked to provide the memory capacity required and the additional wiring added in a relatively simple operation. This is indicated schematically in Fig. 33, and a completed memory module having capacity of 2,048 words of 24 bits each is shown in Fig. 34.

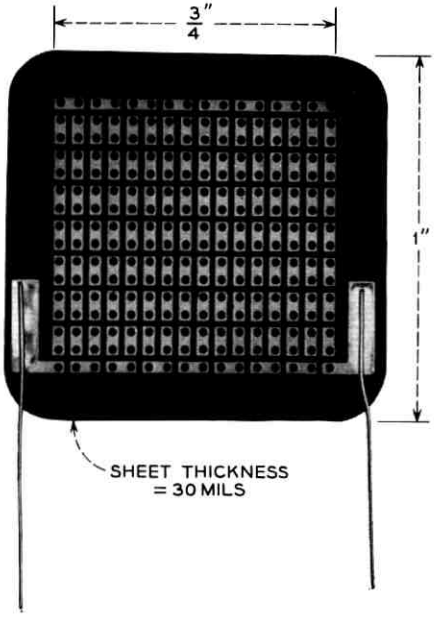


Fig. 32 — Ferrite sheet for No. 1 ESS temporary memory.

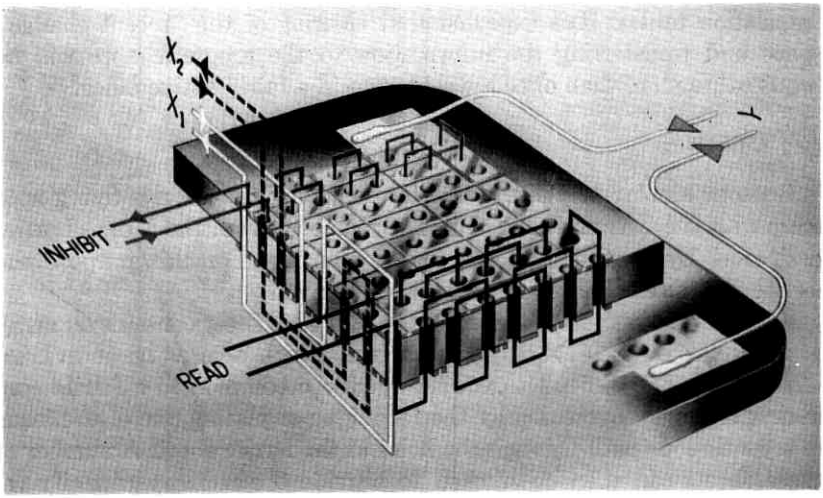


Fig. 33 — Access wiring for ferrite sheet memory.

As indicated in the caption for this figure, four such modules are used in each call store. They are mounted behind the blank panel shown in the photograph of a call store, illustrated in Fig. 35. The electronic packages associated with call store operation may also be seen in this photograph.

The number of call stores required in a particular office varies with size and traffic but will never be less than two for the smallest office because of the duplication requirement.

4.13 *Master Control Center*

The interface between man and machine in No. 1 ESS is the master control center, two portions of which are shown in Figs. 36 and 37. The alarm and display section on the right-hand panel of Fig. 36 indicates which of the duplicated common control units are currently in charge of the office as well as the condition of the off-line units. As already noted, switching between these units is normally made under automatic control of the system. However, push buttons provide for manual intervention. On the test panel at the left are various keys and lamp indications from which line-load control can be exercised under unusual traffic conditions. Facilities are also provided for performing certain system tests.

The main interface with the machine is the teletypewriter shown in

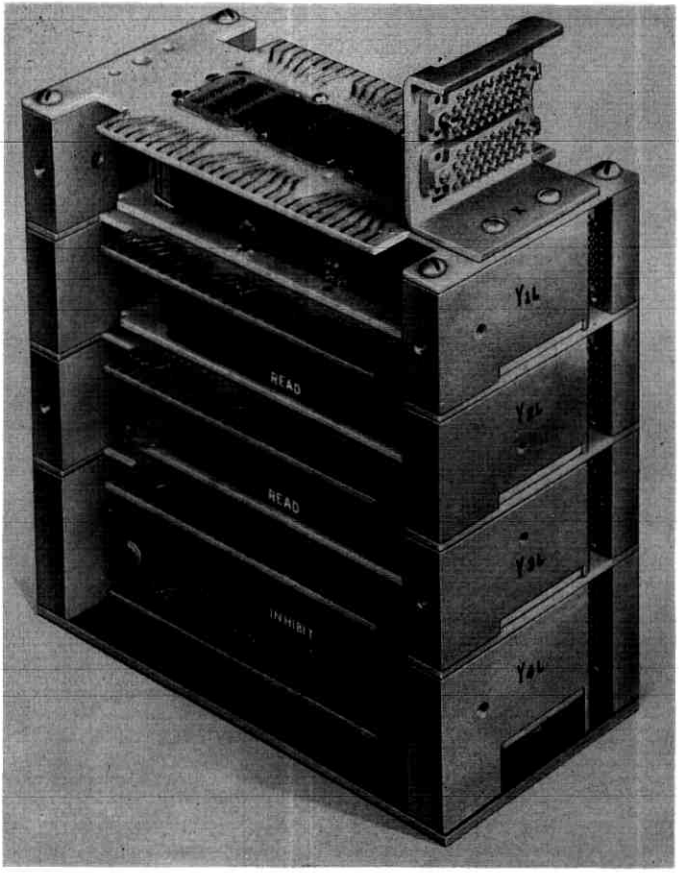


Fig. 34 — The No. 1 ESS ferrite sheet module has a capability of 2048 words of 24 bits each or a total of 49,152 bits. Four such modules are used in each call store.

Fig. 37. It can be used by the operating personnel to request the machine to perform a variety of functions and is also used to print out messages which the machine wishes to give to the maintenance man. Examples of the former are the use of the teletypewriter to update translation information or to insert special service changes such as customer abbreviated dialing lists. The teletypewriter may also be used to request a print-out of traffic data or to perform certain maintenance test sequences.

Under normal circumstances, it is anticipated that No. 1 ESS offices will be unattended. Provisions are therefore made for operation with remote teletypewriters. One might be provided for the service order

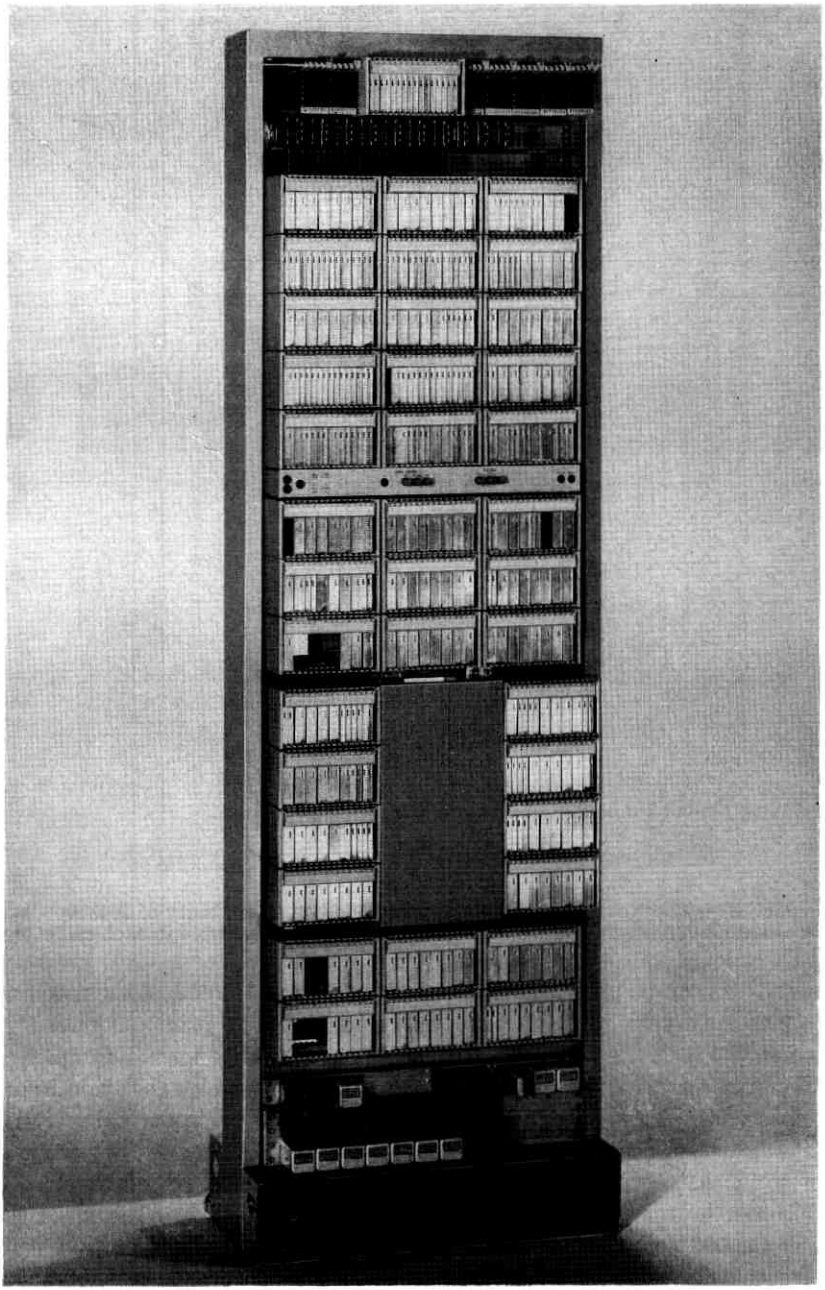


Fig. 35 — No. 1 ESS call store.

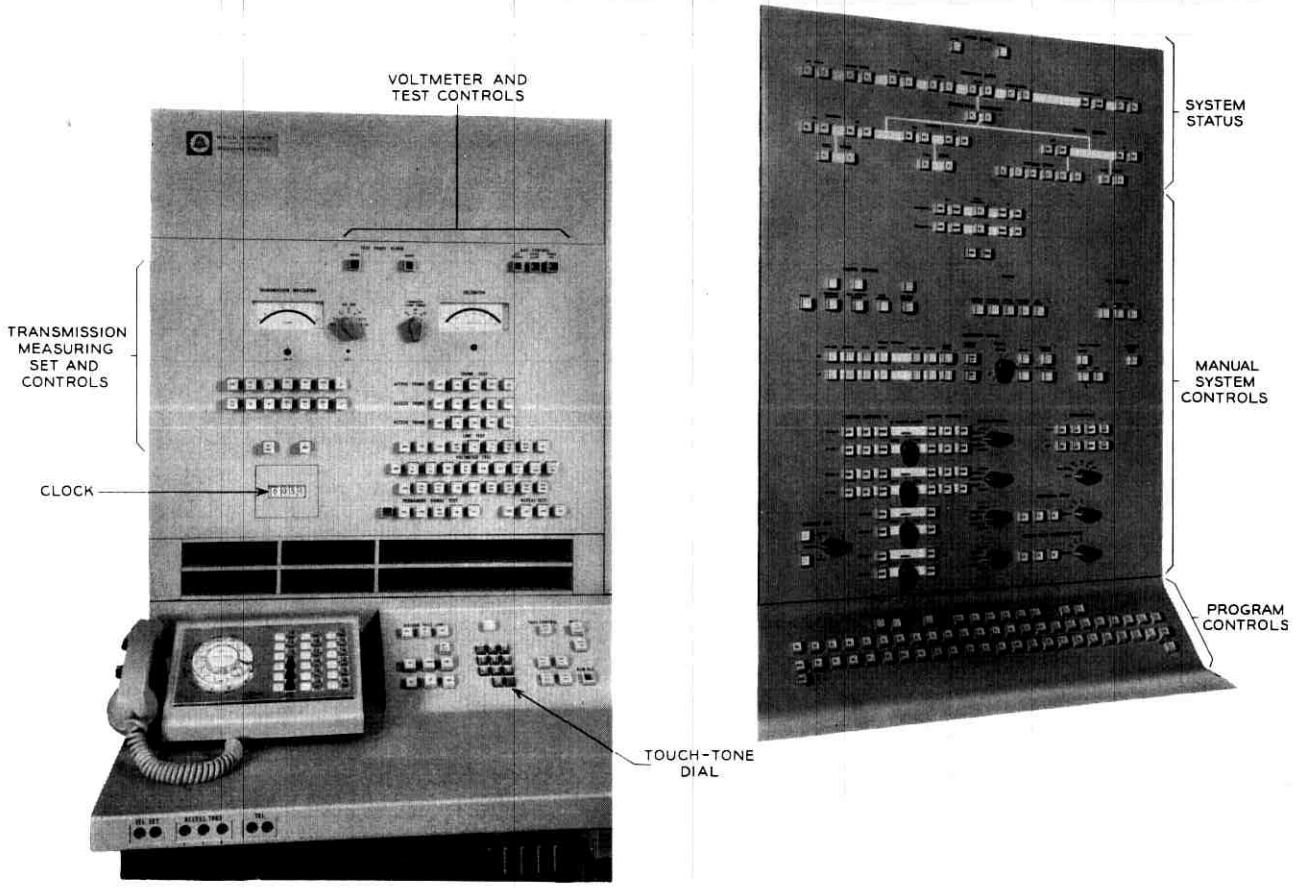


Fig. 36 — No. 1 ESS master control center — test panel and alarm and display section.

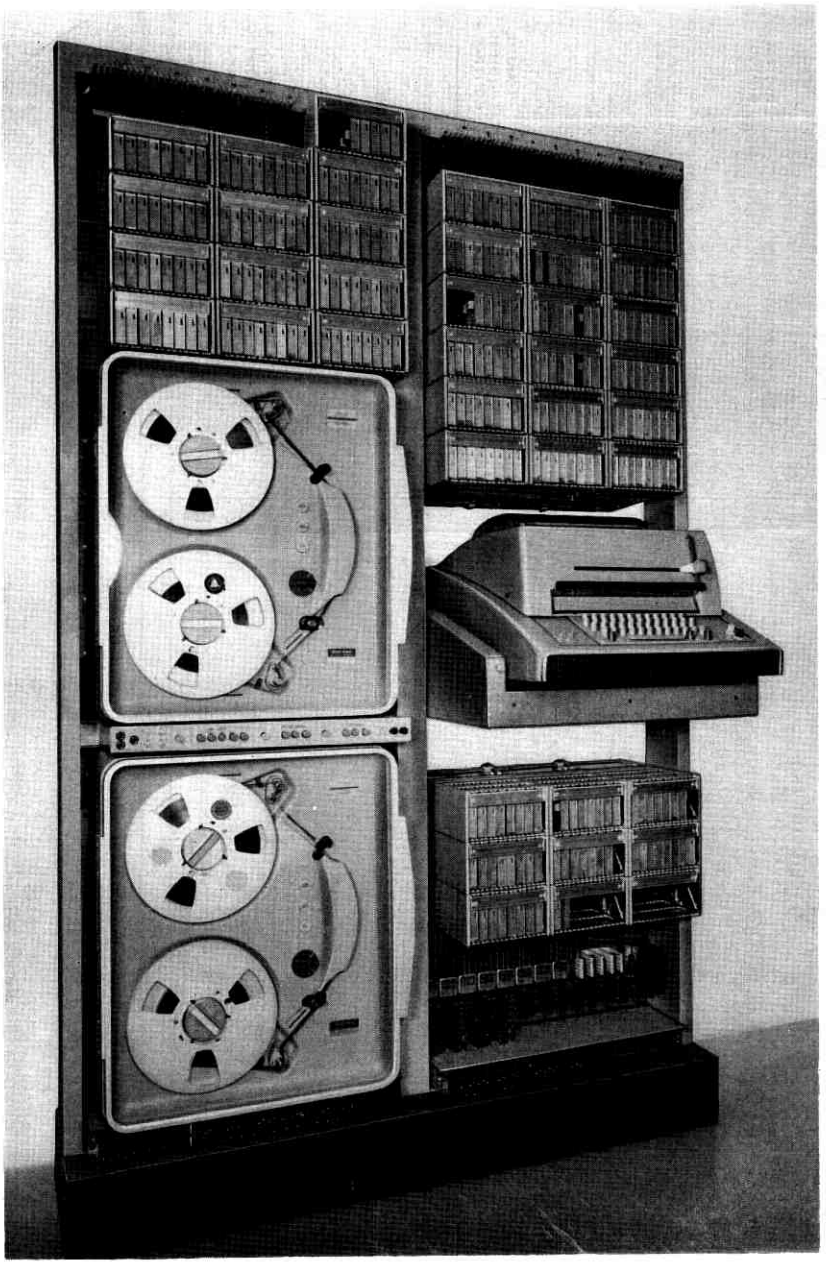


Fig. 37 — No. 1 ESS master control center — TTY and AMA recording.

clerk while another could be located in an attended office where a maintenance man would have access to maintenance information generated at the unattended office.

Next to the teletypewriter in Fig. 37 are two magnetic tape recorders used to record information for automatic message accounting. These are "write only" recorders and operate only when sufficient information has been accumulated in the call store to justify recorder operation. Information on the magnetic tape is recorded in blocks and in a format suitable for processing in a centralized message accounting center. Two recorders are provided for redundancy as well as to permit tape changes without interrupting recording operation.

The remaining portion of the administration center is the twistor magnet card writer, described earlier in connection with Fig. 31.

4.14 *Size and Power Requirements*

In the design of No. 1 ESS, the height of frames was limited to 7 feet rather than the 11-foot, 6-inch height generally found in electromechanical offices. This height eases the maintenance problem and also permits installation in conventional ceiling-height buildings. In spite of the reduced frame height, the floor space requirements are less than one-half the floor space required for an equivalent No. 5 crossbar office.

No. 1 ESS derives its power from +24-volt and −48-volt storage battery plants continuously charged from commercial power with diesel engine generator back-up. Circuits are designed to operate over the full discharge range of the batteries, and there is no requirement for end-cells or counter-cells. A standard ringing generator is used with programmed selection of ringing phase to provide immediate ring on customer lines.

Various tones required in the system, such as dial tone, audible ring, high and low tones, re-order tone and the like, are supplied by solid-state oscillators with electromechanical interruption at the appropriate rates. These tones are made available to the system from a balanced terminated impedance to avoid transmission impairment during tone application.

4.15 *Programming*

The collection of equipment frames comprising a No. 1 ESS office cannot process a single telephone call without the program which defines the myriad steps required to carry out the appropriate system operations. As already noted, this program for a typical office will contain 100,000

or more 44-bit words for telephone operating and maintenance routines and perhaps 30,000 words of translation information. The problem of writing this program is a major one indeed and occupies the time of a large staff of engineers and programmers.

A major part of the programming activity is devoted to defining the various features which the office is intended to provide. A second portion of the problem is to convert this design information into the symbolic language developed for No. 1 ESS and to process the resulting symbolic program into a form suitable for use by the memory card writer. A third and important activity is the testing of these programs on the actual No. 1 ESS to locate and correct errors which may occur during the first two steps.

To simplify and expedite the conversion of symbolic programs into binary information on a magnetic tape for the card writer, a special compiler program has been written. This compiler, known as PROCESS III,* is designed for use with an IBM 7094 scientific computer. It converts symbolic information (punched onto cards) into binary words to which the compiler automatically assigns absolute memory addresses for use in the twistor store. The compiler also supplies a binary tape which can be run with a simulation program in the general-purpose computer for initial program "debugging." A schematic representation of the flow of information in this process is given in Fig. 38.

The program supplied with each central office must uniquely define both the service features to be provided by that office and a programmed definition of the hardware available in that office. In the latter category, for example, would be a programming statement of the number of originating registers in call store, number and types of trunks available to the office, network concentration ratio, etc. However, the preparation of a complete specialized program for each Central Office would be impractical. Fortunately, many of the operating and maintenance characteristics are common to a large class of offices, and only a small part of the program need be produced uniquely for each office. The common portion of the program has been called the generic program. This includes the maintenance program suitable for all offices of the class and the operating program, which includes all service features and operating characteristics anticipated for offices of that class. A small portion of the program, called a "parameter table" is specially prepared to meet the operating company requirements for the individual office. With the use of general-purpose computers to mechanize the conversion of operating

* *PRO*gram for *C*ompiling *ESS*.

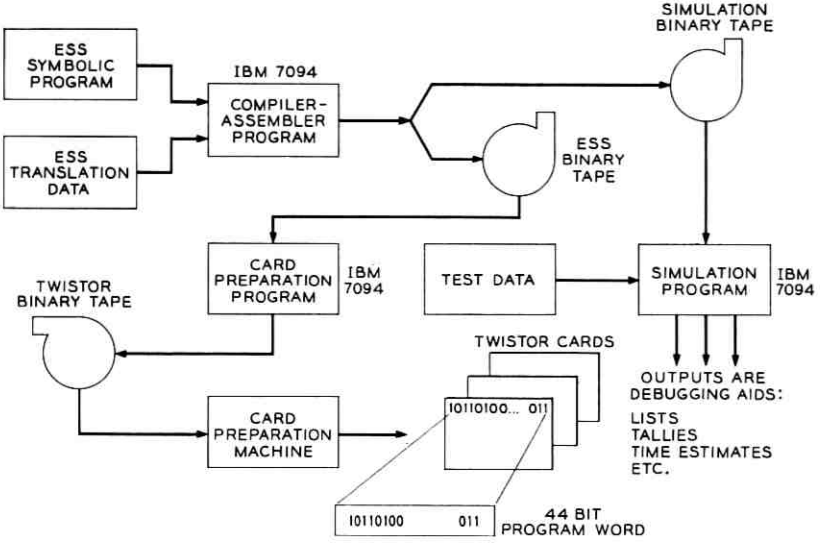


Fig. 38 — No. 1 ESS program information processing.

company requirements into specific office programs, it is possible to generate the twistor information for each office in a very short time.

Special programs are also provided for the twistor memory to serve as both a laboratory tool and by Western Electric Company installation crews. These are called "X-ray programs." When inserted in the program store, they are used to exercise the No. 1 ESS circuits to insure that all installation interconnections have been properly made and to locate any troubles which may have occurred as a result of shipping damage.

V. PROGRESS IN PRODUCTION

During the development of No. 1 ESS, close liaison was maintained with engineers of the Western Electric Company to take advantage of their production experience in the initial designs. This close collaboration resulted in apparatus and equipment designs compatible with high-volume, low-cost manufacture. It also permitted Western Electric to develop special production machines during the development interval. As a result, production systems were available at a much earlier date than would have been possible otherwise.

The Western Electric Company's plant at Columbus, Ohio, produced the first No. 1 ESS during 1962 to serve as a test model at Bell Telephone Laboratories, Holmdel, New Jersey. The frames for the first

commercial office at Succasunna, New Jersey, were produced in 1963 and by the end of 1964 some 1470 frames had been shipped by Columbus for installation at nine more central office locations. Five of these were equipped with four-wire switching networks to serve military users.

The production rate at Columbus will increase rapidly during 1965, and deliveries will also be made from Western's Hawthorne Works in Chicago to meet the rapidly growing demand for this new system. Within the next eight years the combined output of the two Western Electric plants is expected to reach a level of 3,000,000 lines per year.

Several views of early manufacturing operations at Columbus are shown in Figs. 39, 40, and 41.

VI. NO. 101 ESS[9,10,11,12]

6.1 *Design Considerations*

No. 1 ESS will provide modern switching services not only to residence telephones but also to the business community. However, it will be many years before the existing electromechanical switching plant will be superseded by this new central office system. In the meantime, it

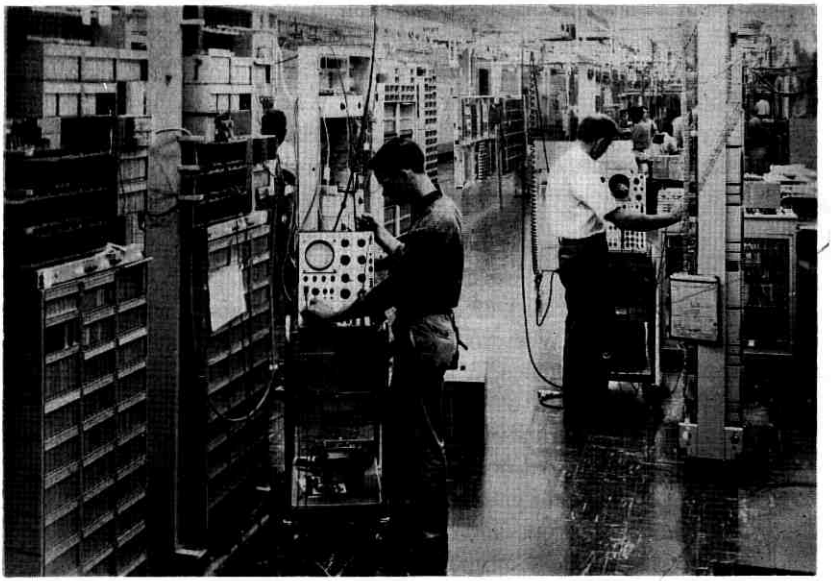


Fig. 39 — Western Electric Company manufacturing operations: monorail area — frame testing.

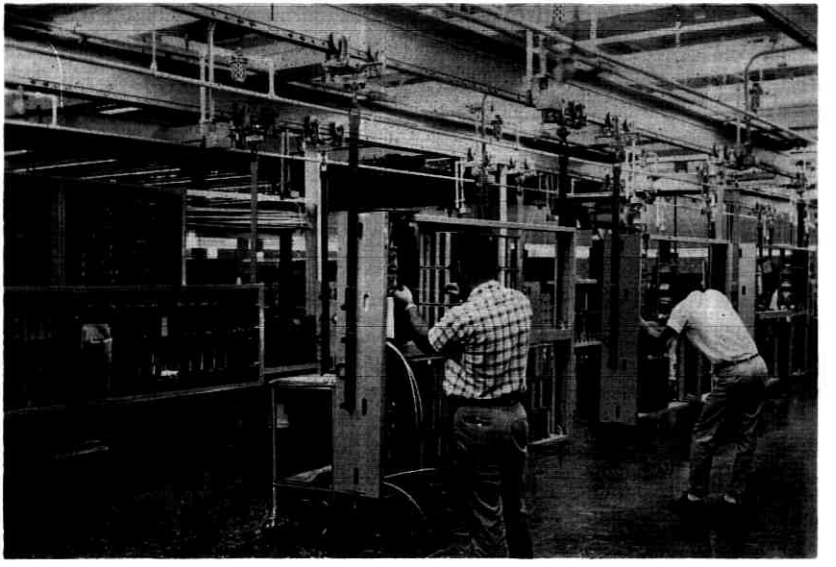


Fig. 40 — Western Electric Company manufacturing operations: monorail area — frame assembly and wiring.

Fig. 41 — Western Electric Company manufacturing operations: unit surface wiring.

seemed appropriate to provide business customers with the new services, made economically attractive with electronics, by supplementing the standard electromechanical central office equipment. No. 101 ESS was designed to fill this need.

An analysis of PBX customers being served by the Bell System indicated that some 80 per cent of existing electromechanical PBX's serve less than 200 extensions. As an initial offering, it was therefore decided to develop the system in this size range. It was also decided that the attractive features of stored program control already discussed in connection with No. 1 ESS should be provided with the PBX design.

It turns out, however, that stored program control systems are currently economical only in large sizes — much larger than the 200-line capacity envisioned for initial service. This led to a concept in which a stored program group control located in the central office would serve a number of outlying PBX switch units on the business customers' premises. This is the concept used in No. 101 ESS. It has the further advantage that most of the maintenance and administration activity for the several PBX's can be performed in the central office, thus reducing servicing costs.

In No. 101 ESS, switching at the customer's premises is performed by the use of a time-division switching network. One of the considerations which led to the choice of time division for this application was a desire to minimize floor space requirements at the customer location. A second consideration was that a time-division switch operates silently and can be installed in any available space without considering acoustic noise interference to customer activities. To minimize installation time on the customer's premises, the switching unit is contained in a single cabinet provided with plug-in connectors.

6.2 *System Organization*

Fig. 42 illustrates the system plan chosen on the basis of the considerations outlined above. A group control unit of the stored program variety is located in the central office building near the electromechanical central office with which it is to be associated. Central office trunks and control data links interconnect this group control to outlying time-division switch units located on the premises of a number of business customers. In this system the group control is designed to handle a maximum of 3200 extensions divided among as many as thirty-two switch units. The maximum capacity of each switch unit in this first offering is 200 lines although larger switch units are under development. By sharing the

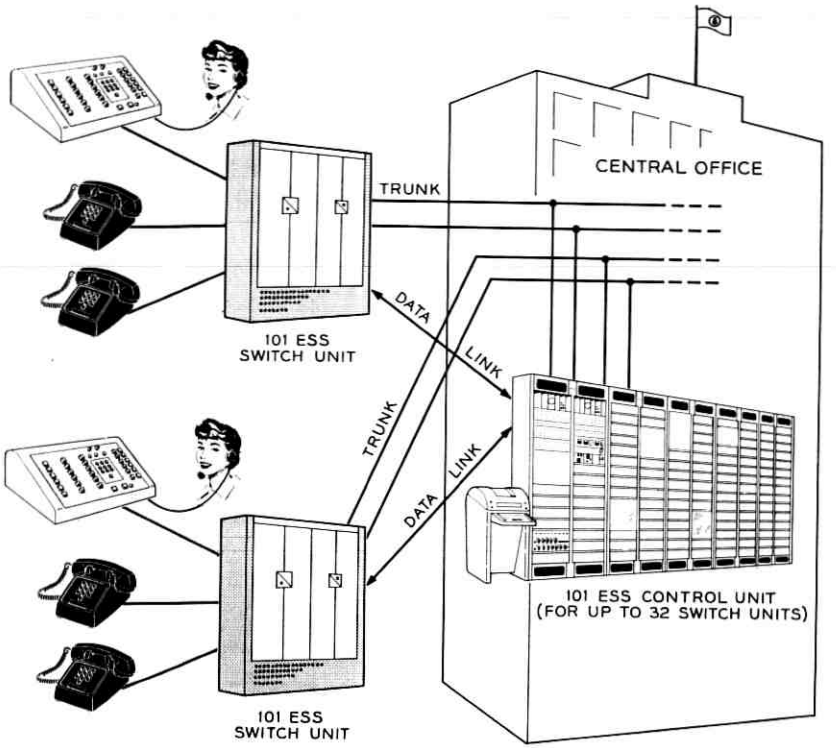


Fig. 42 — 101 Electronic Switching System.

control among a number of business customers, the advantages of stored program flexibility can be obtained economically.

The trunks shown on Fig. 42 between the switch unit and central office provide access between the PBX and the Bell System network for outgoing and incoming calls. Connections shown between those trunks and the control unit are for trunk seizure and control only, and do not provide a voice-frequency transmission path.

The data links shown between the switch unit and control unit are of two types. One is a 4-wire, two-way data link for interchange of digital control information, and the second provides a transmission path for dialed or TOUCH-TONE digits. The data links are ordinary voice-frequency pairs and provide for a data transmission rate of about 750 bits per second. There is no technical limit to the distance between switch units and control unit.

A simplified block diagram of an individual switch unit is shown in Fig. 43. All extensions and trunks to the central office are multipled through electronic gates to two time-division highways or busses. Each of these busses is equipped with its own memory and control to provide system redundancy. The use of two busses doubles the number of time slots available to the customers, provides redundancy in the case of bus failure, and makes possible a very convenient means for establishing conference calls.

6.3 *Group Control*

A photograph of the group control unit is shown in Fig. 44. The four center bays in this equipment line-up contain two stored program call processing units which are essentially mirror images of each other. The system program is contained in the twistor memory module mounted in the lower portion of the frames. A third twistor memory module in one of the frames stores line information, abbreviated dialing lists, class of service marks, and the like. This store is not duplicated since its

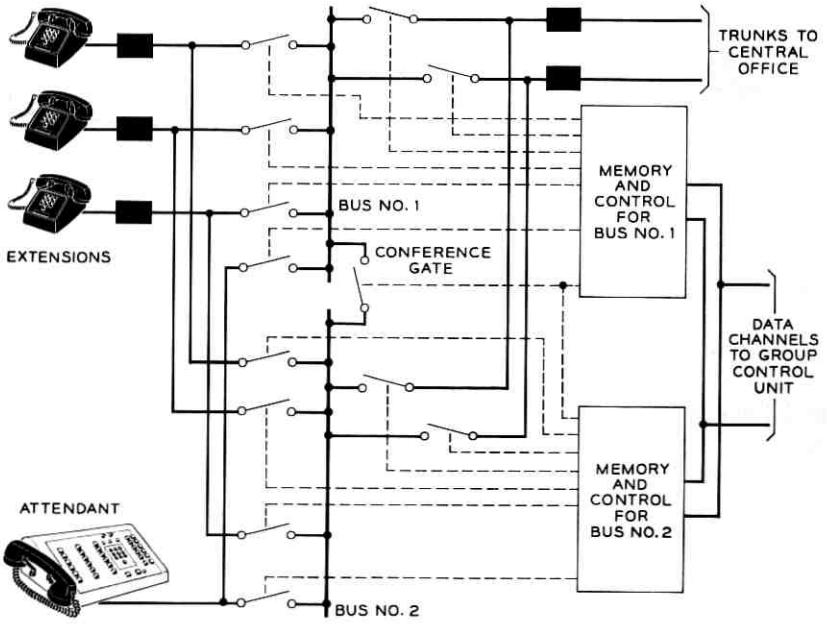


Fig. 43 — 101 ESS switch unit.

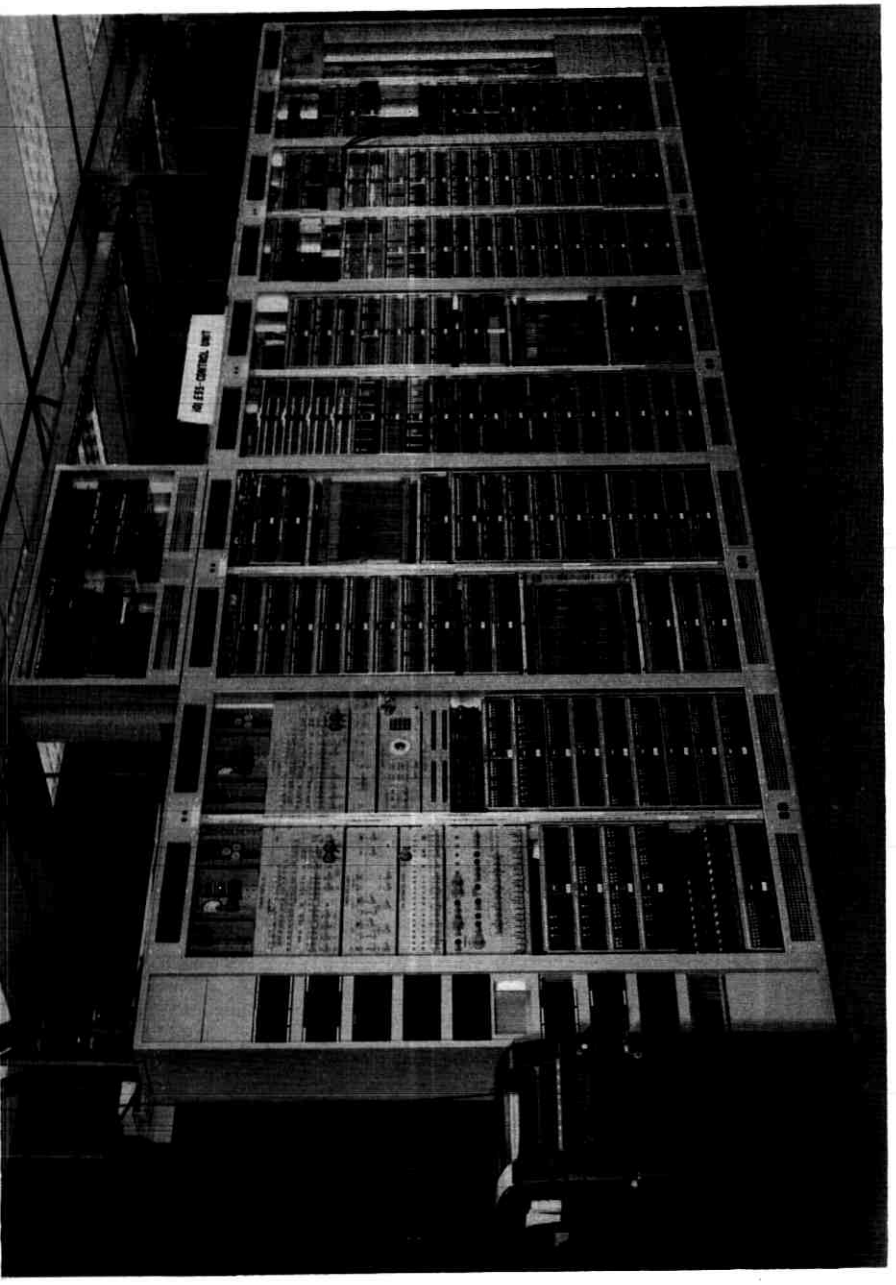


Fig. 44 — No. 101 ESS group control unit.

failure would only deny certain special services. A triplicated system clock is located adjacent to the line information store in the next frame.

To the right of the call processor are two bays of input/output equipment which provide buffering between the outlying switch units and the high-speed call processing equipment. These bays also include data transmission equipment to convert data messages to an appropriate form for storage in a ferrite sheet buffer store. At the far right, and in other bays not shown in this picture, are various trunk circuit interfaces with the electromechanical office as well as TOUCH-TONE and dial pulse receivers.

The bays at the far left, including the teletypewriter, provide for system maintenance in a manner analogous to that of No. 1 ESS. The frame mounted on top of the line-up houses special equipment for laboratory test and is not a part of a normal system.

6.4 *Switch Unit*

A photograph of one 200-line switch unit with the doors open is shown in Fig. 45. Part of the electronics is mounted on swinging gates, which provide access to individual line packages inserted in a matrix behind them. Equipment on these gates consists of the duplicated memory and control units as well as certain equipment associated with attendant console operation. Also included are transfer relays used to connect certain office telephones to central office trunks in the event of failure of commercial power at the customer location. The power supplies, which operate from local commercial power, may be seen at the bottom of the cabinet and are of the solid-state variety.

Individual customer line circuits are mounted on plug-in packages behind the swinging gates and provide access to the factory-wired time-division busses. Growth in the number of extensions on a particular switch unit may be easily accommodated by supplying additional plug-in line packages and the appropriate telephone instrument and interconnecting line.

A simplified schematic of a typical line package is shown in Fig. 46. A pair of pnpn diodes provide the time-division gates for each of the two busses. These are connected through a low-pass filter to an input transformer from the station line. Appropriate circuitry is provided for scanning lines for service requests, and a high-power pnpn triode is used for applying ringing current to a standard telephone ringer by-passing the time-division switch.

Sampling of speech on the telephone lines by the time-division switches

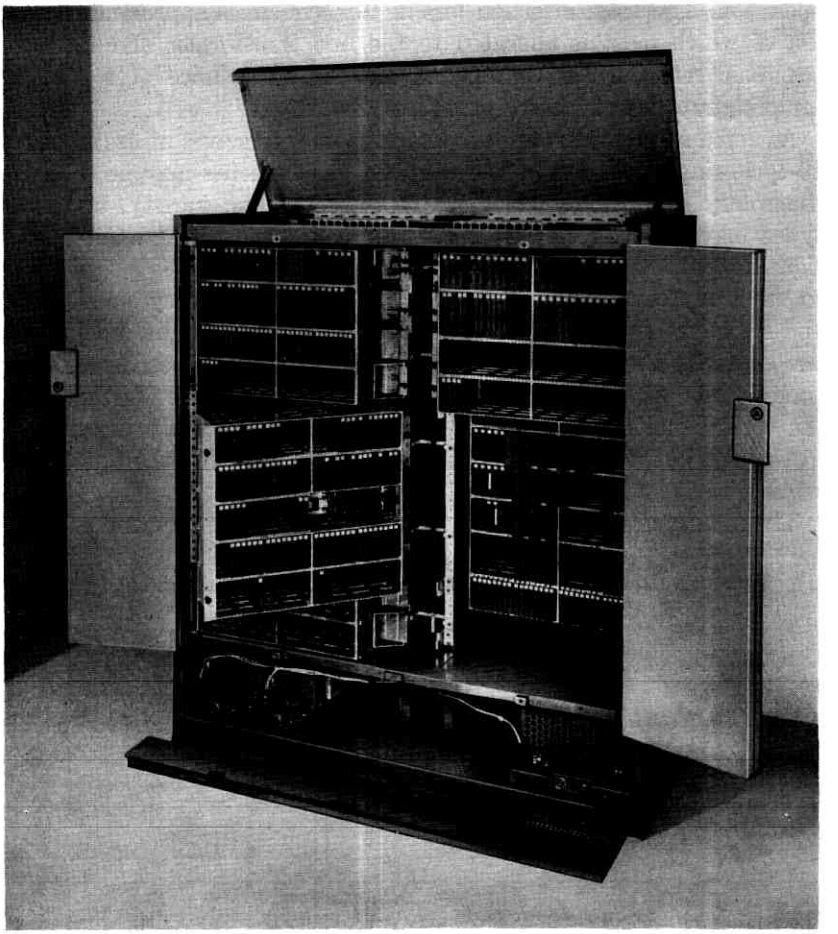


Fig. 45 — No. 101 ESS switch unit.

is carried out at a rate of 12.5 kilocycles per second. The duration of the gated signal is approximately 2 microseconds with a total guard interval of 1.2 microseconds. Thus each bus can provide 25 independent time slots in the 80 microseconds between samples of a particular line. The two busses provide 50 time slots for the maximum 200-line capacity of the switch unit. This provides considerably more traffic-handling capacity than is normally encountered in PBX's of this size.

Transmission loss through a pair of line packages is a combination of loss in the line transformer, the low-pass filter, and the resonant transfer

operation. The total insertion loss in this system is approximately 1.5 db, of which most is allocated to the line transformer and filter for economic reasons. Similar reasons dictated the choice of a 12.5 kc sampling rate to reduce low-pass filter cost.

As noted above, the telephone instruments themselves utilize standard 20-cycle ringers. However, the telephone instruments may be of either the rotary dial variety or TOUCH-TONE calling variety, and both types may be connected to a single line if desired. TOUCH-TONE signals are transmitted through the time-division switch to a digit trunk to the group control unit at the central office, where they are detected and registered in memory at that location. For rotary dial telephones, the dc dial pulses are converted to transients which pass through the time-division switch and control a burst of tone over the digit trunk to the control unit. Digit receivers in the central office are designed to distinguish between TOUCH-TONE calling signals and the tone representing dc dial pulses.

The memory in the switch unit is of the circulating type and employs

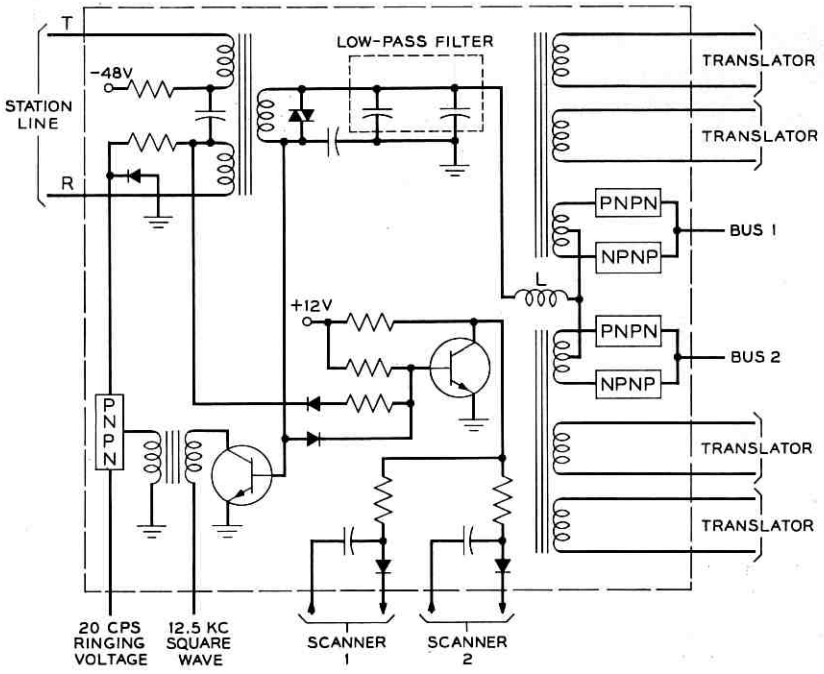


Fig. 46 — No. 101 ESS line circuit.

ferrite cores. When set by an appropriate message from the control unit for interconnecting two customers, that connection remains established until a new order is received from the control unit.

A photograph of the 200-line switch unit mounted in a reception area with an attendant and TOUCH-TONE calling console is shown in Fig. 47. As noted earlier, larger switch units are under development. These will be supplied to serve up to 340, 800, or 3000 extensions.

6.5 *Field Trial*

To gain field experience with this system, a trial was conducted of a prototype at New Brunswick, New Jersey, during 1963. The group control unit was located in the New Jersey Bell Telephone office in that city, and switch units were provided for two business customers a few miles away. A third switch unit, located in Bell Telephone Laboratories at Holmdel, New Jersey, about 30 miles from New Brunswick, was connected to the control unit via N-carrier transmission circuits. The trial was carried out over the period from March, 1963, to the end of December, 1963. Equipment used in the trial was essentially identical to the production system now being manufactured by Western Electric Company at its Hawthorne plant in Chicago.

Fig. 47 — No. 101 ESS 200-line switch unit and attendant's console.

Some of the features incorporated in the system for this trial are worthy of note. The user is first struck by the speed of response. Provisions for immediate ring make the system very attractive to the business customer, even though the time saving may be only a few seconds.

The features most used by our customers in the trial were add-on conference or dial transfer. To establish a conference after two parties were connected required only that one of them momentarily flash his switchhook and dial or key the number of a third conferee. A total of four conferees, one of which could be on an outside trunk, was possible in the trial. The limitation was imposed primarily for transmission reasons, since all customers are effectively connected in parallel. Dial transfer was done in the same way as establishing a conference except that the transferring party would simply hang up after adding the third party.

Another feature, called compressed dialing, was also very popular. Seven- or ten-digit outside numbers could be called by dialing three digits. The identification is made by appropriate magnetic patterns in the twistor module of the line information store in the group control. When the three digits are dialed by the customer, the call processor performs the necessary translation and outpulses the appropriate digits to the distant office. One of the trial customers had a repertoire of 89 compressed numbers with which he could reach all of his sales and service offices throughout the United States as well as a number of suppliers with which he frequently conducted business. This list of numbers is common to a switch unit and can be reached from any extension on that switch unit.

For intra-PBX calling, another service provided abbreviated dialing in which a 1X code could be used to reach six frequently called extensions. Such numbers were provided as a separate list for each telephone extension. In spite of the fact that this code merely reduces the dialing from three digits to two digits, only the second digit had to be remembered and not the full extension number. This may account for its very high usage during the trial.

Only six codes were provided for abbreviated dialing, since the codes 17, 18, 19, and 10 were reserved for other purposes. Code 17 was used for reroute. As described in connection with the Morris Trial, this code permitted an extension user to route his incoming calls to another extension at which he might be reached when away from his desk. After receiving dial tone and dialing the code 17, the extension user dialed the number of the phone to which he wished his calls to be routed. The system acknowledged the receipt of this information by returning a special tone. Thereafter, all calls to that extension would reach the one

designated. When the user returned to his office, he restored service by again dialing the 17 code and receiving the special acknowledgment tone.

The code 18 was used for dial hold. This permitted holding an incoming call without the necessity of providing special buttons on the telephone instrument and key equipment normally required in existing PBX's. If the customer wished to hold a call, he would flash his switchhook, dial 18 to hold the incoming call, and then dial the number of a person with whom he wished to consult. Transfer back and forth between the two parties could be accomplished by a switchhook flash and dialing of the 18 code. If in this process a held party should be forgotten by the original caller, the originating phone was rung back following his disconnect to remind him of this fact. This dial hold feature is attractive in that it does not require a multibutton telephone set, special key equipment and extra wiring as with electromechanical systems.

Code 19 provided a dial pick-up service. In a group office equipped with a number of telephones, it is frequently convenient to provide an arrangement whereby any telephone can pickup any other one in the room. By placing the appropriate pattern of magnetic spots in the line information store, a group of phones may be designated as a pick-up group. When any phone in that group rings, it may be picked up from any other phone by dialing the code 19. Here again, savings result from the use of standard telephone instruments without special key equipment and extra line connections.

The code 10 was used to provide trunk answering from any station when no attendant was present at the console. For example, a night watchman, upon hearing a ring of a special night service bell, could answer the incoming call from the nearest telephone by simply dialing the code 10.

The trial also provided for Direct Inward Dialing to extensions without going through the attendant and Direct Outward Dialing with Automatic Number Identification to distant offices. In the latter case, restrictions could be placed on various lines to prevent direct out-dialing, restrict out-dialing to a specified local area, or provide full access to the Bell System network. The ability to administer this type of restriction at the central office by appropriately magnetizing the twistor cards is another example of stored program convenience.

6.6 *Production and Installation Progress*

Production of No. 101 ESS has been under way at Western Electric's Hawthorne plant in Chicago since 1963. At the beginning of this article,

reference was made to the cut-over of the first system at Brown Engineering and Chrysler Corporation in Cape Kennedy, Florida. The group control unit to serve these customers is located in Southern Bell Telephone Company's office at Cocoa Beach, Florida, where it is associated with a No. 5 crossbar switching office. Four businesses in that area are now (early 1965) being served by that group control. Since the major installation interval for No. 101 ESS is associated with the group control, additional switch units can be added on very short notice and with only a few hours' installation time if the necessary transmission circuits are available.

The second group control produced by Hawthorne was installed for the New York Telephone Company in their Fifty-sixth Street office in Manhattan. It is serving telephone extensions from a switch unit located in the A.T.&T. Co. exhibit at the New York World's Fair. A second switch unit controlled from New York is serving a group of extensions in the New York Telephone Company's headquarters, and a third provides service to about 180 extensions at Bell Telephone Laboratories, Holmdel.

As noted in the Introduction, installations of group control units have also been completed in Chicago, Cleveland, Los Angeles, and Washington, D. C.

VII. SUMMARY

This article has presented a survey of progress being made by the Bell System in introducing electronic switching into the telephone plant and has described two systems developed by Bell Telephone Laboratories for this purpose. Present orders indicate that electronic switching is being favorably received by the operating telephone companies, and customer reactions to the new services have been very encouraging. As production capacity builds up, it can be expected that more and more customers throughout the United States will find these new features available to enhance the value of their telephone service.

REFERENCES

1. Malthaner, W. A., and Vaughan, H. E., An Experimental Electronically Controlled Automatic Switching System, B.S.T.J., *37*, September, 1958, pp. 1091–1124.
2. Malthaner, W. A., and Vaughan, H. E., DIAD — An Experimental Telephone Office, Bell Laboratories Record, *32*, October 1954, pp. 361–365.
3. Keister, W., Ketchledge, R. W., and Lovell, C. A., Morris Electronic Telephone Exchange, Proc. IRE (London) B Suppl. 107, 1960, pp. 257–263.
4. Most of the diagrams used in this paper were prepared by Bell Telephone Laboratories for a Symposium held for Western Electric Company patent

licensees at Holmdel, N. J., January 21 to 24, 1963. Credit for these diagrams and for much of the material on which this article is based belongs to the many engineers who have carried the detailed design responsibility for the systems described.

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. Haugk, G., and Yokelson, B. J., Experience with the Morris Electronic Switching System, IEEE Transactions — Part 1, Comm. & Electronics, *64*, January, 1963, pp. 605–610.
7. Feiner, A., The Ferreed, B.S.T.J., *43*, January 1964, pp. 1–14.
8. Bobeck, A. H., A New Storage Memory Suitable for Large-Sized Memory Arrays — The Twistor, B.S.T.J., *36*, November 1957, pp. 1319–1340.
9. Depp, W. A., and Townsend, M. A., An Electronic Private Branch Exchange Telephone Switching System, IEEE Conference Paper CP 63-580.
10. Irland, E. A., and Vogelsong, J. H., Memory and Logic System for an Electronic Private Branch Exchange, IEEE Conference Paper CP 63-460.
11. Herndon, J. A., and Tendick, F. H., Jr., A Time Division Switch for an Electronic PBX, IEEE Conference Paper CP 65-577.
12. Seley, E. L., and Vigliante, F. S., Common Control for an Electronic Private Branch Exchange, IEEE Conference Paper CP 63-583.

Core Materials for Magnetic Latching Wire Spring Relays

By T. G. GRAU and A. K. SPIEGLER

(Manuscript received July 15, 1964)

The magnetic characteristics of medium carbon steels were examined to determine whether these steels can be used as core materials for magnetic latching wire spring relays. The analysis of the data show that steels with a carbon content ranging from 0.35 per cent to 0.50 per cent, heat treated to a Rockwell hardness ranging from 30 to 45 points on the "C" scale, are satisfactory core materials. The analysis further shows that the relays are latched more securely if the carbon content of the steel used for the core is high, and that the hardness in the above-mentioned range has very little effect on the latching characteristics.

I. INTRODUCTION

In telephone switching circuits it is often necessary to hold a relay operated for a long period of time. An example of such a case is a relay which remains operated during a telephone conversation. In order to hold a conventional relay operated for a long period of time, it is usually locked electrically through one of its own make contacts. This method of latching requires that a continuous current be supplied to the relay coil. In the past, relays with mechanical locking features or auxiliary permanent magnets also have been used. These relays, however, frequently require two electromagnets, one to operate the relay and activate the locking mechanism and the other to deactivate the locking mechanism and allow the relay to release. A relay which requires a mechanical locking mechanism, since it needs two electromagnets, is equivalent to two relays and is therefore uneconomical.

The latching force, which holds the magnetic latching relay operated, depends upon the residual magnetic induction and coercive force of the magnetic materials used in the structure of the electromagnet. The magnetic latching relay is operated by a short current pulse. After the removal of the pulse the relay remains operated. A current pulse of

opposite polarity to, and lower magnitude than, that of the operate pulse will release the relay. Once released, the relay will remain in that state until another operate pulse is supplied.

It is apparent that this type of relay is particularly useful when power is at a premium. Once the relay is operated, no more power is required to hold it in that state. Moreover, the latching relay has another useful feature: memory. It "remembers" the command "operate" or the command "release" and remains in one of these states until ordered into the other state.

Earlier studies have shown that high carbon steels can be used to obtain the magnetic characteristics needed in the magnetic structure of latching relays. An example is the hold magnet of the magnetic latching crossbar switch.[1] High carbon steels, however, are not suitable for the magnetic structures of wire spring relays because the geometry of this structure would make its manufacture very difficult and expensive. Therefore, a study was started to determine whether medium carbon steels could be used.

II. REQUIREMENTS

The magnetic latching relay has two basic requirements. (1) When the relay is in its released state it must operate upon the application of an operate current pulse and remain in the operated state after the pulse is removed and until a release pulse is applied. (2) When the relay is in its operated state it must release upon the application of a release current pulse and remain in the released state after the pulse is removed and until an operate pulse is applied.

As shown by these two requirements, the relay must have two stable states. No outside influence shall falsely operate or release the relay. In general, relays are mounted on frames on which mechanical vibrations occur. The two states, therefore, have to be such that they will not be influenced by mechanical vibrations occurring on these frames. The latching force holding the relay operated has to be large enough to hold the relay operated and to withstand the vibrations on the frame. Also, after the relay has been released, the magnetic structure must be in a magnetic state such that the mechanical back tension can securely hold the relay in its released state. Induced electrical noise in the coil must also be considered since it could change the magnetic state of the core and release the relay when it is latched.

To the basic requirements discussed, two more requirements were added: (3) The magnetic latching relay shall be of the wire spring type. (4) The magnetic latching relay shall use the same mechanical pile-up as

the conventional wire spring relay and it shall be possible to manufacture the magnetic parts with the same tools that produce the magnetic parts of the conventional wire spring relay.

The third requirement was added because of the consistently good and reliable performance of wire spring relays. The fourth requirement was added to hold the cost of manufacture to a minimum. The last two requirements show that it was necessary to select a magnetic material which, when used as the magnetic structure of the wire spring relay, will satisfy the first two requirements.

In a magnetic latching relay the parameters which will determine the latching force and the security of the latched state are the residual induction and the coercive force of the materials used in the magnetic parts. Both parameters will effect the suitability of a material for this application. Fig. 1 shows typical demagnetizing hysteresis loops for two different materials. Materials (1) and (2) have residual inductions B_{r1} and B_{r2} and coercive forces H_{c1} and H_{c2}, respectively. If an air gap is introduced in the magnetic circuit, the flux density will be reduced to the points B_1 and B_2, called the remanent induction. Points B_1 and B_2 are determined by the line given by

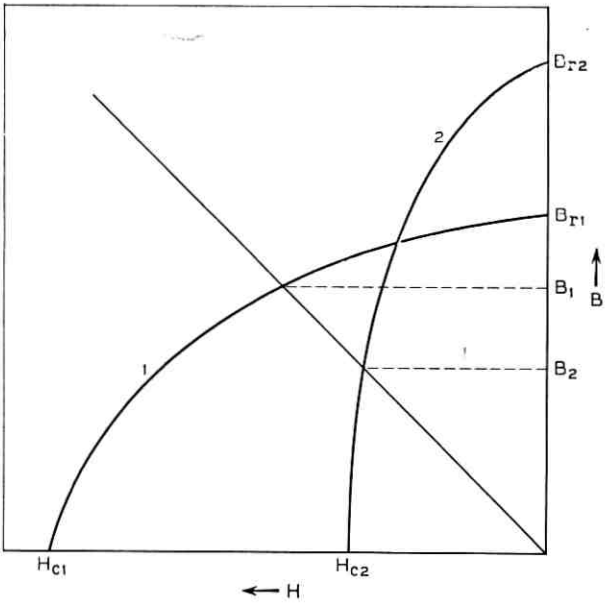


Fig. 1 — Generalized demagnetization curve.

$$B_g = H_M l_M / l_g \tag{1}$$

where:
B_g = magnetic induction in the air gap
l_g = length of the air gap
H_M = field strength in the magnetic material
l_M = length of the magnet.

Equation (1) provides only an approximation for the change in residual induction that occurs in a closed magnetic circuit when an air gap is introduced into the circuit. However, it does indicate these changes.

The latching force is directly proportional to the square of the remanent induction. As shown in Fig. 1, material (2) has a higher residual induction than material (1). However, when an air gap is introduced in the magnetic circuit, the remanent induction B_2 of material (2) is lower than B_1, because H_{c2} is smaller than H_{c1}. Therefore, the latching force that can be obtained from material (2) will be smaller than the one that can be obtained from material (1) for the particular air gap shown.

The third and fourth requirements dictate the geometry of the magnetic circuit. Therefore, to obtain a magnetic latching wire spring relay it is necessary to find a material which has values of residual induction and coercive force suitable for the magnetic circuit of the presently manufactured wire spring relay.

III. MAGNETIC CAPABILITY

Since the geometry of the magnetic structure is fixed by the fourth requirement of the last section, it is necessary to calculate values for the maximum latching force that can be obtained. Fig. 2 shows the individual parts of the magnetic structure of a wire spring relay.

The maximum mechanical load which must be held operated will determine the minimum latching force, which in turn determines the minimum remanent magnetic induction or remanent flux. The relation be-

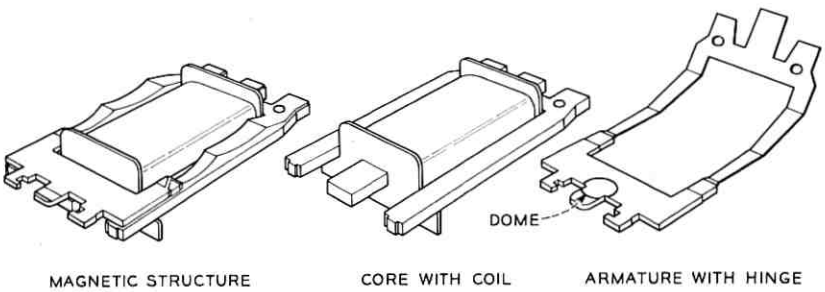


Fig. 2 — Parts of the wire spring relay magnetic structure.

tween the latching force F and the remanent flux φ_G is given by:

$$F = \frac{\varphi_G^2}{8\pi A_G}\left(\frac{1}{980}\right)k \tag{2}$$

where:

F = latching force in grams
φ_G = remanent flux through the air gap
A_G = effective pole face area at the air gap
k = constant correcting for the nonperpendicularity of φ_G between the mating areas.

The maximum remanent flux φ_G through the air gap is determined by the flux saturation level of the armature. If the semipermanent magnet core has a remanent magnetic induction which is large enough to keep the armature saturated, the maximum φ_G and, therefore, the maximum latching force will be obtained. The attainment of this state will depend on the values of the residual induction and coercive force of the core material.

If we assume that the remanent flux of the core is large enough to keep the armature saturated, an estimate for the maximum latching force can be obtained. The armature is made of 1 per cent silicon steel and has a cross-sectional area of 0.346 cm². The nominal value for the operating flux density of the armature is 15,000 gauss; the calculated flux of the armature is, therefore, 5190 maxwells. As shown in Fig. 2, the armature has a dome with an area of 0.712 cm². When the relay is latched, the largest part of φ_G will go through the dome of the armature; therefore, A_G in equation (2) can be set equal to the dome area. The estimated value of k is about 0.8. Substituting the calculated flux of the armature, the dome area, and the constant k into equation (2), we obtain the force at the dome as 1315 grams. This force is shown in Fig. 3

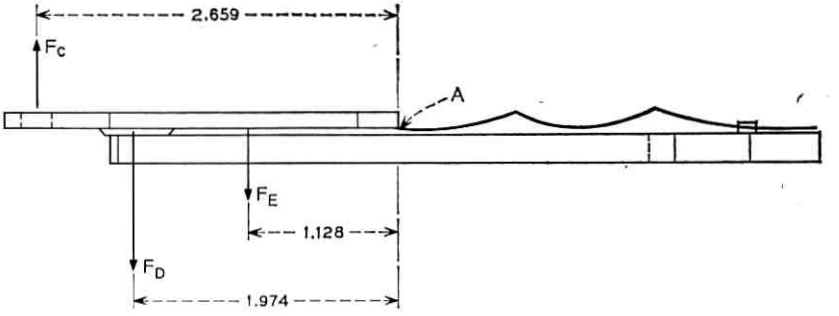


Fig. 3 — Magnetic and mechanical forces.

as F_D. As shown in this figure, we also have a force on each side leg to take into account. The flux through each side leg is half of the flux φ_G through the dome, and the geometrical area of each side leg is 1.25 cm². Using these values and again a constant of 0.8 in equation (2), we obtain a force of 168 grams at each side leg. This force is shown as F_E in Fig. 3. The force F_C in this figure represents the mechanical load, which is opposite to the latching force. Taking the moments about point A with the dimensions given in Fig. 3, we find that the maximum load F_C which can be held in equilibrium by the latching forces F_D and F_E is 1120 grams. This is well above the maximum mechanical load of the relay, which is in the neighborhood of 580 grams. Therefore, based on the assumption that the remanent flux is large enough to produce a flux of 5190 maxwells, the latching force will be about twice as large as the maximum mechanical load. This means that the structure is capable of latching the maximum number of contacts, or 24.

The residual magnetic induction of the core will now be examined to see if it could be large enough to keep the armature saturated when the relay is latched. If we neglect all leakage flux, then the remanent magnetic induction of the core has to be only large enough so as to saturate the armature. Assuming no leakage flux, we have

$$\begin{aligned} \varphi_A &= \varphi_C \\ \varphi_C &= B_C A_C = \varphi_A \end{aligned} \tag{3}$$

where:

φ_A = saturation flux of the armature
φ_C = remanent flux of the core
B_C = remanent magnetic induction of the core
A_C = smallest cross-sectional area of the core.

As mentioned previously, the saturation flux φ_A of the armature is 5190 maxwells. The core cross-sectional area A_C is 0.533 cm². Using these values in equation (3), we obtain a remanent magnetic induction of 9737 gauss for B_C. This value for B_C corresponds to a value which is lower than either of the values B_{r1} or B_{r2} in Fig. 1.

Since the core has to be manufactured with tools presently used to manufacture the cores of conventional relays, only medium carbon steels can be considered. The residual induction of these steels is about 13,000 gauss. This corresponds to points such as B_{r1} or B_{r2} in Fig. 1. Since air gaps and leakage flux are present in the magnetic circuit of the relay, the operating flux level will be lower, as previously discussed

and illustrated in Fig. 1. However, only 9737 gauss are required for the operating flux density of the armature. If the coercive force of the core material is large enough, it is very likely that an operating flux density of 5190 maxwells of the armature can be maintained when the relay is latched.

So far we have only looked at the maximum obtainable latching force. This was necessary in order to see if medium carbon steels should be considered. As shown, the latching force that can be obtained is about twice as large as the maximum mechanical load. However, a relay can be considered securely latched if the latching force exceeds the mechanical load by 50 per cent. Since the maximum mechanical load is 580 grams, a latching force of 770 grams minimum is needed. Using equation (2) with a latching force, F, of 770 grams, we obtain 4727 maxwells for flux φ_G. The remanent flux of the magnetic circuit has to be larger because the leakage flux was not taken into account. For the magnetic circuit under consideration, the estimated leakage flux is approximately 15 per cent of the flux through the core.

Therefore, to obtain a latching force of 770 grams minimum, the remanent flux of the core should be at least 5500 maxwells when the relay is operated.

In the above discussion, it was assumed that the coercive force of medium carbon steel is large enough so that the required remanent flux can be obtained. Since no exact relationship exists between magnetic induction and coercive force, an experimental study is necessary to see if values for these characteristics can be obtained with medium carbon steels so that these steels can be used as core materials for latching relays.

IV. STEEL STUDY

At this point in the development of the latching wire spring relay, the minimum remanent magnetic induction needed is 5500 maxwells for a full complement of contacts. The coercive force is not known, but some considerations such as the need for a securely latched relay and the need for a reasonably high dc release current with a maximum contact load seem to dictate a rather high coercive force in the magnetic structure. The only further restriction for the core material stems from an economic consideration; i.e., the steel must be commercially available and it must be soft enough to be used in present punch press tools.

A steel investigation was made to find a particular steel which should be used for the relay core and to determine how this steel should be heat treated to obtain the necessary magnetic properties. Four grades of

steel containing different amounts of carbon ranging from 0.25 per cent to 0.50 per cent were obtained. The steels were commercial grades of C-1025, C-1035, C-1040 and C-1050. Results of an analysis of each type of steel are shown in Table I.

General-purpose wire spring relay cores and ring samples were made from the four lots of steel. The relay cores and ring samples in each steel group were divided into five subgroups and were heat treated to obtain different values of hardness on the Rockwell "C" scale ranging from 25 points to 45 points in 5-point steps. The C-1025 steel required a water quench to obtain the necessary hardness; this makes it unfit for a material for relay cores because it results in severe core leg twisting and misalignment. Therefore, the C-1025 steel was eliminated from the steel study.

The magnetic characteristics of the ring samples were measured with a Cioffi recording fluxmeter. The resulting measured values of coercive force and residual magnetic induction correspond to such points as H_{c1} and B_{r1} in Fig. 1 and form the raw data which was then analyzed.

Each of these magnetic characteristics was then plotted as the dependent variable against the ring sample hardness as the independent variable. The plots of the residual flux versus hardness and the coercive force versus hardness can be approximated by a straight line for a limited range of hardness. This can be done with reasonable accuracy if the "true" long-range curve has a slowly changing first derivative over the hardness range. The straight-line approximations of the data of the coercive force, as fitted by the least squares technique, are plotted in Fig. 4(a). The equations for the coercive force of the steels are shown in Table II.

The results are somewhat erratic, although they lead to the expected curve.[2] The cause of such erratic results is partly due to the impurities in the steel samples and partly due to variations in the heat treatment.

It should be pointed out that these results represent only one group of data points and that another group could lead to slightly different equations. However, these equations serve to illustrate the trends present in this range of steels.

TABLE I — PERCENTAGE OF ELEMENTS PRESENT

Type of Steel	C	S	P	Mn	Mo	Si	Cr	V	Ni
C-1025	0.24	0.028	0.024	0.39	0.02	0.002	<0.03	<0.02	<0.02
C-1035	0.38	0.024	0.024	0.83	0.04	0.258	0.07	<0.02	0.06
C-1040	0.42	0.023	0.016	0.79	0.02	0.207	0.10	<0.02	0.04
C-1050	0.49	0.019	0.018	0.79	0.02	0.172	0.08	<0.02	0.03

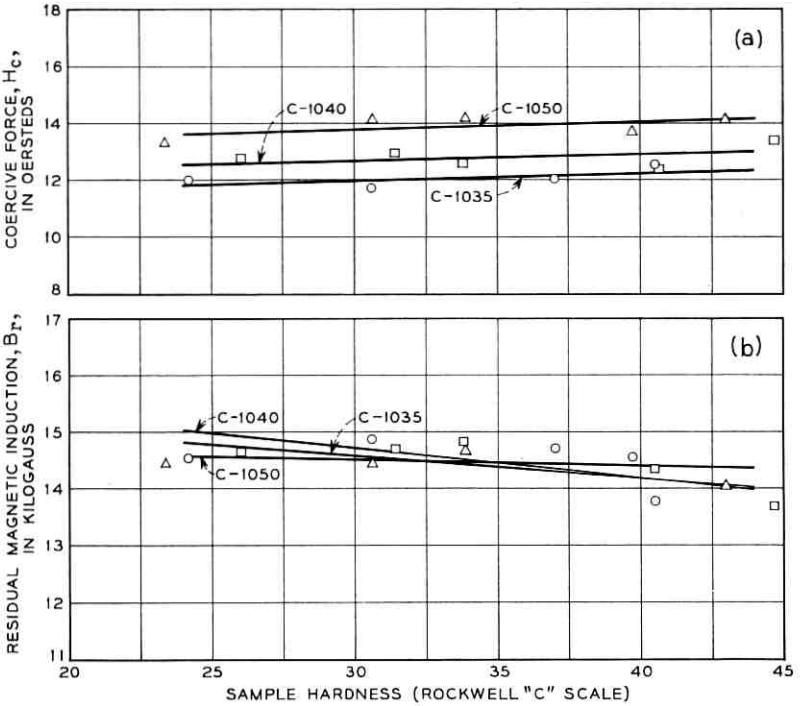


Fig. 4 — Magnetic characteristics of ring samples.

These coercive force curves show three important characteristics: (1) the coercive force increases slightly as the hardness increases, (2) the slopes of the lines are very small and are identical within experimental error, and (3) the lines are displaced from one another in such a manner that a definite trend is evident; i.e., the coercive force increases as the carbon content of the steel increases.

Straight-line approximations as obtained with the least squares tech-

TABLE II — EQUATIONS FOR COERCIVE FORCE

Steel Type	Equations
C-1035	$H_{c1} = 11.16 + (0.0268)X$
C-1040	$H_{c2} = 11.87 + (0.0266)X$
C-1050	$H_{c3} = 12.85 + (0.0305)X$

where: H_{ci} = the coercive force in oersteds, $i = 1, 2, 3$ depending on the type of steel, and X = the hardness in points on the Rockwell "C" scale and in the range of 24 points to 44 points.

nique were found from the data for the residual magnetic induction of the ring samples. The results of this analysis are plotted in Fig. 4(b) and the equations for these curves are given in Table III.

These three curves show that within experimental error the residual flux: (1) decreases as the hardness increases, (2) varies slowly with hardness, and (3) has about the same magnitude for the three different steel groups.

Another analysis of these data was tried. The second method was to fit a curve of the form, $H_{ci} = a_{i0} + a_{i1}X + a_{12}X^2$ to the data by solving three equations for the constants a_{i0}, a_{i1}, and a_{i2} (where $i = 1, 2, 3$ for the different steels). Results of this analysis are shown in Fig. 5(a). The same trends that were observed in Fig. 4(a) also can be seen in this figure. This analysis of the residual magnetic induction data was tried, but it produced about the same results as that of the linear analysis. The method was to fit a curve of the form $R_i = a_{i0} + a_{i1}X + a_{i2}X^2$ to the data by solving three simultaneous equations for the constants a_{i0}, a_{i1}, and a_{12} (where $i = 1, 2, 3$ for the different types of steel). The results of this analysis are plotted in Fig. 5(b).

The test results of the coercive force and residual induction for the ring samples show that in the hardness range between 24 and 44 points on the Rockwell "C" scale, variations in the coercive force between steels seem to be larger than variations in the residual induction. Also since all of the values for the residual flux fall in a relatively narrow band (between 13.98 K gauss and 15.0 K gauss), all of the three steels considered here should provide latching forces with very similar magnitudes. Therefore, latching relays should have cores manufactured from materials which provide the largest possible coercive force.

With these preliminary results, relay cores were manufactured from the three grades of medium carbon steel: C-1035, C-1040 and C-1050. Relays were assembled from these cores and the magnetic characteristics of the relays were then measured. Again, a straight-line approximation of the data for the coercive force and the remanent magnetic

TABLE III — EQUATIONS FOR RESIDUAL MAGNETIC INDUCTION

Steel Type	Equations
C-1035	$R_1 = 15.74 - (0.039)X$
C-1040	$R_2 = 16.30 - (0.053)X$
C-1050	$R_3 = 14.76 - (0.0079)X$

where: R_i = the residual flux in gauss, $i = 1, 2, 3$ depending on the type of steel, and X = the hardness in points on the Rockwell "C" scale in the range of 24 points to 44 points.

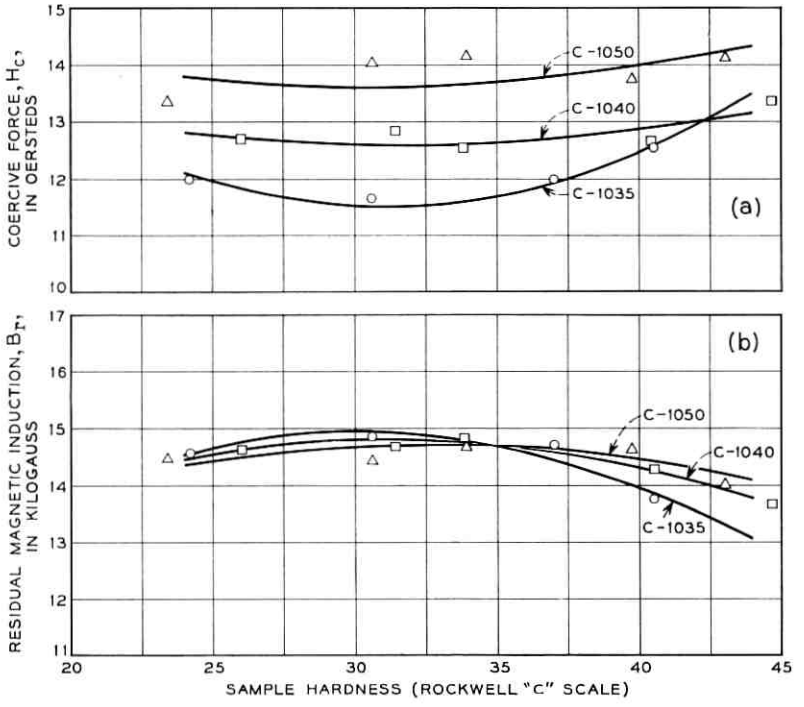


Fig. 5 — Magnetic characteristics of ring samples (cont.).

flux was used. The results for the coercive force are plotted in Fig. 6(a) and the results for the remanent magnetic flux are plotted in Fig. 6(b). Table IV gives the equations of the curves for these two figures.

Examination of coercive force and remanent flux versus hardness shows that for any small range of hardness, say five points on the Rockwell "C" scale, the remanent flux varies less than 8 per cent between the different samples. That is, the value of remanent flux, while it is a function of the carbon content of the steel, is insensitive to large variations of the carbon content. Since the value of the ultimate latching force depends on the amount of remanent flux which can be established in the magnetic circuit of the relay, this result means that the latching force will be insensitive to minor carbon variations in a particular steel. This is an important result for the economical mass production of a latching relay.

Variations in coercive force which are larger than variations of the remanent flux within the same hardness range are observed between

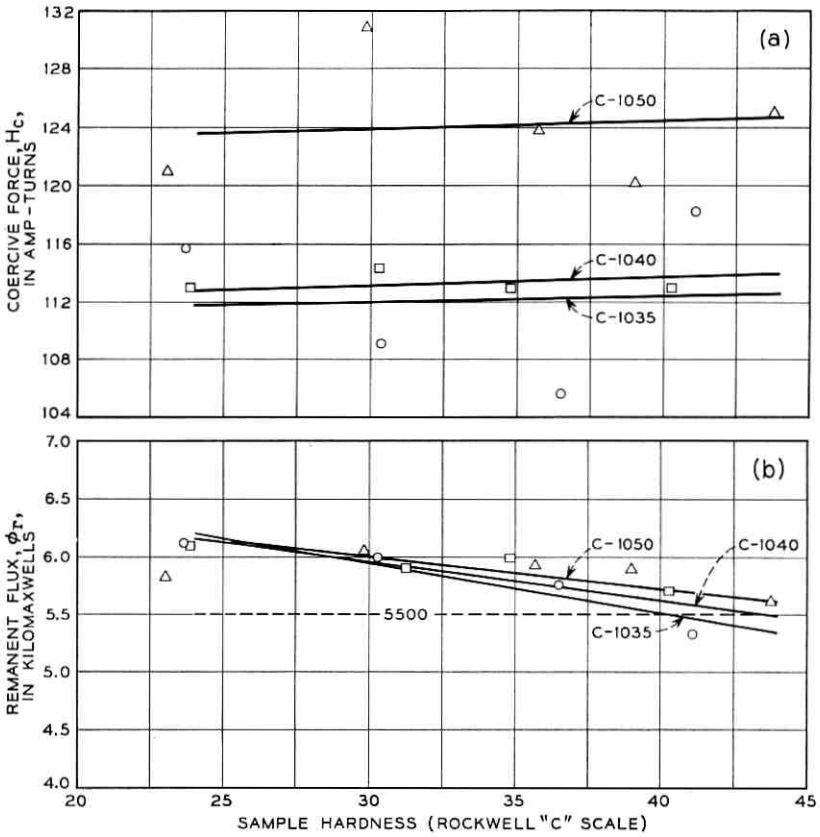


Fig. 6 — Magnetic characteristics of relay samples.

the different steel groups. However, the ability to maintain the level of the latching force once it has been established, and not the magnitude of the force itself, is affected by changes in coercive force. Therefore, large variations of this parameter can be tolerated in manufacturing situations provided that enough margin of force has been designed into

TABLE IV — EQUATIONS FOR COERCIVE FORCE AND REMANENT FLUX

Steel Type	Equations for Coercive Force	Equations for Remanent Flux
C-1035	$H_1 = 110.9 + (0.040)X$	$\varphi_1 = 7.23 - (0.043)X$
C-1040	$H_2 = 111.4 + (0.060)X$	$\varphi_2 = 6.99 - (0.034)X$
C-1050	$H_3 = 122.5 + (0.050)X$	$\varphi_3 = 6.48 - (0.028)X$

where: H_i = coercive force in ampere-turns, X = the hardness in points on the Rockwell "C" scale in the range of 24 points to 44 points, φ_i = remanent flux in kilo-maxwells, and i = 1, 2, 3 depending on type of steel.

the latched relay. Since the stability of the latching force is principally a function of the coercive force and varies directly as the coercive force, it is reasonable to attempt to heat treat the core to produce a high coercive force. Furthermore, any increase in coercive force which can be obtained by increasing the carbon content of the core material should be pursued.

From theoretical calculations a minimum value of remanent flux was found to be 5500 maxwells. Since this value is below all predicted values of remanent flux for the three group of relays, plotted in Fig. 6(b), more emphasis should be placed on choosing a steel to give the maximum security of the latching force and obtaining the best core hardness. Since steels which have carbon percentages above 0.50 per cent are difficult to machine on a mass production basis, this percentage of carbon represents a maximum allowable value. Therefore, Fig. 6(a) shows that the best steel for magnetic latching wire spring relays based on the magnetic characteristics is the C-1050 steel. The carbon content of this steel is 0.5 per cent, +0.05 per cent, −0.02 per cent.

Until now, the remanent flux and the coercive force have been treated as separate and distinct quantities which vary with the carbon content and with the hardness of the steel. In the magnetic circuit of the relay, these parameters each influence the dynamic characteristics in such a manner that measurements of the latching force can show the validity of the choice of the core material. Results of a straight-line approximation of the data for the latching force versus hardness are shown in Fig. 7, and the associated equations are given in Table V.

Several conclusions can be drawn from Fig. 7: (1) Within the group of steels that were tested, the best core material is the C-1050 steel as indicated previously. (2) Since theoretical calculations, based on the

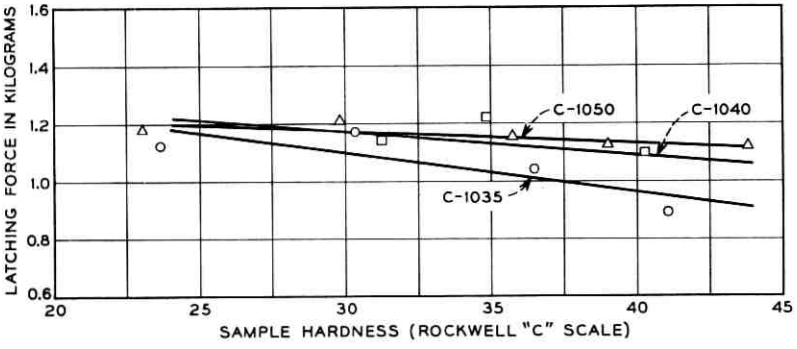


Fig. 7 — Total latching force of relay samples.

TABLE V — EQUATIONS FOR LATCHING FORCE

Steel Type	Equations
C-1035	$F_1 = 1499 - (13.46)X$
C-1040	$F_2 = 1407 - (7.95)X$
C-1050	$F_3 = 1297 - (3.96)X$

where: F_i = the total latching force in grams, X = the sample hardness in points on the Rockwell "C" scale in the range of 24 points to 44 points, and i = 1, 2, 3 depending on type of steel.

assumption that the armature would be saturated, produced a force of 1120 grams, it appears that the C-1050 and the C-1040 groups are producing saturation in the armature. (3) Since it appears that there is armature saturation in the two groups, the latching force to latch the full complement of 24 contacts has been obtained.

The optimum hardness is determined from curves showing the magnetic characteristics of the sample relays — i.e., Figs. 6(a) and (b). It was calculated that a remanent flux of 5500 maxwells should be a minimum value (this allows a 50 per cent operated load margin). Using 5500 maxwells as a lower bound in Fig. 6(b) and considering an experimental error of ±2 per cent, the highest value of hardness would be 40 points on the Rockwell "C" scale. Allowing a manufacturing variation of ±2.5 points, the hardness to be specified in manufacture is Rockwell "C" 37.5 ± 2.5. With this hardness, the relays should have a total latching force greater than 770 grams and they should have good margin with this force.

V. RELAY STUDY

Latching relays, after being operated and latched, can be falsely released by extraneous effects such as random electrical noise induced in coil leads and mounting plate vibrations, if these effects are severe enough. To obtain release parameters in the form of load versus release ampere-turns for different steels, relays having known loads were operated, latched magnetically and released by applying a reverse current to the coil. Fig. 8 shows the results of this test. Two important conclusions can be drawn: (1) The C-1050 steel again is the best steel for the magnetic latching application, and (2) The slope at any point along the curve is very steep. A steep slope indicates that the relay should be capable of accepting various values of loads and still release correctly.

Referring to Fig. 8, a third conclusion can be drawn: the release ampere-turns for the C-1050 steel is larger than 70 NI with a 600-gram

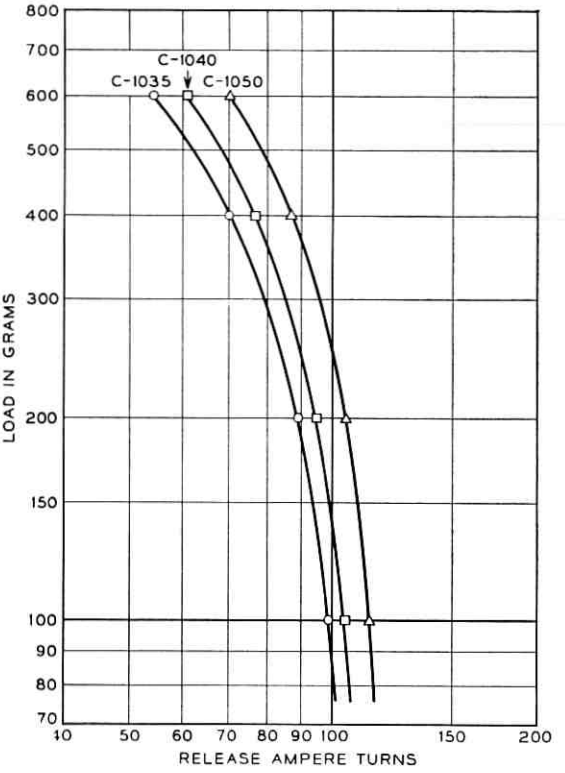


Fig. 8 — Release with a known load for relay samples.

load. The significance of the load value of 600 grams is that this represents a higher value of operated contact load than the maximum operated load (580 grams). A release ampere-turn value of 70 NI represents a large amount of power in a relay coil; i.e., this amount of power is much larger than the expected power which a random electrical pulse could provide by stray coupling to the relay coil. (The coil represents a high impedance for this type of energy transfer.) Therefore, there is adequate latching force so that induced random noise pulses should not release the relay from its latched state.

The test relays were studied to determine dc operate parameters. The resulting curves, pull versus ampere-turns at constant gaps, are shown in Fig. 9 for the C-1050 steel. Similar curves for conventional relays usually display a linear region which has a slope of 2. This region corresponds to a fairly linear section of the magnetization curve below

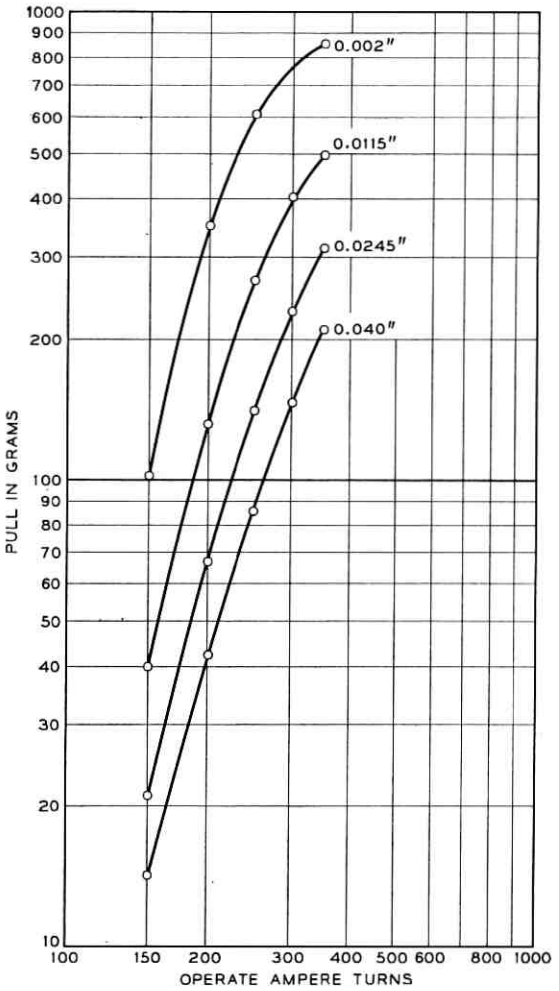


Fig. 9 — Constant-gap pull curves for relay samples made from C-1050 steel.

the knee and above the region of easy magnetization (toe to instep). It is this linear portion of the magnetization curve that produces the linear section of the pull versus ampere-turns on a log-log plot. In higher carbon steels that have been magnetically hardened, the magnetization curve is characterized by the absence of a linear section. For this reason and due to armature saturation, Fig. 9 has no linear section.

The dc operate current for a magnetic latching relay must be found

from the constant gap curves, shown in Fig. 9. The use of these curves assumes: (1) The relay has received enough power from the last release pulse so that the core can be considered in the demagnetized state. (2) The critical load point is found from values of load for the conventional wire spring relay. (3) The minimum operate saturation current will be exceeded.

At this point it is helpful to define the operate flux direction as the positive flux direction and the release flux direction as the negative flux direction.

It is not absolutely necessary for the release pulse to return the core to its zero flux state; however, the pulse must return the core close to that state and must not allow the core to return to a negative flux state. If the core is left in such a state, the operate current will be higher than the current which is found from the pull curves. The operate current which is found from the pull curves represents a current value which is useful only as a readjust current for the relay. However, correct dynamic operation of the relay requires that the operate current supply: (1) sufficient magneto motive force to saturate the magnetic structure, and (2) the saturation flux for a minimum pulse time of at least 15 per cent longer than the operate time of the relay.

Therefore, the minimum dynamic operate current is a current capable of saturating the structure, and higher operate currents can be used to obtain faster operate times. A further requirement for operating the relay is that the current shall not reverse when the current to the coil is turned off.

A contact protection network used across the coil forms an L-R-C circuit. If the protection network is not chosen correctly, oscillations will result (the under-damped case). Then, if the amplitude and period are large enough, the relay will unlatch because of the current reversal.

The dc release current is obtained from a load versus release ampere-turns curve. Again, this current is useful only as a relay readjust current because there is no assurance that the core will return to its zero flux state after the release pulse is turned off.

A N.F. current, meaning "No Flux" current, is specified for magnetic latching relays. This value of current, when used to release the relay from its latched state, returns the core very close to its neutral magnetic state.

VI. CONCLUSION

The data presented show that C-1050 steel, heat treated to a hardness of 37.5 $\pm$ 2.5 points on the Rockwell "C" scale is a satisfactory material

for the core of a magnetic latching wire spring relay. It was the best choice of the medium carbon steels investigated. Satisfactory latching forces can be obtained with medium carbon steels ranging from C-1035 to C-1050; however, as the carbon content increases, more secure latching is obtained.

Magnetic latching wire spring relays presently manufactured are shown in Fig. 10. The cores of these relays are made with C-1050 steel, heat treated to a Rockwell "C" 37.5 ± 2.5 hardness. These latching relays are used in the No. 1 ESS system. Because of the characteristics of magnetic latching relays, it is anticipated that the demand for them will increase in the future.

VII. ACKNOWLEDGMENTS

Grateful acknowledgment is given to H. M. Knapp and C. B. Brown for their valuable guidance and helpful discussions. The authors wish to express their appreciation to V. L. Marsh and J. A. Cooper for their assistance in making laboratory measurements for the preparation of this article.

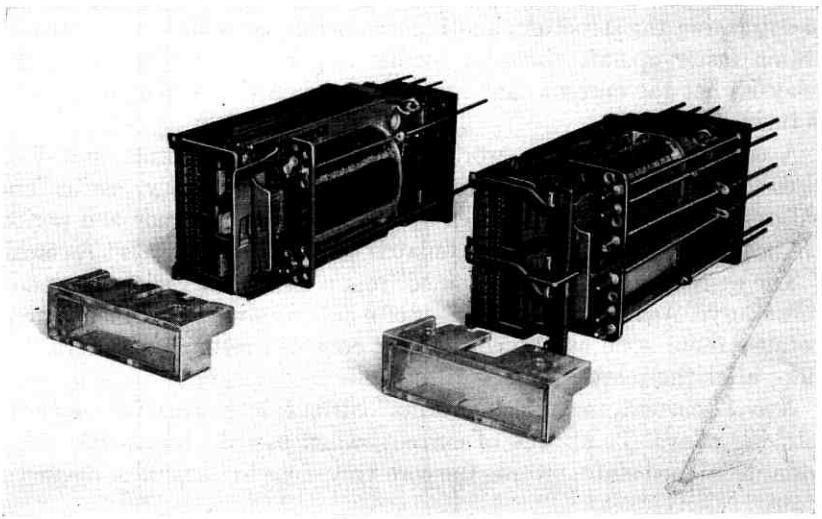


Fig. 10 — Magnetic latching wire spring relays: AL type on left and AM type on right.

REFERENCES

1. Zupa, F. A., Magnetic Latching Crossbar Switches, B.S.T.J., *39*, September, 1960, p. 1351.
2. Bozorth, R. M., *Ferromagnetism*, D. Van Nostrand Co., New York, 1956, Fig. 9–22, p. 370.
3. Peek, Jr., R. L., and Wagar, H. N., *Switching Relay Design*, D. Van Nostrand Co., New York, 1955.
4. Parker, R. J., and Studders, R. J., *Permanent Magnets and Their Application*, John Wiley and Sons, Inc., New York, 1962.

A Precise Measurement of the Gain of a Large Horn-Reflector Antenna

By D. C. HOGG and R. W. WILSON

(Manuscript received March 24, 1965)

The gain of a horn-reflector antenna with an aperture area of about 400 square feet has been measured with a probable error of 2 per cent at a frequency of 4080 mc. *Errors and fluctuations normally introduced into gain measurements by terrain and other environment were obviated by mounting the source on a helicopter which maintained a position about 2500 ft. above ground at a distance of one mile from the antenna under test. It is concluded that high precision can be obtained in measurement of gain of large antennas using such methods.*

I. INTRODUCTION

The gains of horn-reflector antennas have been measured many times in the past at Bell Telephone Laboratories, usually with the result that the effective area is about 1.5 db below full area. In other words, the measured aperture efficiencies run between seventy and seventy-five per cent.*

Traditionally, such gain measurements employ a source located in the Fraunhoffer region of the antenna to be measured. The field radiated by the source is then sampled at the antenna by taking "height runs" with a standard (or reference) horn. A thorough job involves several height runs at various lateral positions to examine the field over the entire area occupied by the aperture of the antenna. Inevitably, due to the presence of the ground and other environment, variations exist in the field that illuminate the antenna under test. If the measurement is made over the flat ground of an antenna test range, it is possible to apply corrections to the measured gain. However, when large, high-gain antennas are involved, it is difficult to find a sufficiently long range over flat ground. Antenna sites are often surrounded by terrain covered with vegetation, and the radiation from the source located on a tower

* A precise measurement of a horn-reflector antenna recently was made elsewhere.[1]

a mile or so away, being in part scattered from such environment, results in a spatially rough and time-varying field at the antenna under test. The time-varying effect is usually more evident in the relatively small reference (standard) horn because its beamwidth is much larger than that of the antenna under test.

Many of these objectionable features are overcome if the source can be located at an elevation angle of about 20°. Thus, neither the main lobe of the reference horn nor that of the antenna under test intercept the environment, and the measurement proceeds under more or less "free-space" conditions as it must do if one wishes to evaluate the absolute gain with confidence.

For the measurement to be discussed here, a source mounted on a helicopter was used to measure the gain of the 20-foot horn reflector[2] on Crawford Hill, Holmdel, N. J., at the frequency 4080 mc. The principle reason for making this measurement was to provide a reliable value of the effective area which could be used, in turn, for absolute measurement of the flux of extra terrestrial radio sources. Once the flux is known such sources can be used as radiators of known power for evaluating the effective areas of other large antennas at four kmc.

The measurement has resulted in a determination of the gain to within a probable error of 2 per cent. The measured gains for transverse and longitudinal polarization (the principal polarizations of the antenna) are 47.73 and 47.57 db while the calculated full area gain, assuming uniform amplitude and phase over the aperture, is 49.27 db at 4080 mc. Thus, the measured gains are 1.54 and 1.70 db below full area gain for transverse and longitudinal polarization, respectively.

II. DESCRIPTION OF THE METHOD

A block diagram of the equipment used in the measurement is shown in Fig. 1. The 4080-mc source shown schematically in the figure was flown on a helicopter. Two observers, with the aid of a TV camera mounted and boresighted along the beam of the horn-reflector antenna, accurately tracked the source antenna. A reference (standard) horn mounted on the horn reflector, was of course, also automatically beamed toward the source. When the tracking was precise (within $\pm.05°$), the receiver was switched from the horn reflector to the reference horn and the difference in the received levels was read on an output meter. When this result was combined with the measured constants of the system, the gain of the horn-reflector relative to that of the standard horn was obtained.

The 4080-mc signal source consisted of a battery operated crystal

controlled 1 mw transmitter in the cab of the helicopter connected by a length of waveguide to a special low-back-lobe radiator suspended in front of the aircraft. The surface of the helicopter nearest the radiator was covered with hair flex absorber so that any backward radiation from the source antenna would not be reflected to interfere with the forward radiation; this effect would result in a ripple in the pattern of the source

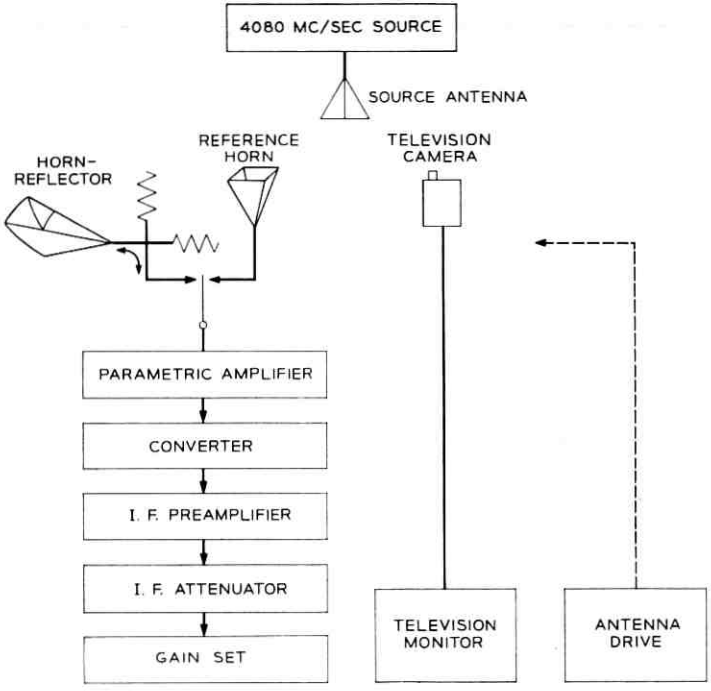


Fig. 1 — Block diagram of measuring system.

horn. The source "horn" consisted of an open ended waveguide near the apex of a pyramidal wooden horn lined with absorbent material and covered on the outside with a fine mesh brass screen. This "lossy horn" was moved along the waveguide to a position which produced the flattest pattern over the central portion of the main beam.* In a test (before installation on the aircraft), four by four-foot sheets of metal placed anywhere behind the lossy horn arrangement were found to produce total changes of only 0.2 db in its gain.

* A helicopter can not maintain an absolutely steady orientation; thus it is necessary to have an essentially flat source pattern.

The reference horn attached to the horn-reflector antenna was a (nominally) 20-db gain pyramidal horn; its performance is described in detail in Ref. 3. The reference horn was connected by a long run of waveguide to a waveguide switch inside the cab and could be mounted to receive either of the two polarizations used for the measurement. The output of the horn reflector was also connected to the waveguide switch through appropriate waveguide including a 31-db directional coupler used as an attenuator. The output of the switch was connected to the receiver by cable. The loss of the various waveguide runs was obtained by measuring the VSWR with a short circuit at the end of the line; the attenuation of the directional coupler was measured carefully by several substitution methods.

The receiver used a 2-stage parametric amplifier for its front end which provided a signal to noise ratio of more than 20 db. The paramp was followed by a converter and IF amplifier which fed through an IF attenuator into a measuring set. One 3-db step of this IF attenuator was carefully calibrated with precision attenuators. By switching in this attenuator at the same time that the input to the receiver was switched from the horn reflector to the reference horn, the signal level at the gain set remained essentially constant, and the difference could be read on the expanded scale of an output meter to within 0.01 db; it was recorded as the nearest tenth db.

III. RESULTS

The distributions of measured level differences between the signals received by the horn-reflector and the reference horn are shown in Fig. 2 for both planes of polarization. A measurement with a higher level for the horn reflector is plotted with a positive abscissa, R. The meaning of the terms "transverse polarization" and "longitudinal polarization" is given in Fig. 3.

The medians of both distributions have been found as well as the range which has a 99 per cent chance of including the true value. These results are given in Table I.

To convert the numbers in Table I to the corresponding gains of the horn reflector we need the following additional constants of the system, given in decibels in Table II. (For an explanation of the last two columns, see the next section, which discusses accuracy and corrections.) From the equation:

$$G_{HR} = G_{RH} + L_{HR} + L_{DC} - L_{RH} - L_{IF} + \bar{R} - C_{SNR} + C_{NF}$$

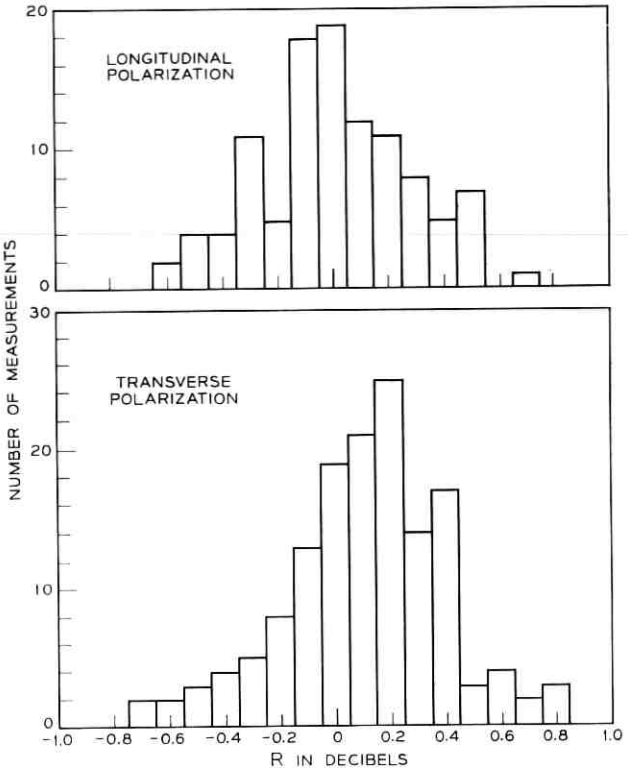


Fig. 2 — Distribution of measured data.

where all quantities are in db and $\bar{R}$ is the median value given in Table I, the measured gain of the horn-reflector is obtained:

Longitudinal Pol. $G_{HR} = 47.57$ db

Transverse Pol. $G_{HR} = 47.73$ db

IV. ACCURACY OF RESULTS AND CORRECTIONS

The known sources of error in the gain measurement are listed in Table III along with the corresponding maximum error for each.

The first three errors are lumped together in this list because their combined effect gives the scatter in the observed data (Fig. 3). The maximum error given in Table III is the average of the deviations of the 99 per cent confidence limits from their corresponding median as given in Table I.

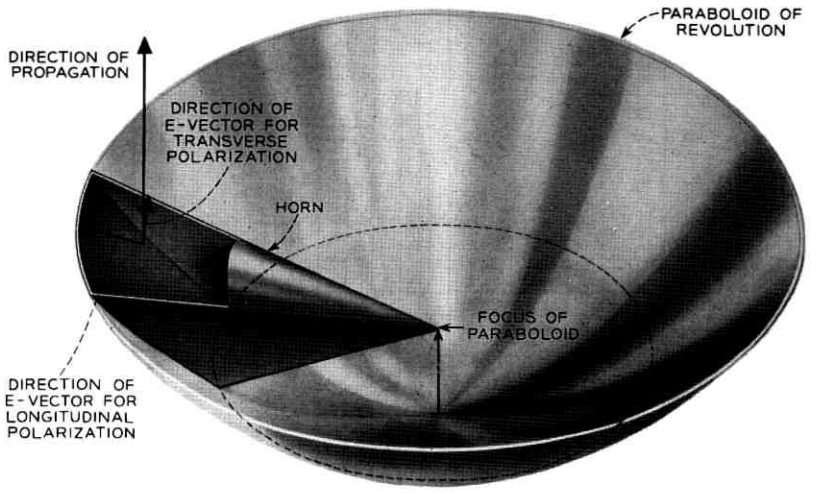


Fig. 3 — Directions of the two planes of polarization for a horn-reflector antenna.

The gain measurements were made with the helicopter at a slant range of about 5000 feet or 3 D^2/λ, D being the aperture dimension. The unusual shape and illumination of the aperture of the 20-foot horn-reflector make the usual calculations of the effect of phase error inapplicable, so the gain reduction at this range was calculated using formulas similar to those used in Ref. 2. This reduction amounts to 0.037 db in longitudinal polarization and 0.036 db in transverse. These corrections have been entered as C_{NF} in the calculation of gain above, and the uncertainty in the values due to changes in distance of the helicopter, and phase errors in the antenna is shown in Table III.

During the gain measurement the signal to noise ratio when the receiver was switched to the horn reflector was about 20 db. In this condition the noise of the parametric amplifier ($\approx$1.7 db noise figure including the input cable) was increased by the noise from the room temperature load (300°K) in the 31-db directional coupler (Fig. 1) used to approximately equalize the received signal levels. When switched to

TABLE I

Polarization	Number of Measurements	Median	99% Confidence Limits	
			Upper	Lower
Longitudinal	107	0.00 db	+0.092 db	−0.078 db
Transverse	145	+0.129 db	+0.206 db	+0.040 db

TABLE II

Polarization	Gain of Ref Horn G_{RH} (db)	Trans. Line & Switch Loss Horn Refl. L_{HR} (db)	Directional Coupler in Horn Refl. Line L_{DC} (db)	Trans. Line & Switch Loss Ref. Horn L_{RH} (db)	I.F. Attenuation Used While Switched to Reference Horn L_{IF} (db)	Median Level Difference $\bar{R}$ (db)	Correction for Change in Signal to Noise Ratio C_{SNR} (db)	Near Field Correction C_{NF} (db)
Longitudinal	20.11	0.11	31.10	0.75	3.00	0.00	0.04	0.04
Transverse	20.11	0.15	31.10	0.76	3.00	0.129	0.04	0.04

the reference horn, however, the noise added to that of the parametric amplifier was only 120°K, rather than 300°K, since that horn looks toward the cool sky. In addition the signal from the reference horn was about 3 db stronger than the signal from the horn reflector. These considerations result in a 0.04 db correction to the measured gain (C_{SNR}) with an uncertainty as listed in Table III.

The directional coupler used to equalize the signals from the two antennas was measured at Bell Telephone Laboratories by the Calibration Service of the National Bureau of Standards and by Weinschel Engineering Co. The results were 31.17 $\pm$ 0.1 db, 35.10 $\pm$ 0.3 db, and 35.04 $\pm$ 0.03 db respectively. In all cases the ranges given are "limit of error". The mean of 31.10 will be used with a limit error of ± 2 per cent.

Line and switch losses (L_{HR} and L_{RH}) were obtained by measuring the standing wave ratio with a short circuit at the end of the line. The uncertainty quoted in Table III allows for errors in the calibration of the IF attenuator used in making the SWR measurements and random errors in the measurement.

TABLE III

Maximum Error	Source of Error
1.9%	Reading Meter Re-orientation of source during measurement Inaccurate pointing of the horn-reflector
1.0%	Uncertainty in near field correction
1.0%	Signal to noise ratio uncertainty
	Uncertainty in the attenuation of components
2.0%	Directional Coupler
0.7%	Line and switch losses
0.5%	IF attenuator
0.7%	Uncertainty in gain of reference horn
2.8%	Mismatch of parametric amplifier

The IF attenuator has been compared with standard attenuators and the error listed in Table III is the uncertainty in the standards.

The measurement of the gain of the reference horn is discussed in Ref. 3. The pertinent error is made up of two parts: random errors and the error in a standard attenuator. This same attenuator was used in the Bell Telephone Laboratories measurement of the 31-db directional coupler. Therefore a correlation between the error in the standard horn measurement and the error in the Bell Telephone Laboratories value of the attenuation of the directional coupler used in the horn-reflector gain measurement is expected. The correlation, however, is in the sense of reducing the total error, so by treating the errors as random we are being conservative.

The question of correlation of errors might be asked in relation to all of our attenuation measurements since, except for the directional coupler, they are based entirely on Bell Laboratories attenuation standards which may have common origins of error of which the present authors are unaware. We therefore show in Table IV the effect of a 1 per cent error in attenuation scale for each of the terms involved in measurement of the gain of the horn reflector. However, it is not suggested that such large errors exist. What is shown in the table is measured gain minus true gain and the error assumed is in the sense that a standard attenuator labeled 20 db would actually have 20.2-db attenuation. It is seen from the table that the errors tend to cancel and it is acceptable to treat them as independent.

The last source of error shown in Table III is an unknown interaction of the mismatch of the reference horn with the input impedance of the receiver. The magnitude of this effect was not realized until after the measurement had been made and it was not possible to correct for it precisely. The maximum error of 2.8 per cent is derived from measurements made on the parametric amplifier just prior to the antenna gain measurement, taking into consideration the impedance of its connection to the switch and the impedance of the reference horn. The match of the components in the horn-reflector line was good enough that no significant uncertainty was introduced by them.

Taking the square root of the sum of the squares of all but the last error in Table III, one obtains a total (99 per cent confidence) limit

TABLE IV

Transmission Lines	Directional Coupler	IF Atten.	Reference Horn
−.013	−0.31	+0.03	+0.05

error of 3.3 per cent. Adding 2.8 per cent for the mismatch error we have a total (maximum) uncertainty in the gain of 6.1 per cent which corresponds to a probable error of about 2 per cent.

V. COMPARISON OF THE MEASUREMENT WITH THEORY AND EVALUATION OF SPILLOVER LOSS

The gain of the horn-reflector antenna at 4080 mc, when calculated by the method discussed in Ref. 2, results in gains of 48.16 db and 48.23 db for longitudinal and transverse polarization respectively. This calculation assumes that the dominant mode in the feed waveguide is preserved in the horn and is based on the projection of that mode into the aperture plane, i.e. the gain degradation due to spillover (significant only in longitudinal polarization) is neglected.

The calculated gain values can be corrected for the loss of energy in the spillover lobe (discussed in Ref. 2). Consider the equation for conservation of energy in an antenna pattern. If $G(\theta,\varphi)$ is the gain of the antenna in the direction specified by the angles θ,φ and $d\Omega$ is the differential of solid angle:

$$\iint_{\text{sphere}} G(\theta,\varphi)\ d\Omega = 1$$

Since the main-lobe region of the antenna pattern is distinct from the spillover region for horn-reflector antennas, this integral can be broken into two parts:

$$\iint_{\substack{\text{main-lobe}\\ \text{region}}} G(\theta,\varphi)\ d\Omega = 1 - \iint_{\substack{\text{spillover}\\ \text{region}}} G(\theta,\varphi)\ d\Omega$$

Because of the geometry of the antenna, the shape of the pattern in the region of the main lobe is essentially undisturbed by the presence of spillover, so the maximum gain will be reduced by the factor

$$1 - \iint G(\theta,\varphi)\ d\Omega$$

In order to make this correction to the calculated gain, the antenna pattern (longitudinal polarization) was measured in the spillover region. A source was mounted on a tower about 2 miles away ($\approx 6D^2/\lambda$) and the antenna was swept in azimuth at constant elevations. Ten unequally spaced scans were made covering 35° in elevation. On each scan the response was averaged over a degree or two and a contour map was drawn from the scans (Fig. 4). The lower elevation scans were

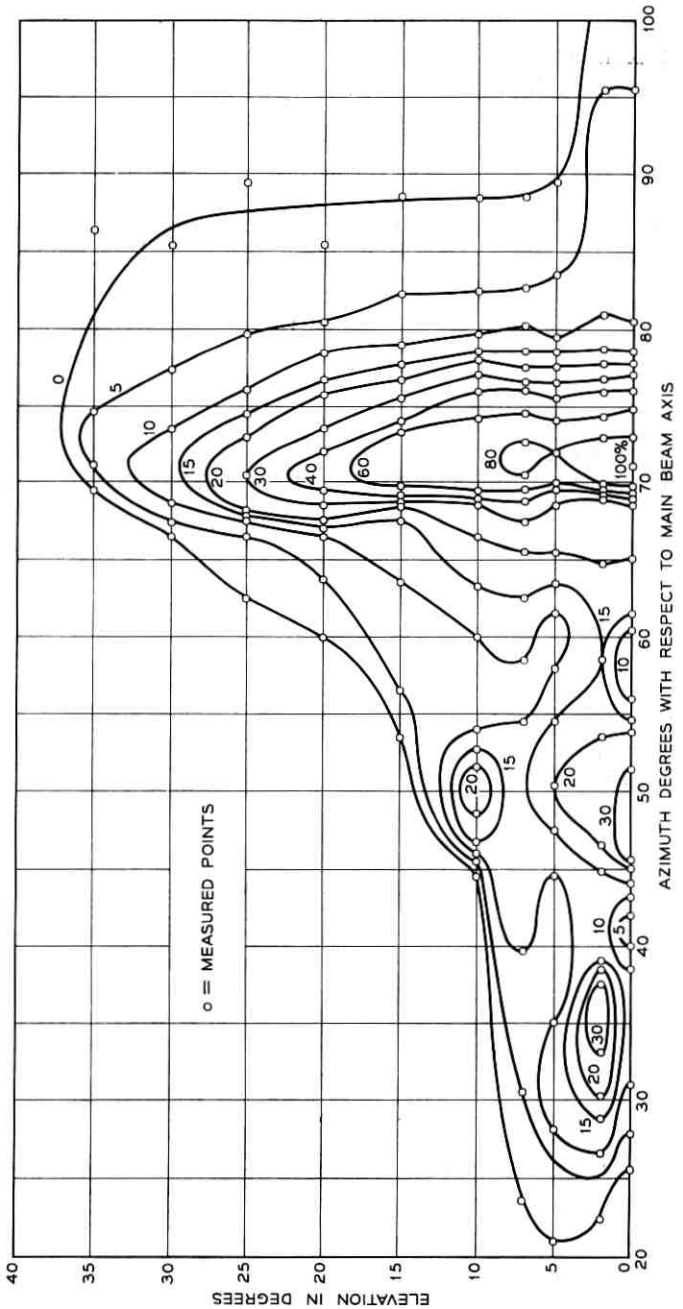


Fig. 4 — Contour map of spillover region.

checked for reflections from the environment into the main beam by inverting the antenna (which orients the main beam in a different direction) and repeating the scan through the same section of the spillover region; no significant difference was found.

The contour map was then divided into elevation zones and integrated with a planimeter. The resultant value of the integral over the full spillover lobe was 5.4 per cent of total power; this corresponds to a 0.23-db decrease in gain in longitudinal polarization.*

These results are summarized in Table V.

VI. REMARKS

A method of measuring the gain of a moderately large antenna (dimension $\approx 80\lambda$) at 4 gc, using a source mounted on a helicopter in order to minimize environmental effects, has proven accurate to within

TABLE V

	Longitudinal	Transverse
Full area gain	49.27	49.27
Computed gain	48.23	48.16
Spillover correction	0.23	0.00
Theoretical gain	48.00	48.16
Measured gain	47.57	47.73
Difference	0.43	0.43

a probable error of 2 per cent (≈ 0.09 db) and, with knowledge gained from this experience, could now be repeated with a probable error of about 1 per cent. The method would, however, be expected to prove somewhat less accurate for much larger antennas due to the increased altitude required and resultant instability of the aircraft under such conditions. Specifically, the gain of the 20-foot horn-reflector antenna has been found to be 1.70 and 1.54 db below full area gain (efficiencies 68 and 70 per cent) for longitudinal and transverse polarizations respectively. Recent measurements on radio sources (Ref. 4) have resulted in a value of 0.21 db for the difference in gain for the two polarizations; this compares favorably with the 0.16 db obtained using the aircraft-borne source.

VII. ACKNOWLEDGMENT

We thank the numerous people who actively participated in the design, calibration and operation of the experiment.

* In transverse polarization the spillover lobe is found to be negligible.

REFERENCES

1. Jull, E. V., and Deloli, E. P., IEEE Trans., *AP-12*, July, 1964, p. 439.
2. Crawford, A. B., Hogg, D. C., and Hunt, L. E., B.S.T.J., *40*, July, 1961, p. 1095.
3. Chu, T. S., and Semplak, R. A., B.S.T.J., *44*, March, 1965, p. 527.
4. Penzias, A. A., and Wilson, R. W., The Astrophysical Journal, to be published.

A General Statistical Determination of Transmission Characteristics Applied to L Multiplex

By H. G. SUYDERHOUD

(Manuscript received October 29, 1964)

There is a growing demand for modern communication systems capable of transmitting high-speed data signals over Bell System facilities designed primarily to provide telephone service. Much of this data traffic can be handled within telephone channels in the switched network. Somewhat higher data rates are feasible on private lines by selection of facilities and special treatment of the voice-grade circuits. The need for still higher data bit rates requires bandwidths equivalent to many message channels, e.g., the 12-channel group, the 60-channel supergroup, and the 600-channel mastergroup.

Presentation of basic measured data to statistically characterize transmission in the broader bands in terms of frequency domain specifications is the object of this paper. Data reduction is covered in detail. Characteristics of built-up connections are predicted from knowledge of the characteristics of subunits, including inherent variability. Such variability is a basic limitation on the degree of equalization that can be achieved with a small set of fixed networks.

I. INTRODUCTION

The demand for new services offered by the Bell System is increasing rapidly. Many of the services have performance requirements that are more critical than those of voice message telephone. They include data services over voice-band private lines and several types of DATA-PHONE service over the switched network, and wideband data services over groups (48-kc wide) and supergroups (240-kc wide) provided by L-multiplexed carrier facilities of the L-carrier plant. Services covering even wider bandwidths are under development. They are more critical in that they tolerate considerably less impairment from a number of sources. Two such impairments are amplitude and envelope delay

distortion that are present in the L-type terminals which provide our long-haul facilities. Equalization of the terminals carrying these services seems the obvious solution for reducing distortion to tolerable limits.

Equalization in the frequency domain is the process of designing networks which introduce distortions of equal magnitude but in the opposite sense to those of the system characteristic to be equalized. The success of such equalization hinges critically on the precise statistical knowledge of existing characteristics actually encountered in the plant and is limited by their variability.

Data acquision on transmission of a network as vast and complex as the L carrier plant can be accomplished only by having a judiciously chosen sampling plan. The plan discussed here consists of dividing the plant into representative subunits; many combinations of types of subunits in tandem make up a complete transmission system. Data have been obtained by way of accurate measurements of loss and envelope delay of a random sample of each category of subunits. A general method for processing and statistically determining transmission characteristics from the measurements is discussed. The method includes the problem of synthesizing the over-all system characteristic in statistical terms from separate knowledge of subunit characteristics, given any system make-up. However, the statistics of system make-up would require another study of comparable complexity.

The subunits chosen are back-to-back group, supergroup, and mastergroup modems* and interconnecting equipment of the LMX 1 carrier plant. Examples are shown of both predicted and measured multiple-link characteristics of a complex system make-up.

Comparable work is planned on data acquisition and statistical presentation for the newly designed, transistorized LMX 2 carrier terminals.[1] The methods described in this paper are therefore believed to be of general usage for presentation of transmission characteristics.

II. BASIC DATA AND THE METHOD OF ANALYSIS

2.1 *Transmission Characteristics*

Transmission through any equipment unit can be characterized in general terms by its frequency transfer function in complex form:

$$H(\omega) = \exp \{\alpha(\omega) - j\beta(\omega)\}$$

with α representing amplitude and β phase.

* The word "modem" stands for modulator/demodulator and is used here for equipment being interconnected at modulated frequencies; thus, input and output of a modem are at identical frequencies.

For the purpose of this study we are concerned with accurate knowledge of α in the frequency band of interest. For measurements at like level points or measurements corrected to such points, α is ideally equal to 0. This value is arbitrarily assigned to some reference frequency, generally near the center of the band. The amplitude characteristic of interest is then given in terms of deviation from 0 for all other frequencies. Such deviations are expressed in db, with 0 db at the reference frequency.

For distortionless transmission, β should be a linear function of angular frequency ω. In other words, we are concerned with accurate knowledge of $d\beta/d\omega$ versus frequency. $d\beta/d\omega$ is called envelope delay, and the frequency at which there is minimum envelope delay is generally used as reference. The time units of envelope delay are typically microseconds or milliseconds. Test sets measure envelope delay over a *finite* frequency difference Δf, and results are not precisely $d\beta/d\omega$; but the error is negligible for low-order distortion.

2.2 *Data Acquisition*

The quantities amplitude and envelope delay should be known continuously over the frequency bandwidth to be transmitted. For practical reasons, however, one needs only to measure at discrete intervals such that the actual variation between measurement points is of the same order of magnitude as measurement accuracy. From previous experience with or knowledge of the physics of the equipment, the necessary number and frequency spacing of measurements can be determined. Frequency spacing for group modems was 5 kc, for supergroup modems 20 kc,* and for mastergroup modems 100 kc.

Equipment of a certain type, such as group, supergroup, or mastergroup equipment of the L-carrier terminal, will not exhibit identical transmission characteristics at each installation. Manufacturing tolerances and cabling between the actual equipment and access points are the main contributors to variability. The measurements have shown, however, that this variability is substantially less than that of the quantity of interest to be estimated. In statistical terms, we may assume the coefficient of variation σ_μ/μ to be $\ll 1$, where σ_μ is the standard deviation of the quantity of interest, μ.

Basic data are thus comprised of a representative sample of characteristics of like equipments measured point by point at successive frequencies. In general, we need to make n measurements for each quantity

* Except for supergroups 1 and 3, where ripples at the band edges proved of generally higher order and intervals smaller than 20 kc where measured.

(loss or envelope delay) across the frequency band of interest and repeat them on a number of like equipments, k, to be determined by the desired degree of statistical accuracy. For instance, supergroup modem characteristics have been measured at 20-kc intervals, if variability was low enough to allow that wide a spacing. Thus, for the 240-kc band, n would generally be 12, and k was selected to equal 10.*

Since the quantity of interest for each type of equipment is the amount of distortion relative to some frequency in the band of interest, all measurements at other frequencies are normalized in terms of deviation from the value at that frequency. For amplitude, the normalizing frequency is usually the lineup or maintenance frequency. For envelope delay, it is usually the frequency of minimum envelope delay.

Measurement accuracy has an important bearing on the results and should be an order of magnitude better than the quantity to be measured. Equally important is the fact that the measurement error is random with zero mean, so that it will not bias the outcome of the experiments. By using laboratory-type equipment and exercising care during measurement and calibration procedures, we can usually fulfill these requirements. For the data reported here, a frequency accuracy of 1 part in 10^6 was ensured. Loss measurements were accurate to within 0.03 db, and for envelope delay measurements the accuracy was ± 1 microsecond.

2.3 *Data Reduction*

The statistician R. A. Fisher wrote that the object of statistical methods is the reduction of data.† In the problem at hand, some 15,000 measurements were taken to represent only 20 pairs of characteristics on amplitude and envelope delay distortion.

For each type of equipment (subunit), the quantity of interest (loss or envelope delay) was measured at a number of selected frequencies. The average and standard deviation at each frequency was then computed. The data, when suitably corrected for the mean, are assumed normal with $\mu = 0$. The validity was proven by plotting the residuals on probability paper, an example of which is shown in Section V.

Where possible, data were pooled to obtain a more precise estimate of the standard deviations. For instance, 5 group modems are numbered

* A sample of 10 generally is not regarded to yield high accuracy, but as will be explained later, an estimate of the standard deviation in this case was obtained on a "pooled" basis, increasing its accuracy by about a factor 3.

† R. A. Fisher, *On the Mathematical Foundations of Theoretical Statistics*, Phil. Trans. Royal Soc., April 1922.

1 through 5. The average loss at some frequency of a sample of each numbered group in general differs from that of another numbered group. However, the variability within each sample proves the same, which allows pooling the estimates of the variance. In the case of supergroups, 10 separate estimates could be pooled at each frequency. Since the data were taken in such a methodical and even manner, the "gamma-plot routine" was used to ensure the validity of pooling.[2] This routine compares an ordered sample of variances obtained from the (normal) data with a mathematically expected set of values. If they are plotted against each other on linear coordinates, one expects a more or less straight line for a good fit of the assumption of equality. The word "gamma" is related to the function of that name because equal-sample variances for Gaussian data are distributed as

$$f(s^2,n)\,ds^2 = \frac{\left\{\frac{s^2(n-1)}{\sigma^2}\right\}^{[(n-1)/2]-1}}{\Gamma\left(\frac{n-1}{2}\right)\cdot 2^{(n-1)/2}} \cdot \exp\left[-\frac{1}{2}\left(\frac{s}{\sigma}\right)^2 (n-1)\right] d\left\{\frac{s^2(n-1)}{\sigma^2}\right\} \tag{1}$$

which is a gamma density function for the variable $\{[(n-1)s^2]/\sigma^2\}$, also known as a χ^2 variate with $n-1$ degrees of freedom. In (1), s^2 is the sample variance and n the sample size. An example of the use of the "gamma-plot routine" in relation with this distribution is illustrated in Section V.

Cross products of sets of data at any two frequencies were computed for any subunit to obtain the sample covariance between the sets. As will be shown in Section 2.4, covariance computations are essential for unambiguous interpretations of presented graphs.

2.4 *Statistical Characterization*

Having acquired the data on amplitude and envelope delay and having performed calculations to reduce them to a few significant numbers, we are now in a position to statistically characterize transmission of the equipment subunits under study.

Over the bandwidth of interest, we have the average loss or envelope delay at a number of frequencies. A smooth curve connecting these points represents the regression of the quantity of interest on frequency, or to express this another way, the curves shown are "least squared"

estimates of loss or envelope delay versus frequency. This curve is the "most likely" characteristic to be encountered and should be used as basic data for trend equalization. However, the degree of variability calculated from the standard deviation indicates the residual variation to be expected after fixed trend equalization.

This variability is expressed in terms of population percentile. Using standard tables, limits are calculated within which a certain percentage of all probable curves is expected to fall for a certain degree of confidence. The coherence of the loss measurements and their limits will now be discussed.

Consider Fig. 1, which shows an average amplitude characteristic of a supergroup modem of the L carrier. Measurements were made at 20-kc

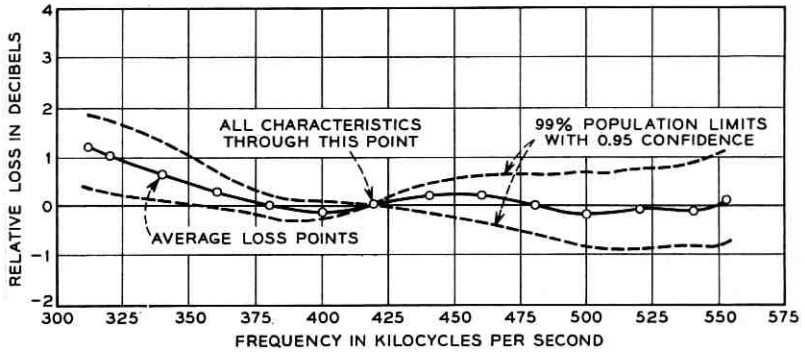


Fig. 1 — Example of average supergroup amplitude characteristic (solid line) and 99 per cent population limits (0.95 confidence).

intervals, and at each measurement frequency the average loss and 99 per cent population limits are indicated. The essential idea here is that the curve connecting the average loss points is representative of the general *shape* of the characteristic. To clarify this point statistically, one should consider sets (x_i, x_j) of pairs of data at two test frequencies, f_i and f_j for any numbered supergroup. *Any* such set may be considered a sample of a bivariate normal distribution whose general density function is

$$f(x_i, x_j)\,dx_i dx_j = \frac{dx_i dx_j}{2\pi\sigma_i\sigma_j\sqrt{1-\rho^2}} \exp -\left\{\frac{1}{2(1-\rho^2)} \cdot\left[\left(\frac{x_i-\mu_i}{\sigma_i}\right)^2 - 2\rho\,\frac{(x_i-\mu_i)\,(x_j-\mu_j)}{\sigma_i\sigma_j} + \left(\frac{x_j-\mu_j}{\sigma_j}\right)^2\right]\right\}, \tag{2}$$

where

$$\rho^{(i,j)} = \frac{\text{cov } (x_i, x_j)}{\sqrt{\text{var } (x_i) \text{ var } (x_j)}}, \tag{3}$$

the correlation coefficient of the variates x_i and x_j. For uncorrelated variates $\rho^{(i,j)} = 0$, and for perfect correlation $\rho^{(i,j)} = \pm 1$. The general form of this distribution is a bell-shaped surface, more oblong as ρ becomes closer to ± 1 (see Fig. 2). The probability of the variates lying in some specified *area* is given by the volume integral of $f(x_i, x_j)$ over that area. Tables of the bivariate normal distribution are published in Ref. 3.

The value of ρ can be estimated from the data at pairs of *adjacent* test frequencies and at pairs successively *further* apart. A typical plot of $\hat{\rho}$, an estimate of ρ, versus measurement interval is shown for supergroups in Fig. 3(*a*) and for groups in Fig. 3(*b*).

This "$\hat{\rho}$-function" illustrates very clearly two aspects of the graphs. One is that adjacent test frequencies have high *positive* correlation. This means that if loss at one frequency goes up, so does the loss in the range of ± 20 kc (for supergroups) or ± 10 kc (for groups) about this frequency. The other aspect is that losses at test frequencies more than 120 kc apart for supergroups and more than 20 kc apart for groups have very little correlation. The "threshold" of signifiance taken from significance tables of ρ (see Ref. 4), is given by the straight horizontal dotted lines in Fig. 3(a) and (b).

We are now in a firm position to make the following statements about the presented graphs: (a) each average curve is representative of the universe of like curves, and (b) ripples of higher order than those shown are unlikely. However, pivoting of the curve around the zero reference

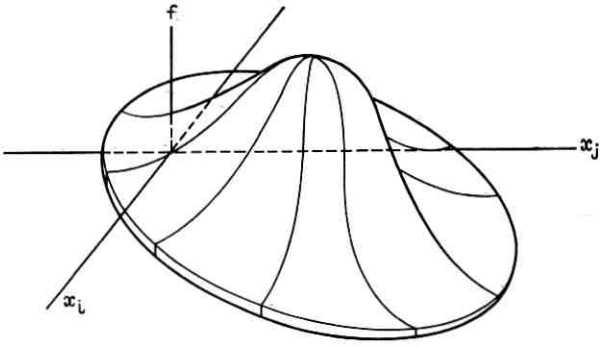
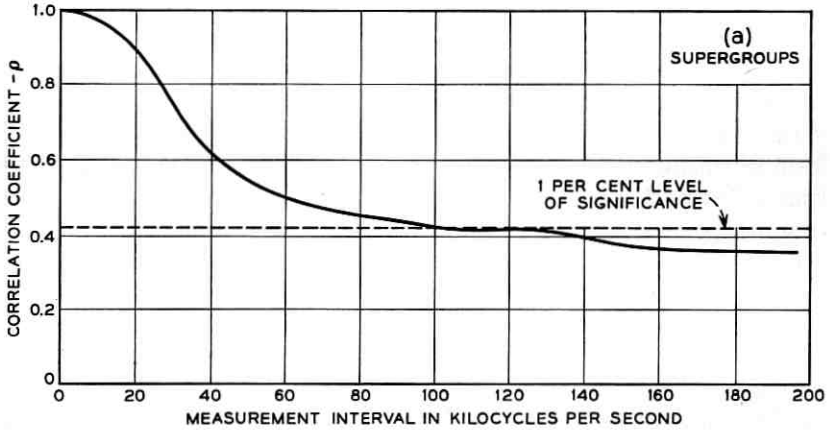


Fig. 2 — General shape of bivariate normal distribution.



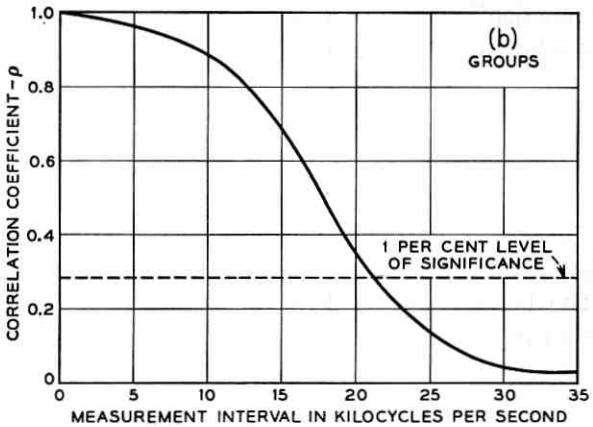


Fig. 3(a) — Correlation coefficient as a function of measurement interval for supergroups.
Fig. 3(b) — Correlation coefficient as a function of measurement interval for groups.

loss point is as likely one way as it is the other. By using the tables of the bivariate normal,[3] one can calculate the likelihood of deviations from the average characteristic.

Although these points may seem intuitively obvious, they are not. For instance, a variably poor impedance match across the band may cause ripples over short-frequency intervals. Yet, an average of such curves may still be smooth and have considerably less ripple. The ρ-function of the type shown in Fig. 3 for such a case, however, would rapidly fall off to statistically insignificant levels as the test frequency spread is increased.

2.5 *Impedance Interactions*

For accurate data, equipments were measured with laboratory-type instruments having precisely known, constant impedances. Actually these equipments are used for operation between less precisely known and controlled impedances. Since the equipments under consideration contain networks sensitive to the value of terminating impedances, departure from design values of these impedances may have an effect on the transmission characteristics.

Thus it may be that what was measured departs from the actual characteristic of the equipment in operation. It is important therefore to have knowledge of this departure, again in statistical form. Usually these departures are small and more or less random. As such, they may be incorporated in measurement errors.

Impedance interaction effects were determined by measuring amplitude characteristics of a built-up tandem connection of equipments, and then comparing it with a synthesized characteristic from appropriate addition of those measured for individual equipments. The results shown in Fig. 4 indicate a good match when the measurement accuracy is noted. Only at the band edges are some departures noticeable, but for wideband equalization purposes these regions are in general precluded due to excessive delay distortion.

III. MULTILINK PREDICTION

3.1 *General*

The problem considered here is the synthesis of a transmission characteristic of a circuit comprised of several subunits in tandem. The individual characteristics of these subunits are known in terms discussed in Section II. Again, the result would be in the form of an average and expected variation for a given degree of confidence.

In general, transmission characteristics of subunits in tandem which behave like independent variables are additive if expressed in db and if their impedance levels are identical at the interfaces. Statistically this means that the best estimate of the average characteristic of a sum of subunits is the sum of the subunit averages. Thus if the estimated average of the quantity of interest for the ith subunit at some frequency f is $\hat{\mu}_i$, then for n subunits in tandem

$$\hat{\mu} = \sum_{i=1}^{n} \hat{\mu}_i .$$

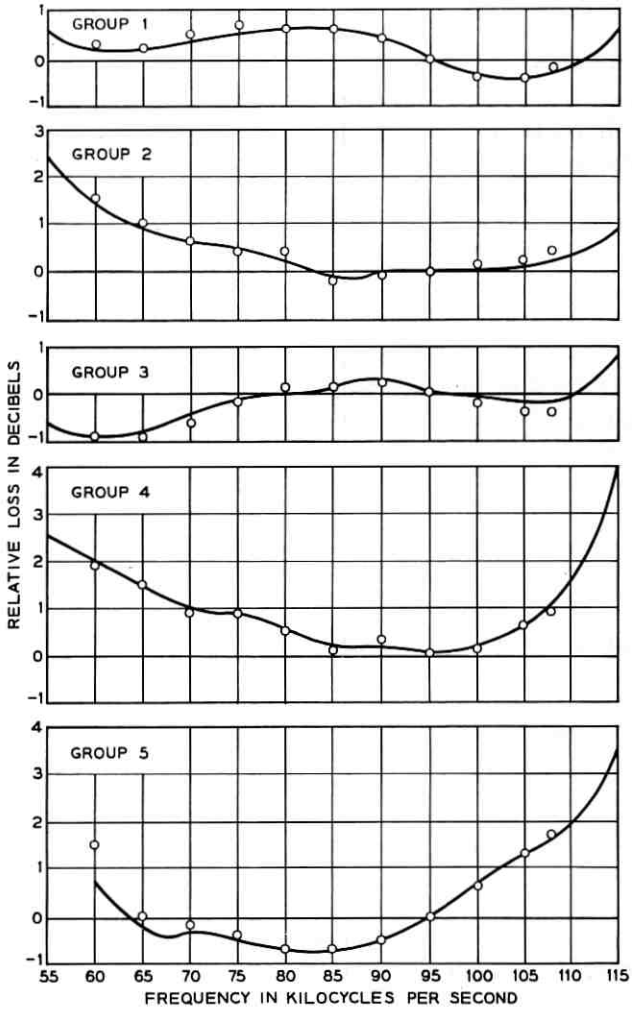


Fig. 4 — Typical example of added versus measured group characteristics. Solid lines are measured.

Similarly, for the variances (σ_i^2)

$$\hat{\sigma}^2 = \sum_{i=1}^{n} \hat{\sigma}_i^2.$$

These expressions should be used with caution if junction loss distortion

is appreciable. One may be forced in such cases to actually measure the characteristic, or else appropriately enlarge $\hat{\sigma}^2$ to reflect the effect of impedance interactions. For our purposes we showed in Section 2.4 that impedance interactions resulted in negligible junction loss distortion.

The ρ-function as discussed in Section 2.4 for a composite characteristic can be estimated from knowledge of subunit ρ-functions. The general expression for estimating ρ for n subunits in tandem between two frequencies, f_i and f_j, is

$$\hat{\rho}^{(i,j)} = \frac{\sum_{l=1}^{n} \hat{\rho}_l^{(i,j)}}{n},$$

since we may assume the variance within subunits at a given frequency to be equal. For instance, the estimation of values of ρ for a composite group characteristic traversing supergroups would result in a weighted average of ρ for groups and ρ for supergroups according to the above expression. An example will be shown in Section 3.2 of a ρ-function so derived.

3.2 *Some Applications*

A major portion of the long-distance communication facilities of the Bell System is frequency multiplexed by L-carrier equipment at terminal offices. The technique of frequency-multiplexing employs numerous filters, although the number of different filter designs is relatively few. For instance, to multiplex 600 channels only 27 different filter designs are employed.[5]

To complete transmission from one terminal office to another, a transmitting terminal is necessary at one end and a receiving terminal at the other. To characterize transmission from the transmitting to the receiving offices, one needs transmission characteristics of a transmitting and receiving terminal only, interconnected back to back. The influence of the high-frequency medium, such as radio or coaxial cable, is negligible compared with the terminal over the fraction of bandwidth considered here. That is, a supergroup occupies a 240-kc band, and an equalized coaxial cable has a maximum residual ripple of the order of only 0.2 db per 240 kc.[6]

Having gathered and statistically analyzed back-to-back terminal characteristics, one is in a position to estimate the characteristics of circuits consisting of a given number of links in tandem. Each interconnection may take place at basic group frequencies (60 to 108 kc),

basic supergroup frequencies (312 to 552 kc), or basic mastergroup frequencies (564 to 3088 kc). An appropriate connector is used for each interconnection of incoming receiving to outgoing transmitting equipment, and its transmission characteristic should be included in the overall prediction.

Fig. 5 shows the characteristic for an actual group circuit representing a complex layout between New York City and Phoenix, Arizona. This

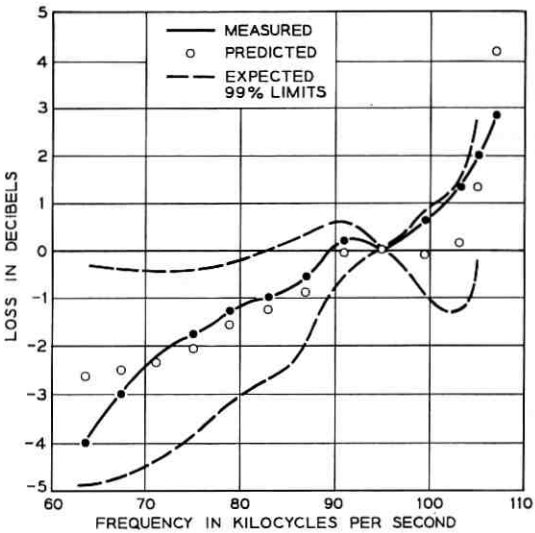


Fig. 5 — New York City-Phoenix group circuit, 8 SG connectors, 4 GR connectors.

group circuit traversed 9 supergroup modems, 8 supergroup connectors, 5 group modems, and 4 group connectors to give a total of 26 subunits. The measured characteristic* is indicated by the solid line, and it will be observed that it falls within the expected limits indicated by the dashed lines. Using the expression given in Section 3.1 for $\hat{\rho}$ and the data in Fig. 3 (*a*) and (*b*), the ρ-function for this kind of a built-up connection was calculated and is shown in Fig. 6.

Another important application of multilink prediction is that of estimating group amplitude slopes. Slope is defined here as loss difference between 63-kc and 103-kc points. From knowledge of the individual characteristics of group and supergroup modems, it is possible to lay out

* Courtesy of the Long Lines Department of the American Telephone and Telegraph Company.

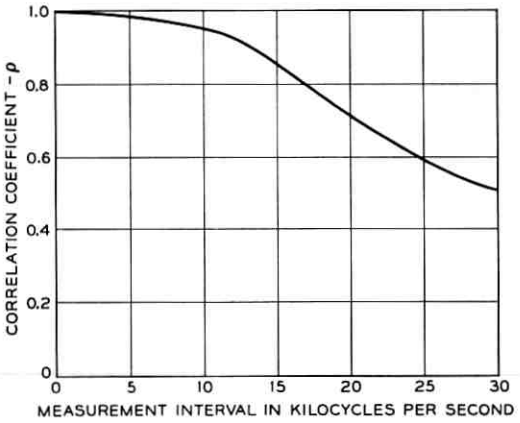


Fig. 6 — Correlation coefficient as function of measurement interval for composite group characteristic of Fig. 5.

all "paths" that a basic group band might take in being transmitted from office A to office B, given the number of links in between. For three links interconnected at basic supergroup frequencies there are 5000* equally probable paths, thus 5000 equally probable slopes. Not all the slopes will have different values, but together they will constitute a *distribution* of average slopes. If the variability within subunits is included, two more distributions of extreme (positive and negative) slopes can be numerically calculated. These three distributions are derived from composite modem characteristics of groups, supergroups, and supergroup connectors and are shown in Fig. 7.

As a third example, consider the problem of lineup of system loss. At present, L multiplex facilities in the Bell System use 92 kc for group circuits, and 424 kc for supergroup circuits as the lineup frequencies. Signals at these frequencies are permanently present for monitoring and adjustment purposes and are called pilots. However, to clear the band of interfering signals for wideband data transmission, the new standard pilot frequencies will become 104.08 kc for the group and 315.92 kc for the supergroup. The new pilot frequencies are thus located approximately 4 kc from the band edge. Maintaining the same system loss as the old pilots at these new frequencies means a general *increase* of load delivered to the high-frequency medium. This comes about because loss at new pilots in general is higher than at old pilots, a fact that could be de-

* In the first link, a set of five possible groups has a "choice" of a set of ten supergroups, and similarly ten choices in the second and third links, totaling 5×10^3.

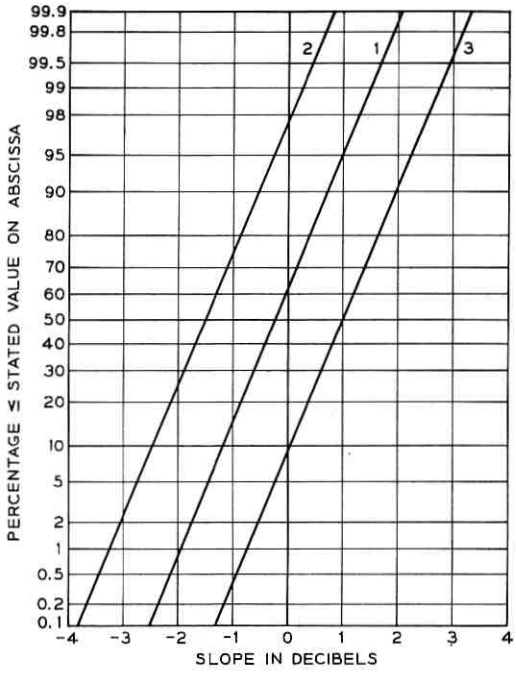


Fig. 7 — Distributions of (1) average slopes, (2) extreme negative slopes, and (3) extreme positive slopes across the group bank in a 3-link system.

termined accurately only from precise knowledge of the amplitude characteristics. On complex multilink circuits, the consequences of change in system misalignment could be determined. This, in turn, has led to work on the design of equalizers to overcome these difficulties, and also has helped to establish requirements controlling the design of new multiplex equipment.

IV. EQUALIZATION

4.1 *Introduction and Requirements*

A "theory of equalization" is very concisely presented in Ref. 6 where the two modes of frequency domain equalization, dynamic and fixed, are considered in detail and a basic set of rules is postulated. Since we assume time-varying distortion to be present only outside the terminals under consideration, it is natural to assume also that dynamic equaliza-

tion is likewise implemented. Therefore only *fixed* equalization will be considered briefly here.

As was explained in the previous sections, we deal here with characteristics that vary from equipment to equipment, although the frequency band is identical among like units. Equipments of a certain category, such as groups or supergroups, are used interchangeably and very often are operated in tandem. Yet the characteristics of, say, Group No. 4 in Supergroup No. 9* between cities A and B are not exactly the same as the same allocation between cities C and D. Moreover, this particular group circuit may be operated in tandem with, say, Group No. 1 in Supergroup No. 3 interconnected by a group connector equipment unit. Each of these have different distortions to be included in the overall distortion of the particular group circuit.

Thus, the question may be raised of how to equalize a varied plant on a *fixed* basis. The answer to a large degree hinges on the requirements for allowable distortion *after* equalization. Moreover, the requirement for residual distortion is highly dependent on the type of signal to be transmitted. In addition, the choice of equalization depends on the manner in which it is to be administered, to ease the burden of the operating telephone companies.

This paper is not primarily concerned with what plan of equalization should be used under given circumstances. However, the method of data analyzation described here is considered a basic tool to arrive at some plan of equalization with fixed networks. The assessment of residual variability and the method of correlating its limits over the frequency band of interest will subsequently be used to determine the maximum obtainable benefit in applying such networks. Work at Bell Laboratories is in progress to formulate effective equalization plans based on data gathered and analyzed in the way described here, to meet present day service requirements.

V. CHARACTERISTICS OF PRESENT PLANT

5.1 *Purpose*

At present, L-multiplex facilities generally are used for long-distance communication transmission in the Bell System. In order to gain precise knowledge of transmission characteristics, particularly to enable engineering of wideband data communications, numerous point-by-point

* Group and supergroup numbers refer to their frequency allocation *after* modulation or *before* demodulation; see also Ref. 5, p. 34.

amplitude and envelope delay measurements have been made of L-terminal equipment.

The data have been reduced and analyzed in the manner described and are presented in this section for future reference. As the plant grows and new designs like LMX 2[1] are added, more data will be taken, similarly processed, and analyzed to keep a running account of the transmission facilities. The data presented here pertain to LMX 1 terminal equipment.

5.2 *Groups and the Group Connector*

The basic group frequency band is from 60 to 108 kc. Groups numbered 1 to 5, each with different carrier frequencies, are assembled in a group *bank* the output of which, transmitting, constitutes the basic supergroup band. When looped back directly into a receiving group bank, characteristics can be measured on a back-to-back basis for each measured group.

Results of amplitude characteristics of 17 banks are shown in Figs. 8 to 12. Figure 13 shows a typical example of envelope delay for groups. The sample of groups for which envelope delay was measured was rather

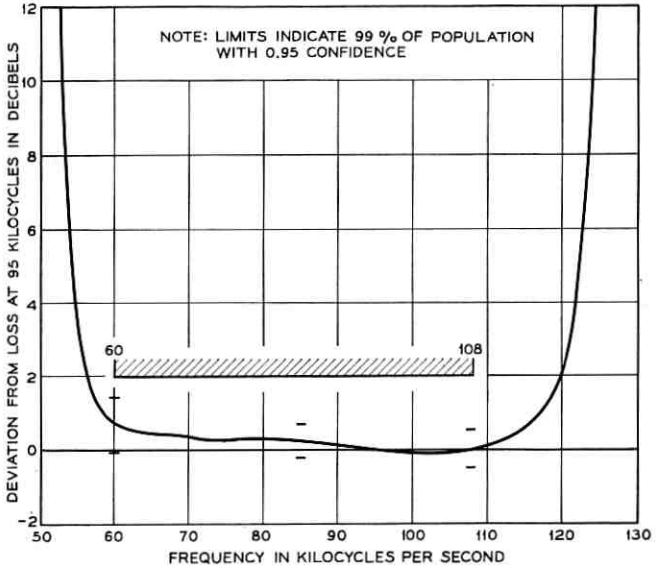


Fig. 8 — Typical modem amplitude-frequency characteristic of group No. 1.

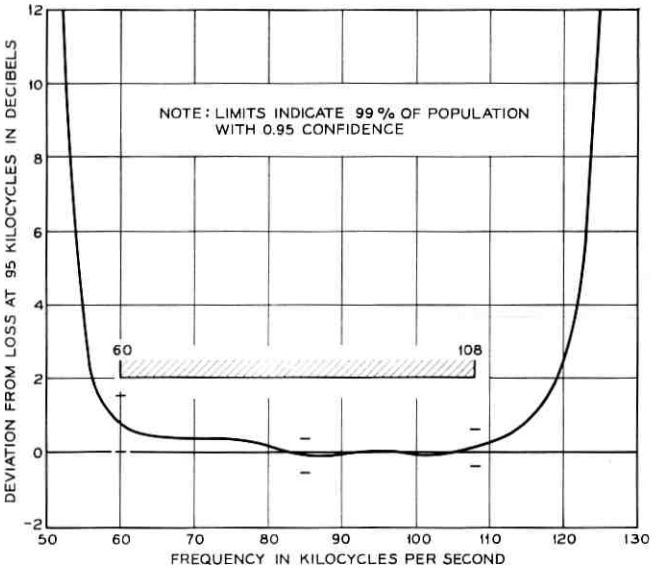


Fig. 9 — Typical modem amplitude-frequency characteristic of group No. 2.

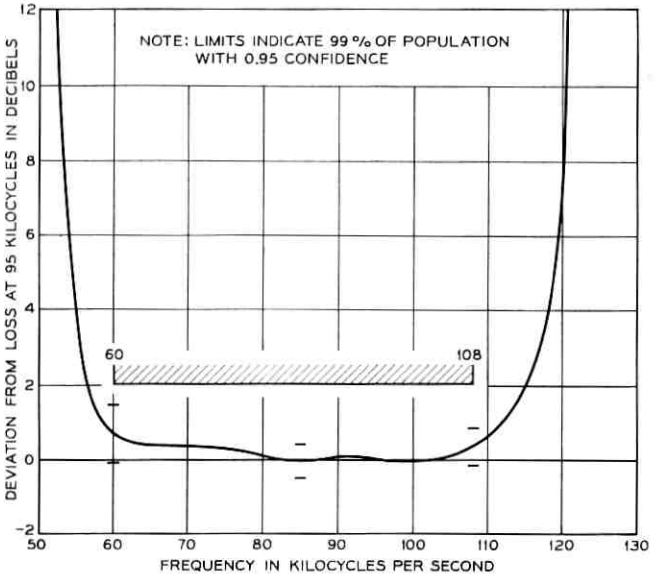


Fig. 10 — Typical modem amplitude-frequency characteristic of group No. 3.

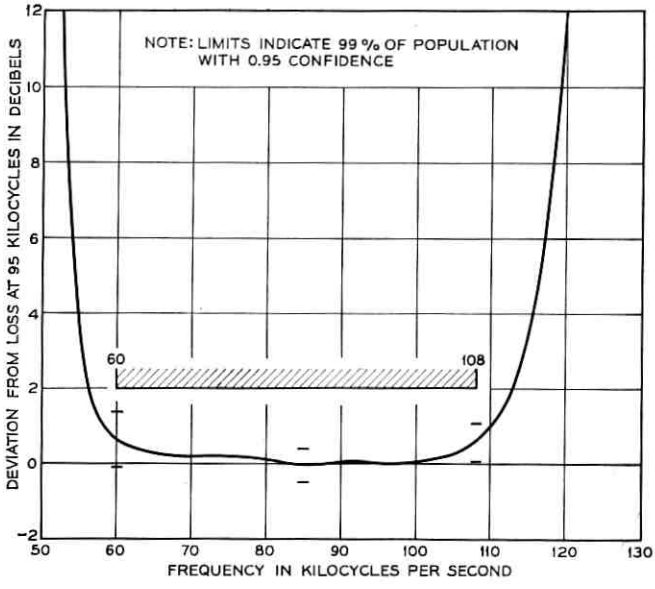


Fig. 11 — Typical modem amplitude-frequency characteristic of group No. 4.

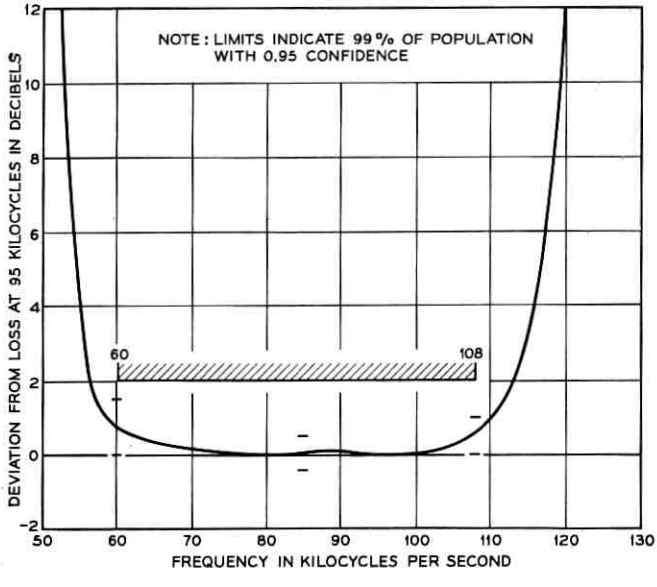


Fig. 12 — Typical modem amplitude-frequency characteristic of group No. 5.

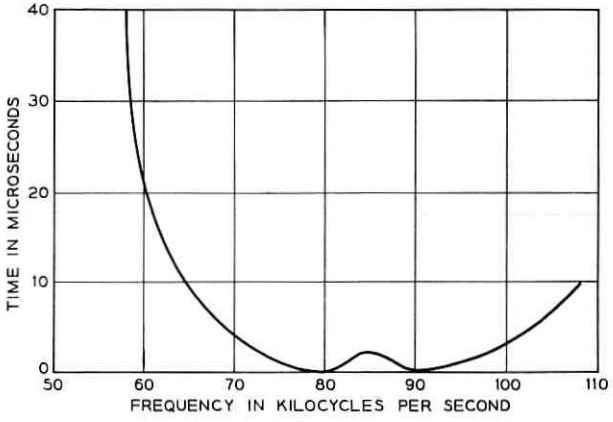


Fig. 13 — Envelope delay distortion for back-to-back group circuits of the LMX 1.

small, and variability within numbered groups proved of the same order of magnitude as variability between groups of a different number, and thus no separate presentation was warranted. The method used for precise checking of this statement is that of the Analysis of Variance where "treatments" are represented by numbered groups. At several frequencies such an analysis was made and the above conclusion confirmed. Population limits are omitted on envelope delay curves because the measurement variability proved to be comparable with equipment variability.

At terminal offices receiving groups are often retransmitted without demodulation to voice frequency. For such cases a *group connector* is used to interconnect the output of one receiving group with the input of a transmitting group. Very sharp cutoff accompanied by severe delay distortion at the band edges is the main characteristic of a group connector. The average and dispersion of nine such characteristics are shown in Fig. 14. As was the case for group bank delay curves, this figure shows no dispersion for the delay characteristics.

5.3 *Supergroups and the Supergroup Connector*

As was done for groups, the average of ten supergroup bank characteristics is presented in Figs. 15 to 24. The basic supergroup frequency band is from 312 to 552 kc, and in LMX 1 carrier systems there are ten numbered supergroups, again each with a different carrier frequency.

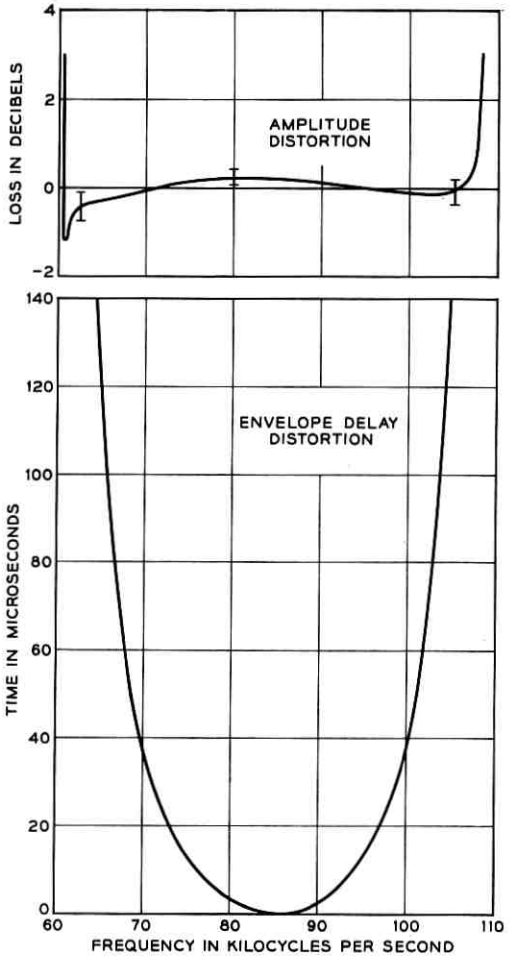


Fig. 14 — Average of nine characteristics of the group connector (L-L).

The envelope delay characteristics cannot all be lumped together as was done for groups. Here, Supergroups No. 1 and 3 are markedly different from each other and from the remaining eight. Thus, Figs. 25 and 26 show the envelope delay for Supergroups No. 1 and 3, respectively. Fig. 27 shows the characteristic for the combined measurements of Supergroups No. 2 and 4 through 10.

Similarly, supergroups are also interconnected at terminal offices, and a supergroup connector is used for this purpose. It is likewise characterized by sharp edge cutoffs, as shown in Fig. 28.

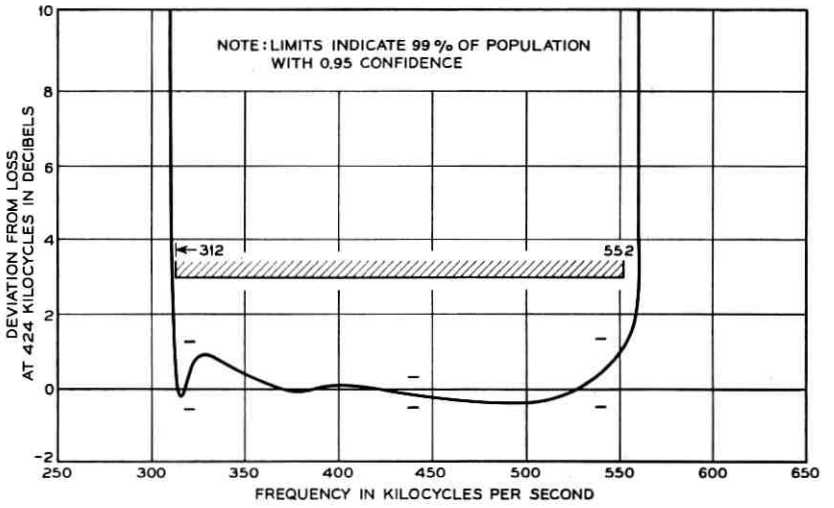


Fig. 15 — Typical modem amplitude-frequency characteristic of supergroup No. 1.

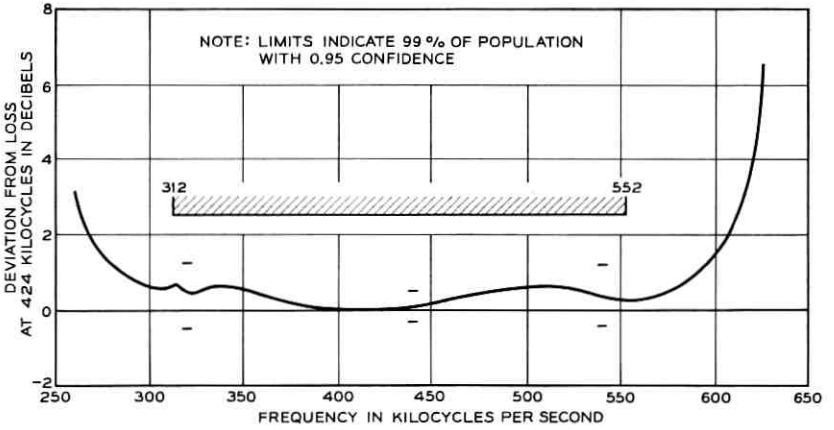


Fig. 16 — Typical modem amplitude-frequency characteristic of supergroup No. 2.

The "gamma-plot routine" mentioned in Section 2.3 was successfully applied to justify pooling of the sample loss variances for each numbered supergroup at each measurement frequency. The ten sample variances at any test frequency for each numbered supergroup multiplied by $(n - 1)/\sigma^2$ form by themsleves a sample of ten of the gamma

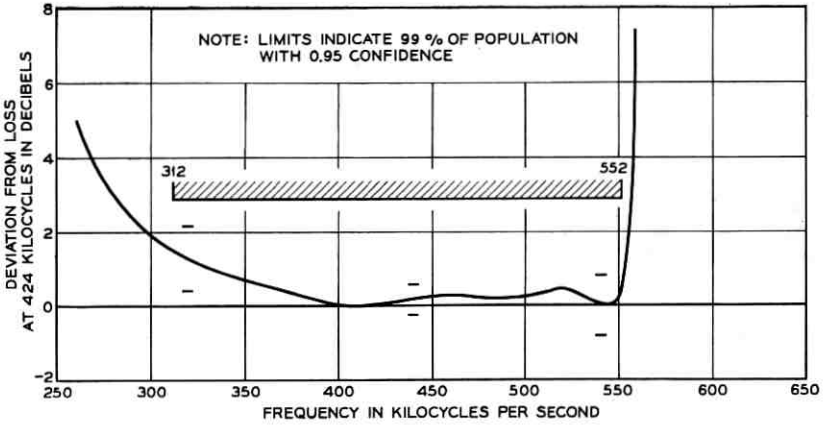


Fig. 17 — Typical modem amplitude-frequency characteristic of supergroup No. 3.

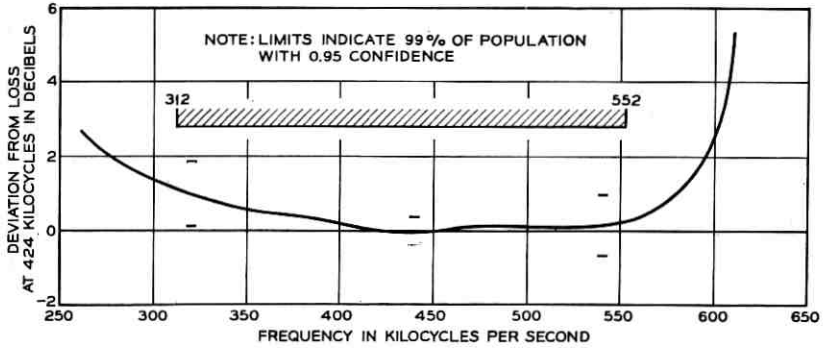


Fig. 18 — Typical modem amplitude-frequency characteristic of supergroup No. 4.

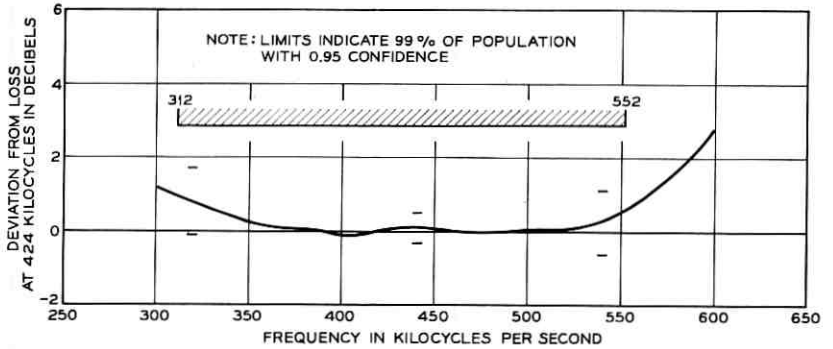


Fig. 19 — Typical modem amplitude-frequency characteristic of supergroup No. 5.

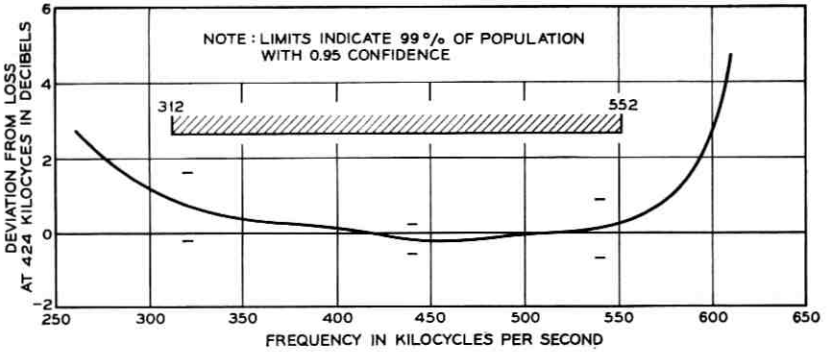


Fig. 20 — Typical modem amplitude-frequency characteristic of supergroup No. 6.

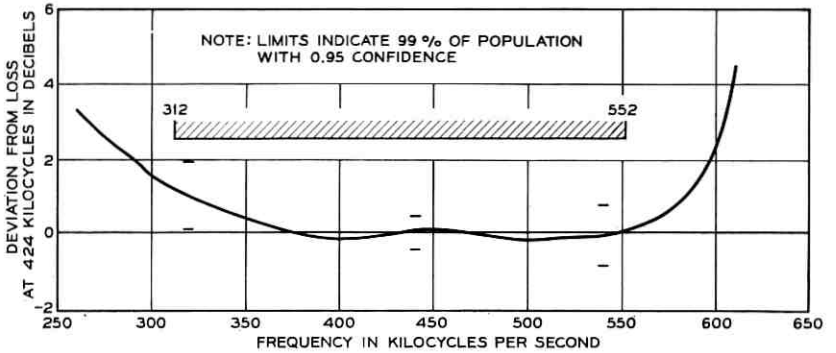


Fig. 21 — Typical modem amplitude-frequency characteristic of supergroup No. 7.

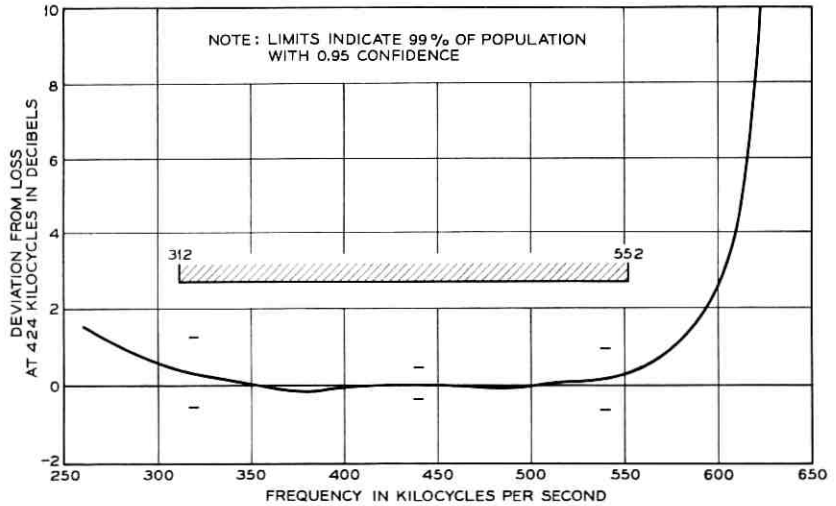


Fig. 22 — Typical modem amplitude-frequency characteristic of supergroup No. 8.

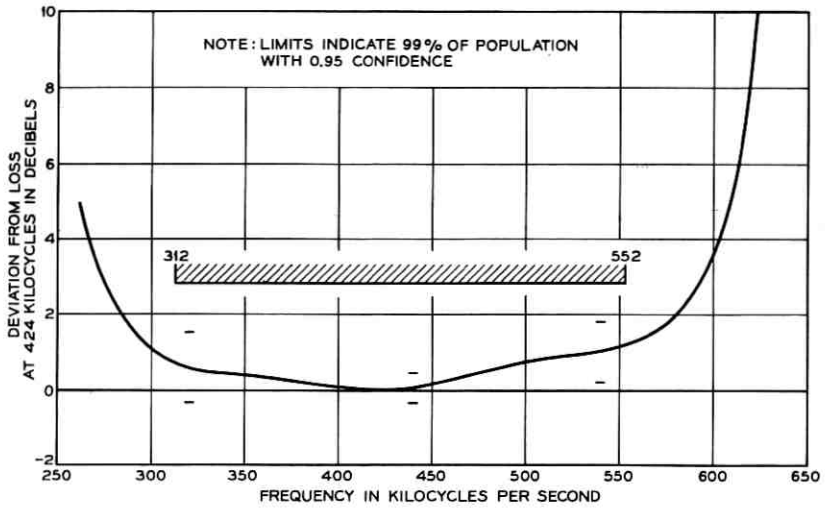


FIG. 23 — Typical modem amplitude-frequency characteristic of supergroup No. 9.

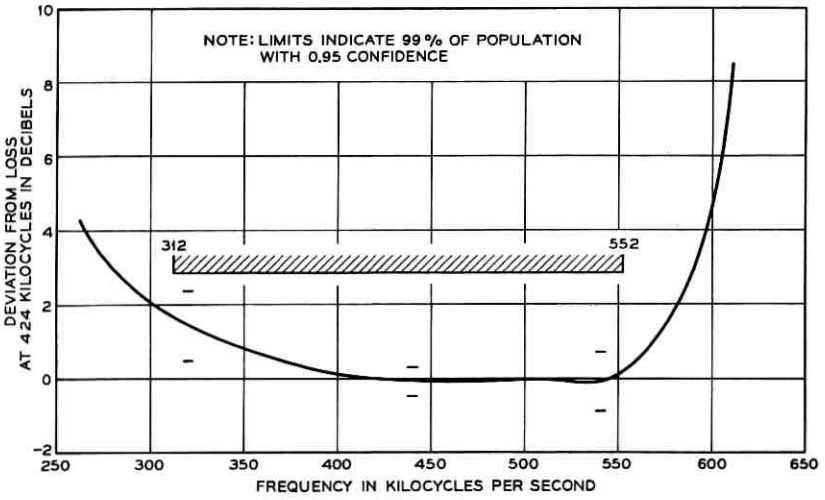


Fig. 24 — Typical modem amplitude-frequency characteristic of supergroup No. 10.

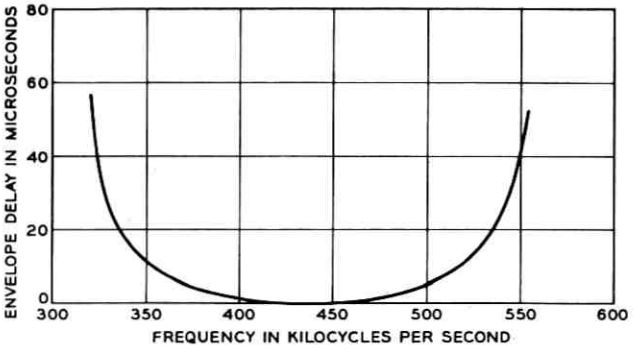


Fig. 25 — Envelope delay distortion for supergroup No. 1 of the LMX 1.

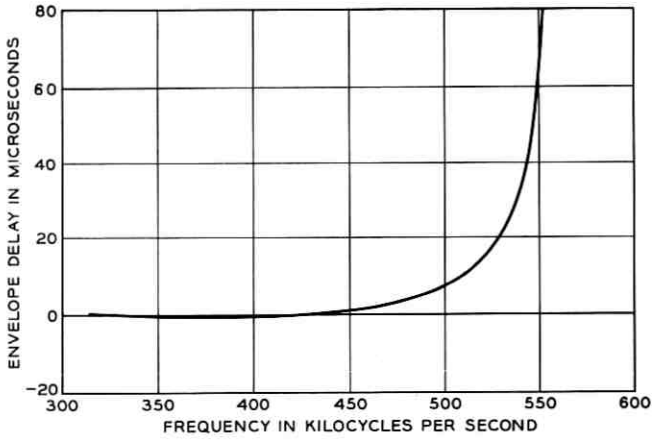


Fig. 26 — Envelope delay distortion for supergroup No. 3 of the LMX 1.

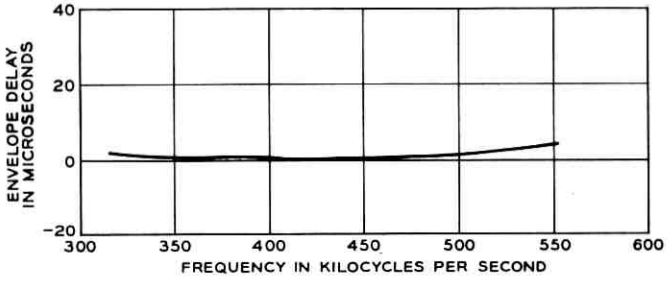


Fig. 27 — Envelope delay distortion for supergroups No. 2 and 4 through 10 of the LMX 1.

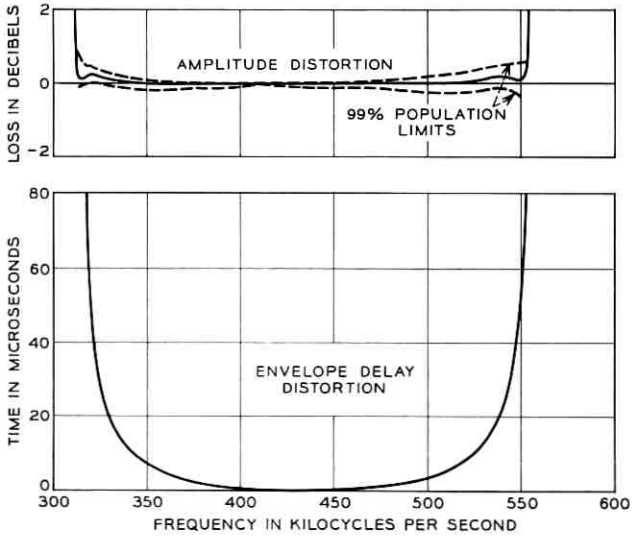


Fig. 28 — Typical amplitude and envelope delay distortion characteristics of the SG connector.

distribution (1), under the assumption that they represent only one variance, σ^2. If these ten points more or less fall on a straight line in the gamma plot, the assumption is adopted. More precisely, if the differences between the straight line values and the observations are less than, say, the 95 per cent confidence intervals of the observations, the fit may be considered good. Fig. 29 shows a typical example of the gamma plot for variances at 400 kc.

In the program of the gamma plot, σ^2 is an unknown constant. In general, it may be taken to equal one. Then the slope of the line fitted through the points would be an estimate of the pooled variance, $\hat{\sigma}^2$. Or vice versa, if another value than one is chosen and the slope of the straight line equals one, that chosen value for σ^2 would prove to be a good estimate. In the example shown in Fig. 29, the value for σ^2 inserted was 0.053 which proved to be a good estimate.

5.4 *Mastergroups and the Mastergroup Connector*

Basic mastergroup frequencies are in the band from 564 to 3084 kc. At present, the L-3 terminal combines three such mastergroups so they occupy line frequencies from 564 to 8284 kc.[7] Again, the three master-

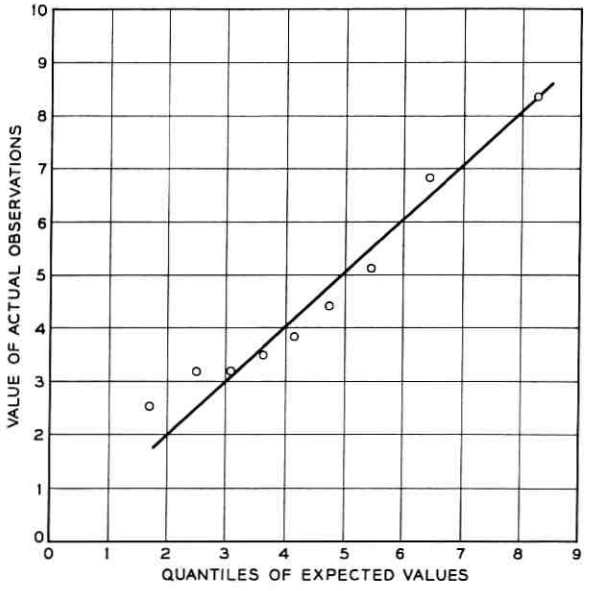


Fig. 29 — Typical gamma plot for supergroup variances at 400 kc.

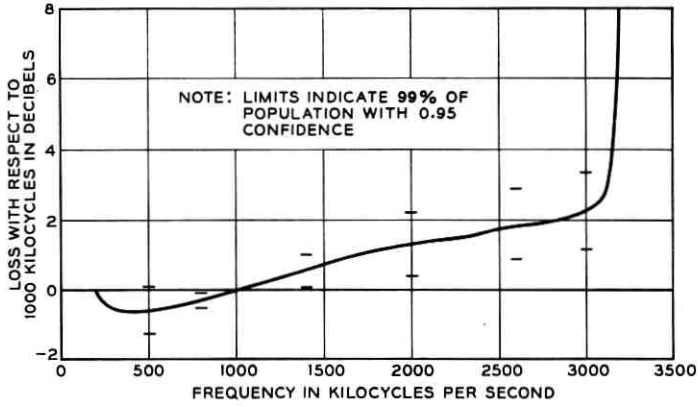


Fig. 30 — Typical modem amplitude-frequency characteristic of mastergroup No. 1.

groups are assembled into a bank, and the average characteristics shown in Figs. 30, 31, and 32 are taken of back-to-back transmitting and receiving mastergroups numbered 1, 2, and 3.

The mastergroup connector serves the same basic purpose as its group and supergroup counterparts. As this connector is relatively new however, with no sample of significant size yet measured, a statistical characterization has not yet been possible.

As was mentioned in Section 2.3, all the data have been assumed

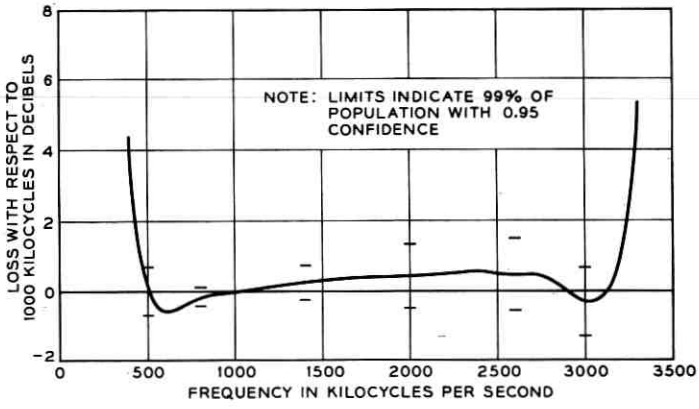


Fig. 31 — Typical modem amplitude-frequency characteristic of mastergroup No. 2.

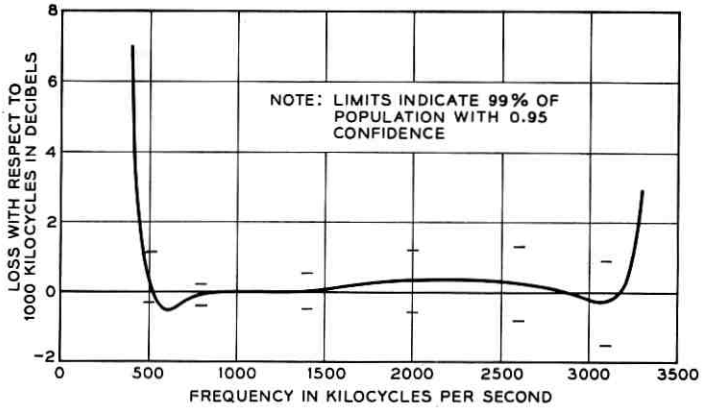


Fig. 32 — Typical modem amplitude-frequency characteristic of mastergroup No. 3.

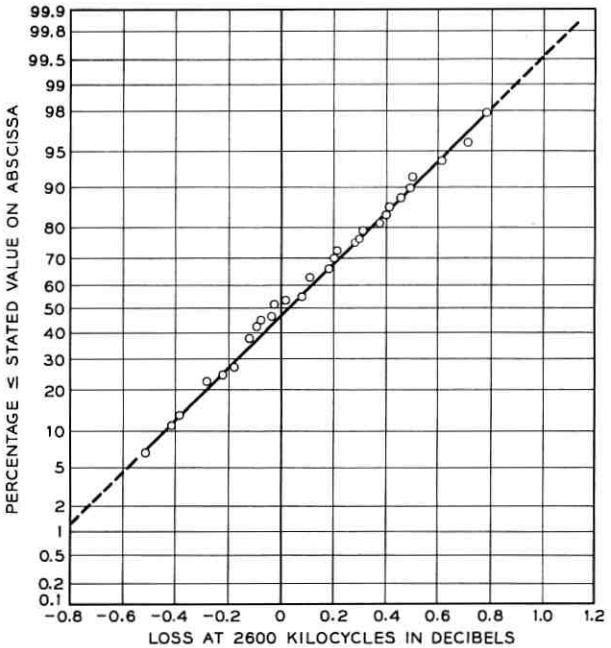


Fig. 33 — Typical example of distribution of back-to-back loss with respect to 1000 kc of a mastergroup (corrected for the mean).

normal. To justify this assumption, the sample distributions of residuals have been plotted on probability paper.* An example is shown in Fig. 33 for mastergroups where residual loss at 2600 kc relative to 1000 kc has been plotted in cumulative form. Similar plots have been made for groups and supergroups with similar results.

ACKNOWLEDGMENTS

The author is indebted to Mr. J. J. Mahoney, Jr. and Mr. R. C. Boyd for their comments and encouragement and Messrs. D. J. Chirico and A. J. Resch for their assistance in carefully obtaining the numerous data.

REFERENCES

1. (a) Hallenbeck, F. J., and Mahoney, J. J., Jr., The New L Multiplex-System Description and Design Objectives, B.S.T.J., *42*, Mar., 1963, p. 207 (also in Monograph 4509).
   (b) Graham, R. S., Adams, W. E., Powers, R. E., and Bies, F. R., New Group and Supergroup Terminals for L Multiplex, p. 223.
2. Wilk, M. B., Gnanadesikan, R., and Huyett, M. J., Probability Plots for the Gamma Distribution, Technometrics, *4*, No. 1, Feb., 1962, p. 1 (also in Monograph 4132).
3. Owen, D. B., Tables for Computing Bivariate Normal Probabilities, Ann. Math. Statistics, *27*, No. 4, Dec., 1956, p. 1075.
4. Davies, O. L., *Statistical Methods in Research and Production*, Third Revised Edition, Hafner Publishing Company, New York, 1957.
5. Mahoney, J. J., Jr., New Multiplex for Long Distance Service, Bell Laboratories Record, Jan., 1964, p. 33.
6. Ketchledge, R. W., and Finch, T. R., The L3 Coaxial System — Equalization and Regulation, B.S.T.J., *32*, July, 1953, p. 833.
7. Elmendorf, C. H., Ehrbar, R. D., Klie, R. H., and Grossman, A. J., The L3 Coaxial System — System Design, B.S.T.J., *32*, July, 1953, p. 781.
8. Fraser, D. A. S., *Statistics, An Introduction*, John Wiley and Sons, New York, 1957, p. 197.

* These residuals are not independent but could be made so by an orthogonal transformation of the observations. See also Ref. 8.

Capabilities of Bounded Discrepancy Decoding

By A. D. WYNER

(Manuscript received February 23, 1965)

The following four channels are considered: (A) a class of discrete memoryless channels with q inputs and outputs, (B) the time-discrete, amplitude-continuous memoryless channel with additive Gaussian noise and amplitude constraint, (C) the same as channel B but with energy instead of amplitude constraint, (D) a class of time-discrete, amplitude-continuous memoryless channels with amplitude constraint and non-Gaussian noise. For each channel the theoretical capabilities of "bounded discrepancy decoding" are studied.

The "discrepancy" between two vectors is a distance or distance-like quantity defined such that the optimal decoder is a "minimum discrepancy decoder." For example, for channel A the discrepancy is the Hamming distance, and for channel B the discrepancy is the Euclidean distance. Bounded discrepancy decoding is a nonoptimal decoding scheme in which disjoint regions in the space of possible received vectors are constructed about each code word, each region consisting of those vectors within a fixed discrepancy of that code word. For example, in channels A and B the regions are spheres with centers at the code words and radius $d/2$ where d is the minimum distance between code words. If the received vector is in the region about code word i, it is decoded as code word i; otherwise the decoder announces an error.

For all four classes of channels the following is shown to hold: There exists a fixed positive rate C_B below which it is possible (asymptotically in n) to obtain exponentially small error probability using bounded discrepancy decoding. In many cases C_B is shown to be strictly less than the channel capacity.

TABLE OF CONTENTS

page

I. INTRODUCTION 1062
II. SUMMARY OF RESULTS 1065
III. CHANNEL A (DISCRETE CHANNEL) 1075

page

3.1 *Lower Bound on* $M(n,d)$ ... 1075
3.2 *Asymptotic Estimates of* $M(n,d)$ ... 1078
3.3 *Bounded Discrepancy Decoding Channel Capacity* ... 1079
3.4 *Exponential Behavior of* P_{eB} ... 1080
IV. CHANNEL B (GAUSSIAN CHANNEL WITH AMPLITUDE CONSTRAINT) ... 1081
4.1 *Lower Bound on* $R(\beta)$ ... 1081
4.2 *Upper Bound on* $R(\beta)$ ... 1083
4.3 *Bounded Discrepancy Decoding Channel Capacity* ... 1087
4.4 *Exponential Behavior of* P_{eB} ... 1089
V. CHANNEL C (GAUSSIAN CHANNEL WITH ENERGY CONSTRAINT) ... 1089
5.1 *Lower Bound on* $M(n,\theta)$ ... 1089
5.2 *Asymptotic Estimates of* $M(n,\theta)$ ... 1091
5.3 *Bounded Discrepancy Decoding Channel Capacity* ... 1092
5.4 *Exponential Behavior of* P_{eB} ... 1094
VI. CHANNEL D (CONTINUOUS CHANNEL WITH AMPLITUDE CONSTRAINT) ... 1096
6.1 *Upper Bound on* $R(\beta)$ ... 1097
6.2 *Lower Bound on* $R(\beta)$ ... 1097
6.3 *Bounded Discrepancy Decoding Channel Capacity* ... 1098
6.4 *The Quadratic Discrepancy* ... 1099
APPENDIX A ... 1107
APPENDIX B ... 1109
APPENDIX C ... 1111
APPENDIX D ... 1113
APPENDIX E ... 1115
APPENDIX F ... 1115
APPENDIX G ... 1116
APPENDIX H ... 1118
GLOSSARY OF SYMBOLS ... 1119
REFERENCES ... 1122

I. INTRODUCTION

To fix ideas, let us consider first the special case of coding for the binary symmetric channel. A *code* is defined as a set of M binary n-vectors $\mathbf{x} = (x_1, x_2, \ldots, x_n)$ where $x_k = 0$ or 1 $(k = 1, 2, \ldots, n)$. The individual vectors are called code words. The transmission rate R is defined by $M = 2^{nR}$. The *Hamming distance* between two binary n-vectors is the number of positions in which they differ.

The code words are transmitted through a noisy channel. The received vector $\mathbf{y}$ is a binary n-vector whose kth coordinate is

$$y_k = x_k + z_k \pmod 2, \qquad k = 1, 2, \ldots, n, \tag{1}$$

where x_k is the kth coordinate of the transmitted code vector, and the z_k $(k = 1, 2, \ldots, n)$ are statistically independent random variables which assume the value 1 with probability $p_o (0 \leqq p_o \leqq \frac{1}{2})$, and the value 0 with probability $1 - p_o$. Thus p_o is the probability that a given bit is received in error. This channel is the "binary symmetric channel." It is assumed that each of the M code words is equally likely to be transmitted, and it is the task of the decoder to examine the received vector $\mathbf{y}$ and decide which code word was actually transmitted. We are interested in two types of decoding.

The first is termed *minimum distance decoding* or *minimum discrepancy decoding* (MDD), and here the decoder selects that code word which has the smallest Hamming distance from the received vector, and announces that word as the one which was transmitted. It is not hard to show that MDD is optimum in the sense that it minimizes the average probability of error for a given code. Let us denote by P_{eM} the average probability of error using MDD. The Fundamental Theorem of Information Theory[1] states that for any rate R less than the channel capacity $C = 1 + p_o \log_2 p_o + (1 - p_o) \log_2 (1 - p_o)$, there exists a sequence of n-dimensional codes (one for each n) with rate R such that $P_{eM} \to 0$, as $n \to \infty$. Further we may write $P_{eM} = 2^{-nE(R)+o(n)}$ where $E(R) > 0$ when $R < C$. Estimates of the exponent $E(R)$ have been found.[2,3,4]

In order to construct specific codes many workers (for example, see Refs. 5 and 6) have considered codes in which the minimum Hamming distance between code words is d. Such codes are capable of correcting errors affecting $e = (d - 1)/2$ or fewer coordinates. Suppose that the code under consideration has minimum distance d and that the decoder corrects *only* errors corrupting $e = (d - 1)/2$ or fewer coordinates (and announces an error if the received vector is not within Hamming distance e of some code word). We term this type of decoding *bounded distance decoding* or *bounded discrepancy decoding* (BDD) and the resulting error probability P_{eB}.† Since BDD does not exploit the full error-correcting potential of the code (an error may corrupt more than $e = (d - 1)/2$ coordinates and still be correctable using MDD) it is clear that $P_{eB} \geqq P_{eM}$. In this paper we shall study the theoretical capabilities of BDD, and show quantitatively what is lost by using BDD instead of MDD.

For the binary symmetric channel the following will be shown to hold:

Theorem A: There exists a fixed rate C_B (*called the* bounded distance decoding channel capacity) *below which it is possible* (*asymptotically in* n) *to obtain exponentially small error probability using BDD. In other words, for every* $R < C_B$, *there exists a sequence of n-dimensional codes* (*one for each* n) *with rate* R *such that* $P_{eB} = 2^{-nE_B(R)+o(n)}$ *where* $E_B(R) > 0$ *if* $R < C_B$). *Further if* $R > C_B$, $P_{eB} \to 1$ *as* $n \to \infty$.

Although C_B is not known exactly, it can be shown to satisfy

† The Peterson-Chien algorithm for decoding Bose-Chaudhuri-Hocquenghem codes is an example of BDD. (See Chien, R. T., Cycling Decoding Procedures for Bose-Chaudhuri-Hocquenghem Codes, IEEE Trans. on Information Theory, *IT-10*, 1964, pp. 357–363).

$$1 - H(2p_o) \leqq C_B \leqq 1 - H(\tfrac{1}{2} - \tfrac{1}{2}\sqrt{1 - 4p_o}), \qquad (2)$$

where $H(\rho) = -\rho \log_2 \rho - (1 - \rho) \log_2 (1 - \rho)$, and p_o is the bit error probability of the binary symmetric channel. These upper and lower bounds on C_B are plotted vs p_o in Fig. 1. It is clear that C_B is bounded below the channel capacity C, the maximum "error-free rate" obtainable using (optimum) MDD. The exponent $E_B(R)$ can also be estimated by upper and lower bounds.

In this paper we shall study a number of different channels (continuous as well as discrete). For each channel we shall define a distance-like function called the "discrepancy" which will be chosen so that the optimum decoder is a "minimum discrepancy decoder." (For the binary symmetric channel the discrepancy is the Hamming distance, and in most of the cases to be considered the discrepancy is a metric.) We then define a "bounded discrepancy decoder" and compare the capabilities of BDD to those of optimal MDD. In all cases we will deduce the existence of a "bounded distance decoding channel capacity" C_B for which Theorem A holds. In many of these cases we will show that C_B is strictly less than the channel capacity.

A glossary of symbols is included at the end of the paper.

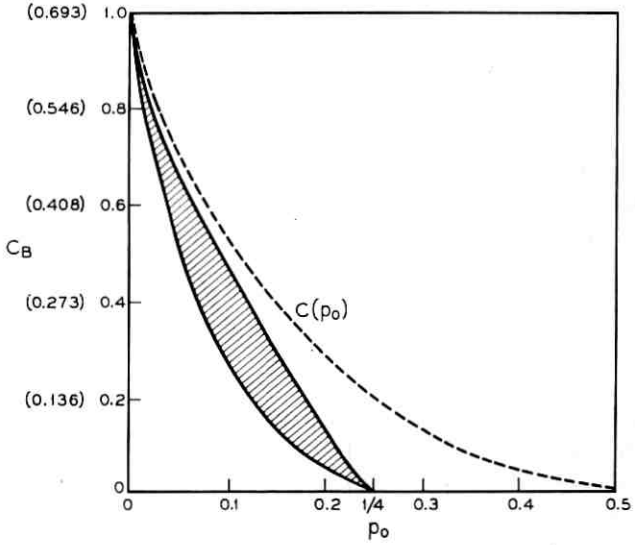


Fig. 1 — Upper and lower bounds on C_B (2) (in bits) for binary symmetric channel (solid lines). Thus C_B lies in the shaded area. The channel capacity C is the dotted line. The equivalent value of C_B corresponding to natural logarithms is given in parenthesis.

II. SUMMARY OF RESULTS

We shall consider four classes of channels. In each case the input and output are n-vectors $\mathbf{x} = (x_1, x_2, \cdots, x_n)$ and $\mathbf{y} = (y_1, y_2, \cdots, y_n)$ respectively, related by

$$y_k = x_k + z_k, \qquad k = 1, 2, \cdots n. \tag{3}$$

The symbols x_k, y_k and the noise digits z_k are as follows:

Channel A (Discrete Channel): The input digits x_k ($k = 1, 2, \cdots, n$), the output digits y_k ($k = 1, 2, \cdots, n$), and the noise digits z_k ($k = 1, 2, \cdots, n$) are members of the finite alphabet of q symbols, $0, 1, \cdots, q-1$. The addition in (3) is modulo q. The z_k are independent random variables assuming the value 0 with probability $1 - p_o$, and the values $1, 2, \cdots, q-1$ with probability $p_o/(q-1)$. Thus the channel transmits each symbol correctly with probability $1 - p_o$, and makes an error with probability p_o, all errors being equally likely. The *Hamming distance* $d_H(\mathbf{u},\mathbf{v})$ between two n-vectors $\mathbf{u}$ and $\mathbf{v}$ with entries from the alphabet of q symbols is the number of positions in which $\mathbf{u}$ and $\mathbf{v}$ differ.

Channel B (Gaussian Channel with Amplitude Constraint): The digits x_k, y_k, z_k ($k = 1, 2, \cdots, n$) are real numbers. The input vector $\mathbf{x}$ satisfies an amplitude constraint:

$$-A \leqq x_k \leqq +A, \qquad k = 1, 2, \cdots, n. \tag{4}$$

The noise digits z_k ($k = 1, 2, \cdots, n$) are independent Gaussian random variables with mean zero and variance N. The Euclidean distance between two vectors $\mathbf{u}$ and $\mathbf{v}$ is denoted by $d_E(\mathbf{u},\mathbf{v})$.

Channel C (Gaussian Channel with Energy Constraint): The digits x_k, y_k, z_k ($k = 1, 2, \cdots, n$) are real numbers. The input vector $\mathbf{x}$ lies on the surface of the n-dimensional hypersphere with center at the origin and radius $\sqrt{nP}$. Thus

$$\sum_{k=1}^{n} x_k^2 = nP. \tag{5}$$

As in channel B, the noise digits z_k ($k = 1, 2, \cdots, n$) are independent Gaussian random variables with mean zero and variance N. The signal "energy" is $\sum x_k^2 = nP$, and the expected noise "energy" is

$$E\left(\sum_k z_k^2\right) = nN,$$

so that the signal-to-noise energy ratio is P/N. This quantity is also the ratio of signal-to-noise "average power."

Channel C is of course closely related to the bandlimited channel with white Gaussian noise.[1] Such an identification, however, must be made with care, and we shall sidestep the issue here.

Channel D (Continuous Channel with Amplitude Constraint): The vectors $\mathbf{x}$, $\mathbf{y}$ and $\mathbf{z}$ are members of $\mathcal{C}_n$ defined as the set of n-vectors $\mathbf{u} = (u_1, u_2, \cdots, u_n)$ where the coordinates u_k $(k = 1, 2, \cdots, n)$ are real numbers satisfying

$$-A \leqq u_k \leqq A. \tag{6}$$

We shall assume that the symbols "$\dotplus$" and "$\dot{-}$" when applied to coordinates of vectors in $\mathcal{C}_n$ denote addition and substraction modulo $2A$, with the result reduced into the interval $[-A,A]$. Equation (3) will thus be rewritten as

$$y_k = x_k \dotplus z_k, \qquad k = 1, 2, \cdots, n. \tag{7}$$

The noise coordinates z_k are assumed to be independent identically distributed random variables with probability density function $p(u)$ which satisfies:

(a) $p(u) = 0, \quad |u| > A.$
(b) $p(u) > 0, \quad |u| \leqq A.$
(c) $p(u)$ is an even function of u. (8)
(d) $p(u)$ is a continuous, strictly monotone decreasing function of u for $0 \leqq u \leqq A$.
(e) There exists an $\alpha > 0$ such that for small u we may write

$$p(u) = p(0)[1 + O(u^{\alpha})].$$

Thus what we have done is to wrap the interval $[-A,+A]$ onto the circumference of a circle, and assume that the noise perturbs each coordinate along the circumference a distance z_k $(-A \leqq z_k \leqq A)$. Such a channel is reasonable for the case where the x_k correspond to the phase of a fixed waveform,† and also as an approximation to other channels.

For each channel we define a *code* as a set of M n-vectors $\mathbf{x}$ satisfying the above constraints. The transmission rate R is defined by $R = (1/n) \ln M$‡ so that $M = e^{nR}$ We assume that each of the M code words is equally likely to be transmitted. It is the task of the decoder to examine the received vector $\mathbf{y}$ and to decide which code word was actually

† An example in which this model is applicable may be found in A. J. Viterbi, "On a Class of Polyphase Codes for the Coherent Gaussian Channel," *IEEE International Convention Record*, part 7, 1965, pp. 209–213.

‡ For the remainder of this paper all logarithms will be taken to the base e.

transmitted. If P_{ei} is the probability that the decoder makes an incorrect choice when code word i is transmitted ($i = 1, 2, 3, \cdots, M$), and if each of the M code words is equally likely to be transmitted, then the overall probability of a decoding error is

$$P_e = (1/M)\sum_{i=1} P_{ei}. \tag{9}$$

The optimal decoder is defined as the decoding system which minimizes P_e for a given code.

As was done for the binary symmetric channel in Section I we shall consider two types of decoding.

Channel A: The optimal decoder may be shown to be the one which selects that code word $\mathbf{x}$ which minimizes the Hamming distance, $d_H(\mathbf{x},\mathbf{y})$ between $\mathbf{x}$ and the received vector $\mathbf{y}$. Accordingly, we define the "discrepancy" as the Hamming distance, and the optimal decoder is the *minimum discrepancy decoder* (or *minimum distance decoder*) denoted by MDD. Let us denote by P_{eM} the probability of error (P_e) using MDD.

The channel capacity of Channel A is readily shown to be

$$C = C(p_o) = \ln q - H(p_o) - p_o \ln (q - 1), \tag{10}$$

where

$$H(\rho) = -\rho \ln \rho - (1 - \rho) \ln (1 - \rho). \tag{11}$$

The Fundamental Theorem of Information Theory[1,7] states that for any $R < C$ there exists a sequence of n-dimensional codes (one for each n) such that $P_{eM} = e^{-nE(R)+o(n)}$ (where $E(R) > 0$ when $R < C$). Further if $R > C$, $P_{eM} \xrightarrow{n} 1$ so that C is the supremum of those rates for which it is possible to obtain vanishingly small error probability using MDD.

The second type of decoding is described as follows: For the code being used, let d be the minimum Hamming distance between pairs of code words. About each of the M code words we construct a "sphere" in the space of q^n n-vectors, consisting of those vectors not more than Hamming distance $(d - 1)/2$ from that code word. All these spheres are disjoint. If the received vector is in the sphere about code word i, it is decoded as code word i. If the received vector is in no sphere, then the decoder announces an error. We term this type of decoding *bounded discrepancy decoding* (BDD), and denote the resulting error probability by P_{eB}. (i.e., P_{eB} is the probability that the received vector is not in the sphere about the transmitted code word.) Alternately, the bounded discrepancy decoder corrects errors affecting up to $e = (d - 1)/2$ positions and no more. Clearly $P_{eB} \geqq P_{eM}$.

In connection with BDD we are interested in the quantity $M(n,d)$, the maximum number of code words in an n-dimensional code with minimum distance d. The corresponding transmission rate is $R(n,d) = (1/n) \ln M(n,d)$. The following bounds hold:

For $d/2n > (q-1)/2q$:

$$M(n,d) \leqq \frac{\beta}{\beta - \left(\frac{q-1}{2q}\right)}. \tag{12a}$$

For $d/2n < (q-1)/2q$:

$$\frac{q^n}{\sum_{r=0}^{d-2} \binom{n}{r} (q-1)^r} \leqq M(n,d) \leqq \begin{cases} q^{n[1-(2q/q-1)\beta]} qd \\ \dfrac{nq^n K(\beta)}{\sum_{r=0}^{[td/2]} \binom{n}{r} (q-1)^r \left(\frac{td}{2} - r\right)} \end{cases}, \tag{12b}$$

where

$$\beta = d/2n, \tag{12c}$$

$$t = \frac{q-1}{q\beta}\left[1 - \sqrt{1 - \frac{2q}{q-1}\beta}\right], \tag{12d}$$

$$K(\beta) = \beta/[1 - t\beta q/(q-1)], \tag{12e}$$

and where $[x]$ denotes the largest integer not greater than x. The upper bound (12a) and the first upper bound of (12b) are the well known Plotkin bounds,[8,9] and the lower bound of (12b) is the well known Varshamov-Gilbert-Sacks bound as given in Ref. 8; the second upper bound of (12b) is established in Section III.

Now let us let n and d become large while the ratio $\beta = d/2n$ is held fixed, and define $R(\beta) = \lim_{n\to\infty} R(n,d) = \lim_{n\to\infty} R(n,2\beta n)$. We obtain from (12a):

$$R(\beta) = 0, \qquad \beta > (q-1)/2q, \tag{13a}$$

and from (12b):

$$\ln q - H(2\beta) - 2\beta \ln (q-1) \leqq R(\beta) \leqq \begin{cases} \left(1 - \dfrac{2q\beta}{q-1}\right) \ln q, \\ \ln q - H(t\beta) - t\beta \ln (q-1), \end{cases} \tag{13b}$$

where $H(\rho)$ is defined by (11). The second upper bound in (13b) is the same as the Elias bound[4] which was obtained independently. Although

the bounds of (12) and (13) are of interest in themselves, we make use of them here to demonstrate the following:

Theorem A: There exists a fixed rate C_B, called the "bounded discrepancy decoding channel capacity," such that for any rate $R < C_B$, there exists a sequence of n-dimensional codes (one for each n) such that $P_{eB} = \exp [-nE_B(R) + o(n)]$ (where $E_B(R) > 0$ for $R < C_B$). Further if $R > C_B$, $P_{eB} \xrightarrow{n} 1$, so that C_B is the supremum of those rates for which it is possible to obtain vanishingly small error probability using BDD.

For channel A we shall show

$$C(2p_o) \leqq C_B \leqq C(tp_o) < C(p_o) = C, \tag{14}$$

so that C_B is strictly less than C the "maximum error free" rate using MDD.

Finally we can estimate $E_B(R)$ by

$$\alpha\left(\frac{s}{2}, p_o\right) \leqq E_B(R) \leqq \begin{cases} \alpha\left(s\left[1 - \dfrac{q}{2(q-1)}\, s\right], p_o\right), \\ \alpha\left(\dfrac{q-1}{2q}\left[\dfrac{H(s) + s \ln (q-1)}{\ln q}\right], p_o\right), \end{cases} \tag{15}$$

where $s = s(R)$ is defined by

$$R = C(s) = \ln q - H(s) - s \ln (q-1), \tag{15a}$$

and

$$\alpha(\rho,p_o) = \rho \ln \frac{\rho}{p_o} + (1-\rho) \ln \frac{(1-\rho)}{(1-p_o)}. \tag{15b}$$

Channel B: For this channel it may be shown that the optimum decoder minimizes the Euclidean distance $d_E(\mathbf{x},\mathbf{y})$ between the received vector $\mathbf{y}$ and the code word $\mathbf{x}$. Accordingly, we define the discrepancy as the Euclidean distance d_E, so that the optimal decoder is the minimum discrepancy decoder (MDD). Here too the channel capacity C is the maximum rate below which it is possible to obtain vanishingly small P_{eM}. An exact expression for C is not known but it has been estimated by upper and lower bounds by Shannon[1] and a method for computing C is outlined by Wolfowitz.[10] Bounded discrepancy decoding (BDD) is defined exactly as for Channel A with the Euclidean distance used instead of the Hamming distance.

Let $M(n,d^2)$ be the maximum number of points in an n-dimensional code with minimum distance d, and let $R(n,d^2) = (1/n) \ln M(n,d^2)$ be

the corresponding transmission rate. We let n and d become large while the ratio $\beta = (d/2)^2/n = d^2/(4n)$ is held fixed, and define $R(\beta) = \lim_{n\to\infty} R(n,d^2) = \lim_{n\to\infty} R(n,4\beta n)$. Let $\hat{\beta} = \beta/A^2$. The following estimate of $R(\beta)$ is obtained:

$$R_L(\beta) \leqq R(\beta) \leqq R_U(\beta) \tag{16}$$

where

$$R_U(\beta) = \begin{cases} 0 & \hat{\beta} \geqq \frac{1}{2} \\ (2 \ln 2)(1 - 2\hat{\beta}), & \frac{1}{4} \leqq \hat{\beta} < \frac{1}{2} \\ \dfrac{k^2}{k^2 - 2} \ln k(1 - 2\hat{\beta}), & \dfrac{1}{k^2} \leqq \hat{\beta} < \dfrac{1}{(k-1)^2} \\ & (k = 3, 4, 5, \cdots) \end{cases} \tag{17}$$

and

$$R_L(\beta) = \max \begin{cases} \ln 2 + \hat{\beta} \ln (\hat{\beta}) + (1 - \hat{\beta}) \ln (1 - \hat{\beta}) = R_{L1} \\ C_o(4\beta) = R_{L2} \end{cases} \tag{18}$$

where $C_o(\xi)$ is defined by

$$C_o(\xi) = \ln 2AK_o(\xi) - \xi\lambda(\xi), \tag{19}$$

and where $\lambda(\xi)$ is defined by

$$\int_0^A r(u) e^{-\lambda(\xi) r(u)} du = \xi \int_0^A e^{-\lambda(\xi) r(u)} du, \tag{19a}$$

where

$$r(u) = u^2, \tag{19b}$$

and

$$K(\xi) = \left[\int_{-A}^{A} e^{-\lambda(\xi) r(u)} du \right]^{-1}. \tag{19c}$$

It is verified in Appendix A that for $0 < \xi/A^2 \leqq \frac{1}{3}$, there exists a unique $\lambda(\xi)$ satisfying (19a). The first value of the lower bound R_{L1} is dominant for $0.02 \leqq \hat{\beta} < 0.5$, and the second R_{L2} for $0 < \hat{\beta} \leqq 0.02$.

We make use of the estimate of $R(\beta)$ (16) to establish Theorem A for Channel B. Here we have

$$R_L(N) \leqq C_B \leqq R_U(N). \tag{20}$$

For large values of A^2/N,

$$C_B \geqq C - \ln 2 + \epsilon(A^2/N), \tag{21}$$

where C is the channel capacity, the maximum "error free" rate using MDD, and $\epsilon \to 0$ as $A^2/N \to \infty$. Thus for large values of the "signal-to-noise ratio" A^2/N, C_B is within a constant of C, so that the ratio $C_B/C \to 1$ as $A^2/N \to \infty$. An estimate of $E_B(R)$ is also obtained.

Channel C: As with Channel B, the optimal decoder is the decoder which selects that code word $\mathbf{x}$ which has the smallest Euclidean distance from the received vector $\mathbf{y}$. Thus if $\mathbf{y} = (y_1, y_2, \cdots, y_n)$, the decoder announces that code word $\mathbf{x}$ which minimizes (with respect to $\mathbf{x}$)

$$d_E(\mathbf{x},\mathbf{y}) = \sum_{k=1}^{n} (x_k - y_k)^2 = \sum_k x_k^2 + \sum_k y_k^2 - 2\sum_k x_k y_k . \tag{22}$$

Since $\sum_k x_k^2 = nP$, $d_E(\mathbf{x},\mathbf{y})$ is minimized when $\sum_k x_k y_k$ is maximized. Hence optimal decoding is equivalent to selection of that code word $\mathbf{x}$ which minimizes the angle $a(\mathbf{x},\mathbf{y})$ between $\mathbf{x}$ and $\mathbf{y}$, where

$$\cos a(\mathbf{x},\mathbf{y}) = \frac{\sum_k x_k y_k}{\left[\sum_k x_k^2 \cdot \sum_k y_k^2\right]^{\frac{1}{2}}}. \tag{23}$$

Thus if we define the discrepancy between $\mathbf{x}$ and $\mathbf{y}$ as the angle $a(\mathbf{x},\mathbf{y})$, the optimal decoder is the minimum discrepancy decoder (MDD). Let us denote by P_{eM} the error probability using MDD.

The channel capacity is $C = \frac{1}{2}\ln[1 + (P/N)]$, and is the maximum rate below which it is possible to obtain vanishingly small P_{eM}. Further for any $R < C$, there exists a sequence of n-dimensional codes such that $P_{eM} = e^{-nE(R)+o(n)}$. Estimates of $E(R)$ are obtained in Refs. 11 and 12.

The bounded discrepancy decoder (and P_{eB}) is defined exactly as for Channels A and B but with the angle $a(\mathbf{x},\mathbf{y})$ used instead of the Hamming or the Euclidean distance.

In connection with BDD we consider $M(n,\theta)$, the maximum number of points in an n-dimensional code with minimum angle θ, and the corresponding rate $R(n,\theta) = (1/n)\ln M(n,\theta)$. The following bounds hold for $\theta < \pi/2$:

$$\frac{n}{n-1}\sqrt{\pi}\,\frac{\Gamma\left(\frac{n+1}{2}\right)}{\Gamma\left(\frac{n+2}{2}\right)}\left[\int_0^\theta \sin^{n-2}\varphi\, d\varphi\right]^{-1}$$

$$\leqq M(n,\theta) \tag{24}$$

$$\leqq \frac{\sqrt{\pi}\,\Gamma\left(\frac{n-1}{2}\right)\sin\psi\,\tan\psi}{2\Gamma\left(\frac{n}{2}\right)\int_0^{\psi}(\sin\varphi)^{n-2}\,(\cos\varphi-\cos\psi)\,d\varphi}$$

where $\psi = \sin^{-1}\sqrt{2}\,\sin\,(\theta/2)$. The upper bound was obtained by Rankin,[13] and the lower bound is obtained in Section V. If we let n become large while θ is held fixed and let $R(\theta) = \lim_{n\to\infty} R(n,\theta)$ we can obtain from (24)

$$-\ln\sin\theta \leqq R(\theta) \leqq -\ln\sqrt{2}\,\sin\,(\theta/2). \tag{25}$$

Inequalities (25) will be used to establish Theorem A for Channel C. Here we have

$$C - \ln 2 - \tfrac{1}{2}\ln\,(1-e^{-2C}) \leqq C_B \leqq C - \tfrac{1}{2}\ln 2. \tag{26}$$

Estimates will also be obtained for the exponent $E_B(R)$ and comparisons to the estimates of $E(R)$ will be made.

Channel D: For this channel we shall find the optimal decoding scheme by proceding as follows. Define the function $r(u)$ by

$$r(u) = \frac{1}{\lambda}\ln\frac{p(0)}{p(u)}, \qquad -A \leqq u \leqq +A \tag{27}$$

where $p(u)$ is the noise probability density function which satisfies assumptions (8), and λ is a constant to be specified later. Equation (27) is meaningful since by (8b), $p(u) \neq 0$. From (27) we see that

$$p(u) = p(0)\,\exp\,[-\lambda r(u)] = K_o\,\exp\,[-\lambda r(u)], \tag{28}$$

where $K_o = p(0)$. The n-fold joint probability density for the n independent noise coordinates is

$$p_n(u_1, u_2, \cdots, u_n) = \prod_{k=1}^{n} p(u_k) = K_o{}^n\,\exp\,[-\lambda\sum_{k=1}^{n} r(u_k)]. \tag{29}$$

Let us now consider the decoder. Suppose that the received vector is $\mathbf{y}$. It is not hard to show that the probability of incorrect decoding is minimized when the decoder selects that code word $\mathbf{x}$ which maximizes $p(\mathbf{y}|\mathbf{x})$, the conditional probability density of receiving $\mathbf{y}$ given that $\mathbf{x}$ is transmitted. This quantity is

$$p(\mathbf{y} \mid \mathbf{x}) = p_n(y_1 \dot{-} x_1, y_2 \dot{-} x_2, \cdots, y_n \dot{-} x_n) = K_o^n \exp\left[-\lambda \sum_{k=1}^{n} r(y_k \dot{-} x_k)\right]. \tag{30}$$

The subtraction of coordinates $y_k \dot{-} x_k$ in (30) is performed modulo $2A$ with the result reduced into the interval $[-A,+A]$. Thus for a given $\mathbf{y}$, the optimal decoder selects that code word $\mathbf{x}$ which minimizes

$$d_o(\mathbf{x},\mathbf{y}) = \sum_{k=1}^{n} r(y_k \dot{-} x_k). \tag{31}$$

The function $d_o(\mathbf{x},\mathbf{y})$ defined on $\mathcal{C}_n \times \mathcal{C}_n$ will be defined as the discrepancy function, so that the optimal decoder is the minimum discrepancy decoder (MDD). Denote the resulting error probability by P_{eM}.

Let us remark at this point that the discrepancy has the following properties:

(a) $d_o(\mathbf{x},\mathbf{y}) \geqq 0$, with equality if and only if $\mathbf{x} = \mathbf{y}$.
(b) $d_o(\mathbf{x},\mathbf{y}) = d_o(\mathbf{y},\mathbf{x})$.

It is not necessarily a metric, however, since the triangle inequality need not hold.

For any vector $\boldsymbol{\alpha} \in \mathcal{C}_n$, let the "region" $S_n(\boldsymbol{\alpha},\rho)$ be the set of vectors $\boldsymbol{\beta} \in \mathcal{C}_n$ satisfying $d_o(\boldsymbol{\alpha},\boldsymbol{\beta}) < \rho$. We say that a code has *discrepancy* ρ if the regions $S_n(\mathbf{x},\rho)$ about all M code words $\mathbf{x}$ are disjoint.

We now describe another, though nonoptimum decoding technique. Let ρ_o be the largest number such that the code under consideration has discrepancy ρ_o. Hence the regions $S_n(\mathbf{x},\rho_o)$ for all code words $\mathbf{x}$ are disjoint. If the received vector $\mathbf{y} \in S_n(\mathbf{x},\rho_o)$ for some code word $\mathbf{x}$, then it is decoded as $\mathbf{x}$. If $\mathbf{y}$ belongs to no region, an error is announced. We term this type of decoding bounded discrepancy decoding (BDD), and denote the resulting error probability by P_{eB}. Clearly $P_{eB} \geqq P_{eM}$.

A case of special interest is that for which $p(u) = K_o \exp(-\lambda u^2)$. This channel is similar to channel B when λ is large (so that the effects of wrapping the interval $[-A,+A]$ onto a circle are minimized]. In this case $r(u) = u^2$.

Suppose we are given a function $r(u)$ defined on $[-A,+A]$. This function defines a discrepancy which is appropriate for the class of noise densities $p(u) = K_o \exp[-\lambda r(u)]$. Now a given member of the class could be specified by the parameter λ. (K_o is then determined by setting the total mass of $p(u)$ equal to unity.) It is convenient instead to specify a given member of the class by the parameter N defined by

$$N = E[r(z)] = \int_{-A}^{+A} r(u)\,p(u)\,du = \int_{-A}^{+A} r(u)\,K_o e^{-\lambda r(u)}\,du. \quad (32)$$

That is, given the parameter N corresponding to a $\lambda \geqq 0$, and the function $r(u)$, one can solve (32) for λ and K_o. Thus $r(u)$ and N specify the channel. For example if $r(u) = u^2$, then $N = E[z^2]$ is the average noise "power." It is shown in Appendix H that the channel capacity, the maximum "error free" rate using MDD is

$$C = C_o(N) \quad (33)$$

where the function $C_o(\xi)$ is defined by (19) with the appropriate function $r(u)$. It is shown in Appendix A, that $C_o(\xi)$ is well defined for

$$0 < \xi \leqq \frac{1}{A}\int_0^A r(u)\,du.$$

Let us now consider BDD. Two important quantities here are $M(n,\rho)$ the largest number of code points in an n-dimensional code with discrepancy ρ, and the corresponding rate $R(n,\rho) = (1/n)\ln M(n,\rho)$. We let n and ρ become large while the ratio $\beta = \rho/n$ is held fixed, and then define $R(\beta) = \lim_{n\to\infty} R(n,\rho) = \lim_{n\to\infty} R(n,\beta n)$. It is shown in Section VI that

$$C_o(2\eta\beta) \leqq R(\beta) \leqq C_o(\beta), \quad (34)$$

where $C_o(\xi)$ is defined by (19), and η is defined by

$$\eta = \sup_{-A\leqq u_1,u_2\leqq +A} \frac{r(u_1 \dot{+} u_2)}{r(u_1) + r(u_2)}. \quad (35)$$

The addition in (35), $u_1 \dot{+} u_2$, is modulo $2A$, with the result reduced into the interval $[-A,+A]$. It is shown in Appendix B that η is finite.

The estimate (34) of $R(\beta)$ is used to establish Theorem A for channel D. Here C_B can be estimated by

$$C_o(2\eta N) \leqq C_B \leqq C_o(N) = C. \quad (36)$$

For the special case of the quadratic discrepancy $r(u) = u^2$, the quantity $\eta = 2$, so that the left-hand member of (36) is $C(4N)$. It will be shown that in this case C_B is bounded above by $C_o(2N)$ so that

$$C_o(4N) \leqq C_B \leqq C_o(2N) < C_o(N) = C. \quad (37)$$

Hence in this case C_B is strictly less than C. Further, both the upper and lower bounds of (37) will be refined for small values of the "signal-to-noise" ratio A^2/N.

For large values of "signal-to-noise ratio" A^2/N, (37) becomes

$$\frac{1}{2}\ln\frac{1}{2\pi e}\frac{A^2}{N}+\epsilon_1\left(\frac{A^2}{N}\right)\leqq C_B\leqq\frac{1}{2}\ln\frac{1}{\pi e}\frac{A^2}{N}+\epsilon_2\left(\frac{A^2}{N}\right), \tag{38}$$

where ϵ_1, $\epsilon_2\to 0$ as $A^2/N\to\infty$. It will follow that C_B is within a constant of C, so that the ratio $C_B/C\to 1$ as $A^2/N\to\infty$.

III. CHANNEL A (DISCRETE CHANNEL):

3.1 *Lower Bound on $M(n,d)$*

Our first task is to obtain the second upper bound on $M(n,d)$ of inequality (12b). We need the following lemmas:

Lemma 3.1: Let $g_1, g_2, \cdots, g_n$ be real numbers. Then

$$\sum_{k=1}^{n} g_k^2\geqq\frac{1}{n}\left(\sum g_k\right)^2. \tag{39}$$

Proof: From the Schwarz inequality

$$\left(\sum_{k=1}^{n}1\cdot g_k\right)^2\leqq\left(\sum_{k=1}^{n}1^2\right)\left(\sum_{k=1}^{n}g_k^2\right)=n\sum_{k=1}^{n}g_k^2.$$

Lemma 3.2: Given a code with minimum distance d, let $\mathbf{x}_i=(x_{i1}, x_{i2}, \cdots, x_{in})$, $i=1, 2, \cdots, m$ be any set of m points from the code. Let $\mathbf{z}$ be any n-vector and r_i $(i=1, 2, \cdots, m)$ the Hamming distance $d_H(\mathbf{x}_i, \mathbf{z})$ from $\mathbf{x}_i$ to $\mathbf{z}$. Then

$$\left(\frac{\sum_{i=1}^{m}r_i}{n}\right)^2-\frac{2(q-1)m}{q}\left(\frac{\sum_{i=1}^{m}r_i}{n}\right)+\frac{(q-1)}{q}m(m-1)\frac{d}{n}\leqq 0. \tag{40}$$

Proof: Without loss of generality assume $\mathbf{z}=0$. Arrange the m code words in an array

$$\begin{aligned}\mathbf{x}_1&=x_{11}, x_{12}, \cdots, x_{1n}\\ &\vdots\\ \mathbf{x}_m&=x_{m1}, x_{m2}, \cdots, x_{mn}.\end{aligned}$$

Denote by s_{jk} $(j=0, 1, \cdots, q-1;\ k=1, 2, \cdots, n)$ the number of times symbol j appears in column k. Then, since the code has minimum distance d,

$$\binom{m}{2}d\leqq\sum_{1\leqq r<i\leqq m}d_H(\mathbf{x}_r, \mathbf{x}_i)=\sum_{k=1}^{n}\sum_{j=0}^{q-1}\tfrac{1}{2}s_{jk}(m-s_{jk})$$

$$= \sum_k \sum_j \tfrac{1}{2} m s_{jk} - \sum_k \sum_j \tfrac{1}{2} s_{jk}^2.$$

Now $\sum_k \sum_j s_{jk} = mn$ so that

$$\binom{m}{2} d \leqq \frac{m^2 n}{2} - \frac{1}{2} \sum_k s_{0k}^2 - \frac{1}{2} \sum_k \sum_{j>0} s_{jk}^2. \tag{41}$$

Since $s_{0k} = m - \sum_{j>0} s_{jk}$, by Lemma 3.1,

$$\sum_k s_{0k}^2 \leqq \frac{1}{n} \left[\sum_{k=1}^{n} \left(m - \sum_{j>0} s_{jk} \right) \right]^2 = \frac{1}{n} \left[mn - \sum_k \sum_{j>0} s_{jk} \right]^2. \tag{42}$$

Also by Lemma 3.1,

$$\sum_{k=1}^{n} \sum_{j=1}^{q-1} s_{jk}^2 \geqq \frac{1}{(q-1)n} \left[\sum_k \sum_{j>0} s_{jk} \right]^2. \tag{43}$$

Observing that $\sum_k \sum_{j>0} s_{jk} = \sum_{i=1}^{m} r_i$, and substituting (42) and (43) into (41), we obtain

$$\begin{aligned} \binom{m}{2} d &\leqq \frac{m^2 n}{2} - \frac{1}{2n} \left(nm - \sum_i r_i \right)^2 \\ &\quad - \frac{1}{2n(q-1)} \left(\sum_i r_i \right)^2 = m \sum_i r_i - \frac{q}{2(q-1)n} \left(\sum_i r_i \right)^2. \end{aligned} \tag{44}$$

On dividing (44) by $nq/2(q-1)$, the lemma follows.

Derivation of the Bound:

Let us assume that we have an n-dimensional code with minimum distance $d(d/2n < (q-1)/2q)$ with $M = M(n,d)$ code points. Consider the "sphere" of radius $t(d/2)$ in the space of n-vectors about each code point where

$$t = \frac{q-1}{q\beta} \left(1 - \sqrt{1 - \frac{2q}{(q-1)} \beta} \right) \tag{45}$$

and $\beta = d/2n$. (Since $t \geqq 1$, these spheres are not necessarily disjoint.) To each point of the sphere at Hamming distance r from the center assign weight $\omega(r) = td/2 - r$. The "mass" μ of each sphere is therefore

$$\mu = \sum_{r=0}^{[td/2]} \binom{n}{r_i} (q-1)^r \left(\frac{td}{2} - r \right). \tag{46}$$

If an n-vector $\mathbf{z}$ is simultaneously in the sphere about the m code words $\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_m$, then we assign a weight $\omega_{\mathbf{z}}$ to $\mathbf{z}$ given by the sum of its weights in each sphere, i.e.,

$$\omega_{\mathbf{z}} = \sum_{i=1}^{m} \omega(r_i) = \frac{mtd}{2} - \sum_{i=1}^{m} r_i, \tag{47}$$

where $r_i = d_H(\mathbf{x}_i, \mathbf{z})$. If $\mathbf{z}$ lies in no sphere $\omega_{\mathbf{z}} = 0$. Consequently, we have

$$\text{mass of all } n\text{-vectors} = \sum_{\substack{\text{all } n\text{-vectors} \\ \mathbf{z}}} \omega_{\mathbf{z}} = M(n,d) \cdot \mu. \tag{48}$$

We will bound M by finding a bound on $\sum \omega_{\mathbf{z}}$. Letting $s = s_{\mathbf{z}} = \omega_{\mathbf{z}}/n$, (47) becomes

$$\frac{\sum_i r_i}{n} = \frac{mtd}{2n} - s = mt\beta - s. \tag{49}$$

Substituting (49) into (40) we get

$$\begin{aligned} m^2t^2\beta^2 - 2mt\beta s + s^2 - 2\frac{(q-1)}{q} m^2t\beta \\ + 2\frac{(q-1)}{q} ms + 2\frac{(q-1)}{q} m^2\beta - \frac{2(q-1)}{q} \beta m \leqq 0. \end{aligned} \tag{50}$$

Rewriting (50)

$$\begin{aligned} 0 \leqq s^2 \leqq m \Bigg[& 2\frac{(q-1)}{q} \beta \\ & - m\beta \left(t^2\beta - 2\frac{(q-1)}{q} t + 2\frac{(q-1)}{q} \right) \\ & - s\left(2\frac{(q-1)}{q} - 2t\beta \right) \Bigg]. \end{aligned} \tag{51}$$

Since by choice of t (45),

$$t^2\beta - 2\frac{(q-1)}{q} t + 2\frac{(q-1)}{q} = 0,$$

and

$$2\frac{(q-1)}{q} - 2\beta t > 0 \quad \text{when} \quad \beta < \frac{q-1}{2q},$$

(51) can be satisfied only when

$$s \leqq \frac{\beta}{1 - t\beta q/(q-1)} \triangleq K(\beta). \tag{52}$$

Thus

$$\sum_{\substack{\text{all } n\text{-vectors} \\ \mathbf{z}}} \omega_{\mathbf{z}} = \sum s \cdot n \leqq K(\beta) n q^n. \tag{53}$$

Hence from (53), (48) and (46) we have

$$M(n,d) \leqq \frac{K(\beta) n q^n}{\sum_{r=0}^{[t\beta n]} \binom{n}{r} (t\beta n - r)(q-1)^r}, \tag{54}$$

where

$$t = \frac{q-1}{q\beta}\left(1 - \sqrt{1 - \frac{2q\beta}{q-1}}\right), \qquad K(\beta) = \frac{\beta}{1 - t\beta q/(q-1)},$$

and $\beta < (q-1)/2q$.

3.2 *Asymptotic Estimates of* $M(n,d)$ (13)

Equation (13a) and the first upper bound of (13b) follow directly from (12a) and the first upper bound of (12b) by writing $R(n,2\beta n) = (1/n) \ln M(n,2\beta n)$ and letting n tend to infinity. The lower bound on $R(\beta)$ of (13b) follows from the lower bound on $M(n,d)$ of (12b) and the fact that (Ref. 8, Appendix A)

$$\lim_{n\to\infty} \frac{1}{n} \ln \sum_{r=0}^{\xi n} \binom{n}{r} (q-1)^r = H(\xi) - \xi \ln (q-1). \tag{55}$$

The second upper bound of (13b) follows from (54) and (55) and the fact that

$$\sum_{r=0}^{[t\beta n]} \binom{n}{r} \left(\frac{td}{2} - r\right) (q-1)^r \geqq \sum_{r=0}^{[t\beta n]-1} \binom{n}{r} (q-1)^r. \tag{56}$$

In the important special case of binary codes ($q = 2$), the second upper bound of (13b) is always sharper than the first upper bound. Thus for $q = 2$ we have for $\beta < \frac{1}{4}$:

$$1 - H(2\beta) \leqq R(\beta) \leqq 1 - H(\tfrac{1}{2} - \tfrac{1}{2}\sqrt{1 - 4\beta}). \tag{57}$$

These upper and lower bounds converge at $\beta = \frac{1}{4}$ yielding $R(\beta) = 0$, $\beta \geqq \frac{1}{4}$. Inequality (57), is plotted in Fig. 2.

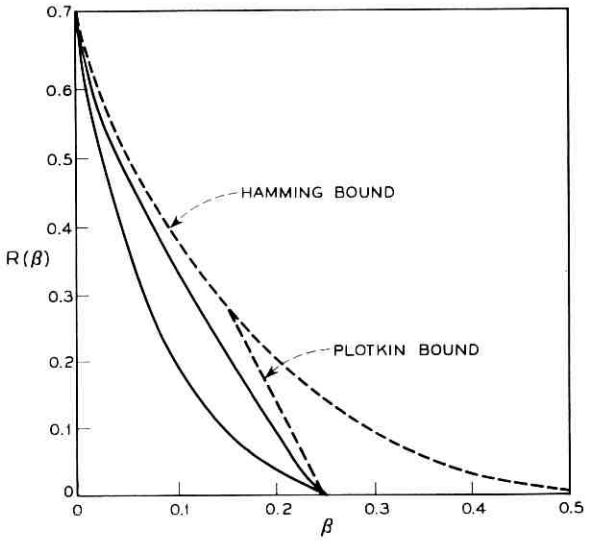


Fig. 2 (Channel A) — Upper and lower bounds on $R(\beta)$ for the binary symmetric channel (57). The dotted lines are the best bounds given in Ref. 8.

3.3 *Bounded Discrepancy Decoding Channel Capacity*

When Shannon's Fundamental Coding Theorem is applied to channel A, one finds that for every R less than the channel capacity $C = 1 - H(p_o) - p_o \ln (q - 1)$, there exists a sequence of codes (one for each n) with transmission rate R such that the error probability using MDD, $P_{eM} \overset{n}{\rightarrow} 0$. Further $R > C$, $P_{eM} \overset{n}{\rightarrow} 1$. Thus the channel capacity C is the supremum of those rates R for which it is possible (asymptotically in n) to obtain vanishingly small error probability using (optimal) MDD. We now ask what is the largest rate for which it is possible to obtain asymptotically vanishingly small error probability using BDD?

Let us suppose that for every n, an n-dimensional code is available with $d/2n = \beta$. Using BDD we have error probability

$$P_{eB} = \Pr [\text{number of errors} \geqq d/2 = \beta n]. \tag{58}$$

Since the errors in each digit occur independently with probability p_o, we have by the weak law of large numbers that $\lim_{n \to \infty} P_{eB} = 0$ or 1 according as $\beta > p_o$ or $\beta < p_o$.

If we define the *bounded discrepancy decoding channel capacity*, de-

noted by C_B, as the supremum of the rates for which it is possible (asymptotically in n) to obtain vanishingly small P_{eB}, we have from the foregoing that $C_B = R(p_o)$. Making use of the second upper bound on $R(\beta)$ of (13b) and the fact that $t > 1$ for $\beta > 0$, we have for $p_o > 0$

$$C_B = R(p_o) \leqq 1 - H(tp_o) - tp_o \ln (q-1) = C(tp_o) < C. \quad (59)$$

Thus C_B is bounded away from C.

In the binary case ($q = 2$), we make use of (13b) or (57) to obtain

$$1 - H(2p_o) \leqq C_B \leqq 1 - H(\tfrac{1}{2} - \tfrac{1}{2}\sqrt{1 - 4p_o}). \quad (60)$$

Inequality (60) is plotted in Fig. 1.

3.4 *Exponential Behavior of P_{eB}*

For a fixed $R < C_B$, denote by P_{eB}^* the smallest attainable value of P_{eB}. It was shown above that $P_{eB}^* \xrightarrow{n} 0$. We shall now show that $P_{eB}^* = e^{-nE_B(R)+o(n)}$, where $E_B > 0$ and obtain estimates of $E_B(R)$.

Given an n and R, denote by $\beta_n(R)$ the largest value of β attainable for an n-dimensional code with transmission rate R. With R held fixed, let $\beta(R) = \lim_{n\to\infty} \beta_n(R)$. Then $\beta(R)$ satisfies

$$R_L(\beta(R)) \leqq R \leqq R_U(\beta(R)), \quad (61)$$

where $R_L(\beta)$ and $R_U(\beta)$ are the upper and lower bounds of (13b). If we define the parameter $s = s(R)$ by

$$R = \ln q - H(s) - s \ln (q-1), \quad (62)$$

we obtain from (61) and (13b)

$$\frac{s}{2} \leqq \beta(R) \leqq \begin{cases} \dfrac{q-1}{2q}\left[\dfrac{H(s) + s\ln(q-1)}{\ln q}\right] \\ s\left(1 - \dfrac{qs}{2(q-1)}\right). \end{cases} \quad (63)$$

Thus for any R there exists a code (for n sufficiently large) with minimum distance $d = \beta(R)\cdot 2n$. With R fixed, this code minimizes P_{eB}. Thus from (58)

$$P_{eB}^* = \Pr[\text{no. of errors} \geqq n\beta(R)]$$
$$= \sum_{r=[\beta(R)n]}^{n} \binom{n}{r} p_o^r (1-p_o)^{n-r}. \quad (64)$$

Making use of the fact that (Ref. 8, Appendix A)

$$\lim_{n\to\infty} -\frac{1}{n} \ln \sum_{r=\rho n}^{n} \binom{n}{r} p_o^{\,r} (1 - p_o)^{n-r} = \alpha(\rho,p_o), \tag{65}$$

where $\alpha(\rho,p_o) = \rho \ln (\rho/p_o) + (1 - \rho) \ln [(1 - \rho)/(1 - p_o)]$, we have from (64):

$$E_B(R) = -\lim_{n\to\infty} (1/n)\, P_{eB}{}^{*} = \alpha(\beta(R),p_o). \tag{66}$$

Inequality (63) provides bounds on $\beta(R)$ and hence an estimate of $E_B(R)$. Let us observe that for $R = 0$ $(s = (q - 1)/q)$ the upper and lower bounds on $\beta(R)$ (63) converge yielding $\beta(R) = (q - 1)/2q$ so that

$$E_B(0) = \alpha\left(\frac{q - 1}{2q}, p_o\right). \tag{67}$$

Further since $\alpha(p_o , p_o) = 0$, $E_B(R)$ vanishes when $R = R(p_o) = C_B$.

In the binary case, the second upper bound on $\beta(R)$ (63) is always sharper than the first, so that (66) yields

$$\alpha\left(\frac{s}{2}, p_o\right) \leqq E_B(R) \leqq \alpha(s(1 - s), p_o). \tag{68}$$

Inequality (68) is plotted in Figs. 3(a) and 3(b) for $p_o = 5 \times 10^{-2}$ and $p_o = 10^{-4}$ respectively. It can be seen from Fig. 3(b) that for certain values of R the upper bound on $E_B(R)$ is greater than the lower bound on $E(R)$ (the best exponent for MDD). Thus although $E \geqq E_B$ (since $P_{eM} \leqq P_{eB}$), there is nothing to indicate that the strict inequality always holds.

IV. CHANNEL B (GAUSSIAN CHANNEL WITH AMPLITUDE CONSTRAINT)

Our first task is to establish the bounds on $R(\beta)$ given in Section II.

4.1 *Lower Bound on* $R(\beta)$

4.1.1 *Bound for Large* β

It would not violate the code constraints if the coordinates of the code words were further restricted to be $\pm A$. In this case the code is a binary code and the Hamming distance $d_H(\mathbf{x},\mathbf{y})$ between two vectors $\mathbf{x}$ and $\mathbf{y}$ is related to the Euclidean distance $d_E(\mathbf{x},\mathbf{y})$ by

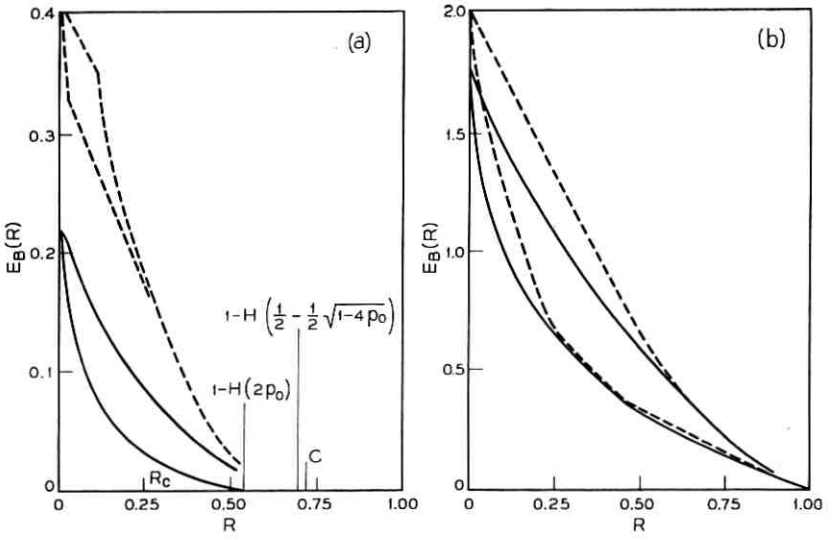


Fig. 3(a) (Channel A) — Upper and lower bounds on the exponent $E_B(R)$ for the case $q = 2$ with $p_o = 0.05$ (solid lines). Upper and lower bounds on $E(R)$ are in dotted lines.

Fig. 3(b) (Channel A) — Upper and lower bounds on the exponent $E_B(R)$ for the case $q = 2$ with $p_o = 10^{-4}$ (solid lines). Upper and lower bounds on $E(R)$ are in dotted lines.

$$d_H(\mathbf{x},\mathbf{y}) = d_E^2(\mathbf{x},\mathbf{y})/4A^2. \tag{69}$$

Thus if a code (with coordinates restricted to $\pm A$) has minimum Hamming distance $d_H = d/4A^2$, the minimum Euclidean distance is d. Thus $\hat{\beta} \triangleq \beta/A^2 = d/4A^2n = d_H/n$.

Now let $R_B(n,d_H)$ be the maximum rate for which a binary n-dimensional code with minimum Hamming distance d_H exists. We let n, and d_H become large while the ratio $\alpha = d_H/n$ is held fixed, and define $R_B(\alpha) = \lim_{n\to\infty} R_B(n,\alpha n)$. In the light of the above $R(\beta) \geqq R_B(\hat{\beta})$. The Gilbert bound (13b) (Ref. 8, p. 52) tells us that

$$R_B(\hat{\beta}) \geqq \ln 2 + \hat{\beta} \ln \hat{\beta} + (1 - \hat{\beta}) \ln (1 - \hat{\beta}) \text{ for } 0 \leqq \hat{\beta} \leqq \tfrac{1}{2}.$$

Thus we have

$$R(\beta) \geqq \ln 2 + \hat{\beta} \ln \hat{\beta} + (1 - \hat{\beta}) \ln (1 - \hat{\beta}) = R_{L_1}. \tag{70}$$

4.1.2 *Bound for Small β*

Consider a maximum size n-dimensional code with minimum distance d, and with $M = M(n,d^2)$ code points $\mathbf{x}_1, \cdots, \mathbf{x}_M$. About each code

point $\mathbf{x}_\mu$ construct an open hypersphere in n-space of radius d. Let V_μ denote the volume of the intersection of this sphere with the n-cube $[-A,+A]^n$. Now the union of these M spheres must cover the n-cube: for if $\mathbf{x}_\nu \in [-A,+A]^n$ is not contained in one of the spheres, $d(\mathbf{x}_\nu, \mathbf{x}_\mu) \geqq d$ for all μ, so that $\mathbf{x}_\nu$ may be added to the code destroying maximality. Thus

$$\sum_{\mu=1}^{M} V_\mu \geqq (2A)^n. \tag{71}$$

Now let S be the n-dimensional hypersphere of radius d with center at the origin, and $V_n(d)$ the volume of $S \cap [-A,+A]^n$. It is not hard to show that

$$V_\mu \leqq V_n(d), \qquad \mu = 1, 2, \cdots, M.$$

Consequently from (71)

$$MV_n(d) \geqq \sum_{\mu=1}^{M} V_\mu \geqq (2A)^n,$$

so that

$$M(n,d^2) \geqq [(2A)^n/V_n(d)]. \tag{72}$$

Applying the result of Appendix C to (72) yields

$$R(\beta) = \lim_{n\to\infty} (1/n) \ln M(n,4\beta n) \geqq C_o(4\beta). \tag{73}$$

It is shown in Appendix D that for small β

$$R_L(\beta) = R_{L_2}(\beta) = \tfrac{1}{2} \ln (A^2/2\pi e\beta) + \epsilon(\beta), \tag{74}$$

where $\epsilon(\beta) \to 0$ as $\beta \to 0$.

$R_L(\beta)$ is plotted in Figs. 4(a) and 4(b).

4.2 *Upper Bound on* $R(\beta)$

The approach used in this derivation is similar to Plotkin's technique for binary codes. We begin by obtaining the following:

Lemma 4.1: If $n < d^2/2A^2 (\hat{\beta} = d^2/4A^2n > \tfrac{1}{2})$,

$$M(n,d^2) \leqq \frac{d^2}{d^2 - 2A^2n} = \frac{2\hat{\beta}}{2\hat{\beta} - 1}. \tag{75}$$

Proof: Consider the maximum size n-dimensional code with minimum

distance d with $M = M(n,d^2)$ code points $\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_M$. Let $\mathbf{x}_\nu = (x_{\nu 1}, x_{\nu 2}, \cdots, x_{\nu n})$. Then

$$\binom{M}{2} d^2 \leqq \sum_{1 \leqq \mu < \nu \leqq M} d^2(\mathbf{x}_\mu, \mathbf{x}_\nu) = \sum_{k=1}^{n} \sum_{\mu<\nu} (x_{\nu k} - x_{\mu k})^2$$

$$= \sum_{k=1}^{n} \left\{ M \sum_{\nu} x_{\nu k}^2 - \left(\sum_{\nu} x_{\nu k} \right)^2 \right\}$$

$$\leqq M \sum_{k=1}^{n} \sum_{\nu=1}^{M} x_{\nu k}^2 .$$

Since $x_{\nu k}^2 \leqq A^2$,

$$M(M-1)\frac{d^2}{2} = \binom{M}{2} d^2 \leqq M^2 n A^2 ,$$

from which

$$M(n,d^2)(d^2 - 2A^2 n) \leqq d^2,$$

and if $n < d^2/2A^2$,

$$M(n,d^2) \leqq \frac{d^2}{d^2 - 2A^2 n}, \tag{76}$$

completing the proof.

Lemma 4.2: Let α be an integer not less than two. Then

$$M(n,d) \leqq \alpha M[n-1, d^2 - (2A/\alpha)^2].$$

Proof: Again consider the maximum size code with $M(n,d)$ points. Partition the code into α classes $S_1, S_2, \cdots, S_\alpha$, where S_i consists of those code points $\mathbf{x}_\nu = (x_{\nu 1}, x_{\nu 2}, \cdots, x_{\nu n})$ such that

$$-A + (i-1)(2A/\alpha) \leqq x_{\nu 1} < -A + i(2A/\alpha), \qquad i = 1, 2, \cdots, \alpha.$$

In other words we partition the interval $[-A,+A]$ into α subintervals of length $2A/\alpha$, and assign $\mathbf{x}_\nu$ to class S_i according as its first coordinate $x_{\nu 1}$ is in the ith subinterval. (To be precise we must close the last subinterval ($i = \alpha$) at both ends to make the α subintervals cover $[-A,+A]$.)

Now delete the first coordinate from each point in the code. Each class S_i is now a code of length $n-1$ with minimum distance not less than

$$\sqrt{d^2 - \left(\frac{2A}{\alpha}\right)^2},$$

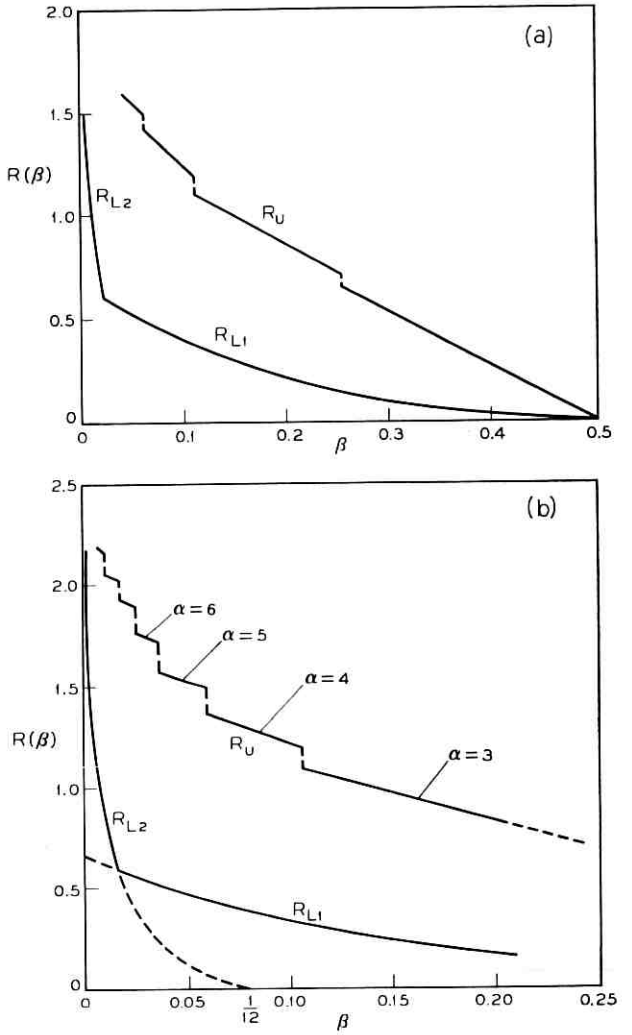


Fig. 4(a) (Channel B) — Upper and lower bounds on $R(\beta)$ vs β $(0 \leqq \beta \leqq 0.5)$.
Fig. 4(b) (Channel B) — Upper and lower bounds on $R(\beta)$ vs β $(0 \leqq \beta \leqq 0.2)$.

since the first coordinates of two code words in class S_i do not differ by more than $2A/\alpha$.

Further some class S_i has at least $M(n,d^2)/\alpha$ points, so that

$$M[n - 1, d^2 - (2A/\alpha)^2] \geqq (1/\alpha)\, M(n,d^2),$$

and the lemma follows.

Corollary: Let $a < (d\alpha/2A)^2$ be an integer. Then

$$M(n,d^2) \leqq \alpha^a M[n - a, d^2 - a(2A/\alpha)^2]. \tag{77}$$

Proof: Inequality (77) follows by repeated application of Lemma 4.2. Since by hypothesis $d^2 - a(2A/\alpha)^2 > 0$, the expression $M[n - a, d^2 - a(2A/\alpha)^2]$ is meaningful.

Derivation of the Bound

Let n_o be the greatest integer satisfying

$$n_o < \frac{1}{2A^2}\left[d^2 - (n - n_o)\left(\frac{2A}{\alpha}\right)^2\right], \tag{78}$$

where $\alpha \geqq 2$. Rearranging (78) we obtain

$$n_o < n\left[\frac{\alpha^2 \dfrac{d^2}{2A^2 n} - 2}{\alpha^2 - 2}\right] = n\left[\frac{2\hat{\beta}\alpha^2 - 2}{\alpha^2 - 2}\right]. \tag{79}$$

Let us also assume as an additional constraint on α that $\alpha^2 > 1/\hat{\beta}$, so that $[(2\hat{\beta}\alpha^2 - 2)/(\alpha^2 - 2)] > 0$, and for sufficiently large n, $n_o \geqq 1$. In fact for large n we may approximate n_o by

$$n_o = n\left[\frac{2\alpha^2\hat{\beta} - 2}{\alpha^2 - 2}\right]. \tag{80}$$

Now by choice of n_o (78), $0 < 2A^2 n_o < [d^2 - (n - n_o)(2A/\alpha)^2]$. Hence the Corollary to Lemma 4.2 applies with $a = n - n_o$ yielding

$$M(n,d^2) \leqq \alpha^{n-n_o} M[n_o, d^2 - (n - n_o)(2A/\alpha^2)]. \tag{81}$$

Also by (78), we may apply Lemma 4.1 to get

$$M\left(n_o, d^2 - (n - n_o)\left(\frac{2A}{\alpha^2}\right)^2\right) \leqq \frac{d^2 - (n - n_o)\left(\frac{2A}{\alpha}\right)^2}{d^2 - (n - n_o)\left(\frac{2A}{\alpha}\right)^2 - 2A^2 n_o} \triangleq Q(\alpha, d, n). \tag{82}$$

Thus from (81) and (82) we have

$$M(n,d^2) \leqq \alpha^{n-n_o} Q(\alpha,d,n). \tag{83}$$

Taking logarithms yields:

$$R(n,d^2) \leqq \ln \alpha \left[1 - \frac{n_o}{n}\right] + \frac{1}{n} \ln Q(\alpha,d,n). \tag{84}$$

We now let n and d become large while holding the ratio $\hat{\beta} = d^2/4A^2n$ fixed. It is easy to show that

$$\frac{1}{n} \ln Q(\alpha, 2A\sqrt{\hat{\beta}n},n) \xrightarrow{n} 0, \tag{85}$$

so that using (80) we obtain

$$R(\beta) \leqq \ln \alpha \left[1 - \frac{2\hat{\beta}\alpha^2 - 2}{\alpha^2 - 2}\right] = [\ln \alpha]\left[\frac{\alpha^2}{\alpha^2 - 2}\right][1 - 2\hat{\beta}], \tag{86}$$

where α is an arbitrary integer satisfying $\alpha \geqq 2$, and $\alpha^2 > 1/\hat{\beta}$. Using the choice of α indicated in Appendix E we obtain $R(\beta) \leqq R_U(\beta)$ where

$$R_U(\beta) = \begin{cases} 2(\ln 2)(1 - 2\hat{\beta}), & \frac{1}{2} \geqq \beta \geqq \frac{1}{4} \\ \dfrac{k^2}{k^2 - 2}(\ln k)(1 - 2\hat{\beta}) & \dfrac{1}{(k-1)^2} > \hat{\beta} \geqq \dfrac{1}{k^2} \\ & (k = 3, 4, \cdots). \end{cases} \tag{87}$$

$R_U(\beta)$ is plotted in Figs. 4(a) and 4(b). For small values of $\hat{\beta}$, $\alpha \approx 1/\sqrt{\hat{\beta}}$ so we obtain

$$R_U(\beta) = -\tfrac{1}{2} \ln (\hat{\beta}) + \epsilon(\hat{\beta}), \tag{88}$$

where $\epsilon(\hat{\beta}) \to 0$ as $\hat{\beta} \to 0$.

4.3 *Bounded Discrepancy Decoding Channel Capacity*

Suppose that for every n, an n-dimensional code is available with $d^2/4n = \beta$. Using BDD we have error probability

$$P_{eB} = \Pr[d(\mathbf{x},\mathbf{y}) \geqq d/2] = \Pr[d^2(\mathbf{x},\mathbf{y}) \geqq \beta n]. \tag{89}$$

Since $d^2(\mathbf{x},\mathbf{y}) = \sum_{k=1}^{n} z_k^2$, where the z_k are the (normally distributed) noise components we have

$$P_{eB} = \Pr\left[\sum_{k=1}^{n} z^2/n \geqq \beta\right]. \tag{90}$$

By the weak law of large numbers $\sum z_i^2/n$ tends in probability to $N(= E(z_i^2))$. Thus $\lim_{n\to\infty} P_{eB} = 0$ or 1 according as $\beta > N$ or $\beta < N$.

We define the *bounded discrepancy decoding channel capacity* de-

noted by C_B as the supremum of the rates for which it is possible (asymptotically in n) to obtain vanishingly small error probability using BDD. From the foregoing we see that $C_B = R(N)$. Making use of the bounds on $R(\beta)$ established above we have

$$R_L(N) \leqq C_B \leqq R_U(N), \tag{91}$$

where R_L and R_U are defined by (17) and (18) respectively. These bounds on C_B are plotted vs the "signal-to-noise ratio" A^2/N in Fig. 5.

For large values of the ratio A^2/N (91) becomes [using (74) and (88)]

$$\frac{1}{2}\ln\frac{1}{2\pi e}\frac{A^2}{N} + \epsilon_1\left(\frac{A^2}{N}\right) \leqq C_B \leqq \frac{1}{2}\ln\frac{A^2}{N} + \epsilon_2\left(\frac{A^2}{N}\right) \tag{92}$$

where ϵ_1, $\epsilon_2 \to 0$ as $A^2/N \to \infty$.

The channel capacity is the "maximum error free rate" using MDD (clearly $C \geqq C_B$). An exact expression for C is not known, however for large values of the ratio A^2/N Shannon[1] has shown that

$$C = \frac{1}{2}\ln\frac{2}{\pi e}\frac{A^2}{N} + \epsilon_3\left(\frac{A^2}{N}\right) \tag{93}$$

where $\epsilon_3 \to 0$ as $A^2/N \to \infty$. Combining the left inequality of (92) with (93) we obtain

$$C - \ln 2 + \epsilon(A^2/N) \leqq C_B \leqq C \tag{94}$$

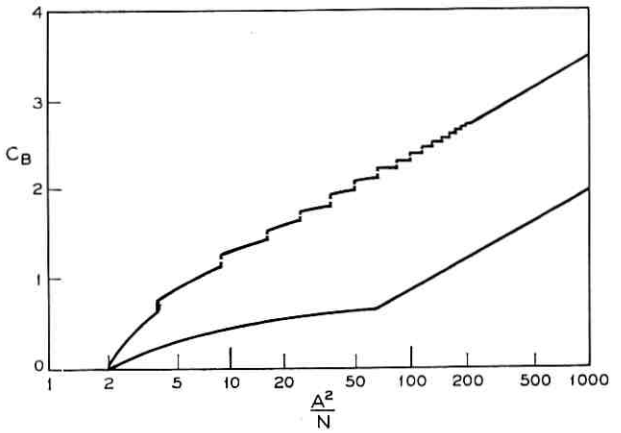


Fig. 5 (Channel B) — Upper and lower bounds on C_B vs A^2/N.

where $\epsilon \to 0$ as $A^2/N \to \infty$. Hence for large signal-to-noise ratio A^2/N, C_B differs from C by no more than a constant. Alternately $C_B/C \to 1$ as $A^2/N \to \infty$.

4.4 *Exponential Behavior of* P_{eB}

For a fixed $R < C_B$, denote by P_{eB}^* the smallest attainable value of P_{eB}, the error probability using BDD. It was shown above that $P_{eB}^* \xrightarrow{n} 0$. In this section we shall show that $P_{eB}^* = \exp[-nE_B(R) + o(n)]$ and obtain estimates of $E_B(R)$.

Given an n and R, denote by $\beta_n(R)$ the largest value of β attainable for a code of length n and with transmission rate R. With R held fixed, let $\beta(R) = \lim_{n\to\infty} \beta_n(R)$. We can estimate $\beta(R)$ in terms of R by

$$R_L(\beta(R)) \leqq R \leqq R_U(\beta(R)), \tag{95}$$

where R_L and R_U are given (17) and (18) respectively. Inequalities (95) result in upper and lower bounds on $\beta(R)$. Thus for any R there exists a code (for n sufficiently large) with minimum distance corresponding to $\beta(R)$ (i.e., $d = 2\sqrt{\beta n}$). With R fixed, this code minimizes P_{eB}. If code word $\mathbf{x}$ is transmitted and $\mathbf{y}$ is received, the error probability is [from (89)]

$$P_{eB}^* = \Pr[d^2(\mathbf{x},\mathbf{y}) \geqq \beta(R)n]. \tag{96}$$

This quantity depends only on the noise and not on $\mathbf{x}$. It is shown in Appendix F that

$$P_{eB}^* = \exp[-nE_B(R) + o(n)], \tag{97a}$$

where

$$E_B(R) = \frac{\hat{\beta}(R)}{2N}A^2 - \frac{1}{2}\ln\frac{A^2}{N}e\hat{\beta}(R) = \frac{\beta(R)}{2N} - \frac{1}{2}\ln\frac{e\beta(R)}{N}, \tag{97b}$$

where $\hat{\beta}(R) = \beta(R)/A^2$. The upper and lower bounds on $\beta(R)$ (95) yield corresponding bounds on $E_B(R)$. These bounds are plotted in Fig. 6 for the case $A^2/N = 10$.

V. CHANNEL C (GAUSSIAN CHANNEL WITH ENERGY CONSTRAINT)

5.1 *Lower Bound on* $M(n,\theta)$

The following bound is similar, though slightly sharper, than the

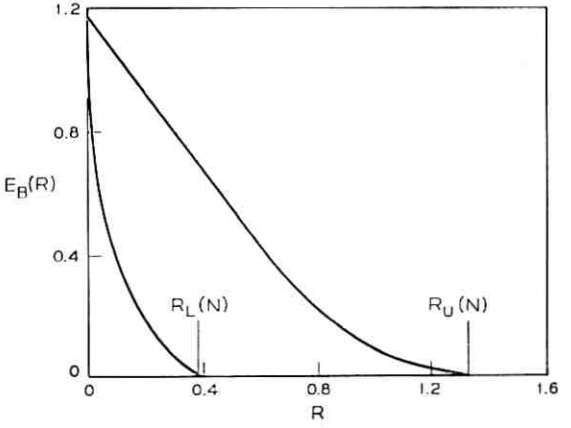


Fig. 6 (Channel B) — Upper and lower bounds on the error exponent $E_B(R)$ vs R (for $A^2/N = 10$).

lower bound on $M(n,\theta)$ obtained by Shannon.[11] The derivation used here is based on a similar argument in Blachman.[14]

Let

$$S_n(r) = \frac{n \cdot \pi^{n/2}}{\Gamma\left(\frac{n+2}{2}\right)} \cdot r^{n-1}$$

be the surface area of a sphere in Euclidean n-space of radius r, and let $A_n(r,\theta)$ be the area of the n-dimensional spherical cap cut from a sphere of radius r about the origin by a right circular cone of half angle θ with apex at the origin and axis the semi-infinite line connecting the origin and the point $(r,0,0,\cdots)$. It is not hard to show that

$$A_n(r,\theta) = \frac{(n-1)\pi^{(n-1)/2}}{\Gamma\left(\frac{n+1}{2}\right)} r^{n-1} \int_0^\theta \sin^{(n-2)} \varphi \, d\varphi.$$

Derivation of the Bound

For a given n and θ consider the maximum size n-dimensional code with minimum angle θ between code points. This code has $M(n,\theta)$ code words. About each code point $\mathbf{x}$, construct the spherical cap cut from the surface of the sphere of radius $\sqrt{nP}$ about the origin by the right circular cone with half angle θ and axis the semi-infinite line

joining the origin and $\mathbf{x}$. Thus the cap is the set of points $\mathbf{y}$ on the surface of the sphere such that the angle $a(\mathbf{x},\mathbf{y}) < \theta$. Now the set of all such caps (about each of the M code points) must cover the entire surface of the sphere. This is so since if $\mathbf{x}_o$ is a point on the surface of the sphere, and $\mathbf{x}_o$ is not on any cap then $a(\mathbf{x}_o, \mathbf{x}) \geqq \theta$ for all code words $\mathbf{x}$, so that $\mathbf{x}_o$ may be added to the code destroying the maximality. Since the area of each of the M caps is $A_n(\sqrt{nP},\theta)$, we have

$$M \cdot A_n(\sqrt{nP},\theta) \geqq S_n(\sqrt{nP})$$

or

$$M(n,\theta) \geqq \frac{S_n(\sqrt{nP})}{A_n(\sqrt{nP},\theta)} = \frac{n}{(n-1)} \sqrt{\pi} \frac{\Gamma\left(\frac{n+1}{2}\right)}{\Gamma\left(\frac{n+2}{2}\right)} \cdot \left[\int_0^\theta \sin^{(n-2)} \varphi \, d\varphi\right]^{-1}. \tag{98}$$

This result taken together with Rankin's upper bound[13] yields the following estimate of $M(n,\theta)$:

$$\frac{n}{n-1} \sqrt{\pi} \frac{\Gamma\left(\frac{n+1}{2}\right)}{\Gamma\left(\frac{n+2}{2}\right)} \cdot \left[\int_0^\theta \sin^{n-2} \varphi \, d\varphi\right]^{-1} \leqq M(n,\theta) \leqq \frac{\sqrt{\pi}\Gamma\left(\frac{n-1}{2}\right) \sin\psi \tan\psi}{2\Gamma\left(\frac{n}{2}\right) \int_0^\psi (\sin\varphi)^{n-2}(\cos\varphi - \cos\beta)\, d\varphi}, \tag{99}$$

where $\psi = \sin^{-1} \sqrt{2} \sin(\theta/2)$.

5.2 *Asymptotic Estimates of* $M(n,\theta)$

For a given n and θ, $M(n,\theta)$ is the number of points in a maximum size n-dimensional code with minimum angle between code points θ. Let the corresponding transmission rate be $R(n,\theta) = (1/n) \ln M(n,\theta)$. Now with θ held fixed, let n become large and let $R(\theta) = \lim_{n\to\infty} R(n,\theta)$.

We shall obtain upper and lower bounds on $R(\theta)$ from the behavior of (99) for large n.

Taking logarithms of both sides of inequality (99) yields

$$\frac{1}{n}\ln\frac{n}{n-1}\sqrt{\pi}+\frac{1}{n}\ln\frac{\Gamma\left(\frac{n+1}{2}\right)}{\Gamma\left(\frac{n+2}{2}\right)}-\frac{1}{n}\ln\int_0^\theta \sin^{n-2}\varphi\,d\varphi$$

$$\leqq R(n,\theta) \leqq \frac{1}{n}\ln\frac{\sin\psi\tan\psi}{2}\cdot\sqrt{\pi}+\frac{1}{n}\ln\left(\frac{\Gamma\left(\frac{n+1}{2}\right)}{\Gamma\left(\frac{n}{2}\right)}\right) \tag{100}$$

$$-\frac{1}{n}\ln\int_0^\psi \sin^{n-2}\varphi(\cos\varphi-\cos\psi)\,d\psi,$$

where $\psi=\sin^{-1}\sqrt{2}\sin(\theta/2)$.

It is shown in Appendix G that

$$\lim_{n\to\infty}\frac{1}{n}\ln\int_0^\theta \sin^{n-2}\varphi\,d\varphi=\ln\sin\theta, \tag{101a}$$

and that

$$\lim_{n\to\infty}\frac{1}{n}\ln\int_0^\psi \sin^{n-2}\varphi(\cos\varphi-\cos\psi)\,d\varphi=\ln\sin\psi, \tag{101b}$$

from which we obtain (by letting $n\to\infty$),

$$-\ln\sin\theta \leqq R(\theta) \leqq -\ln\sqrt{2}\sin(\theta/2). \tag{101}$$

The bounds on $R(\theta)$ are plotted in Fig. 7.

5.3 *Bounded Discrepancy Decoding Channel Capacity*

We now assume that a code with minimum angle θ is employed and a bounded discrepancy decoder is used. We may assume, without loss of generality, that the transmitted word is $\mathbf{x}=(\sqrt{nP},0,\cdots,0)$. The received word $\mathbf{y}=(\sqrt{nP}+z_1,z_2,\cdots,z_n)$ will be correctly decoded if and only if $a(\mathbf{x},\mathbf{y})<\theta/2$. Since

$$\cos a(\mathbf{x},\mathbf{y})=\frac{\sqrt{nP}\,(\sqrt{nP}+z_1)}{\sqrt{nP}\left((\sqrt{nP}+z_1)^2+\sum_{k>1}z_k^2\right)^{\frac{1}{2}}},$$

we have

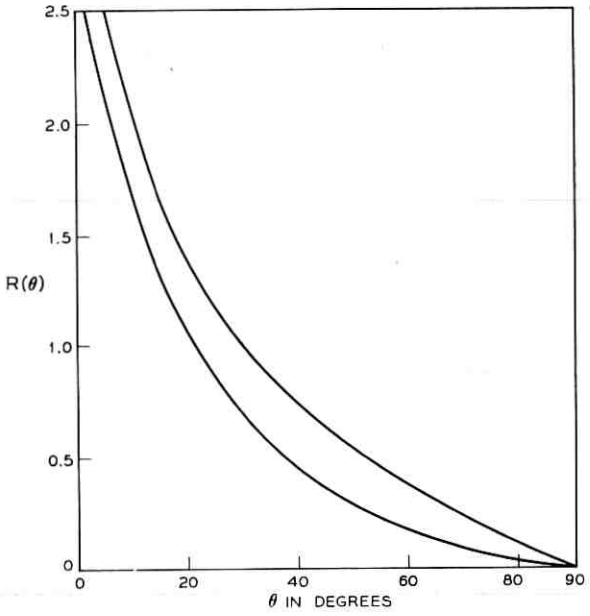


Fig. 7 (Channel C) — Upper and lower bounds on $R(\theta)$ vs θ (101).

$$\cot a = \frac{\sqrt{nP} + z_1}{\left(\sum_{k>1} z_k^2\right)^{\frac{1}{2}}} = \frac{\sqrt{P} + (z_1/\sqrt{n})}{\left(\frac{1}{n}\sum_{k>1} z_k^2\right)^{\frac{1}{2}}}.$$

Hence the probability of error is

$$P_{eB} = \Pr\left[\cot a \leqq \cot\frac{\theta}{2}\right] = \Pr\left[\frac{\sqrt{P} + z_1/\sqrt{n}}{\left(\frac{1}{n}\sum_{k>1} z_k^2\right)^{\frac{1}{2}}} \leqq \cot\frac{\theta}{2}\right]. \quad (102)$$

Now assume that for each n we use a code with minimum angle θ. We shall show that $P_{eB} \xrightarrow{n} 0$ or 1 according as $\cot(\theta/2) < \sqrt{P/N}$ or $\cot(\theta/2) > \sqrt{P/N}$: Recalling that z_k $(k = 1, \cdots, n)$ are independent normally distributed random variables with mean zero and variance N we obtain

$$\sqrt{P} + z_1/\sqrt{n} \xrightarrow{\text{Prob.}} \sqrt{P},$$

and

$$\frac{1}{n}\sum_{k>1} z^2 \xrightarrow{\text{Prob.}} N.$$

Thus the ratio

$$\frac{\sqrt{P} + z_1/\sqrt{n}}{\left(\frac{1}{n}\sum_{k>1} z_k^2\right)^{\frac{1}{2}}} \xrightarrow{\text{Prob.}} \sqrt{\frac{P}{N}}. \tag{103}$$

If $\cot(\theta/2) < \sqrt{P}/N$, then $\sqrt{P}/N - \cot(\theta/2) = \epsilon > 0$ so that from (102) and (103)

$$P_B(e) = \Pr\left[\sqrt{\frac{P}{N}} - \frac{\sqrt{P} + z_1/\sqrt{n}}{\left(\frac{1}{n}\sum_{k>1} z_k^2\right)^{\frac{1}{2}}} > \epsilon\right] \xrightarrow{n} 0.$$

Similarly if $\cot(\theta/2) > \sqrt{P}/N$, $P_B(e) \xrightarrow{n} 1$.

Hence we can obtain vanishingly small error probability by choosing $\theta > 2 \text{ arc cot } \sqrt{P}/N$ or $\theta > 2 \text{ arc sin } (1 + P/N)^{-\frac{1}{2}}$. The *bounded discrepancy decoding channel capacity* C_B is defined as the supremum of rates for which it is possible to achieve $P_{eB} \xrightarrow{n} 0$, or equivalently the largest rate for which $\theta > 2 \text{ arc sin } [1 + (P/N)]^{-\frac{1}{2}}$; i.e., $C_B = R\{2 \text{ arc sin } [1 + (P/N)]^{-\frac{1}{2}}\}$. Since the channel capacity is $C = \frac{1}{2} \ln [1 + (P/N)]$, we may write $[1 + (P/N)]^{-\frac{1}{2}} = e^{-C}$, hence $C_B = R(2 \text{ arc sin } e^{-C})$. We estimate C_B from inequality (101):

$$-\ln \sin (2 \sin^{-1} e^{-C}) \leqq C_B \leqq -\ln \sqrt{2} e^{-C}. \tag{104}$$

Using $\sin 2A = 2 \sin A \cos A$, the left member of (104) becomes $-\ln 2e^{-C} \cos \sin^{-1} e^{-C}$. Since $\cos \sin^{-1} e^{-C} = (1 - e^{-2C})^{\frac{1}{2}}$, inequality (104) becomes

$$C - \ln 2 - \tfrac{1}{2} \ln (1 - e^{-2C}) \leqq C_B \leqq C - \tfrac{1}{2} \ln 2. \tag{105}$$

Inequality (105) is plotted in Figs. 8(a) and 8(b). We see that $C_B = 0$ for $C \leqq \frac{1}{2} \ln 2$ or $P/N \leqq 1$, and $C_B/C \to 1$ as $P/N \to \infty$.

5.4 *Exponential Behavior of* P_{eB}

In this section we show that for a fixed rate $R < C_B$, the smallest attainable probability of error $P_{eB}^* = \exp [-nE_B(R) + o(n)]$, and obtain estimates of $E_B(R)$. Given an n and R, denote by $\theta_n(R)$ the largest minimum angle attainable for an n-dimensional code with transmission rate R. With R held fixed, let $\theta(R) = \lim_{n\to\infty} \theta_n(R)$. From inequality (101)

$$-\ln \sin \theta(R) \leqq R \leqq -\ln \sqrt{2} \sin [\theta(R)/2],$$

from which

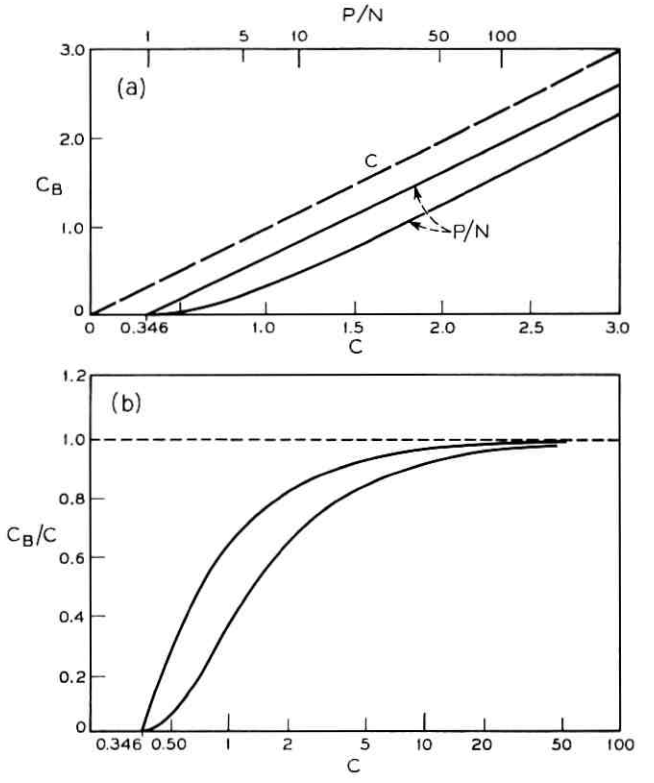


Fig. 8(a) (Channel C) — Upper and lower bounds on C_B vs C and P/N (solid lines).
Fig. 8(b) (Channel C) — Upper and lower bounds on C_B/C vs C.

$$\tfrac{1}{2} \sin^{-1} e^{-R} \leqq \theta(R)/2 \leqq \sin^{-1} (e^{-R}/\sqrt{2}). \tag{106}$$

Thus for every R there exists a code (for n sufficiently large) with minimum angle $\theta(R)$, where $\theta(R)$ is estimated by (106). For such a code P_{eB} is minimized. If code word $\mathbf{x}$ is transmitted and $\mathbf{y}$ received, the error probability is

$$P_{eB}{}^{*} = \Pr\,[a(\mathbf{x},\mathbf{y}) > \theta(R)/2]. \tag{107}$$

This quantity depends only on the noise (and not on $\mathbf{x}$). Shannon [Ref. 11, equation (4)] has obtained an expression for the asymptotic behavior of (107), which shows that

$$P_{eB}{}^{*} = \exp\,[-nE_B(R) + o(n)] \tag{108}$$

where

$$E_B(R) = \frac{P}{2N} - \frac{1}{2}\sqrt{\frac{P}{N}}\, G \cos\frac{\theta}{2} - \ln G \sin\frac{\theta}{2}$$

$$\theta = \theta(R) \tag{109}$$

$$G = \frac{1}{2}\left(\sqrt{\frac{P}{N}}\cos\frac{\theta}{2} + \sqrt{\frac{P}{N}\cos^2\frac{\theta}{2} + 4}\right).$$

The bounds on $\theta(R)$ in (106) yield corresponding bounds on $E_B(R)$. These bounds are plotted in Fig. 9.

VI. CHANNEL D (CONTINUOUS CHANNEL WITH AMPLITUDE CONSTRAINT):

As in the previous sections we begin by obtaining bounds on $M(n,\rho)$, the maximum number of points in an n-dimensional code with discrepancy ρ.

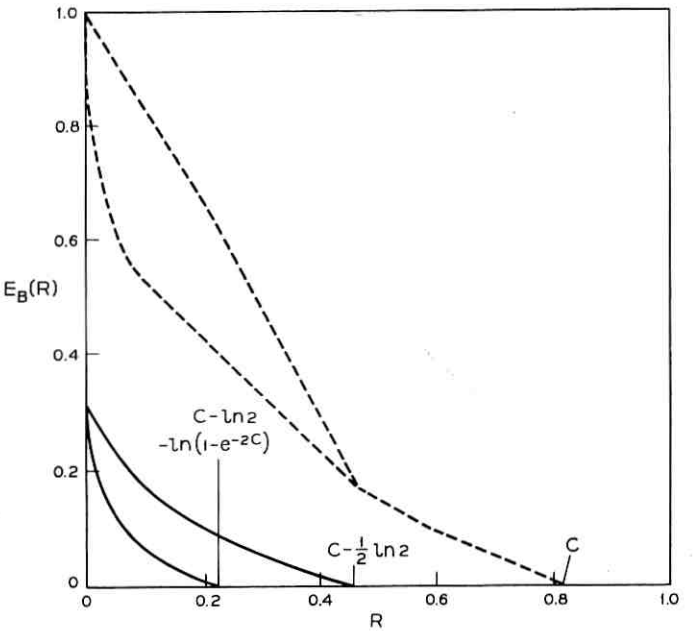


Fig. 9 (Channel C) — Upper and lower bounds on the exponent $E_B(R)$ vs R for $P/N = 4$. Upper and lower bounds on $E(R)$ are in dotted lines.

6.1 *Upper Bound on $R(\beta)$*

We have defined $S_n(\mathbf{x},\rho)$ as the region consisting of those vectors $\boldsymbol{\alpha} \in \mathcal{C}_n$ for which $d_o(\mathbf{x},\boldsymbol{\alpha}) < \rho$. Applying the Euclidean measure to $\mathcal{C}_n$ in the obvious way, we set $V_n(\mathbf{x},\rho)$ equal to the volume of $S_n(\mathbf{x},\rho)$. Now consider a maximum size n-dimensional code with discrepancy ρ consisting of $M = M(n,\rho)$ points $\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_M$. Since the regions $S_n(\mathbf{x}_i, \rho)$ about each of the M code points $\mathbf{x}_i$ are disjoint,

$$\sum_{i=1}^{M} V_n(\mathbf{x}_i, \rho) \leqq \text{volume of } \mathcal{C}_n = (2A)^n. \tag{110}$$

Since $V_n(\mathbf{x}_i, \rho)$ is independent of $\mathbf{x}_i$ (due to the homogeneity of $\mathcal{C}_n$ brought about by wrapping the interval onto the circumference of a circle) we set $V_n(\mathbf{x}_i, \rho) = V_n(\rho)$, and (110) yields

$$M(n,\rho) \leqq (2A)^n / V_n(\rho), \tag{111}$$

thus

$$R(n,\rho) = \frac{1}{n} \ln M(n,\rho) \leqq \frac{1}{n} \ln \frac{(2A)^n}{V_n(\rho)}. \tag{112}$$

If we set $\rho = \beta n$ and let $n \to \infty$ while β is held fixed we obtain

$$R(\beta) \leqq \lim \frac{1}{n} \ln \frac{(2A)^n}{V_n(\rho n)} = R_U(\beta). \tag{113}$$

It is shown in Appendix C that $R_U(\beta) = C_o(\beta)$ which establishes our upper bound.

6.2 *Lower Bound on $R(\beta)$*

Again let us consider a maximum size code with discrepancy ρ and $M = M(n,\rho)$ code words. About each of the code words $\mathbf{x}_i$ ($i = 1, 2, \cdots, M$) consider the region $S_n(\mathbf{x}_i, 2\eta\rho)$ where

$$\eta = \sup_{-A \leqq u_1, u_2 \leqq +A} \frac{r(u_1 \dot{+} u_2)}{r(u_1) + r(u_2)}. \tag{114}$$

We claim that the union of these regions $\bigcup_{i=1}^{M} S_n(\mathbf{x}_i, 2\eta\rho)$ contains $\mathcal{C}_n$. First let us observe that by definition of η,

$$r(u_1 \dot{+} u_2) \leqq \eta[r(u_1) + r(u_2)],$$

so that for $\mathbf{x},\mathbf{y},\mathbf{z} \in \mathcal{C}_n$,

$$d_o(\mathbf{x},\mathbf{y}) = \sum_{k=1}^{n} r(x_k \dot{-} y_k) \leqq \eta\Big[\sum_k r(x_k \dot{-} z_k) + \sum_k r(z_k \dot{-} y_k)\Big]$$
$$= \eta d_o(\mathbf{x},\mathbf{z}) + \eta d_o(\mathbf{z},\mathbf{y}). \tag{115}$$

Now suppose there existed a vector $\mathbf{x}_o \in \mathcal{C}_n$ such that $\mathbf{x}_o \notin \bigcup_{i=1}^{M} S_n(\mathbf{x}_i, 2\eta\rho)$. Then

$$d_o(\mathbf{x}_o, \mathbf{x}_i) \geqq 2\eta\rho \tag{116}$$

for each code word $\mathbf{x}_i$ $(i = 1, 2, \cdots, M)$. Let $\boldsymbol{\alpha} \in S_n(\mathbf{x}_i, \rho)$ for some code word $\mathbf{x}_i$ so that $d_o(\boldsymbol{\alpha},\mathbf{x}_i) < \rho$, hence from (115) and (116) we have

$$d_o(\mathbf{x}_o, \boldsymbol{\alpha}) \geqq (1/\eta) d_o(\mathbf{x}_o, \mathbf{x}_i) - d_o(\mathbf{x}_i, \boldsymbol{\alpha}) > (1/\eta)(2\eta\rho) - \rho = \rho. \tag{117}$$

We conclude from (117) that $\boldsymbol{\alpha} \notin S_n(\mathbf{x}_o, \rho)$, so that $S_n(\mathbf{x}_o, \rho) \cap S_n(\mathbf{x}_i, \rho)$ is empty for all code words $\mathbf{x}_i$, and $\mathbf{x}_o$ may be added to the code destroying the maximality. Thus we conclude that

$$\mathcal{C}_n \subseteq \bigcup_{i=1}^{M} S_n(\mathbf{x}_i, 2\eta\rho). \tag{118}$$

As in the previous section, let $V_n(2\eta\rho)$ be the volume of $S_n(\mathbf{x}_i, 2\eta\rho)$. From (118) we have

$$\text{volume of } \mathcal{C}_n = (2A)^n \leqq M \cdot V_n(2\eta\rho),$$

or

$$M(n,\rho) \geqq \frac{(2A)^n}{V_n(2\eta\rho)}. \tag{119}$$

Again as in the previous section,

$$R(\beta) \geqq \lim \frac{1}{n} \ln \frac{(2A)^n}{V_n(2\eta\beta n)} = R_L(\beta). \tag{120}$$

It is shown in Appendix C that $R_L(\beta) = C_o(2\eta\beta)$, establishing our lower bound.

6.3 *Bounded Discrepancy Decoding Channel Capacity*

Suppose that for every n, an n-dimensional code is available with

discrepancy $\rho = \beta n$ (β fixed). Using bounded discrepancy decoding we have error probability

$$P_{eB} = \Pr [d_o(\mathbf{x},\mathbf{y}) \geqq \rho] = \Pr [d_o(\mathbf{x},\mathbf{y}) \geqq \beta n], \tag{121}$$

where $\mathbf{x}$ is the transmitted word and $\mathbf{y}$ is the received vector. Since $d_o(\mathbf{x},\mathbf{y}) = \sum_{k=1}^{n} r(z_k)$, where the z_k are the statistically independent noise components, we have

$$P_{eB} = \Pr\left[\sum_{k=1}^{n} r(z_k)/n > \beta\right]. \tag{122}$$

By the weak law of large numbers, $\sum_{k=1}^{n} r(z_k)/n$ tends in probability to $N(= E(r(z_k)))$. Thus $\lim_{n\to\infty} P_{eB} = 0$ or 1 according as $\beta < N$ or $\beta > N$.

We have defined the *bounded discrepancy decoding channel capacity* denoted by C_B as the supremum of the rates for which it is possible (asymptotically in n) to obtain vanishingly small error probability using bounded discrepancy decoding. From the foregoing we see that $C_B = R(N)$. Making use of the bounds on $R(\beta)$ established above we have

$$C_o(2\eta N) \leqq C_B \leqq C_o(N) = C, \tag{123}$$

where C is the channel capacity (the supremum of those rates for which it is possible (asymptotically in n) to obtain vanishing small error probability using (optimum) minimum discrepancy decoding). The error exponent $E_B(R)$ could be estimated exactly as for channel B in Section IV.

Thus it is an open question whether C_B is *strictly* less than the channel capacity. In the special case of the quadratic discrepancy where $r(u) = u^2$, i.e., the case where $p(u) = K_o \exp(-\lambda u^2)$, it is possible to show that $C_B < C$. This is done in the following section.

6.4 *The Quadratic Discrepancy*

We now consider the case of the quadratic discrepancy where $r(u) = u^2$, which corresponds to a noise probability density function $p(u) = K_o \exp(-\lambda u^2)$, and a discrepancy function

$$d_o(\mathbf{x},\mathbf{y}) = \sum_{k=1}^{n} (x_k \dot{-} y_k)^2.$$

Note that the subtraction $x_k \dot{-} y_k$ is performed modulo $2A$ with the

difference reduced into the interval $[-A,A]$, but the squaring and summing operations are ordinary arithmetic.

Let us first observe that

$$\frac{r(u_1 \dot{+} u_2)}{r(u_1) + r(u_2)} \leqq \frac{(u_1 + u_2)^2}{u_1^2 + u_2^2} = 1 + \frac{u_1 u_2}{\left(\frac{u_1^2 + u_2^2}{2}\right)}.$$

Since for any two numbers u_1^2 and u_2^2, the algebraic mean $(u_1^2 + u_2^2)/2$ is not less than the geometric mean $u_1 u_2$,

$$\frac{r(u_1 \dot{+} u_2)}{r(u_1) + r(u_2)} \leqq 1 + 1 = 2.$$

Thus, since this value is achieved when $u_1 = u_2 \leqq A/2$, $\eta = 2$. The lower bound on $R(\beta)$ (34) is therefore

$$R(\beta) \geqq C_o(4\beta). \tag{124}$$

6.4.1 *Upper Bound on $R(\beta)$*

Now we establish a new upper bound on $R(\beta)$ for this special case. First we need the following

Lemma: Let $\mathbf{x}_\nu = (x_{\nu 1}, x_{\nu 2}, \cdots, x_{\nu n})$, $\nu = 1, 2, \cdots, m$, *be any m points selected from a code with discrepancy* $\rho = \beta n$. *Let* $\mathbf{y}$ *be any n-vector and let* $d_\nu = d_o(\mathbf{x}_\nu, \mathbf{y})$, $\nu = 1, 2, \cdots, m$. *Then*

$$\sum_{\nu=1}^{m} d_\nu \geqq 2(m-1)\rho.$$

Proof: First we show that for $1 \leqq \mu < \nu \leqq m$ that

$$d_o(\mathbf{x}_\nu, \mathbf{x}_\mu) \geqq 4\rho. \tag{125}$$

To show this consider the vector $\mathbf{z} \in \mathcal{C}_n$:

$$\mathbf{z} = \left(\frac{x_{\nu 1} \dot{+} x_{\mu 1}}{2}, \frac{x_{\nu 2} \dot{+} x_{\mu 2}}{2}, \cdots, \frac{x_{\nu n} \dot{+} x_{\mu n}}{2}\right).$$

The addition $x_{\nu k} \dot{+} x_{\mu k}$ is, as always, modulo $2A$. Clearly

$$d_o(\mathbf{x}_\nu, \mathbf{z}) = d_o(\mathbf{x}_\mu, \mathbf{z}) = \sum_{k=1}^{n} \frac{(x_{\nu k} \dot{-} x_{\mu k})^2}{4} = \frac{d_o(\mathbf{x}_\nu, \mathbf{x}_\mu)}{4}. \tag{126}$$

Since the regions $S_n(\mathbf{x}_\nu, \rho)$ and $S_n(\mathbf{x}_\mu, \rho)$ are disjoint, $d_o(\mathbf{x}_\nu, \mathbf{z})$, $d_o(\mathbf{x}_\mu, \mathbf{z}) \geqq \rho$. Thus (126) yields $d_o(\mathbf{x}_\nu, \mathbf{x}_\mu) \geqq 4\rho$. We now continue with the proof of the lemma.

Without loss of generality take $\mathbf{y} = 0$ so that $d_\nu = \sum_{k=1}^{n} x_{\nu k}^2$. Since $d_o(\mathbf{x}_\nu, \mathbf{x}_\mu) \geqq 4\rho$ $(\mu < \nu)$,

$$\begin{aligned}\binom{m}{2} 4\rho &\leqq \sum_{1 \leqq \mu < \nu \leqq m} d_o(\mathbf{x}_\mu, \mathbf{x}_\nu) = \sum_{k=1} \sum_{\mu<\nu} (x_{\mu k} - x_{\nu k})^2 \\ &\leqq \sum_k \sum_{\mu<\nu} (x_{\mu k} - x_{\nu k})^2 \\ &= m \sum_\nu \sum_k x_{\nu k}^2 - \sum_k \Big(\sum_\nu x_{\nu k}\Big)^2 \leqq m \sum_\nu d_\nu .\end{aligned} \tag{127}$$

The lemma follows on dividing through by m. We now obtain the upper bound on $R(\beta)$.

Consider again a maximum size n-dimensional code with discrepancy ρ and with $M = M(n,\rho)$ code words $\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_M$. Consider the regions $S_n(\mathbf{x}_\nu, 2\rho)$ about each of the code words $\mathbf{x}_\nu$ $(\nu = 1, 2, \cdots, M)$. These regions are not necessarily disjoint. At each point $\mathbf{y}$ in $S_n(\mathbf{x}_\nu, 2\rho)$ define a density $\sigma(d)$:

$$\sigma(d) = 2\rho - d, \tag{128}$$

where d is the discrepancy $d_o(\mathbf{x}_\nu, \mathbf{y})$. The mass of each region is

$$\mu = \int_{d<2\rho} \sigma(d)\, dV. \tag{129}$$

If a vector $\mathbf{y} \in \mathcal{C}_n$ belongs simultaneously to the regions about the m code points $\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_m$, we assign to $\mathbf{y}$ a density equal to the sum of the densities contributed by each region; i.e.,

$$\sigma_{\mathbf{y}} = \sum_{\nu=1}^{m} \sigma(d_\nu) = 2m\rho - \sum_{\nu=1}^{m} d_\nu , \tag{130}$$

where $d_\nu = d(\mathbf{x}_\nu, \mathbf{y})$. Thus we have

$$\text{mass of } \mathcal{C}_n = \int_{\mathcal{C}_n} \sigma_{\mathbf{y}} dV = M(n,\rho) \cdot \mu. \tag{131}$$

We will bound $M(n,\rho)$ by finding an upper bound on the mass of $\mathcal{C}_n$.

By applying the above lemma to (130) we obtain

$$\sigma_{\mathbf{y}} \leqq 2m\rho - 2(m-1)\rho = 2\rho. \tag{132}$$

Thus

$$\text{mass of } \mathcal{C}_n \leqq (2\rho)\ (\text{volume of } \mathcal{C}_n) = 2\rho(2A)^n. \tag{133}$$

Applying (133) to (131) yields

$$M(n,\rho) \leqq \frac{2\rho(2A)^n}{\mu}. \tag{134}$$

Now,

$$\mu = \int_{d<2\rho} (2\rho - d)\,dV \geqq \int_{d<2\rho-1} dV = V_n(2\rho - 1) \tag{135}$$

where $V_n(2\rho - 1)$ is the volume of the region $S_n(\mathbf{x},2\rho - 1)$ (which is independent of $\mathbf{x}$). Applying (135) to (134) yields

$$M(n,\beta n) \leqq \frac{2\beta n(2A)^n}{V_n(2\beta n - 1)} \tag{136}$$

where $\rho = \beta n$. Applying the result of Appendix C to (136) yields

$$R(\beta) = \lim_{n\to\infty} (1/n) \ln M(n,\beta n) \leqq C_o(2\beta). \tag{137}$$

This is our upper bound.

6.4.2 *Refinements of Bounds for Large β/A^2*

The upper and lower bounds on $R(\beta)$ obtained above are plotted vs. β/A^2 in Fig. 10. It can be seen that these bounds diverge for large

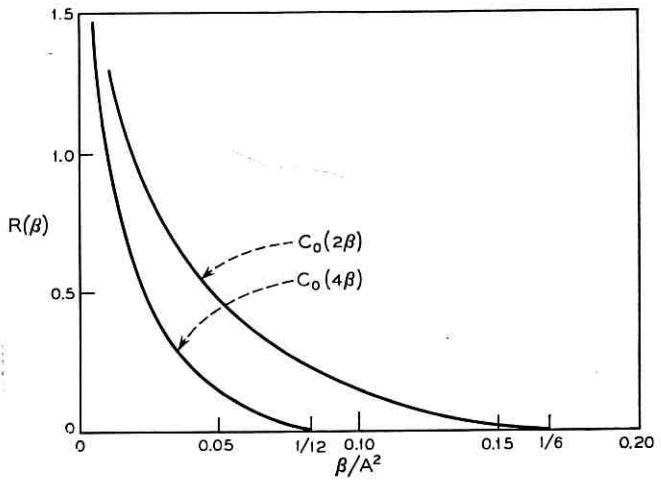


FIG. 10 (Channel D) — Upper and lower bounds on $R(\beta)$ vs β/A^2 for quadratic discrepancy.

values of β/A^2. We will now obtain new upper and lower bounds on $R(\beta)$ which in fact converge at $\beta/A^2 = \frac{1}{8}$.

6.4.2.1 *Upper Bound*

A new upper bound on $R(\beta)$ will be obtained which will tell us that $R(\beta) = 0, \beta/A^2 > \frac{1}{8}$. First we need the following:

Lemma: Let $a_1, a_2, \cdots, a_m$ *be a set of real numbers such that* $-A \leqq a_j \leqq +A, j = 1, 2, \cdots, m$. *Then*

$$\sum_{1 \leqq i < j \leqq m} (a_i \dot{-} a_j)^2 \leqq A^2 m^2/4.$$

Note that, as usual, the difference $(a_i \dot{-} a_j)$ is performed modulo $2A$ with the result reduced into the interval $[-A,+A]$, and the squaring and summing operations are ordinary arithmetic.

Proof: Let us wrap the interval $[-A,+A]$ onto the circumference of a circle of radius A/π (so that the circumference is $2A$). Denote by

$$d_c(a_i, a_j) = |(a_i \dot{-} a_j)|,$$

the circumferential distance between a_i and a_j, and by $d_E(a_i, a_j)$ the Euclidean distance between a_i and a_j (see Fig. 11). It is easy to see that

$$d_E(a_i, a_j) = 2\frac{A}{\pi} \sin \frac{1}{2} \frac{d_c(a_i, a_j)}{(A/\pi)}. \tag{138}$$

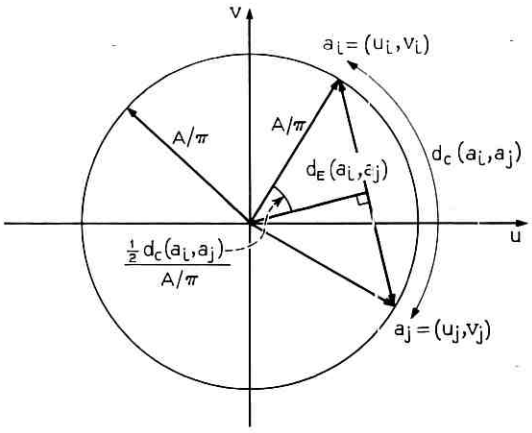


Fig. 11 — Diagram illustrating proof of lemma.

Since, for $0 \leqq x \leqq \pi/2$, $\sin x \geqq (2/\pi)x$, (138) yields

$$d_E(a_i, a_j) \geqq (2/\pi)\, d_c(a_i, a_j). \tag{139}$$

Now taking the origin to be the center of the circle, we may assign Cartesian coordinates (u_j, v_j) to the point a_j, where

$$u_j^2 + v_j^2 = A^2/\pi^2. \tag{140}$$

Thus from (139) we have

$$\sum_{i<j} (a_i \dot{-} a_j)^2 = \sum_{i<j} d_c^2(a_j, a_j) \leqq \frac{\pi^2}{4} \sum_{i<j} d_E^2(a_i, a_j). \tag{141}$$

Since $d_E^2(a_i, a_j) = (u_i - u_j)^2 + (v_i - v_j)^2$, we have

$$\begin{aligned} \sum_{i<j} (a_i \dot{-} a_j)^2 &\leqq \frac{\pi^2}{4} \sum_{i<j} \{(u_i - u_j)^2 + (v_i - v_j)^2\} \\ &= \frac{\pi^2}{4} \left\{ \sum_{j=1}^{m} m(u_j^2 + v_j^2) - \left(\sum_j u_j\right)^2 - \left(\sum_j v_j\right)^2 \right\} \\ &\leqq \frac{\pi^2}{4} \left\{ \sum_{j=1}^{m} m\left(\frac{A^2}{\pi^2}\right) \right\} \\ &= \frac{A^2 m^2}{4}. \end{aligned} \tag{142}$$

Hence the lemma.

Derivation of the Bound

Suppose we have a maximum size code with discrepancy ρ and $M = M(n,\rho)$ code words $\mathbf{x}_\nu = (x_{\nu 1}, x_{\nu 2}, \cdots, x_{\nu n})$, $\nu = 1, 2, \cdots, M$. We have shown [inequality (125)] that $d(x_\mu, x_\nu) \geqq 4\rho$ $(\mu \neq \nu)$. Thus, making use of the above lemma, we have

$$\begin{aligned} \binom{M}{2} 4\rho &\leqq \sum_{1 \leqq \mu < \nu \leqq M} d_c(\mathbf{x}_\nu, \mathbf{x}_\mu) = \sum_{k=1}^{n} \sum_{\mu<\nu} (x_{\mu k} \dot{-} x_{\nu k})^2 \\ &\leqq \sum_{k=1}^{n} \frac{A^2 M^2}{4} = \frac{A^2 M^2 n}{4}, \end{aligned}$$

so that for $\beta = \rho/n > A^2/8$,

$$M = M(n,\rho) \leqq \frac{8\rho}{8\rho - A^2 n} = \frac{\beta}{\beta - \frac{A^2}{8}}. \tag{143}$$

Hence,

$$R(n,\rho) = \frac{1}{n} \ln M(n,\rho) \leqq \frac{1}{n} \ln \frac{\beta}{\beta - \frac{A^2}{8}}. \tag{144}$$

Letting $n \to \infty$ with β held fixed we obtain for $\beta/A^2 > \frac{1}{8}$,

$$R(\beta) = \lim_{n \to \infty} R(n,\beta n) = 0. \tag{145}$$

In a manner similar to that used in Section IV we can use (143) to obtain the following bound on $R(\beta)$ valid for $\beta/A^2 < \frac{1}{8}$:

$$R(\beta) \leqq 9 \ (\ln 3) \ [1 - (8\beta/A^2)]. \tag{146}$$

As is evident from Fig. 12, inequality (146) does not yield much improvement in our upper bound, hence the derivation is omitted.

6.4.2.2 *Lower Bound*

A new lower bound on $R(\beta)$ will now be obtained. This bound is always sharper than the previously obtained bound $R(\beta) \geqq C_o(4N)$, however the best improvement is for large β/A^2.

Suppose that we require that x_k be one of the following m points on the interval $[-A,+A]$, where m is an even integer:

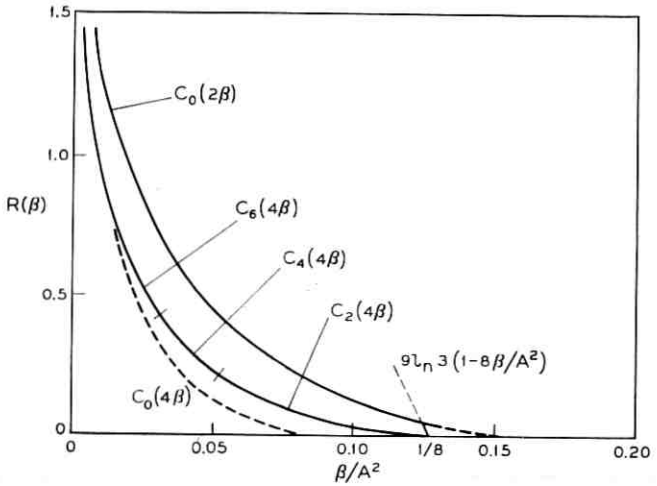


Fig. 12 (Channel D) — Refined upper and lower bounds on $R(\beta)$ vs β/A^2 for quadratic discrepancy.

$$0, \pm\frac{2A}{m}, \pm 2\left(\frac{2A}{m}\right), \pm 3\left(\frac{2A}{m}\right), \cdots, \pm\left(\frac{m-2}{2}\right)\left(\frac{2A}{m}\right), A. \quad (147)$$

Such a code certainly satisfies the requirements set forth in Section II. In the exactly the same manner that the previously derived lower bound $R(\beta) \geqq C_o(4\beta)$ was obtained, we can show that

$$R(\beta) \geqq C_m(4\beta)$$

where

$$C_m(\xi) = \ln AK_m(\xi) - \xi\lambda_m(\xi), \quad (148)$$

where $\lambda_m(\xi)$ is defined by

$$\sum_k u_k^2 e^{-\lambda u_k^2} = \xi \sum_k e^{-\lambda u_k^2}, \quad (149a)$$

and $K_m(\xi)$ is

$$K_m(\xi) = [\sum_k e^{-\lambda(\xi) u_k^2}]^{-1}, \quad (149b)$$

where the u_k are the m points of (147).

Since no value of m yields a uniformly strongest bound we write

$$R(\beta) \geqq \max_{m \text{ even}} C_m(4\beta). \quad (150)$$

This new bound is plotted in Fig. 12. Let us observe that the lower bound $R(\beta) \geqq C_2(4\beta)$, and the upper bound $R(\beta) \leqq 9 \ln 3[1 - (8\beta/A^2)]$ agree when $\beta/A^2 = \frac{1}{8}$. Thus $C_B = 0$ for $\beta/A^2 \geqq \frac{1}{8}$.

6.4.3 *Estimation of C_B*

We now obtain an estimate of C_B for the case of a quadratic discrepancy function. As discussed above $C_B = R(N)$. The bounds on $R(\beta)$ of (137), (146), (150) yield

$$\max_{m \text{ even}} C_m(4N) \leqq C_B \leqq \min \begin{cases} C_o(2N), \\ 9 \ln 3\left(1 - \dfrac{8N}{A^2}\right). \end{cases} \quad (151)$$

Since the channel capacity $C = C_o(N) > C_o(2N)$, the first upper bound of (151) implies that C_B is strictly less than the channel capacity C. For large values of the "signal-to-noise" ratio A^2/N, the left side of (151) may be approximated by $C_o(4N)$. We can make use of the asymptotic form of $C_o(\xi)$ obtained in Appendix D:

$$C_o(\xi) = \tfrac{1}{2} \ln (2A^2/\pi e\xi) + \epsilon(\xi), \quad (152)$$

where $\epsilon(\xi) \to 0$ as $\xi \to 0$. Applying (152) to (151), we obtain

$$\tfrac{1}{2} \ln \frac{A^2}{2\pi eN} + \epsilon_1\left(\frac{A^2}{N}\right) \leqq C_B \leqq \tfrac{1}{2} \ln \frac{A^2}{\pi eN} + \epsilon_2\left(\frac{A^2}{N}\right), \tag{153}$$

where ϵ_1, $\epsilon_2 \to 0$ as $A^2/N \to \infty$. Further since the channel capacity C is (for large A^2/N)

$$C = C_o(N) = \tfrac{1}{2} \ln \frac{2}{\pi e} \frac{A^2}{N} + \epsilon_3\left(\frac{A^2}{N}\right), \tag{154}$$

where $\epsilon_3 \to 0$ as $A^2/N \to \infty$, (153) may be rewritten as

$$C - \ln 2 + \epsilon_5(A^2/N) \leqq C_B \leqq C - \tfrac{1}{2} \ln 2 + \epsilon_6(A^2/N), \tag{155}$$

where ϵ_5, $\epsilon_6 \to 0$ as $A^2/N \to \infty$. Thus for large values of A^2/N (and hence C), the bounded discrepancy channel capacity C_B differs by no more than a constant (ln 2) from the channel capacity C. Thus the ratio $C_B/C \to 1$ as $A^2/N \to \infty$.

Let us remark at this point that the channel capacity of the Gaussian channel with amplitude constraint has been shown by Shannon[1] to be approximately $C_o(N)$ (for large A^2/N), which is the same as the capacity of the present channel. This fact lends plausibility to the claim that the present channel is an approximation to the Gaussian amplitude constrained channel for large values of A^2/N.

APPENDIX A

In this appendix we show that for any function $r(u)$ for which $r(u) \to 0$ as $u \to 0$, and for any ξ satisfying

$$0 < \xi \leqq \frac{1}{A} \int_0^A r(u)\, du, \tag{156}$$

there exists a unique $\lambda(\xi)$ which satisfies

$$\int_0^A r(u) e^{-\lambda(\xi) r(u)} du = \xi \int_0^A e^{-\lambda(\xi) r(u)} du. \tag{157}$$

For channel B we are interested in the case $r(u) = u^2$, however for channel D we need this proposition for arbitrary $r(u)$. If we define the function $\xi(\lambda)$ by

$$\xi(\lambda) = \frac{\int_0^A r(u) e^{-\lambda r(u)} du}{\int_0^A e^{-\lambda r(u)} du}, \qquad 0 \leqq \lambda < \infty, \tag{158}$$

it will suffice to show that

$$\text{(a) } \xi(\lambda) \text{ is strictly monotone decreasing,}$$

$$\text{(b) } \xi(0) = \frac{1}{A}\int_0^A r(u)\,du,$$

$$\text{(c) } \lim_{\lambda\to\infty} \xi(\lambda) = 0.$$

If (a), (b) and (c) are true, $\xi(\lambda)$ is a one-to-one mapping of the half line $[0,\infty)$ onto the interval

$$\left(0, \frac{1}{A}\int_0^A r(u)\,du\right].$$

(a) To show that $\xi(\lambda)$ is monotone decreasing, consider

$$\frac{d\xi(\lambda)}{d\lambda} = \frac{-\left(\int_0^A e^{-\lambda r}du\right)\left(\int_0^A r^2e^{-\lambda r}du\right) + \left(\int_0^A re^{-\lambda r}du\right)^2}{\left(\int_0^A e^{-\lambda r}du\right)^2}, \tag{159}$$

by the Schwarz inequality,

$$\left(\int_0^A re^{-\lambda r}du\right)^2 < \left(\int_0^A r^2e^{-\lambda r}du\right)\left(\int_0^A e^{-\lambda r}du\right), \tag{160}$$

(the strict inequality holding). Thus $d\xi(\lambda)/d\lambda < 0$ and (a) is established.

$$\text{(b) } \xi(0) = \int_0^A r(u)\,du \Big/ \int_0^A du = \frac{1}{A}\int_0^A r(u)\,du.$$

(c) (due to H. O. Pollak†) since $\xi(\lambda)$ is monotone decreasing and positive for $\lambda < \infty$, we know that $\lim_{\lambda\to\infty} \xi(\lambda) = \beta \geqq 0$. If $\beta = 0$ (c) is established. Thus we assume the contrary, i.e., $\beta > 0$. Since $\xi(\lambda)$ is monotone decreasing we have $\xi(\lambda) \geqq \beta$, all $\lambda < \infty$. Thus for any Λ,

$$\int_0^\Lambda \xi(\lambda)\,d\lambda \geqq \beta\Lambda. \tag{161}$$

Now let us observe that $\xi(\lambda)$ may be written

$$\xi(\lambda) = -\frac{d}{d\lambda}\left(\ln \int_0^A e^{-\lambda r(u)}du\right). \tag{162}$$

† An alternate proof was given to the author by L. A. Shepp.

Substituting (162) into (161) we obtain

$$\int_0^\Lambda \xi(\lambda)d\lambda = -\ln \int_0^\Lambda e^{-\Lambda r(u)}du + \ln A \geqq \beta\Lambda. \tag{163}$$

Or,

$$\frac{1}{A}\int_0^A e^{-\Lambda r(u)}du \leqq e^{-\beta\Lambda}. \tag{164}$$

Dividing through by $e^{-\beta\Lambda}$ we have

$$\frac{1}{A}\int_0^A e^{+\Lambda(\beta - r(u))}du \leqq 1. \tag{165}$$

Now since $r(u) \to 0$ as $u \to 0$, choose δ sufficiently small so that $r(u) < \beta/2$ whenever $0 \leqq u \leqq \delta$. Equation (165) now becomes

$$\begin{aligned} 1 &\geqq \frac{1}{A}\int_0^A e^{+\Lambda(\beta - r)}du \\ &\geqq \frac{1}{A}\int_0^\delta e^{+\Lambda(\beta - r(u))}du \\ &\geqq \frac{1}{A}e^{\Lambda\beta/2}\int_0^\delta du = \frac{\delta}{A}e^{\Lambda\beta/2}. \end{aligned} \tag{166}$$

Now (166) holds for all $\Lambda < \infty$. Thus we need only choose Λ large enough so that

$$\frac{\delta}{A}e^{\Lambda\beta/2} > 1$$

to deduce a contradiction. Thus (c) follows.

APPENDIX B

Proof That η is Finite

Define the function

$$g(u_1, u_2) = \frac{r(u_1 \dot{+} u_2)}{r(u_1) + r(u_2)}, \tag{167}$$

where $-A \leqq u_1, u_2 \leqq +A$ and $(u_1, u_2) \neq (0,0)$, and the function $r(u)$ is given by (27). Note that the addition $u_1 \dot{+} u_2$ is performed modulo $2A$. We must show that $\eta = \sup g(u_1, u_2)$ is finite, or that $g(u_1, u_2)$ is bounded.

By assumption (8d), $r(u)$ is continuous, and by assumptions (8c) and (8d) $r(u) > 0$ when $u \neq 0$. Thus $g(u_1, u_2)$ is continuous over its domain. If g is unbounded, let $(u_1^{(n)}, u_2^{(n)})_{n=1}^{\infty}$ be a sequence such that $g(u_1^{(n)}, u_2^{(n)})_n \underset{n}{\rightarrow} \infty$. Then it is easy to see that $(u_1^{(n)}, u_2^{(n)}) \rightarrow (0,0)$. Thus to show that η is finite we need only show that g is bounded in the neighborhood of the origin.

Now let $R_1 = \{(u_1, u_2): u_1, u_2 \geqq 0\}$. We shall show that

$$\eta = \sup_{-A \leqq u_1, u_2 \leqq +A} g(u_1, u_2) = \sup_{(u_1, u_2) \epsilon R_1} g(u_1, u_2). \tag{168}$$

If $(u_1, u_2) \notin R_1$, then either u_1 and u_2 are both negative or u_1 and u_2 have opposite signs. In the first case $g(u_1, u_2) = g(-u_1, -u_2)$, where $(-u_1, -u_2) \in R_1$. In the second case say $|u_1| \geqq |u_2|$, then by assumption (8d) and (8c) $r(u_1 \dot{+} u_2) \leqq r(u_1)$. Thus $g(u_1, u_2) \leqq r(u_1)/[r(u_1) + r(u_2)] \leqq 1 = g(A,0)$ where $(A,0) \in R_1$. Thus we need show only that g is bounded in the neighborhood of the origin where $u_1, u_2 \geqq 0$.

With u_1 and u_2 sufficiently small, the addition $u_1 \dot{+} u_2 = u_1 + u_2$. Also by assumption (8e), we may write

$$r(u) = au^{\alpha}(1 + \epsilon(u)), \tag{169}$$

where $a > 0$, $\alpha > 0$, and $\epsilon(u) \rightarrow 0$ as $u \rightarrow 0$. Thus

$$\begin{aligned} g(u_1, u_2) &= \frac{a(u_1 + u_2)^{\alpha}(1 + \epsilon(u_1 + u_2))}{a(u_1)^{\alpha}(1 + \epsilon(u_1)) + au_2^{\alpha}(1 + \epsilon(u_2))} \\ &= \frac{(u_1 + u_2)^{\alpha}}{u_1^{\alpha} + u_2^{\alpha}} \left[\frac{1 + \epsilon(u_1 + u_2)}{1 + \frac{u_1^{\alpha}}{u_1^{\alpha} + u_2^{\alpha}} \epsilon(u_1) + \frac{u_2^{\alpha}}{u_1^{\alpha} + u_2^{\alpha}} \epsilon(u_2)} \right]. \end{aligned} \tag{170}$$

Now,

$$0 \leqq \frac{u_1^{\alpha}}{u_1^{\alpha} + u_2^{\alpha}}, \qquad \frac{u_2^{\alpha}}{u_1^{\alpha} + u_2^{\alpha}} \leqq 1, \tag{171}$$

so that

$$g(u_1, u_2) = \frac{(u_1 + u_2)^{\alpha}}{u_1^{\alpha} + u_2^{\alpha}} [1 + \epsilon_1(u_1, u_2)], \tag{172}$$

where $\epsilon_1(u_1, u_2) \rightarrow 0$ as $u_1, u_2 \rightarrow 0$. Thus, since

$$0 \leqq \frac{(u_1 + u_2)^\alpha}{u_1^\alpha + u_2^\alpha} = \frac{\left(1 + \frac{u_2}{u_1}\right)^\alpha}{1 + (u_2/u_1)^\alpha} \leqq 2^{\alpha-1}, \tag{173}$$

we conclude that g is bounded in the neighborhood of the origin, and therefore that η is finite.

Let us remark at this point that discrepancies $r(u)$ do exist for which $\eta = \infty$. For example, $r(u) = \exp(-1/u^2)$. If we set $u_1 = u_2$ and let $u_1 \to 0$ we obtain

$$g(u_1, u_2) = \frac{e^{-1/(2u_1)^2}}{2e^{-1/u_1^2}} \to \infty. \tag{174}$$

In this case, of course, $r(u)$ does not satisfy (169) so that $p(u)$ does not satisfy (8e).

APPENDIX C

For channel B, let $V_n(\rho)$ be the volume of the intersection of a sphere in Euclidean n-space of radius ρ and center at the origin with the cube $[-A,A]^n$. For channel D, $V_n(\rho)$ is the volume of $S_n(\mathbf{x},\rho)$ = volume of $S_n(\mathbf{0},\rho)$, where

$$S_n(\mathbf{0},\rho) = \left\{\boldsymbol{\alpha} = (\alpha_1, \alpha_2, \alpha_3, \cdots, \alpha_n) \in \mathcal{C}_n : d_o(\mathbf{0},\boldsymbol{\alpha}) = \sum_{k=1}^{n} r(\alpha_k) < \rho\right\}.$$

In this appendix we evaluate

$$\lim_{n\to\infty} \frac{1}{n} \ln \frac{(2A)^n}{V_n(an)} = E_a. \tag{175}$$

We shall find E_a by solving an equivalent probability problem: Let $X_1, X_2, \cdots$ be a sequence of random variables uniformly distributed on the interval $[-A,+A]$. Let $Y_n = \sum_{k=1}^{n} r(X_k)$. For channel B, $r(u) = u^2$. It is clear that

$$\Pr[Y_n < \rho] = \frac{V_n(\rho)}{(2A)^n}, \tag{176}$$

hence

$$-\lim_{n\to\infty} (1/n) \ln \Pr[Y_n < an] = E_a. \tag{177}$$

We now make use of

Chernoff's Theorem[15]: Let $Z_1, Z_2, \cdots$ be a sequence of independent identically distributed random variables with moment generating function $E(\exp(Z_i t)) = M(t)$. Let $P_n = \Pr\left[\sum_{i=1}^{n} Z_i \leqq an\right]$, where $a \leqq E(Z_i)$. Then

$$\frac{1}{n} \ln P_n \xrightarrow{n} \ln m,$$

where $m = \min_{t \leqq 0} e^{-at} M(t)$.

If we let $Z_i = r(X_i)$ where X_i is the above random variable, then $Y_n = \sum_{i=1} Z_i$. Thus from (177) $E_a = -\ln m$.

The moment generating function of Z_i is

$$M(t) = E[e^{Z_i t}] = \frac{1}{2A} \int_{-A}^{+A} e^{r(x)t} dx,$$

so that

$$f(t) \triangleq e^{-at} M(t) = \frac{1}{2A} e^{-at} \int_{-A}^{+A} e^{r(x)t} dx, \tag{178}$$

and

$$m = \min_{t \leqq 0} f(t).$$

To minimize $f(t)$, let us differentiate (178) with respect to t:

$$\frac{df(t)}{dt} = 0 = \frac{1}{2A} e^{-at} \left[\int_{-A}^{+A} r(x) e^{r(x)t} dx - a \int_{-A}^{+A} e^{r(x)t} dx\right],$$

so that

$$\int_{-A}^{+A} r(x) e^{r(x)t} dx = a \int_{-A}^{+A} e^{r(x)t} dx. \tag{179}$$

The solution of (179) for t is $t = -\lambda(a)$ where $\lambda(\xi)$ is defined by (19a). With t so chosen

$$m = f(t) = \frac{1}{2A} e^{+a\lambda(a)} \int_{-A}^{+A} e^{-r(x)\lambda(a)} dx,$$

so that

$$\ln m = -\ln 2AK_o(a) + a\lambda(a) \tag{180}$$

where $K_o(a)$ is defined by (19b). Thus

$$E_a = -\ln m = \ln 2AK(a) - a\lambda(a) = C_o(a), \tag{181}$$

where $C_0(a)$ is defined by (19).

If we apply this result to (73) (channel B) with $a = 4\beta$, we obtain $R_{L_2}(\beta) = C_o(4\beta)$. If we apply this result to (113) and (120) (channel D) with $a = \beta$ and $2\eta\beta$ respectively, we obtain $R_U(\beta) = C_o(\beta)$ and $R_L(\beta) = C_o(2\eta\beta)$ respectively. Finally, applying this result to (136) (channel D with quadratic discrepancy) with $a = 2\beta$ yields $R(\beta) \leqq C_o(2\beta)$.

APPENDIX D

Estimate of $C_o(\xi)$ for Small ξ with $r(u) = u^2$

We first obtain an estimate of $\lambda(\xi)$ for small ξ and then show how this estimate can be used to estimate $C(\xi)$.

The quantity $\lambda(\xi)$ is defined by (19a):

$$\int_0^A u^2 e^{-u^2\lambda(\xi)} du = \xi \int_0^A e^{-u^2\lambda(\xi)} du. \tag{182}$$

Observe that $\lambda(\xi)$ monotonically approaches infinity as $\xi \to 0$. Changing the variable of integration in (182) we obtain

$$\int_0^{\sqrt{2\lambda}A} x^2 e^{-x^2/2} dx = 2\lambda\xi \int_0^{\sqrt{2\lambda}A} e^{-x^2/2} dx. \tag{183}$$

Integrating the left integral by parts yields:

$$-xe^{-x^2/2}\Big|_0^{\sqrt{2\lambda}A} + \int_0^{\sqrt{2\lambda}A} e^{-x^2/2} dx = 2\lambda\xi \int_0^{\sqrt{2\lambda}A} e^{-x^2/2} dx. \tag{184}$$

Rearranging terms we obtain

$$\lambda = \frac{1}{2\xi}(1 - \mu(\lambda)) \tag{185}$$

where

$$\mu(\lambda) = \frac{\sqrt{2\lambda}Ae^{-\lambda A^2}}{\int_0^{\sqrt{2\pi}A} e^{-x^2/2} dx}. \tag{186}$$

Since $\mu(\lambda) \geqq 0$ we have an upper bound on λ:

$$\lambda \leqq 1/2\xi. \tag{187}$$

To obtain a lower bound on λ set

$$\Delta = (1/2\xi) - \lambda. \tag{188}$$

From (185)

$$\Delta = \frac{\mu(\lambda)}{2\xi} = \frac{\lambda\mu(\lambda)}{1 - \mu(\lambda)} = \frac{A\sqrt{2}\lambda^{\frac{1}{2}}e^{-\lambda A^2}}{\int_0^{\sqrt{2\lambda}A} e^{-x^2/2}dx - A\sqrt{2\lambda}e^{-\lambda A^2}} = \frac{N(\lambda)}{D(\lambda)}. \tag{189}$$

It may be verified by differentiation that for $\lambda \geqq 3/(2A^2)$ the numerator $N(\lambda)$ is monotonically decreasing and the denomination $D(\lambda)$ is monotonically increasing so that Δ is monotonically decreasing. With $\lambda = 3/(2A^2)$ we obtain by substitution into (189) $\Delta = 0.76/A^2$ and by substitution into (185), $\xi = 0.22A^2$. Thus for $\xi \leqq 0.22A^2$

$$\lambda = \frac{1}{2\xi} - \Delta \geqq \frac{1}{2\xi} - \frac{0.76}{A^2}. \tag{190}$$

Returning to (189), we may write

$$\Delta \leqq \frac{N\left(\frac{1}{2\xi} - \frac{0.76}{A^2}\right)}{D(3/2)} \leqq \frac{A(1.12)e^{-A^2/2\xi}\xi^{-\frac{1}{2}}}{0.76} = 1.35A\frac{e^{-A^2/2\xi}}{\xi^{\frac{1}{2}}}. \tag{191}$$

Thus we have for $\xi \leqq 0.22A^2$:

$$\lambda = \frac{1}{2\xi} - \Delta \geqq \frac{1}{2\xi}\left[1 - \frac{2.70Ae^{-A^2/2\xi}}{\xi^{\frac{1}{2}}}\right]. \tag{192}$$

Since the quantity $C_o(\xi)$ is defined by

$$C_o(\xi) = \ln 2AK_o(\xi) - \xi\lambda(\xi), \tag{193}$$

where

$$K_o(\xi) = \left[\int_{-A}^{A} e^{-u^2\lambda(\xi)}du\right]^{-1}, \tag{194}$$

we could then use the upper and lower bounds on $\lambda(\xi)$ of (187) and

(192) to obtain an estimate of $C_o(\xi)$. However, this turns out to be a very cumbersome procedure and we shall side-step this chore. Suffice to observe that $\lambda(\xi)$ approaches $1/(2\xi)$ very rapidly as ξ approaches zero so that for small ξ we could take λ to be $1/(2\xi)$ and obtain

$$C_o(\xi) = \tfrac{1}{2} \ln \frac{2A^2}{\pi e \xi} + \epsilon(\xi), \tag{195}$$

where $\epsilon(\xi) \to 0$ as $\xi \to 0$.

APPENDIX E

Completion of Derivation of Upper Bound on $R(\beta)$ for Channel B

Inequality (86) expresses the fact that

$$R(\beta) \leqq f(\alpha)(1 - 2\hat{\beta}) \tag{196}$$

where

$$f(\alpha) = \frac{\alpha^2}{\alpha^2 - 2} \ln \alpha \tag{197}$$

and α is any integer satisfying $\alpha \geqq 2$, $\alpha^2 > 1/\hat{\beta}$ $(0 \leqq \hat{\beta} < \tfrac{1}{2})$. To obtain the tightest bound we seek to minimize $f(\alpha)$ subject to these constraints. It may be verified by differentiation that $f(\alpha)$ is a monotone increasing function for integer values of α for $\alpha \geqq 2$. Thus to minimize $f(\alpha)$ we choose α as the smallest integer satisfying $\alpha \geqq 2$, $\alpha^2 > 1/\hat{\beta}$. Thus we choose

$$\alpha = 2 \quad \text{when} \quad \tfrac{1}{2} \geqq \hat{\beta} \geqq \tfrac{1}{4},$$

and

$$\alpha = k \quad \text{when} \quad \frac{1}{(k-1)^2} > \hat{\beta} \geqq \frac{1}{k^2}, \qquad (k = 3, 4, 5, \cdots).$$

APPENDIX F

Estimate of $E_B(R)$ for Channel B

Equation (96) expresses the fact that

$$P_{eB}{}^* = \Pr\left[\sum_{i=1}^{n} z_i{}^2 > \alpha n\right], \tag{198}$$

where the z_i are independent normally distributed random variables

with mean zero and variance N, and $\alpha = \beta(R) = A^2\hat{\beta}(R)$. We seek an expression for $E_B(R) = \lim_{n\to\infty} -(1/n) \ln P_{eB}^*$. We again make use of a form of:

Chernoff's Theorem[15]: Let $Y_1, Y_2, \cdots, Y_n$ be independent and identically distributed random variables with moment generating function $E[\exp(Y_i t)] = M(t)$. Let $P_n = \Pr\left[\sum_{i=1}^{n} Y_i \geqq \alpha n\right]$, where $\alpha \geqq E(Y_i)$. Then

$$\lim_{n\to\infty} (1/n) P_n = \ln m,$$

where

$$m = \min_{t\geqq 0} e^{-\alpha t} M(t).$$

If we set $Y_i = z_i^2$ then $E_B = \lim_{n\to\infty} -(1/n) \ln P_n = -\ln m$. The moment generating function is

$$M(t) = \frac{1}{\sqrt{2\pi N}} \int_{-\infty}^{+\infty} e^{x^2 t} e^{-x^2/2N} dx = \frac{1}{(1-2Nt)^{\frac{1}{2}}} \left(t \leqq \frac{1}{2N}\right).$$

It may be verified by differentiation that the quantity $e^{-\alpha t} M(t)$ is minimized at $t = (1/2N) - (1/2\alpha)$ (which is positive if $\alpha > N$). Thus

$$m = \exp\left[-\alpha\left(\frac{1}{2N} - \frac{1}{2\alpha}\right)\right] M\left(\frac{1}{2N} - \frac{1}{2\alpha}\right).$$

Setting $\alpha = A^2\hat{\beta}(R)$ and taking logarithms we obtain

$$\begin{aligned} E_B(R) = -\frac{1}{n} \ln m &= \frac{\hat{\beta}(R)}{2N} \frac{A^2}{N} - \frac{1}{2} \ln \frac{A^2}{N} e\hat{\beta}(R) \\ &= \frac{\beta(R)}{2N} - \frac{1}{2} \ln \frac{e\beta(R)}{N}. \end{aligned} \tag{199}$$

APPENDIX G

Completion of Asymptotic Estimates for Channel C

1. Let $I_n = \int_0^\theta \sin^{n-2}\varphi \, d\varphi$. We must show that

$E = \lim_{n\to\infty} \frac{1}{n} \ln I_n = \ln \sin\theta$. This is (101a).

(a) $I_n \leqq \int_0^\theta \sin^{n-2}\theta\, d\varphi = (\theta)\sin^{n-2}\theta$, so that

$$\frac{1}{n}\ln I_n \leqq \frac{1}{n}\ln\theta + \frac{n-2}{n}\ln\sin\theta \xrightarrow{n} \ln\sin\theta.$$

(b) $I_n \geqq \int_{\theta-\frac{\theta}{n}}^{\theta} \sin^{n-2}\varphi\, d\varphi \geqq \sin^{n-2}\left(\theta - \frac{\theta}{n}\right)\left[\frac{\theta}{n}\right]$, so that

$\frac{1}{n}\ln I_n \geqq \frac{n-2}{n}\ln\sin\left(\theta - \frac{\theta}{n}\right) + \frac{1}{n}\ln\frac{\theta}{n} \to \ln\sin\theta$. This completes the proof.

2. Let $I_n = \int_0^\psi \sin^{n-2}\varphi\,(\cos\varphi - \cos\psi)d\varphi$. We must show that

$E = \lim_{n\to\infty}\frac{1}{n} I_n = \ln\sin\psi$. This is (101b).

(a) $I_n \leqq \int_0^\psi \sin^{n-2}\psi(\cos\varphi - \cos\psi)d\varphi = \sin^{n-2}\psi[\sin\psi - \psi\cos\psi]$,

so that $\frac{1}{n}\ln I_n \leqq \frac{n-2}{n}\ln\sin\psi + \frac{1}{n}\ln[\sin\psi - \psi\cos\psi] \xrightarrow{n} \ln\sin\psi$

(b) $I_n \geqq \int_{\psi-\frac{\psi}{n}}^{\psi} \sin^{n-2}\varphi(\cos\varphi - \cos\psi)d\varphi$

$$\geqq \sin^{n-2}\left(\psi - \frac{\psi}{n}\right)\int_{\psi-\frac{\psi}{n}}^{\psi}(\cos\varphi - \cos\psi)d\varphi \tag{200}$$

Now $I \triangleq \int_{\psi-\frac{\psi}{n}}^{\psi}(\cos\varphi - \cos\psi)d\varphi$

$$= \sin\psi - \sin\left(\psi - \frac{\psi}{n}\right) - \frac{\psi}{n}\cos\psi$$

$$= \sin\psi - \sin\psi\cos\frac{\psi}{n} + \cos\psi\sin\frac{\psi}{n} - \frac{\psi}{n}\cos\psi.$$

Expanding $\sin(\psi/n)$ and $\cos(\psi/n)$ into power series in (ψ/n) we obtain

$$I = \sin\psi\left[\frac{\psi^2}{2n^2} + o\left(\frac{1}{n^2}\right)\right] = \frac{\psi^2}{2n^2}\sin\psi(1 + o(1)).$$

Thus

$$\frac{1}{n} \ln I = \frac{1}{n} \ln \frac{\psi^2}{2n^2} \sin \psi + \frac{1}{n} \ln (1 + o(1)) \xrightarrow{n} 0.$$

Thus from (200) we have

$$\frac{1}{n} \ln I_n \geqq \frac{n-2}{n} \ln \sin \left(\psi - \frac{\psi}{n}\right) + \frac{1}{n} \ln I \xrightarrow{n} \ln \sin \psi.$$

Thus $E = \ln \sin \psi$ which completes the proof.

APPENDIX H

The Capacity of Channel D

The channel capacity is defined[1] by

$$C = \max_{\hat{p}(x)} [H(y) - H(y \mid x)], \tag{201}$$

where x is the input digit, y the output digit and $H(y \mid x)$ the conditional uncertainty of y given x. The maximization is performed over the input distribution $\hat{p}(x)$. Since $y = x \dotplus z$, $H(y \mid x) = H(z)$ so that

$$H(y \mid x) = H(z) = -\int_{-A}^{+A} p(u) \ln p(u)\, du,$$

independent of $\hat{p}(x)$. Now $H(y)$ is maximized when the random variable y is uniformly distributed on $[-A,+A]$. Due to the symmetry of the channel, this occurs when $\hat{p}(x) = 1/(2A)$, $-A \leqq x \leqq +A$. In this case

$$H(y) = -\int_{-A}^{+A} \frac{1}{2A} \ln \frac{1}{2A}\, dy = \ln 2A.$$

Thus the channel capacity is

$$C = \ln 2A + \int_{-A}^{+A} p(u) \ln p(u)\, du. \tag{202}$$

Writing $p(u) = K_o \exp [-\lambda r(u)]$ we obtain

$$\begin{aligned} C &= \ln 2A + \ln K_o \int_{-A}^{+A} p(u)\, du - \lambda \int_{-A}^{+A} r(u) p(u)\, du \\ &= \ln 2AK_o - \lambda N, \end{aligned} \tag{203}$$

where N is defined by (32) or

$$N = \int_{-A}^{+A} r(u) K_o e^{-\lambda r(u)} du. \tag{204}$$

Also, since $p(u)$ integrates to unity,

$$1 = \int_{-A}^{+A} K_o e^{-\lambda r(u)} du, \tag{205}$$

we have

$$\int_{-A}^{+A} r(u) e^{-\lambda r(u)} du = N \int_{-A}^{+A} e^{-\lambda r(u)} du, \tag{206}$$

with N and $r(u)$ specified, λ may be found as the solution to (206). With λ so specified we may find K_o from (205), thus

$$C = \ln 2AK_o(N) - N\lambda(N),$$

where $\lambda(N)$ is the solution of (206) and $K_o(N)$ is the solution to (205). This is the same as $C = C_o(N)$ where $C(\xi)$ is defined by (19).

GLOSSARY OF SYMBOLS

The following symbols are used throughout the paper:

n = dimension of input, output and noise vectors.
$\mathbf{x} = (x_1, x_2, \cdots, x_n)$ = input vector or code word.
$\mathbf{y} = (y_1, y_2, \cdots, y_n)$ = output vector or received vector.
$\mathbf{z} = (z_1, z_2, \cdots, z_n)$ = noise vector.
M = number of words in a code.
$R = (1/n) \ln M$ = transmission rate.
P_{ei} = probability that the receiver makes an incorrect decoding decision when code word i is transmitted ($i = 1, 2, \cdots, M$).
$P_e = (1/M) \sum_{i=1}^{M} P_{ei}$ = over-all error probability.
MDD = minimum discrepancy decoding (always optimum for the channels considered in this paper).
BDD = bounded discrepancy decoding.
P_{eM} = error probability (P_e) using MDD.
P_{eB} = error probability using BDD.
C = channel capacity = "maximum error free rate" using MDD.
$E(R)$ = the best attainable error exponent using MDD, (at rate R).
C_B = bounded distance decoding channel capacity, or "maximum error free rate" using BDD.
$E_B(R)$ = the best attainable error exponent using BDD (at rate R).

The following symbols are used in connection with specific channels:

Channel A

q = the number of symbols in the input, output and noise alphabets.

p_o = the probability that the channel transmits a given symbol correctly.

$d_H(\mathbf{u},\mathbf{v})$ = the Hamming distance between two n-vectors $\mathbf{u}$ and $\mathbf{v}$ = the number of positions in which $\mathbf{u}$ and $\mathbf{v}$ differ.

$C(p_o) = \ln q - H(p_o) - p_o \ln (q-1)$ = channel capacity of channel A with symbol error probability p_o.

$H(\rho) = -\rho \ln \rho - (1-\rho) \ln (1-\rho)$ = the entropy function.

d = the minimum distance between code words.

$e = (d-1)/2$ = number of correctable errors in a code with minimum Hamming distance d.

$M(n,d)$ = maximum number of code words in an n-dimensional code with minimum Hamming distance d.

$R(n,d) = (1/n)M(n,d)$ = rate corresponding to $M(n,d)$.

$\beta = d/2n$, a ratio appearing in our bounds.

$t = [(q-1)/q\beta]\,[1 - \sqrt{1 - [2q/(q-1)]\beta}]$, another quantity appearing in our bounds.

$[x]$ = largest integer not exceeding x.

$R(\beta) = \lim_{n\to\infty} R(n,2\beta n)$, asymptotic form of $R(n,d)$.

$\alpha(\rho,p_o) = \rho \ln (\rho/p_o) + (1-\rho) \ln [(1-\rho)/(1-p_o)]$, a quantity appearing in our error bounds.

s = parameter defined by $R = \ln q - H(s) - s \ln (q-1) = C(s)$

Channel B

A = maximum amplitude of input coordinates.

N = variance of normally distributed noise coordinates.

$d_E(\mathbf{u},\mathbf{v}) = \sum_{i=1}^{n} (u_i - v_i)^2$ = Euclidean distance between the n-vectors $\mathbf{u}$ and $\mathbf{v}$.

d = the minimum distance between code words.

$M(n,d^2)$ = maximum number of code words in an n-dimensional code with minimum Euclidean distance d.

$R(n,d^2) = (1/n) \ln M(n,d^2)$ = rate corresponding to $M(n,d)$.

$\beta = d^2/4n, \hat{\beta} = \beta/A^2$, ratios appearing in our bounds.

$R(\beta) = \lim_{n\to\infty} R(n,4\beta n)$, asymptotic form of $R(n,d^2)$.

$R_L(\beta)$ and $R_U(\beta)$ = lower and upper bounds on $R(\beta)$ given by (18) and (19) respectively.

The function $C_o(\xi)$, $0 < \xi < A^2/3$, is defined as follows: $\lambda(\xi)$ is the quantity defined by

$$\int_0^A r(u)e^{-\lambda(\xi)r(u)}du = \xi\int_0^A e^{-\lambda(\xi)r(u)}du$$

where $r(u) = u^2$, and

$$K(\xi) = \left[\int_{-A}^{A} e^{-\lambda(\xi)r(u)}du\right]^{-1}.$$

Then

$$C_o(\xi) = \ln 2\,AK_o(\xi) - \xi\lambda(\xi).$$

Channel C

P = $(1/n)\times$ the energy of a code word.
N = variance of the normally distributed noise coordinates.
$d_E(\mathbf{u},\mathbf{v})$ = the Euclidean distance between $\mathbf{u}$ and $\mathbf{v}$.
$a(\mathbf{u},\mathbf{v})$ = the angle between n-vectors $\mathbf{u}$ and $\mathbf{v}$.
θ = the minimum angle between code words.
$M(n,\theta)$ = maximum number of code words in an n-dimensional code with minimum angle θ.
$R(n,\theta)$ = $(1/n)\ln M(n,\theta)$ = rate corresponding to $M(n,\theta)$.
ψ = $\sin^{-1}\sqrt{2}\sin(\theta/2)$, a quantity appearing in our bounds.

Channel D

$\mathcal{C}_n$ = set of real n-vectors $\mathbf{u} = (u_1, u_2, \cdots, u_n)$ satisfying $|u_k| \leqq A$.
$\dotplus, \dot{-}$ = addition and subtraction modulo $2A$ (with result reduced into the interval $[-A,+A]$).
$p(u)$ = noise probability density function.
$r(u)$ = $(1/\lambda)\ln[p(0)/p(u)]$ $(-A \leqq u \leqq +A)$, quantity related to the discrepancy.
$d_o(\mathbf{u},\mathbf{v}) = \sum_{k=1}^{n} r(u_k \dot{-} v_k)$ = discrepancy between n-vectors $\mathbf{u}$ and $\mathbf{v}$ belonging to $\mathcal{C}_n$.
N = $E(r(z))$, a parameter associated with the noise density $p(u)$.

$C_o(\xi)$, defined exactly as for channel B but with the appropriate $r(u)$ used instead of u^2.

$\eta = \sup_{-A \leqq u_1, u_2 \leqq A} \frac{r(u_1 \dotplus u_2)}{r(u_1) + r(u_2)}$, a quantity appearing in our bounds.

ACKNOWLEDGMENT

The author wishes to thank L. A. Shepp, E. N. Gilbert, and especially D. Slepian for many stimulating discussions and helpful suggestions.

REFERENCES

1. Shannon, C. E., A Mathematical Theory of Communication, B.S.T.J., *27*, July and October, 1948, pp. 379–423, 623–656.
2. Elias, P., Coding for Two Noisy Channels, in *Information Theory*, Colin Cherry (ed.), Academic Press, New York, 1956, pp. 61–74.
3. Gallager, R. G., *Low Density Parity Check Codes*, MIT Press, Cambridge, 1963.
4. Gramenapoulos, N., An Upper Bound for Error-Correcting Codes, M.S. Thesis, Department of Electrical Engineering, MIT, 1963.
5. Bose, R. C., and Ray-Chaudhuri, D. K., On a Class of Error Correcting Binary Group Codes, Information and Control, *3*, 1960, pp. 68–79.
6. Hamming, R. W., Error Detecting and Error Correcting Codes, B.S.T.J., *29*, April, 1950, pp. 147–160.
7. Shannon, C. E., Certain Results in Coding Theory for Noisy Channels, Information and Control, *1*, 1957, pp. 6–25.
8. Peterson, W. W., *Error Correcting Codes*, MIT Press and John Wiley & Sons, New York, 1961.
9. Plotkin, M., Binary Codes with Specified Minimum Distance, IRE Transactions on Information Theory, *IT-6*, 1960, pp. 445–450.
10. Wolfowitz, J., *Coding Theorems of Information Theory*, Prentice-Hall, Inc., Englewood Cliffs, 1961.
11. Shannon, C. E., Probability of Error for Optimal Codes in a Gaussian Channel, B.S.T.J., *38*, May, 1959, pp. 611–656.
12. Wyner, A. D., An Improved Error Bound for Gaussian Channels, B.S.T.J., *43*, November, 1964, pp. 3070–3075.
13. Rankin, R. A., The Closest Packing of Spherical Caps in n-dimensions, Proceedings of the Glasgow Mathematical Association, *2*, 1955, pp. 139–144.
14. Blachman, N. M., On the Capacity of a Band-Limited Channel Perturbed by Statistically Dependent Interference, IRE Transactions on Information Theory, *IT-8*, 1962, pp. 48–55.
15. Chernoff, H., A Measure of Asymptotic Efficiency for Tests of a Hypothesis Based on the Sum of Observations, Annals of Math. Stat., *23*, 1952, pp. 493–507.

On the Reception of Binary Signals in the Presence of a Small Random Delay*

By M. I. SCHWARTZ

(Manuscript received March 22, 1965)

Receiver design for a binary communication system which operates over a linear channel with a random delay is considered. It is assumed that the variance of the random delay is very small and that the rate of growth of its moments is restricted. Under certain smoothness requirements on the received signal an approximation to the test statistic, which is optimum in the Neyman-Pearson sense, is derived for the case of gaussian receiver noise with covariance $R(\tau) = R(0)e^{-\beta|\tau|}$. *It is found that the test statistic, which in general is nonlinear, assumes the linear form of a crosscorrelator when phase reversal signaling is employed.*

The case where the noise is white and phase reversal signaling is used is investigated. The correlation waveform in this case is found to consist of the expected value of the received signal plus a term dependent on the slope of the signal when the delay is equal to its mean value.

I. INTRODUCTION AND SUMMARY

In any practical communication system the signal arrival time is never exactly known. This results in a degradation of the average system performance. It would be of considerable interest to determine the receiver which minimizes the effect of this uncertainty on system error performance. A special case of this problem will be considered here.

Helstrom[1] has studied the detection of signals of unknown arrival time using the method of maximum likelihood with particular emphasis on the radar problem. Brown and Palermo[2] consider system performance in the presence of random delays with applications including least squares filtering and sampling with time jitter. Balakrishnan[3] and other

* This work is based on a portion of a thesis entitled "Binary Signal and Receiver Design for Linear Time Invariant Channels". This thesis was accepted by the faculty of the Graduate Division of the School of Engineering and Science of New York University in partial fulfillment of the requirements for the degree of Doctor of Engineering Science, Oct. 1964.

researchers have also considered the problem of time jitter in sampling. However, to the best of this authors knowledge, no optimum statistical test, or approximation thereof, has been determined for detection in the presence of a random delay.

In the subsequent analysis we investigate binary communication for the case where the variance of the random delay is "very small." It is assumed that the transmitted signals and the channel impulse response are such that the received signal satisfies appropriate smoothness conditions, and that the statistics of the random delay δ satisfy the relation that $E[\mid \delta - \bar{\delta} \mid^k] \leqq h^k \lambda^k$; where h is some constant, E denotes expectation, $\bar{\delta} \triangleq E\delta$ and λ^2 is the variance of δ. The requirement on the random delay will always be satisfied when δ is restricted to a bounded interval. Our model will also assume that intersymbol interference is negligible, or equivalently, that we are dealing with a single transmission.

Under these assumptions an approximation is obtained for the test statistic which is optimum, in the Neyman-Pearson sense, for the case of gaussian receiver noise with exponential covariance. Generally the test statistic involves a nonlinear operation. However, for the case of phase reversal signals, only linear operations are required.

The form that the test statistic takes for "white noise," which is considered as a limit of the exponential covariance case, is obtained. It is shown that for phase reversal signaling the optimum receiver is a cross-correlator and that a portion of the correlation waveform is the expected value of the received signal itself.

Fig. 1 depicts the communication system under consideration. A signal, $s^{(i)}(t)$ $(i = 1 \text{ or } 2)$, which is non-zero only over an interval $[0,T]$, is transmitted through a channel. The channel consists of a random delay δ and a linear time invariant filter whose output, $x^{(i)}(t - \delta)$, is disturbed by an additive noise source, $n(t)$. It is assumed that the variance of the random delay, denoted by λ^2, is small. Furthermore, the noise, $n(t)$, will

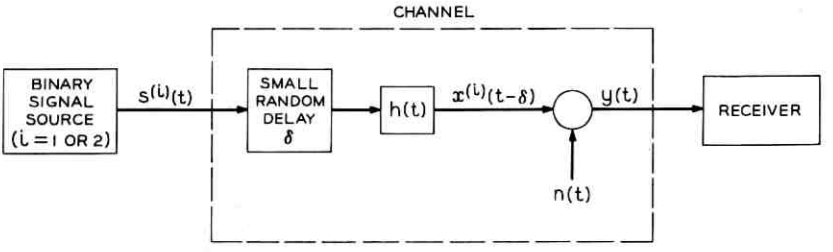


Fig. 1 — Model of Binary Communications System with a Small Random Delay.

be assumed to be a sample function of a stationary gaussian random process with mean zero and covariance $R(\tau) = R(0)e^{-\beta|\tau|}$. For this system we seek the test statistic, based on an observation of the received signal over a fixed interval of length equal to the duration of the transmitted signal, which gives rise to the minimum error probability in the receiver decision process.

II. DERIVATION OF THE TEST STATISTIC

It is known that in testing between two simple hypotheses a Neyman-Pearson test will give rise to minimum error probability. Furthermore, Grenander[4] has shown that in the "regular case" the desired test statistic, which is a random variable called the likelihood function, l, can be obtained as the limit of an N dimensional likelihood ratio. In the subsequent development, in which it is assumed that we deal only with the regular case, the receiver test statistic is obtained as the limit of such an N-dimensional likelihood ratio.

The receiver input, $y(t)$, is given by

$$y(t) = x^{(i)}(t - \delta) + n(t), \qquad i = 1,2 \tag{1}$$

where $x^{(i)}(t - \delta)$ is the portion of the input resulting from sending the signal $s^{(i)}(t)$ when the random delay is δ, and $n(t)$ is the receiver noise. The noise is assumed to be gaussian with covariance $R(\tau) = R(0)e^{-\beta|\tau|}$.

Using a theorem due to Belayev[5] the noise sample functions can be shown to be almost surely continuous. Furthermore, almost all sample functions can be expanded almost surely in a pointwise convergent series in terms of the eigenfunctions of the noise covariance kernel. That is,

$$n(t) = \sum_k n_k \varphi_k(t), \tag{2}$$

with

$$E(n_j n_k) = \sigma_j^2 \delta_{jk},$$

where the φ_k satisfy the integral equation

$$\sigma_k^2 \varphi_k(t) = \int_{t_0}^{t_1} du\, \varphi_k(u) R(t - u), \qquad t_0 < t < t_1 \tag{3}$$

and

$$n_k = \int_{t_0}^{t_1} dt\, \varphi_k(t) n(t).$$

Here t_0 and $t_1 \triangleq t_0 + T$ mark the beginning and end of the receiver processing interval, $\sum_k$ is used to denote $\sum_{k=1}^{\infty}$, the symbol E signifies mathematical expectation, and

$$\delta_{jk} = \begin{cases} 1 & j = k \\ 0 & j \neq k \end{cases}$$

Since the $\{n_k\}$ are themselves gaussian and uncorrelated they are statistically independent. Assuming that the noise has zero mean, the joint density function of the first N coefficients, p_N, can be written as

$$p_N(n_k, k = 1, \cdots, N) \triangleq \prod_{k=1}^{N} \left(\frac{1}{2\pi}\right)^{\frac{1}{2}} \frac{1}{\sigma_k} \exp\left\{-\frac{n_k^2}{2\sigma_k^2}\right\}, \tag{4}$$

where the n_k are ordered corresponding to the relationship that

$$\sigma_1 \geqq \sigma_2 \geqq \cdots .$$

Now consider a formal expansion of the receiver input in terms of the eigenfunctions of (3). One can write

$$\begin{aligned} y(t) \sim \sum_k y_k \varphi_k(t) &= \sum_k \psi_k^{(i)} \varphi_k(t) + \sum_k n_k \varphi_k(t) \\ y_k \triangleq \int_I dt\, \varphi_k(t)\, y(t) &= \int_I dt\, \varphi_k(t)\, [x_\delta^{(i)}(t) + n(t)] \\ \psi_k^{(i)}(\delta) \triangleq \int_I dt\, \varphi_k(t)\, x^{(i)}(t - \delta), & \end{aligned} \tag{5}$$

where by definition $I \triangleq [t_0, t_1]$. Since the series $\sum_k n_k \varphi_k(t)$ is almost surely pointwise convergent to $n(t)$ we need only investigate the sense in which $\sum_k \psi_k^{(i)} \varphi_k(t)$ converges to $x^{(i)}(t - \delta)$. With this in mind we digress to consider some of the properties associated with the eigenfunctions of the integral (3) with $R(t - u) = R(0)e^{-\beta | t-u |}$. It is easy to show that in this case the solutions of the integral equation are identical to those which satisfy the following differential equations and boundary conditions:

$$\begin{aligned} \frac{d^2}{dt^2}\varphi_k(t) - \frac{\beta(\beta\sigma_k^2 - 2R(0))}{\sigma_k^2}\varphi_k(t) &= 0 \\ \beta\varphi_k(t_0) &= \frac{d}{dt}\varphi_k(t)|_{t=t_0} \\ \beta\varphi_k(t_1) &= -\frac{d}{dt}\varphi_k(t)|_{t=t_1}. \end{aligned} \tag{6}$$

The solutions of this system are proportional to

$$\cos \gamma_k \left[t - \left(\frac{t_0 + t_1}{2} \right) \right]$$

for k even, and

$$\sin \gamma_k \left[t - \left(\frac{t_0 + t_1}{2} \right) \right]$$

for k odd. Here γ_k satisfies the relation $(\beta^2 + \gamma_k^2) = 2\beta R(0)/\sigma_k^2$.

The differential equation (6) and the associated boundary conditions together form a Sturm-Liouville eigenvalue problem. The convergence properties of expansions in terms of the resulting eigenfunctions, the $\{\varphi_k(t)\}$, are stronger than those generally associated with expansions in terms of the eigenfunctions of the integral equation (3). An expansion of an integrable function $f(t)$ on the interval (t_0, t_1) in terms of the eigenfunctions of a Sturm-Liouville system, possesses the following property:[6]

In every interval where $f(t)$ is continuous and of bounded variation, the expansion converges uniformly and absolutely to $f(t)$. If at the ends of the interval there are neighborhoods in which $f(t)$ is of bounded variation then the series converges at these points to $f(t_{0+})$ and $f(t_{1-})$.

It will be assumed that the transmitted signal $s^{(i)}(t)$ and $h(t)$ are such that $x^{(i)}(t - \delta)$ is continuous and of bounded variation. In fact to make the succeeding development valid we shall have to impose more stringent requirements on $s^{(i)}(t)$ and $h(t)$. Under this assumption the expansion of $x^{(i)}(t - \delta)$ in terms of eigenfunctions of the integral equation (3) will converge uniformly and absolutely to $x^{(i)}(t - \delta)$.

Returning to (5) we have established that $\sum_k y_k \varphi_k(t)$ converges pointwise to $y(t)$ for almost all sample functions. That is $y(t) = \sum_k y_k \varphi_k(t)$, almost surely.

Choosing the values of the $\{y_k\}$ set as the observable coordinates, the likelihood function, l, can now be determined as the limit of the ratio of two N-dimensional density functions evaluated at the sample values, the y_j, $j = 1, \cdots N$. Here we have used the same symbol, y_j, to represent the sample and the random variable itself.

Thus l can be written as

$$l = \lim_{N \to \infty} \frac{\tilde{p}^{(1)}(y_1, \cdots, y_N)}{\tilde{p}^{(2)}(y_1, \cdots, y_N)}, \tag{7}$$

where $\tilde{p}^{(i)}(y_1, \cdots, y_N)$ is the joint probability density function of the first N members of the $\{y_k\}$ set.

Noting that $y_k = \psi_k^{(i)} + n_k$ and using the fact that the signal and

noise components are statistically independent the joint probability density of the first N of the y_k can be written as a convolution. Thus

$$\begin{aligned} \tilde{p}_N{}^{(i)}(y_1, \cdots, y_N) &= \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} dz_1, \cdots, dz_N \zeta^{(i)}(z_1, \cdots, z_N) \\ &\quad \cdot p_N(y_k - z_k, k = 1, \cdots, N) \\ \tilde{p}_N{}^{(i)}(y_1, \cdots, y_N) &= E_\psi(i)\, p_N(y_k - \psi_k{}^{(i)}, k = 1, \cdots, N) \end{aligned} \tag{8}$$

where

$\zeta^{(i)}(z_1, \cdots, z_N)$ is the joint probability density function of the first N of the $\psi_k{}^{(i)}(\delta)$,

$E_\psi{}^{(i)}$ denotes an expectation with respect to the random vector, $\psi^{(i)} = (\psi_1{}^{(i)}(\delta) \cdots \psi_N{}^{(i)}(\delta))$.

Since the $\psi_k{}^{(i)}(\delta)$ are all functions of the random variable δ, the averaging process can be performed with respect to δ instead of the $\psi_k{}^{(i)}(\delta)$. Thus

$$\tilde{p}_N{}^{(i)}(y_1, \cdots, y_N) = E_\delta p_N(y_k - \psi_k{}^{(i)}(\delta), k = 1, \cdots, N). \tag{9}$$

Using (4), (7) and (9), and cancelling common factors the likelihood function, l, is given by

$$l = \lim_{N\to\infty} \frac{E_\delta \exp\left\{\sum_{k=1}^{N}\left[y_k - \frac{\psi_k{}^{(1)}(\delta)}{2}\right]\frac{\psi_k{}^{(1)}(\delta)}{\sigma_k{}^2}\right\}}{E_\delta \exp\left\{\sum_{k=1}^{N}\left[y_k - \frac{\psi_k{}^{(2)}(\delta)}{2}\right]\frac{\psi_k{}^{(2)}(\delta)}{\sigma_k{}^2}\right\}}. \tag{10}$$

$$\text{Let } Q_N{}^{(i)}[\delta] \triangleq \sum_{k=1}^{N}\left[y_k - \frac{\psi_k{}^{(i)}(\delta)}{2}\right]\frac{\psi_k{}^{(i)}(\delta)}{\sigma_k{}^2}.$$

Then (10) can be rewritten as

$$l = \lim_{N\to\infty} \frac{E_\delta \exp\{Q_N{}^{(1)}[\delta]\}}{E_\delta \exp\{Q_N{}^{(2)}[\delta]\}}. \tag{11}$$

At this point one may develop an integral form for $Q_N{}^{(i)}[\delta]$. The integral will suggest the form of $\lim_{N\to\infty} Q_N{}^{(i)}[\delta]$ which will be required in the subsequent development. Define

$$\begin{aligned} q_N{}^{(i)}(t;\delta) &= \sum_{k=1}^{N} \frac{\psi_k{}^{(i)}(\delta)}{\sigma_k{}^2}\varphi_k(t) \\ x_N{}^{(i)}(t;\delta) &= \sum_{k=1}^{N} \psi_k{}^{(i)}(\delta)\varphi_k(t) \\ y_N(t) &= \sum_{k=1}^{N} y_k\varphi_k(t). \end{aligned} \tag{12}$$

Then using the orthonormality of the $\{\varphi_k(t)\}$ one has the relation

$$Q_N^{(i)}[\delta] = \int_I dt \left[y_N(t) - \frac{x_N(t;\delta)}{2} \right] q_N^{(i)}(t;\delta). \tag{13}$$

It also follows from (3), (12), and the orthonormality relations, that $q_N^{(i)}(t;\delta)$ satisfies the integral equation

$$x_N^{(i)}(t;\delta) = \int_I du\, q_N^{(i)}(u;\delta) R(0) \exp(-\beta \mid t - u \mid). \tag{14}$$

The solution of the integral equation is

$$q_N^{(i)}(t;\delta) = \frac{1}{2\beta R(0)} \left(\beta^2 - \frac{\partial^2}{\partial t^2} \right) x_N(t;\delta). \tag{15}$$

Thus (13) can be rewritten as

$$Q_N^{(i)}[\delta] = \int_I dt \left[y_N(t) - \frac{x_N(t;\delta)}{2} \right] \frac{1}{2\beta R(0)} \left(\beta^2 - \frac{\partial^2}{\partial t^2} \right) x_N(t;\delta). \tag{16}$$

Let us return to the study of the likelihood function of (11). In general it seems that the expectations appearing in (11) cannot be evaluated. However it is possible to evaluate the required expectations when the variance of the random delay δ is very small. With this in mind we drop the superscript and expand $Q_N(\delta)$ in a power series with remainder around $\delta = \bar{\delta}$.

$$\begin{aligned} Q_N[\delta] = Q_N[\bar{\delta}] + (\delta - \bar{\delta}) Q_N'[\bar{\delta}] + \frac{(\delta - \bar{\delta})^2}{2!} Q_N''[\bar{\delta}] \\ + \frac{(\delta - \bar{\delta})^3}{3!} Q_N'''[\bar{\delta} + \theta(\delta - \bar{\delta})], \qquad 0 \leqq \theta \leqq 1 \end{aligned} \tag{17}$$

where we have used the notation

$$Q_N'[\bar{\delta}] \triangleq \frac{d}{d\delta} Q_N(\delta) \mid_{\delta = \bar{\delta}}.$$

It follows that $\exp\{Q_N[\delta]\}$ can be expressed as,

$$\exp\{Q_N[\delta]\} = \exp\{Q_N[\bar{\delta}]\} \exp\left\{ (\delta - \bar{\delta})\, a_N + \frac{(\delta - \bar{\delta})^2}{2!} b_N + \frac{(\delta - \bar{\delta})^3}{3!} c_N(\delta) \right\},$$

where

$$a_N \triangleq Q_N'[\bar{\delta}] \qquad b_N \triangleq Q_N''[\bar{\delta}] \qquad c_N(\delta) \triangleq Q_N'''[\bar{\delta} + \theta(\delta - \bar{\delta})].$$

On expanding the exponential in a powers series and averaging the uniformly convergent series term by term with respect to δ one obtains

$$E_\delta \exp \{Q_N[\delta]\} = \exp \{Q_N[\bar{\delta}]\} \left[1 + \sum_{k=1}^{\infty} \frac{d_{kN}}{k!}\right], \tag{18}$$

where

$$d_{kN} = E_\delta \left\{\left[(\delta - \bar{\delta}) a_N + \frac{(\delta - \bar{\delta})^2}{2!} b_N + \frac{(\delta - \bar{\delta})^3}{3!} c_N(\delta)\right]^k\right\}.$$

The limit of the likelihood ratio of (10), l, can be expressed as

$$l = \lim_{N \to \infty} \exp \{Q_N^{(1)}[\bar{\delta}] - Q_N^{(2)}[\bar{\delta}]\} \frac{\left(1 + \sum_{k=1}^{\infty} \frac{d_{kN}^{(1)}}{k!}\right)}{\left(1 + \sum_{k=1}^{\infty} \frac{d_{kN}^{(2)}}{k!}\right)}. \tag{19}$$

Note that if for some number $D(N)$

$$| d_{kN} | < [D(N)]^k \quad \text{for all } k,$$

then the infinite sums each converge since the series is majorized by $\exp [D(N)]$. It is observed that

$$| z_1 + z_2 + z_3 |^k \leqq [3 \max_i | z_i |]^k$$

which implies

$$| z_1 + z_2 + z_3 |^k \leqq 3^k (| z_1 |^k + | z_2 |^k + | z_3 |^k).$$

Using the definition of d_{kN} and the above inequality in conjunction with the Schwarz inequality yields

$$\begin{aligned} | d_{kN} | \leqq 3^k \Bigg\{ & | a_N |^k E | (\delta - \bar{\delta})^k | + \left| \frac{b_N}{2!} \right|^k E | (\delta - \bar{\delta})^{2k} | \\ & + \left(E \left| \frac{c_N(\delta)}{3!} \right|^{2k} \cdot E | \delta - \bar{\delta} |^{6k} \right)^{\frac{1}{2}} \Bigg\}. \end{aligned} \tag{20}$$

Now we restrict our investigation to the class of random delays whose probability distributions satisfy the relationship that for some number h and all k,

$$E_\delta | (\delta - \bar{\delta})^k | \leqq h^k \lambda^k, \tag{21}$$

where λ^2 is the variance of the distribution. This condition will be satisfied by all probability distribution functions which take on the values

zero and one in a bounded region of the real line. That is when the values of δ are essentially restricted to a bounded region.

For distributions which satisfy (21) one finds

$$| d_{kN} | \leqq 3^k \left\{ | a_N |^k h^k \lambda^k + \left| \frac{b_N}{2!} \right|^k h^{2k} \lambda^{2k} + \frac{h^{3k} \lambda^{3k}}{3!} [E | c_N(\delta) |^{2k}]^{\frac{1}{2}} \right\}.$$

However if $Q_N'''[\delta]$ is bounded, $E | c_N(\delta) |^{2k}. \leqq (\bar{c}_N)^{2k}$ and

$$| d_{kN} | \leqq 3^k \{ | a_N | h\lambda + | b_N | h^2\lambda^2 + (\bar{c}_N)^2 h^3 \lambda^3 \}^k, \qquad | d_{kN} | \leqq A_N{}^k$$

where

$$\bar{c}_N \triangleq \sup Q_N'''[\delta],$$

$$A_N \triangleq 3 \{ | a_N | h\lambda + | b_N | h^2\lambda^2 + (\bar{c}_N)^2 h^3 \lambda^3 \}.$$

Now it can be shown that under certain restrictions on the channel and the class of transmitted signals the following limits exist almost surely:

$$\begin{aligned} \lim_{N \to \infty} a_N &\triangleq a, \\ \lim_{N \to \infty} b_N &\triangleq b, \\ \lim_{N \to \infty} c_N(\delta) &\triangleq c(\delta) \\ \lim_{N \to \infty} E | c_N(\delta) |^{2k} &\leqq \bar{c}^{2k} \end{aligned} \tag{22}$$

for all k.

The convergence of $\lim_{N \to \infty} \sum_{k=1}^{\infty} (d_{kN}/k!)$ can now be demonstrated. In this and the subsequent development we will consider convergence in the almost sure sense. Breaking the sum into a finite sum from k equal 1 through m and a sum from $m + 1$ to ∞ and taking magnitudes gives

$$\left| \lim_{N \to \infty} \sum_{k=1}^{\infty} \frac{d_{kN}}{k!} - \sum_{k=1}^{m} \frac{d_k}{k!} \right| = \left| \lim_{N \to \infty} \sum_{k=m+1}^{\infty} \frac{d_{kN}}{k!} \right|$$

where $d_k \triangleq \lim_{N \to \infty} d_{kN}$. Since $| d_{kN} |$ is less than or equal to some number $A_N{}^k$, then

$$\left| \lim_{N \to \infty} \sum_{k=1}^{\infty} \frac{d_{kN}}{k!} - \sum_{k=1}^{m} \frac{d_k}{k!} \right| \leqq \lim_{N \to \infty} \sum_{k=m+1}^{\infty} \frac{A_N{}^k}{k!}.$$

But $\lim_{N \to \infty} A_N = A$ exists, thus for all $N > N_0$

$$|A - A_N| < \epsilon, \quad |A_N| < |A| + \epsilon,$$

$$\Rightarrow \left| \lim_{N\to\infty} \sum_{k=1}^{\infty} \frac{d_{kN}}{k!} - \sum_{k=1}^{m} \frac{d_k}{k!} \right| \leqq \sum_{k=m+1}^{\infty} \frac{[|A| + \epsilon]^k}{k!}.$$

By choosing m sufficiently large the right hand side of the inequality can be made less than any positive constant. Therefore

$$\lim_{N\to\infty} \sum_{k=1}^{\infty} \frac{d_{kN}}{k!} = \sum_{k=1}^{\infty} \frac{d_k}{k!}. \tag{23}$$

For small λ one makes the approximation

$$\lim_{N\to\infty} \left[1 + \sum_{k=1}^{\infty} \frac{d_{kN}}{k!} \right] \sim 1 + d_1 + \frac{d_2}{2!}, \tag{24}$$

where terms involving λ to powers greater than λ^2 have been neglected, and

$$d_k = \lim_{N\to\infty} d_{kN}.$$

Using the definition of d_{kN}, which follows (18), we find

$$d_1 \sim (\lambda^2/2!)b,$$

$$d_2 \sim \lambda^2 a^2.$$

Restoring the superscript notation and using (22) and (24), (19) becomes

$$l \sim \exp \left[\lim_{N\to\infty} \{Q_N^{(1)}[\bar{\delta}] - Q_N^{(2)}[\bar{\delta}]\} \right]$$

$$\cdot \{1 + (\lambda^2/2)[(a^{(1)})^2 - (a^{(2)})^2 + b^{(1)} - b^{(2)}]\} \tag{25}$$

Equation (25) is the desired approximation for the likelihood function for the case of small λ.

III. STRUCTURE OF THE OPTIMUM RECEIVER

The structure of the approximation to the optimum receiver statistic, the likelihood function, can be determined from (25). The quantities appearing there can all be expressed in terms of the received signal, $y(t)$ and the noise-free filter outputs $x^{(1)}(t)$ and $x^{(2)}(t)$, which are assumed to be known. In the Appendix it is shown that almost surely

$$\lim_{N\to\infty} Q_N[\bar{\delta}] = \frac{1}{2\beta R(o)} \left\{ \int_I \left[y(t) - \frac{x(t-\bar{\delta})}{2} \right] \cdot [\beta^2 x(t-\bar{\delta}) - x''(t-\bar{\delta})]\, dt - [x'(t_0-\bar{\delta}) - \beta x(t_0-\bar{\delta})] \left[y(t_0) - \frac{x(t_0-\bar{\delta})}{2} \right] + [x'(t_1-\bar{\delta}) + \beta x(t_1-\bar{\delta})] \cdot \left[y(t_1) - \frac{x(t_1-\bar{\delta})}{2} \right] \right\}, \tag{26}$$

$$a = \lim_{N\to\infty} a_N = \frac{1}{2\beta R(0)} \left\{ \int_I \left[y(t) - \frac{x(t-\bar{\delta})}{2} \right] \cdot [-\beta^2 x'(t-\bar{\delta}) + x'''(t-\bar{\delta})]\, dt - [x''(t_1-\bar{\delta}) + \beta x'(t_1-\bar{\delta})]\,[y(t_1) - \tfrac{1}{2}x(t_1-\bar{\delta})] + [x''(t_0-\bar{\delta}) - \beta x'(t_0-\bar{\delta})]\,[y(t_0) - \tfrac{1}{2}x(t_0-\bar{\delta})] + \tfrac{1}{2}[x'(t_1-\bar{\delta}) + \beta x(t_1-\bar{\delta})]\,[x'(t_1-\bar{\delta})] - \tfrac{1}{2}[x'(t_0-\bar{\delta}) - \beta x(t_0-\bar{\delta})]\,[x'(t_0-\bar{\delta})] \right\}, \tag{27}$$

$$b = \lim_{N\to\infty} b_N = \frac{1}{2\beta R(0)} \left\{ \int_I [y(t) - x(t-\bar{\delta})] \cdot [\beta^2 x''(t-\bar{\delta}) - x''''(t-\bar{\delta})]\, dt + [\beta x''(t_0-\bar{\delta}) - x'''(t_0-\bar{\delta})]\,[y(t_0) - x(t_0-\bar{\delta})] + [\beta x''(t_1-\bar{\delta}) + x'''(t_1-\bar{\delta})]\,[y(t_1) - x(t_1-\bar{\delta})] - \int_I dt[(\beta x'(t-\bar{\delta}))^2 - x'(t-\bar{\delta})x'''(t-\bar{\delta})] + [x''(t_0-\bar{\delta}) - \beta x'(t_0-\bar{\delta})]x'(t_0-\bar{\delta}) - [x''(t_1-\bar{\delta}) + \beta x'(t_1-\bar{\delta})]x'(t_1-\bar{\delta}) \right\}. \tag{28}$$

It is also found that $c(\delta)$ is a piecewise continuous function of δ.

In obtaining these results it is assumed that the second derivative of $x(t)$ is continuous and that the quantity $[-\beta^2 x'(t-\bar{\delta}) + x'''(t-\bar{\delta})]$ appearing in the integral in (27) is of bounded variation and continuous except at a finite number of points. Similar assumptions are made on

$[\beta^2 x''(t-\bar{\delta}) - x''''(t-\bar{\delta})]$ in the expression for b and a similar term in the expression for $c(\delta)$. One requires that the fifth derivative of $x(t)$ be of bounded variation and be continuous except at a finite number of points. These assumptions allow Sturm-Liouville expansions of the integrands involved in (27) and (28) which, using the orthonormality of the $\{\varphi_k(t)\}$, yield the proof of (27) and (28). Thus a sufficient condition for the validity of these results is that the second derivative of $x(t)$ be continuous and that the fifth derivative of $x(t)$ be of bounded variation and continuous except at a finite number of points.

The desired test statistic, l, is obtained by substituting (26), (27) and (28) in (25) with $x^{(i)}(t)$, $i = 1$ or 2, replacing $x(t)$ in the appropriate places. The resulting expression is nonlinear in $y(t)$ and rather lengthy and will not be explicitly stated here. Observe however from (27) that if phase reversal signaling is used, that is $x^{(2)}(t) = -x^{(1)}(t)$, the quantity $(a^{(1)})^2 - (a^{(2)})^2$ will be linear in $y(t)$. Furthermore under this condition, to first order in λ^2, $\ln l$ will be linear in $y(t)$. Therefore for phase reversal signaling the receiver correlates $y(t)$ with a signal related to $x(t-\delta)$ and its derivatives evaluated at $\delta = \bar{\delta}$.

The "white noise case" will now be obtained as a limit of the exponential covariance case. Setting $R(0) = N_0\beta/2$ and letting $\beta \to \infty$ in the expressions for Q, a and b one obtains

$$\begin{aligned}\ln l \sim \Bigg\{ & \frac{1}{N_0}\int_I dt\left[y(t) - \frac{x(t-\bar{\delta})}{2}\right] x(t-\bar{\delta}) \\ & + \frac{\lambda^2}{2N_0^2}\left(-\int_I dt\left[y(t) - \frac{x(t-\bar{\delta})}{2}\right] x'(t-\bar{\delta})\right)^2 \\ & + \frac{\lambda^2}{2N_0}\left(\int_I dt[y(t) - x(t-\bar{\delta})]\, x''(t-\bar{\delta}) \right. \\ & \left. - \int_I dt[x'(t-\bar{\delta})]^2\right)\Bigg\}_{x=x(2)}^{x=x(1)}\end{aligned} \tag{29}$$

In the above expression the notation $\{\ \}_{x=x(2)}^{x=x(1)}$ is used to indicate that the expression in braces is evaluated with $x(t) = x^{(2)}(t)$ and the result is subtracted from the result which obtains when $x(t) = x^{(1)}(t)$. For phase reversal signalling, $x^{(1)}(t) = -x^{(2)}(t) = x(t)$, (29) becomes

$$\begin{aligned}\ln l \sim \frac{2}{N_0}\int_I dt\, y(t) \Bigg\{ & x(t-\bar{\delta}) - \frac{\lambda^2}{2}\Bigg[\frac{1}{N_0}\,(x^2(t_1-\bar{\delta}) \\ & - x^2(t_0-\bar{\delta}))\, x'(t-\bar{\delta}) - x''(-\bar{\delta})\Bigg]\Bigg\}.\end{aligned} \tag{30}$$

Let us examine the correlation waveform appearing inside the braces in (30). The first term represents the receiver input when the random delay is equal to its mean value, $\delta = \bar{\delta}$. If λ is set equal to zero the remaining terms vanish and one obtains the standard result which is that the receiver input should be correlated with the signal portion of the input, $x(t - \bar{\delta})$. The remaining terms inside the braces in (30) are the perturbations introduced by the random delay.

Let us expand $x(t - \delta)$ in the Taylor series with remainder

$$x(t - \delta) = x(t - \bar{\delta}) + (\delta - \bar{\delta})[-x'(t - \bar{\delta})] + [(\delta - \bar{\delta})^2/2!][x''(t - \bar{\delta})] + \mathcal{R}(\delta - \bar{\delta}, t).$$

If $x'''(t - \bar{\delta})$ is continuous and the moments of δ satisfy

$$E_\delta \mid (\delta - \bar{\delta})^k \mid \leqq h^k \lambda^k,$$

then for small λ, $x(t - \bar{\delta}) + (\lambda^2/2)x''(t - \bar{\delta})$ is the principal part of $E_\delta x(t - \delta)$. Thus, part of the correlation waveform is essentially the expected value of the received signal. The other term in the correlation waveform involves $x'(t - \bar{\delta})$, which is the slope of the received signal when the delay is $\bar{\delta}$. The weight attached to it is proportional to the difference in the squared values of the received signal at time t_0 and at time t_1 when $\delta = \bar{\delta}$. As yet no physical significance has been found for this term.

IV. CONCLUSIONS

An approximation has been obtained for a test statistic that minimizes the error probability of a binary communication system which operates over a linear channel, with a small random delay, in the presence of gaussian noise of covariance $R(0)e^{-\beta|\tau|}$. For the case of phase reversal signaling, the statistic, which in general is nonlinear, is a linear functional of the receiver input. Treating the "white noise" case as a limit of the exponential covariance case, the test statistic is expressed as a cross-correlation operation. The waveform with which the input is correlated is related to the expected value of the received signal plus a term proportional to the slope of the received signal when the delay is equal to its mean value.

V. ACKNOWLEDGMENT

The author wishes to express appreciation to Dr. A. R. Cohen of New York University, Dr. I. Jacobs of Bell Telephone Laboratories and the referees for their valuable suggestions.

APPENDIX

The Evaluation of Terms Arising in the Approximation to the Likelihood Function

The convergence asserted in (22) will now be established and the structure of the terms appearing in the likelihood function, (25), will be exhibited.

First let us note the following property associated with an expansion in terms of the eigenfunctions of a Sturm-Liouville system.

Let $g(t)$ be a piecewise continuous function which is of bounded variation and let $z(t)$ be a piecewise continuous function. If the sets $\{g_k\}$ and $\{z_k\}$ are the expansion coefficients of $g(t)$ and $z(t)$ in terms of the above eigenfunctions, then

$$\int_I dt\, g(t)z(t) = \sum_{k=1}^{\infty} g_k z_k . \tag{31}$$

This expression is obtained by noting that the series expansion of $g(t)$ converges uniformly except at a finite number of points.

One considers the following integral which is suggested by (16),

$$K(\delta) \triangleq \frac{1}{2\beta R(0)} \int_I dt \left[y(t) - \frac{x(t-\delta)}{2} \right] \left(\beta^2 - \frac{\partial^2}{\partial t^2} \right) x(t-\delta), \tag{32}$$

where one requires that $[\beta^2 - (\partial^2/\partial t^2)]x(t-\delta)$ be of bounded variation and piecewise continuous. Let

$$g(t-\delta) \triangleq \left(\beta^2 - \frac{\partial^2}{\partial t^2} \right) x(t-\delta) = \beta^2 x(t-\delta) - x''(t-\delta). \tag{33}$$

$g(t)$ can be expanded in a series in terms of the eigenfunctions of (6). This series converges uniformly to $g(t)$ where $g(t)$ is continuous and to $\frac{1}{2}[g(u_+) + g(u_-)]$ at points where $g(t)$ is not continuous. Thus

$$g(t-\delta) = \sum_{k=1}^{\infty} g_k \varphi_k(t), \tag{34}$$

where

$$g_k = \int_I du\, \varphi_k(u)\, g(t-\delta).$$

By integration by parts and utilizing (6) one can show that

$$\begin{aligned} g_k = (\beta^2 + \gamma_k^2)\psi_k &+ \varphi_k(t_0)[x'(t_0-\delta) - \beta x(t_0-\delta)] \\ &- \varphi_k(t_1)[x'(t_1-\delta) + \beta x(t_1-\delta)], \end{aligned} \tag{35}$$

where

ψ_k is the kth expansion coefficient of $x(t-\delta)$ in terms of the $\{\varphi_k(t)\}$ set,

$(\beta^2+\gamma_k^2)=[2\beta R(0)/\sigma_k^2]$,

σ_k^2 is the kth eigenfunction of (6).

Substituting the expansion for $g(t)$ in (32) and using (35) one finds

$$K(\delta) = \frac{1}{2\beta R(0)}\left\{2\beta R(0)\sum_{k=1}^{\infty}\left(y_k-\frac{\psi_k}{2}\right)\frac{\psi_k}{\sigma_k^2} + [x'(t_0-\delta)-\beta x(t_0-\delta)]\sum_{k=1}^{\infty}\left(y_k-\frac{\psi_k}{2}\right)\varphi_k(t_0) - [x'(t_1-\delta)+\beta x(t_1-\delta)]\sum_{k=1}^{\infty}\left(y_k-\frac{\psi_k}{2}\right)\varphi_k(t_1)\right\}. \tag{36}$$

Noting that $\sum_{k=1}^{\infty}\left(y_k-\frac{\psi_k}{2}\right)\frac{\psi_k}{\sigma_k^2}$ equals $\lim_{N\to\infty} Q_N[\delta]$ one finds

$$\lim_{E\to\infty} Q_N[\delta] = K(\delta) - \frac{[x'(t_0-\delta)-\beta x(t_0-\delta)]}{2\beta R(0)}\sum_{k=1}^{\infty}\left(y_k-\frac{\psi_k}{2}\right)\varphi_k(t_0) + \frac{[x'(t_1-\delta)+\beta x(t_1-\delta)]}{2\beta R(0)}\sum_{k=1}^{\infty}\left(y_k-\frac{\psi_k}{2}\right)\varphi_k(t_1) \tag{37}$$

One now can proceed to evaluate $a=\lim_{N=\infty} a_N$.

From the definitions following (17) and (10) one has

$$a_N = Q_N'[\bar{\delta}] = \sum_{i=1}^{N}\frac{(y_k-\psi_k(\bar{\delta}))}{\sigma_k^2}x_k'(\bar{\delta}). \tag{38}$$

Consider

$$K_1(\delta)\triangleq\int_I dt\left[y(t)-\frac{x(t-\delta)}{2}\right][-\beta^2x'(t-\delta)+x'''(t-\delta)]. \tag{39}$$

Let us assume that $[-\beta^2x'(t-\delta)+x'''(t-\delta)]$ is piecewise continuous, $[\beta^2x(t-\delta)-x''(t-\delta)]$ is continuous and that both are of bounded variation. Define

$$\alpha_k = \int_I dt\,\varphi_k(t)\,[-\beta^2x'(t-\delta)+x'''(t-\delta)]. \tag{40}$$

Then under these assumptions and using (33) and (34) one finds

$$\alpha_k = g_k'(\delta). \tag{41}$$

Applying (31) to $K_1(\delta)$ gives

$$K_1(\delta) = \sum_{k=1}^{\infty} \left(y_k - \frac{\psi_k(\delta)}{2} \right) \alpha_k . \tag{42}$$

Using (41), (42), (35) and (31) yields

$$\begin{aligned} \sum_{k=1}^{\infty} \left(\frac{y_k - \psi_k(\delta)}{\sigma_k^2} \right) \psi_k'(\delta) &= \frac{K_1(\delta)}{2\beta R(0)} \\ &- \frac{1}{2\beta R(0)} [x''(t_1 - \delta) + \beta x'(t_1 - \delta)] \left[y(t_1) - \frac{x(t_1 - \delta)}{2} \right] \\ &+ \frac{1}{2\beta R(0)} [x''(t_0 - \delta) - \beta x'(t_0 - \delta)] \left[y(t_0) - \frac{x(t_0 - \delta)}{2} \right] \\ &- \tfrac{1}{2} [x'(t_0 - \delta) - \beta x(t_0 - \delta)] x'(t_0 - \delta) \\ &\quad + \tfrac{1}{2} [x'(t_1 - \delta) + \beta x(t_1 - \delta)] x'(t_1 - \delta). \end{aligned} \tag{43}$$

The left hand side of (43) evaluated at $\delta = \bar{\delta}$ is $\lim_{N \to \infty} a_N$.

Proceeding in a similar manner the $\lim_{N \to \infty} b_N$ and $\lim_{N \to \infty} c_N(\delta)$ are obtained.

REFERENCES

1. Helstrom, Carl W., *Statistical Theory of Signal Detection*, Chapter IX, Pergamon Press, 1960.
2. Brown, W. M., and Palermo, C. J., System Performance In the Presence of Stochastic Delays, IRE Trans. Info. Theory, *IT-8*, Sept. 1962, No. 5.
3. Balakrishnan, A. V. On the Problem of Time Jitter in Sampling, IRE Trans. Info. Theory, *IT-8*, April 1962, No. 3.
4. Grenander, U., Stochastic Processes and Statistical Inference, Arkiv for Matematik, Band 1 nr 17.
5. Belayev Yu. K., *Continuity and Hölder's Conditions for Sample Functions of Stationary Gaussian Processes*, Fourth Berkeley Symposium on Mathematical Statistics and Probability (1960), University of California Press, 1961.
6. Hobson, E. W., *The Theory of Functions of a Real Variable, 2*, p. 772, Dover Publications, 1956.

On the Accuracy of Loss Estimates

By A. DESCLOUX

(Manuscript received March 30, 1965)

In telephone traffic studies, the observed proportion of unsuccessful attempts over a given time interval is one of the measures commonly used to evaluate the grade of service provided by trunk groups. This paper deals with the derivation of an approximate formula for the variance of this estimate when (i) call arrivals constitute a Poisson process, (ii) service times are independent of each other and identically distributed according to a negative exponential law, and (iii) calls placed when all trunks are busy are either cancelled or sent via some alternate route (loss system). Comparison of simulation data with numerical values computed by means of this formula indicates that the latter is accurate enough for practical purposes.

The observed proportion of time during which all trunks are occupied is also an estimate of the grade of service (defined as the probability that a call will be lost or overflow). It is shown here, that for relatively small loads, this estimate has a smaller variance than the observed proportion of lost or rerouted calls. However, as the load is increased, the inequality between the variances of these two estimates is reversed, the cross-over occurring in the vicinity of the point where the load (in erlangs) is equal to the number of trunks.

For a given observation period, the proportion of time when all trunks are busy can be either measured exactly or estimated by "switch-counting." In the latter case, the group is scanned at regular intervals and one observes, for each scan, whether all trunks are occupied or not. The average number of scans which indicate that all trunks are busy is an estimate of this proportion and is, a fortiori, an estimate of the probability of loss. The effect of the scanning rate on the accuracy of this estimate is investigated.

I. INTRODUCTION

In this paper, we shall consider the simplest type of loss systems, namely full availability groups with Poisson inputs, negative exponential

service times, and cancellation or rerouting of calls finding all trunks occupied. Under these assumptions, we shall obtain an approximate expression for the variance of the measured call congestion, the latter being defined here as the proportion of calls which either are lost or overflow to some alternate group during a given time interval. In the derivation of this expression, use is made of the classical formula for the propagation of errors, whose computation requires the evaluation of the first- and second-order moments of the joint distribution of the number of offered and the number of overflow calls. Since the marginal means and variances of this distribution are known (cf. Ref. 1), the emphasis is placed here on the derivation of the covariance. Computed values of the variance of the measured call congestion are shown to be in good agreement with simulation results (cf. Figs. 1–5).* Charts giving the variance of this ratio for group sizes up to 50 and offered loads (in erlangs) per trunk of 0.1 to 10, are reproduced in Figs. 6–8.*

For a given observation period, the measured call congestion and the observed proportion of time when all servers are busy — here called measured time congestion — provide us with two estimates of the probability that a call will either be lost or overflow to some alternate route. Neither of these two estimates has a uniformly smaller variance than the other. Actually, the following holds: for relatively small loads, the measured time congestion has a smaller variance than the measured call congestion. However, as the load is increased, the direction of the inequality is reversed, the cross-over occurring in the vicinity of the point where the load (in erlangs) is equal to the number of trunks. Thus, the measured time congestion is not always a more efficient estimator of the probability of loss than the measured call congestion (cf. Figs. 9 and 10).

(In what follows, the terms measured time congestion and measured call congestion will always be abbreviated to time and call congestion, respectively. These terms will refer throughout to measurements performed over a given time interval.)

For a given observation period, time congestion can be either measured exactly or estimated by switchcounting. In the latter case, the group is scanned at regular intervals and one observes, for each scan, whether all trunks are busy or not. The proportion of scans which indicate that all trunks are busy is an unbiased estimate of time congestion and is, a fortiori, an estimate of the probability of loss. Clearly, the variance of this estimate increases as the scanning rate decreases. The loss of accuracy due to scanning is depicted in Fig. 11.

* See illustrations placed later in this article.

Under the present assumptions, loss probabilities can also be estimated from carried loads measured either exactly or by scanning. For offered loads (in erlangs) falling short of the number of trunks, simulation has shown that such estimates have smaller variances than the estimates mentioned earlier. This fact is illustrated in Fig. 12. The effect of scanning on the accuracy of loss estimates based on load measurements is sketched in the same figure.

Finally, we note that estimates of loss probabilities based either on observed call congestion or on carried load measurements are biased, respectively, downwards and upwards. These biases are, however, quite small and likely to be negligible in most situations of practical interest.

II. THE COVARIANCE FUNCTION

Consider a group of c trunks which operate in parallel and are fully available to all requests. If a call is placed when a trunk is free, service starts immediately; otherwise the request is either cancelled or routed via some alternate group (loss system). Regarding the input and the service durations, the following assumptions will be made:

(*i*) The time intervals between successive service demands (whether successful or not) are independent of each other and have a common negative exponential distribution with mean equal to $1/a$ (Poisson input).

(*ii*) The service times are independent of each other and have a common negative exponential distribution whose mean will be taken throughout as the unit of time (a is therefore the offered load in erlangs).

The following notation will be used:

$$\begin{aligned} N(t) &= \text{number of busy trunks at time } t, \\ R(t) &= \text{total number of requests offered during } (0,t), \\ S(t) &= \text{total number of unsuccessful requests during } (0,t), \\ P(t,n,r,s) &= \Pr\,[N(t) = n,\, R(t) = r,\, S(t) = s]. \end{aligned}$$

From the definition of $P(\cdot,\cdot,\cdot,\cdot)$, it follows that:

$$\begin{aligned} P(t,n,r,s) &= 0 \text{ for } n > c, \qquad (t \geqq 0) \\ P(t,n,r,s) &= 0 \text{ for } s > r, \qquad (t \geqq 0) \\ P(0,n,r,s) &= 0 \text{ for } r \geqq 1. \end{aligned}$$

It will be convenient to extend the definition of $P(\cdot,\cdot,\cdot,\cdot)$ and to adopt the convention:

pansion of F_1 in powers of x and the summation on the right of (6) is the coefficient of x^c in that same expansion. Therefore

$$G(w,y,z) = -\frac{(c+1)\tau_{c+1}(w,y) - ayz\cdot\tau_c(w,y)}{(c+1)\sigma_{c+1}(w,y) - ayz\cdot\sigma_c(w,y)} \tag{7}$$

where the τ's and σ's are defined by

$$e^{ayz}(1-x)^{a(y-1)-w} = \sum_0^\infty \sigma_n(w,y)x^n \tag{8}$$

$$Ke^{(x-1)ay}(1-x)^{a(y-1)-w}\int_0^{1-x} u^{a(1-y)+w-1}e^{a(y-1)u}\,du = \sum_0^\infty \tau_n(w,y)\,x^n. \tag{9}$$

We note that if $c_n(\cdot,\alpha)$ and $L_n^{(\alpha)}(\cdot)$ stand respectively for the Poisson-Charlier and the Laguerre polynomials of degree n and parameter α; i.e., if (Ref. 2, pp. 34–35 and 101, and Ref. 3, p. 26)

$$c_n(t,\alpha) = \alpha^{n/2}(n!)^{\frac{1}{2}}\sum_{\nu=0}^{n}(-1)^{n-\nu}\binom{n}{\nu}\alpha^{-\nu}t(t-1)\cdots(t-\nu+1)$$

and

$$L_n^{(\alpha)}(t) = \sum_{\nu=0}^{n}\binom{n+\alpha}{n+\nu}\frac{(-t)^\nu}{\nu!}.$$

then:

$$e^{\alpha x}(1-x)^t = \sum_0^\infty c_n(t,\alpha)\,[(\alpha x)^n/n!]$$
$$= \sum_0^\infty (-1)^n L_n^{(t-n)}(\alpha)x^n$$

and

$$\sigma_n(w,y) = [(ay)^n/n!]\,c_n[a(y-1)-w,ay]$$
$$= (-1)^n L_n^{[a(y-1)-w-n]}(ay).$$

[The relation between the σ's and Kosten's φ-functions (cf. Ref. 1) is readily found. Indeed, by definition

$$e^{a(x-1)}(1-x)^z = \sum_0^\infty \varphi_n{}^z x^n$$

so that

$$\sigma_n(w,y) \equiv e^{ay}\varphi_n{}^{a(y-1)-w}.]$$

For later purposes we note that:

$$\sigma_{m+1}(w,y) = \sigma_{m+1}(w+1,y) - \sigma_m(w+1,y), \qquad (m = 0, 1, \cdots) \tag{10}$$

$$[w - a(y-1)]\sigma_m(w+1,y) = (m+1)\sigma_{m+1}(w,y) - ay\cdot\sigma_m(w,y), \qquad (m = 0, 1, \cdots) \tag{11}$$

$$\sum_0^c \sigma_n(w,y) = \sigma_c(w+1,y). \tag{12}$$

These identities are immediate consequences of recurrence relations known to hold for the corresponding Laguerre polynomials (cf. Ref. 2, p. 98).

Substituting (7), (8) and (9) into (5) yields:

$$F_1(w,x,y,z) = -\frac{(c+1)\tau_{c+1}(w,y) - ayz\cdot\tau_c(w,y)}{(c+1)\sigma_{c+1}(w,y) - ayz\cdot\sigma_c(w,y)} \sum \sigma_n(w,y)x^n + \sum \tau_n(w,y)x^n. \tag{13}$$

We can now obtain the generating function, $F(\cdot,\cdot,\cdot)$, of the Laplace transforms of the joint probabilities Pr $[R(t) = r, S(t) = s]$, $r,s = 0, 1, \cdots$, by deleting from (12) all terms of degree higher than c in x and then setting x equal to 1. If we perform these operations and then make use of (12), we find that:

$$F(w,y,z) = -\frac{(c+1)\tau_{c+1}(w,y) - ayz\cdot\tau_c(w,y)}{(c+1)\sigma_{c+1}(w,y) - ayz\cdot\sigma_c(w,y)}\cdot\sigma_c(w+1,y) + \sum_0^c \tau_n(w,y). \tag{14}$$

The moments of the joint distribution of R and S can now be obtained by evaluating the derivatives of $F(w,\cdot,\cdot)$ for $y = z = 1$.

Differentiating (7) with respect to z and making use of (11), we find

$$\left.\frac{\partial F}{\partial z}\right|_{z=1} = \frac{ay\cdot\tau_c(w,y)}{w - a(y-1)} - \frac{ay\cdot\sigma_c(w,y)[(c+1)\tau_{c+1}(w,y) - ay\cdot\tau_c(w,y)]}{[w - a(y-1)]^2\,\sigma_c(w+1,y)}. \tag{15}$$

In particular, for $y = 1$, we have the well-known result

$$\left.\frac{\partial F}{\partial z}\right|_{y=z=1} = \frac{aE_{1,c}(a)}{w^2}$$

so that

$$ES(t) = at\,E_{1,c}(a) \tag{16}$$

where $E_{1,c}(a)$ is Erlang's loss formula.

Taking the derivative of (15) with respect to y and then setting y equal to 1, yields:

$$\begin{aligned}\left.\frac{\partial^2 F}{\partial y \partial z}\right|_{y=z=1} &= \frac{a}{w}\,\tau_c(w)\left[1 + \frac{a}{w} + \frac{a\sigma_c(w)}{w\sigma_c(w+1)}\right] \\ &\quad + \frac{a}{w}\frac{\partial}{\partial y}\,\tau_c(w,y)\mid_{y=1} \\ &\quad - \frac{a\sigma_c(w)}{w^2\sigma_c(w+1)}\left[(c+1)\,\frac{\partial}{\partial y}\,\tau_{c+1}(w,y)\right. \\ &\quad \left. - a\,\frac{\partial}{\partial y}\,\tau_c(w,y)\right]_{y=1}\end{aligned} \tag{17}$$

where

$$\sigma_m(w) \equiv \sigma_m(w,1) \text{ and } \tau_m(w) \equiv \tau_m(w,1), \qquad (m = 0, 1, \cdots).$$

To determine the derivatives of $\tau_c(w,y)$ and $\tau_{c+1}(w,y)$ with respect to y, consider the generating function

$$H(w,x,y) \equiv Ke^{(x-1)ay}\,(1-x)^{a(y-1)-w}\int_0^{1-x} u^{a(1-y)+w-1}\,e^{a(y-1)u}\,du.$$

Differentiating this expression with respect to y and then setting y equal to 1, we find that

$$\begin{aligned}\frac{\partial}{\partial y}\,H(w,x,y)\mid_{y=1} &= \sum_0^\infty \frac{\partial}{\partial y}\,\tau_m(w,y)\,x^m\mid_{y=1} \\ &= K\cdot\frac{ae^{a(x-1)}}{w^2(1+w)}\,(1+wx)\end{aligned}$$

and, therefore:

$$\frac{\partial}{\partial y}\,\tau_m(w,y)\mid_{y=1} = K\,\frac{ae^{-a}}{(m-1)!w(1+w)}\left(1+\frac{a}{mw}\right), \qquad (m = 1, \cdots). \tag{18}$$

Since

$$\tau_m(w) = \frac{Ke^{-a}a^m}{w\cdot m!}$$

we obtain, upon taking (18) into account:

$$\mathfrak{L}\{E[R(t)\cdot S(t)]\} = 2K\frac{e^{-a}a^{c+2}}{c!w^3} + K\frac{e^{-a}a^{c+1}}{c!w^2}\cdot\frac{\sigma_{w+2}(c)}{\sigma_{w+1}(c)} \tag{19}$$

where the notation $\mathfrak{L}\{f\}$ is used to designate the Laplace transform of f.

Since $ER(t) = at$ and $ES(t) = at\,E_{1,c}(a)$, we also have:

$$\mathfrak{L}\{\text{Cov}\,[R(t), S(t)]\} = K\frac{e^{-a}a^{c+1}}{c!w^2}\cdot\frac{\sigma_c(w+2)}{\sigma_c(w+1)} \tag{20}$$

where Cov $[R(t),S(t)]$ stands for the covariance between $R(t)$ and $S(t)$.

For $y = 1$, $m = c$ and w replaced by $w + 1$, (11) reduces to

$$(w+1)\sigma_c(w+2) = (c+1)\sigma_{c+1}(w+1) - a\sigma_c(w+1),$$

and (20) can be rewritten as follows:

$$\mathfrak{L}\{\text{Cov}\,[R(t), S(t)]\} = K\frac{e^{-a}a^{c+1}}{c!w^2(w+1)\sigma_c(w+1)}\,[(c+1)\sigma_{c+1}(w+1) - a\sigma_c(w+1)]. \tag{21}$$

Let w_i, $i = 1, \cdots, c$ be the c roots of $\sigma_c(w+1)$. It is well known that these roots are simple, smaller than -1 and at least one unit apart. Then expanding (21) in partial fractions and making use of the relation $(c+1)\sigma_{c+1}(0) - a\sigma_c(0) = 0$, which is (11) for $y = 1$, $w = 0$, we find:

$$\begin{aligned}\mathfrak{L}\{\text{Cov}\,[R(t), S(t)]\} &= K\frac{e^{-a}a^{c+1}}{c!}\left[\frac{(c+1)\sigma_{c+1}(1) - a\sigma_c(1)}{w^2\,\sigma_c(1)}\right.\\ &\quad + \frac{c+1}{w}\frac{\partial}{\partial w}\frac{\sigma_{c+1}(w+1)}{\sigma_c(w+1)}\bigg|_{w=0} - \frac{(c+1)\sigma_{c+1}(1) - a\sigma_c(1)}{w\sigma_c(1)}\\ &\quad \left. + (c+1)!\sum_{i=1}^{c}\frac{\sigma_{c+1}(w_i+1)}{w_i^2(1+w_i)(w-w_i)\prod_{j\neq i}(w_i-w_j)}\right]\end{aligned}$$

and the covariance between R and S is, therefore, given by

$$\begin{aligned}\text{Cov}\,[R(t), S(t)] &= K\frac{e^{-a}a^{c+1}}{c!}\left[\frac{(c+1)\sigma_{c+1}(1) - a\sigma_c(1)}{\sigma_c(1)}\cdot t\right.\\ &\quad + (c+1)\frac{\partial}{\partial w}\frac{\sigma_{c+1}(w+1)}{\sigma_c(w+1)}\bigg|_{w=0} - \frac{(c+1)\sigma_{c+1}(1) - a\sigma_c(1)}{\sigma_c(1)}\\ &\quad \left. + (c+1)!\sum_{i=1}^{c}\frac{\sigma_{c+1}(w_i+1)e^{w_i t}}{w_i^2(1+w_i)\prod_{j\neq i}(w_i-w_j)}\right].\end{aligned} \tag{22}$$

To determine explicitly the derivative appearing in (22), let us consider Cov $[R(t),S(t)]$ for small values of t. Writing $P(t,r,s)$ for the (equilibrium) probability that, during a time interval of length t, r requests arrived and that, among these r requests, s of them found all the trunks busy, we have (t small):

$$\begin{aligned} P(t,0,0) &= 1 - at + o(t) \\ P(t,1,0) &= at\,[1 - E_{1,c}(a)] + o(t) \\ P(t,1,1) &= at\,E_{1,c}(a) + o(t) \\ P(t,0,s) &= 0, \qquad (s \geqq 1) \\ P(t,r,0) &= o(t), \qquad (r > 1) \end{aligned}$$

and

$$0 \leqq \sum_{r,s=2}^{\infty} sP(t,r,s) < \sum_{r,s=2}^{\infty} rP(t,r,s) < \sum_{r,s=2}^{\infty} rsP(t,r,s)$$

$$< \sum_{r=2}^{\infty} r^2 e^{-at} \frac{(at)^r}{r!} = o(t).$$

Hence Cov $[R(t),S(t)] = at\,E_{1,c}(a) + o(t)$. Letting t tend to 0 in (22), we find that

$$(c+1) \frac{\partial}{\partial w} \frac{\sigma_{c+1}(w+1)}{\sigma_c(w+1)}\bigg|_{w=0}$$

$$= \frac{(c+1)\sigma_{c+1}(1) - a\sigma_c(1)}{\sigma_c(1)}$$

$$- (c+1)! \sum_1^c \frac{\sigma_{c+1}(w_i+1)}{w_i^2(1+w_i) \prod_{j \neq i} (w_i - w_j)}.$$

Substituting this expression in (22) yields

$$\text{Cov}\,[R(t), S(t)] = K \frac{e^{-a} a^{c+1}}{c!} \left[\frac{(c+1)\sigma_{c+1}(1) - a\sigma_c(1)}{\sigma_c(1)} t \right.$$

$$- (c+1)! \sum_1^c \frac{\sigma_{c+1}(w_i+1)}{w_i^2(1+w_i) \prod_{j \neq i} (w_i - w_j)} \qquad (22')$$

$$\left. + (c+1)! \sum_1^c \frac{\sigma_{c+1}(w_i+1) e^{w_i t}}{w_i^2(1+w_i) \prod_{j \neq i} (w_i - w_j)} \right].$$

We shall now determine the constant term appearing in (22′). To this end, we note that

$$\sigma_c(w+1) = \frac{1}{c!}\prod_{i=1}^{c}(w-w_i)$$

and, therefore

$$\frac{\sigma_{c+1}(w+1)}{w(1+w)\sigma_c(w+1)} = \frac{1}{w}\frac{\sigma_{c+1}(1)}{\sigma_c(1)} - \frac{1}{1+w}\frac{\sigma_{c+1}(0)}{\sigma_c(0)} + c!\sum_{i=1}^{c}\frac{1}{(w-w_i)}\frac{\sigma_{c+1}(w_i+1)}{w_i(1+w_i)\prod_{j\neq i}(w_i-w_j)}$$

Hence, for $w = 0$, we have:

$$\begin{aligned}(c+1)!\sum_{i=1}^{c}\frac{\sigma_{c+1}(w_i+1)}{w_i^2(1+w_i)\prod_{j\neq i}(w_i-w_j)} &= (c+1)\lim_{w\to 0}\left[\frac{1}{w}\frac{\sigma_{c+1}(1)}{\sigma_c(1)} - \frac{1}{1+w}\frac{\sigma_{c+1}(0)}{\sigma_c(0)}\right. \\ &\quad \left. - \frac{\sigma_{c+1}(w+1)}{w(1+w)\sigma_c(w+1)}\right] \\ &= -a. \end{aligned} \tag{23}$$

Furthermore, for $w = 1 + w_i$, $y = 1$ and $m = c$, (11) yields:

$$(w_i+1)\sigma_c(w_i+2) = (c+1)\sigma_{c+1}(w_i+1). \tag{24}$$

We also note that the c roots of $\sigma_c(w+2)$ are $w_i - 1$, $i = 1, \cdots, c$, so that

$$\sigma_c(w+2) = \frac{1}{c!}\prod_{1}^{c}(w-w_j+1). \tag{25}$$

Hence, combining (24) and (25), we have:

$$\frac{(c+1)!\sigma_{c+1}(w_i+1)}{1+w_i} = \prod_{j=1}^{c}(w_i-w_j+1). \tag{26}$$

Using (23), (26) and the relations

$$Ke^{-a}a^{c+1} = c!\,aE_{1,c}(a)$$

$$(c+1)\sigma_{c+1}(1) = \sigma_c(1)\{c+1+aE_{1,c}(a)\}$$

(22′) can be simplified as follows:

$$\text{Cov}\,[R(t), S(t)] = aE_{1,c}(a)\left[\{c + 1 - a[1 - E_{1,c}(a)]\}t + a - \sum_{i=1}^{c} \frac{e^{w_i t}}{w_i^2} \prod_{j \neq i} \left(1 + \frac{1}{w_i - w_j}\right)\right]. \tag{27}$$

Since, as pointed out above:

$$\max_{1 \leqq i \leqq c} w_i < -1$$

$$\min_{\substack{1 \leqq i,j \leqq c \\ i \neq j}} |w_i - w_j| > 1$$

we have:

$$\sum_{i=1}^{c} \frac{e^{w_i t}}{w_i^2} \prod_{j \neq i} \left(1 + \frac{1}{w_i - w_j}\right) > 0$$

and

$$\frac{\partial}{\partial t} \sum_{i=1}^{c} \frac{e^{w_i t}}{w_i^2} \prod_{j \neq i} \left(1 + \frac{1}{w_i - w_j}\right) < 0.$$

Hence we have the following inequalities [use is also made here of (23)]:

$$\begin{aligned} &aE_{1,c}(a)\{c + 1 - a[1 - E_{1,c}(a)]\}t \\ &\quad < \text{Cov}\,[R(t), S(t)] \\ &\quad < aE_{1,c}(a)\,[\{c + 1 - a[1 - E_{1,c}(a)]\}t + a] \end{aligned}$$

and, for large values of $t(>0)$:

$$\text{Cov}\,[R(t), S(t)] = aE_{1,c}(a)\,[\{c + 1 - a[1 - E_{1,c}(a)]\}t + a] + o(e^{-t}). \tag{28}$$

III. VARIANCE OF CALL CONGESTION

In the preceding section, exact and asymptotic formulas were obtained for the covariance between the number of offered calls, $R(t)$, and the number of overflow calls, $S(t)$, during a time interval of length t. These expressions can now be combined with known formulas for the means and variances of $R(t)$ and $S(t)$ to obtain an approximate expression for the variance of call congestion. Indeed, according to the classical formula for the propagation of errors, we have:

$$\begin{aligned}
&\text{Var } [S(t)/R(t)] \\
&\quad \sim \{1/ER(t)]^2 \text{ Var } [S(t)] + \{ES(t)/[ER(t)]^2\}^2 \text{ Var } [R(t)] \\
&\qquad - 2\{ES(t)/[ER(t)]^3\} \text{ Cov } [R(t),S(t)] \\
&\quad = (1/at)^2 \text{ Var } [S(t)] + E_{1,c}{}^2(a)/(at) \\
&\qquad - 2[E_{1,c}(a)/(at)^2] \text{ Cov } [R(t),S(t)] \\
&\quad < (1/at)^2 \text{ Var } [S(t)]
\end{aligned} \tag{29}$$

where, assuming t large:

$$\text{Var } [S(t)] \sim at\, E_{1,c}(a) \left[1 + 2a \frac{\partial}{\partial w} \frac{\sigma_c(w)}{\sigma_c(w+1)}\right]_{w=0} \tag{30}$$

and

$$\text{Cov } [R(t),S(t)] \sim at\, E_{1,c}(a)[c + 1 - a + aE_{1,c}(a)]. \tag{31}$$

We note that Var $[S(t)/R(t)]$ is, asymptotically, of the form k/t, where k depends only on a and c.

The exact and asymptotic expressions for Var $[S(t)]$ were first derived by Kosten, Manning and Garwood.[1] These formulas can be obtained in a straightforward manner from the generating function (14) with y set equal to 1. The asymptotic expression (30), however, is rather involved and its use can be avoided as follows. Indeed, we note that, under the present assumptions, the instants at which overflows occur constitute a renewal process (i.e., the intervals between any pair of consecutive overflows are independent of each other and have the same distribution). Then using Smith's extension of a result due to Feller (cf. Ref. 4, pp. 296–298 and Ref. 5, pp. 30–33), we have:

$$\text{Var } [S(t)] \sim [\mu_2(c) - \mu_1{}^2(c)]t/\mu_1{}^3(c) \tag{32}$$

where $\mu_n(c)$ is the nth moment of the interoverflow distribution of a group of c trunks.

The expression on the right-hand side of (32) is rather easy to compute, since we have the following recurrence relations (cf. Ref. 6, p. 388)

$$a\mu_2(n) = 2a\mu_1{}^2(n) + n\mu_2(n-1), \qquad (n = 1,2,\cdots) \tag{33}$$

where $\mu_1{}^{-1}(c) = aE_{1,c}(a)$. Hence

$$\text{Var } [S(t)] \sim \frac{[(c/a)\mu_2(c-1) + \mu_1{}^2(c)]}{\mu_1{}^3(c)}\, t.$$

The second moment $\mu_2(c-1)$ can be computed either by repeated use of (33) or by means of the explicit formula

$$\mu_2(c-1) = 2\sum_{n=0}^{c-1} \frac{(c-1)_n}{a^n}\,\mu_1{}^2(c-1-n)$$

with

$$(c-1)_0 = 1, \qquad (c-1)_n = (c-1)(c-2)\cdots(c-n), \qquad (n \geqq 1).$$

We note that the renewal theorem used above can be applied as long as the input to the system is recurrent, and the other assumptions made here remain the same. In these more general cases, (32) still holds, but the moments $\mu_1(c)$ and $\mu_2(c)$ satisfy less simple recurrence relations. Indeed, we have then:

$$\mu_1(c)\gamma_{c-1}(1) - \mu_1(c-1) = 0$$

and

$$\mu_2(c)\gamma_{c-1}(1) - 2\mu_1(c)[\mu_1(c-1) - \gamma_{c-1}{}'(1)] - \mu_2(c-1) = 0$$

where $\gamma_n(\cdot)$ is the Laplace-Stieltjes transform of the interoverflow distribution of a group of n trunks and $\gamma_m{}'(1)$ stands for the derivative of $\gamma_m(\cdot)$ at 1.

The preceding relations follow immediately from Palm's recurrences (cf. Ref. 3, pp. 36–38, and Ref. 7, pp. 16–22):

$$\gamma_n(s)[1 - \gamma_{n-1}(s) + \gamma_{n-1}(s+1)] = \gamma_{n-1}(s+1), \qquad (n = 1,2,\cdots).$$

The standard deviation of the call congestion computed by means of (29) and (31) to (33) is compared in Figs. 1–5 with simulation results. As may be seen from these graphs, there is good agreement between the theoretical and observed values.

On each one of these charts, two additional curves are also plotted, namely:

(*i*) $\{\text{Var}\,[S(t)]\}^{\frac{1}{2}}/ER(t)$ as a function of the offered load.

This expression is an upper bound for the standard deviation of the call congestion. However, unless the offered load is relatively small, it considerably overestimates this standard deviation.

(*ii*) $\{E_{1,c}(a)[1 - E_{1,c}(a)]\}^{\frac{1}{2}}/\{ER(t)\}^{\frac{1}{2}}$ as a function of the offered load.

This quantity, referred to as the binomial approximation, is a lower bound for the standard deviation of the call congestion. This bound

underestimates the latter to such an extent, however, that it is of little if any value.

In view of the agreement between the observed and theoretical variances of the call congestion, the latter are graphed in Figs. 6–8 for $c = 1(1)10(2)\ 20(5)50$ and $0.1 \leqq a/c \leqq 10$. These values pertain to the case $t = 20$. The asymptotic variance of the call congestion for any (sufficiently large) value of t may be obtained by multiplying the variances of Figs. 6–8 by $20/t$.

Simulation results have shown that (29) — with Var $[S(t)]$ and Cov $[R(t),S(t)]$ replaced by their respective asymptotic expressions — give sufficiently accurate values of the variance of the call congestion whenever the length of the observation period, t, is such that the expected number of offered calls, $ER(t)$, is about 40 or more. When $ER(t)$

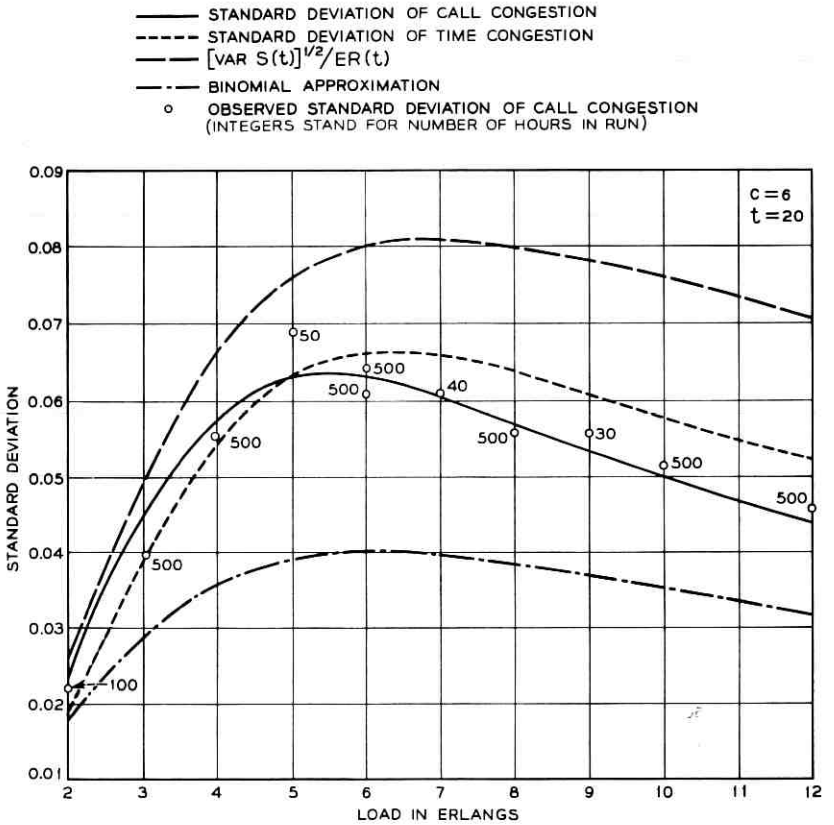


Fig. 1 — Standard deviations of call and time congestions, $c = 6$, $t = 20$.

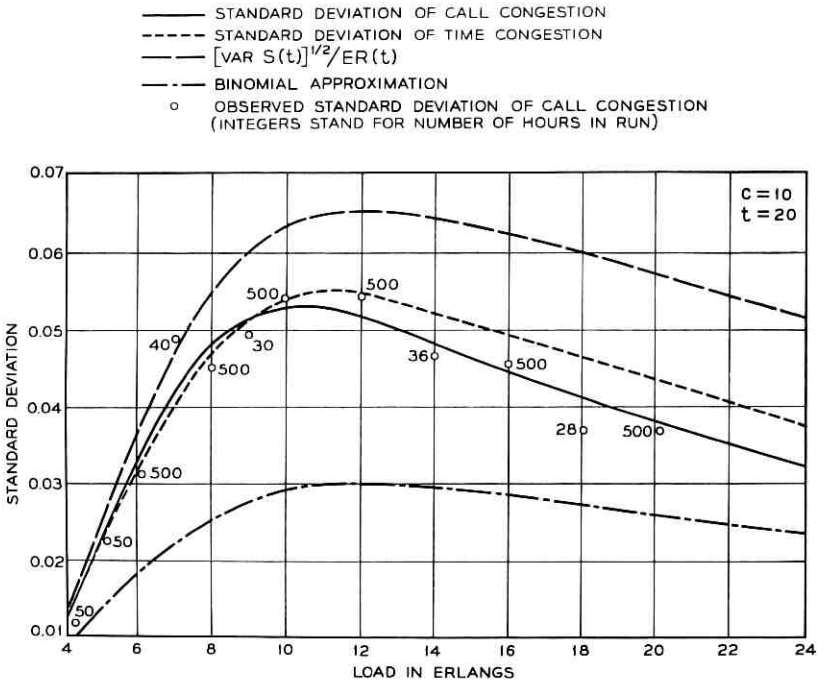


Fig. 2 — Standard deviations of call and time congestions, $c = 10$, $t = 20$.

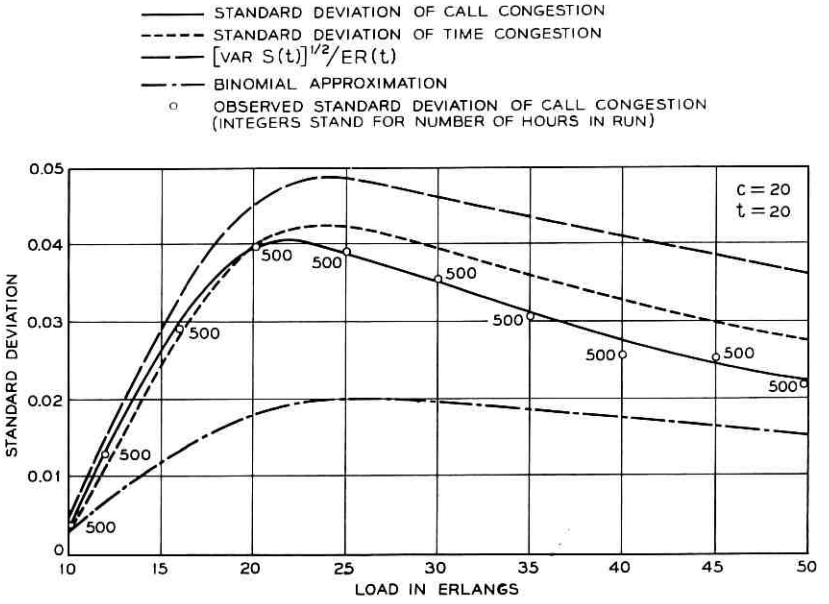


Fig. 3 — Standard deviations of call and time congestions, $c = 20$, $t = 20$.

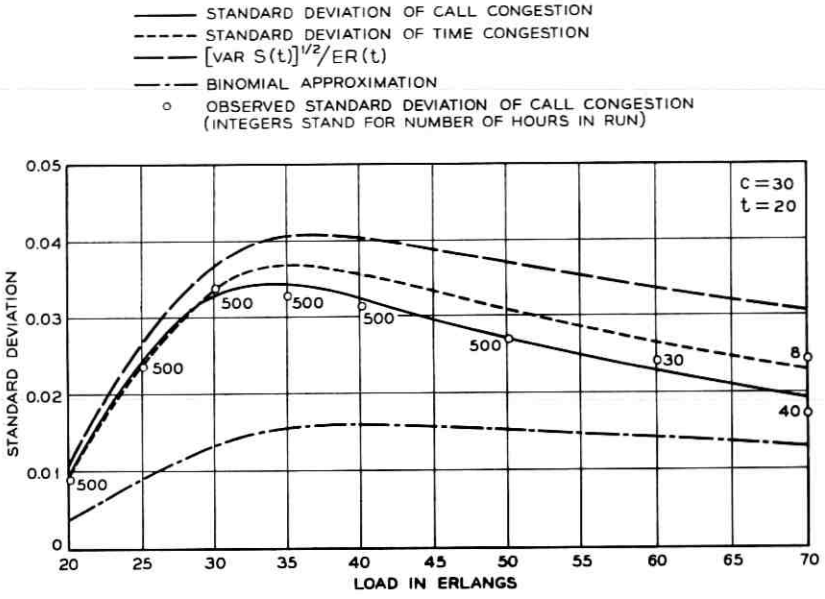


Fig. 4 — Standard deviations of call and time congestions, $c = 30$, $t = 20$.

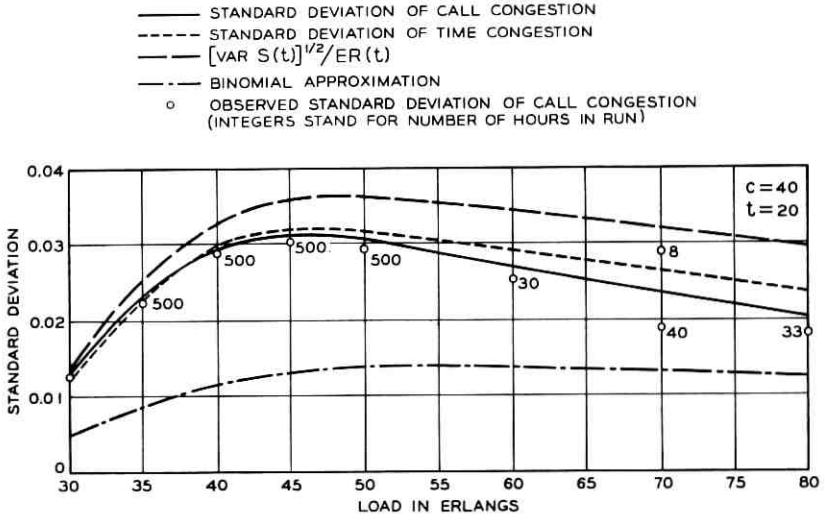


Fig. 5 — Standard deviations of call and time congestions, $c = 40$, $t = 20$.

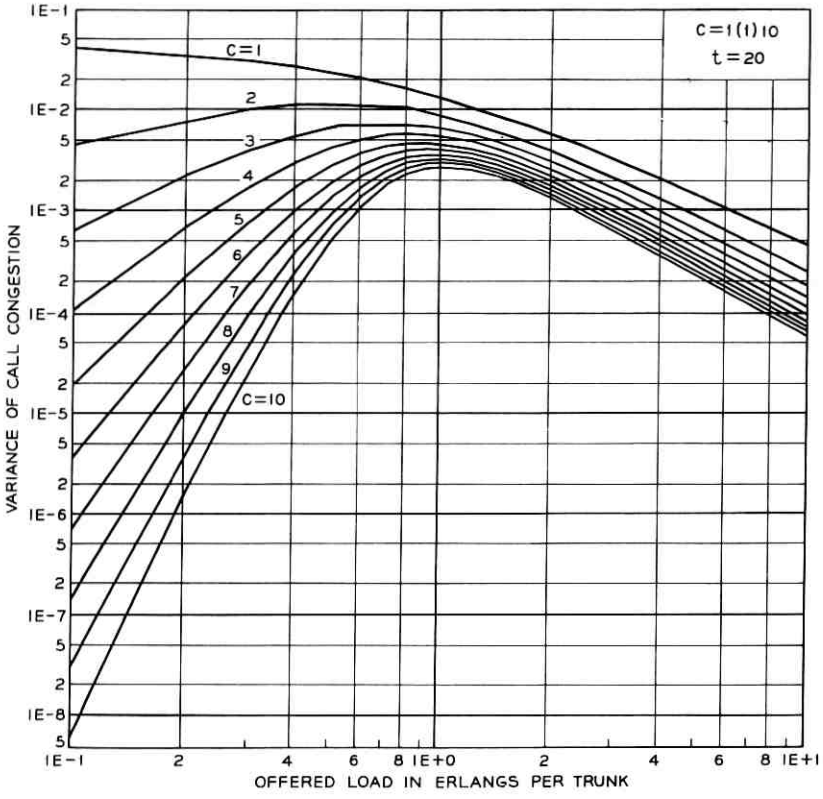


Fig. 6 — Variance of call congestion. Observation period = 20 holding times $c = 1(1)10$.

falls below 40, (29) provides us with an upper bound which becomes increasingly coarse as the expected number of offered calls decreases.

IV. RELATIVE ACCURACY OF LOSS ESTIMATES

Various measurements can be used to estimate the probability of loss. The principal ones are:

(*i*) *The number of offered calls and the number of lost (or overflow) calls.* The ratio of the latter to the former (i.e., the call congestion) is an asymptotically unbiased estimate of the probability of loss.

(*ii*) *The time congestion.* This quantity, which is an unbiased estimate of the probability of loss, may be either measured exactly or estimated by scanning the trunks at regular intervals and observing, at each

scan, how many trunks are busy. The proportion of scans which indicate that all trunks are busy is also an unbiased estimate of the probability of loss.

(*iii*) *The carried load (i.e., the average number of busy trunks) obtained either by continuous observation or by scanning at regular intervals.* This last measurement consists in observing, at regular intervals, the number of busy trunks. The average of these numbers, for a given number of scans, is an unbiased estimate of the carried load. If $\hat{L}$ stands for the carried load measured either exactly or by scanning, then the demand rate, $\hat{a}$, may be estimated by means of the formula

$$\hat{L} = \hat{a}[1 - E_{1,c}(\hat{a})].$$

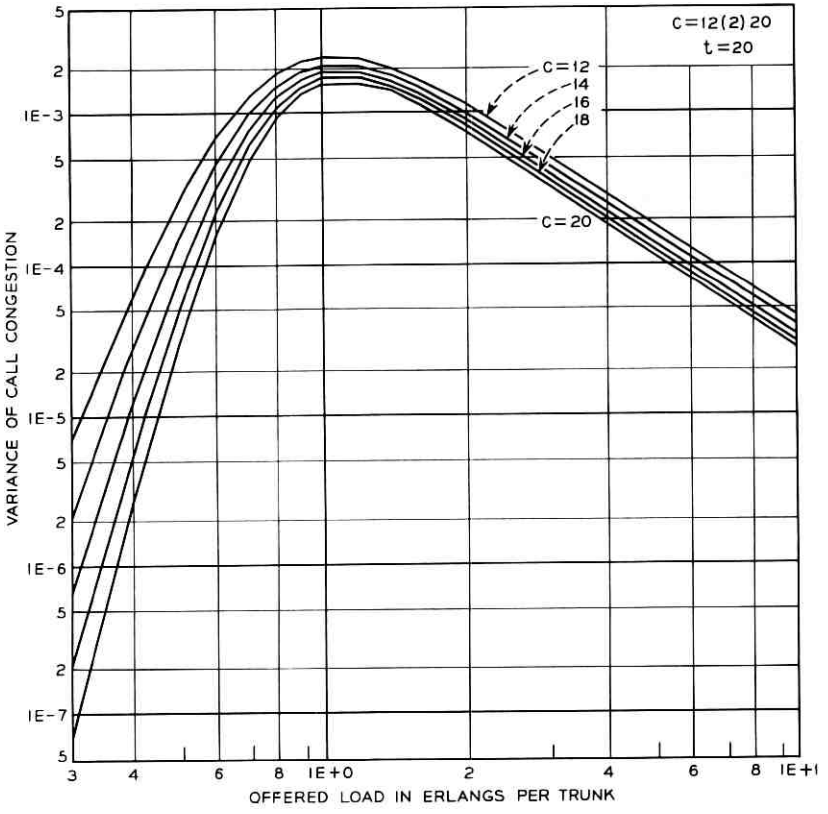


Fig. 7 — Variance of call congestion. Observation period = 20 holding times $c = 12(2)20$.

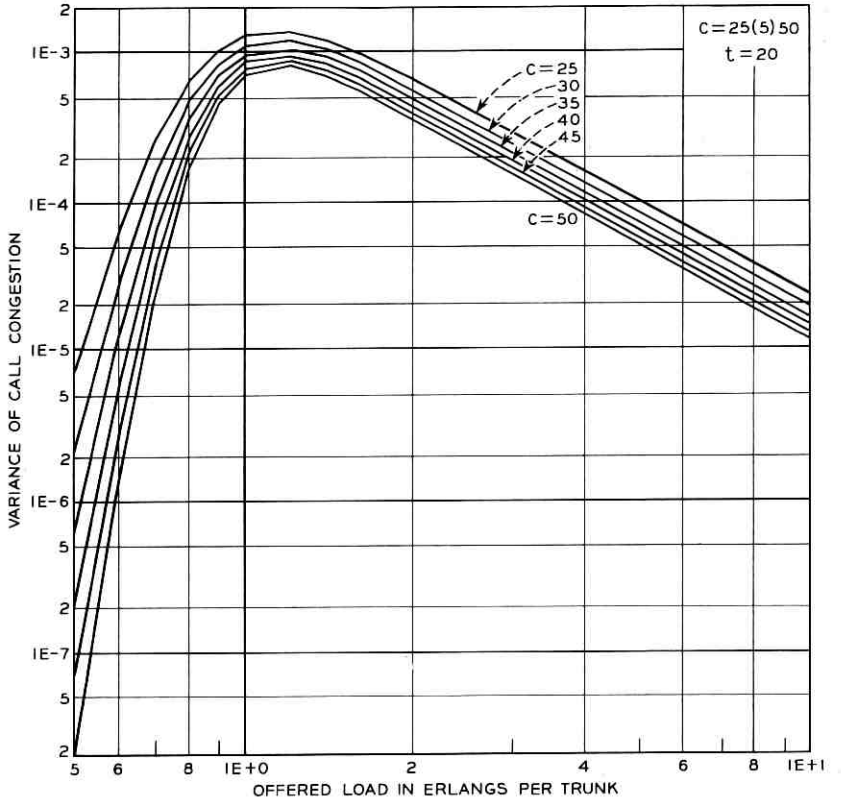


Fig. 8 — Variance of call congestion. Observation period = 20 holding times $c = 25(5)50$.

$E_{1,c}(\hat{a})$ itself is an estimate of the probability of loss. This estimate has a small positive bias which tends to zero as the length of the observation period gets large.

Theoretical as well as observed (simulation) values of the standard deviations of these estimates are plotted in Figs. 9–12. (For each load, the simulation results given in Fig. 12 were computed from a single run of 500 hours. The numerator and the denominator of each ratio appearing in Figs. 9 and 10 were evaluated from a single 500-hour run of simulated traffic.) These graphs reveal typical patterns, namely:

(*i*) When the offered load, in erlangs, falls short of the number of trunks, the loss estimates based on continuous load measurements have smaller standard deviations than both the call and the time

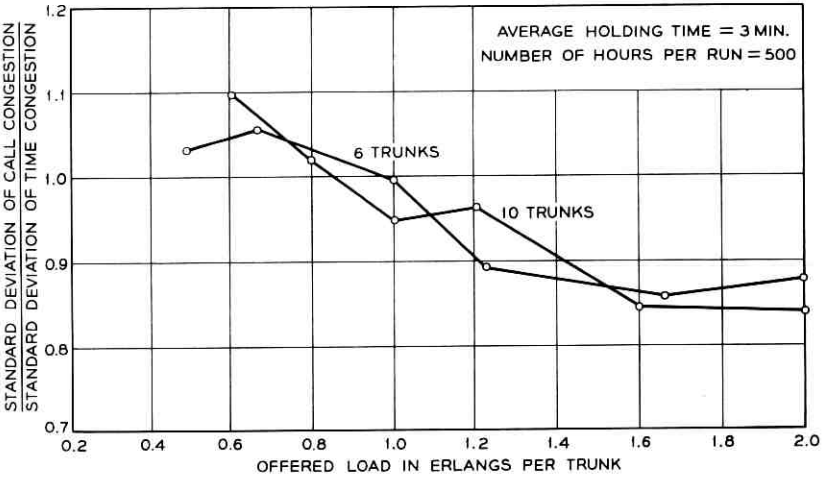


Fig. 9 — Relative accuracy of grade of service estimates based on hourly measurements of call and time congestions — simulation results.

congestions. In the same range, the call congestion has a larger standard deviation than the time congestion.

(*ii*) When the offered load exceeds the number of trunks, the converse situation holds; i.e., the call congestion has a smaller standard deviation than the time congestion, and the standard deviation of the

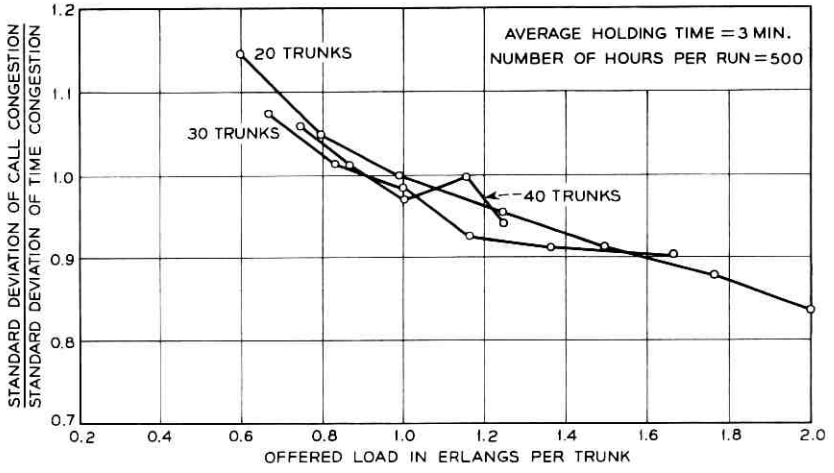


Fig. 10 — Relative accuracy of grade of service estimates based on hourly measurements of call and time congestions — simulation results (cont.).

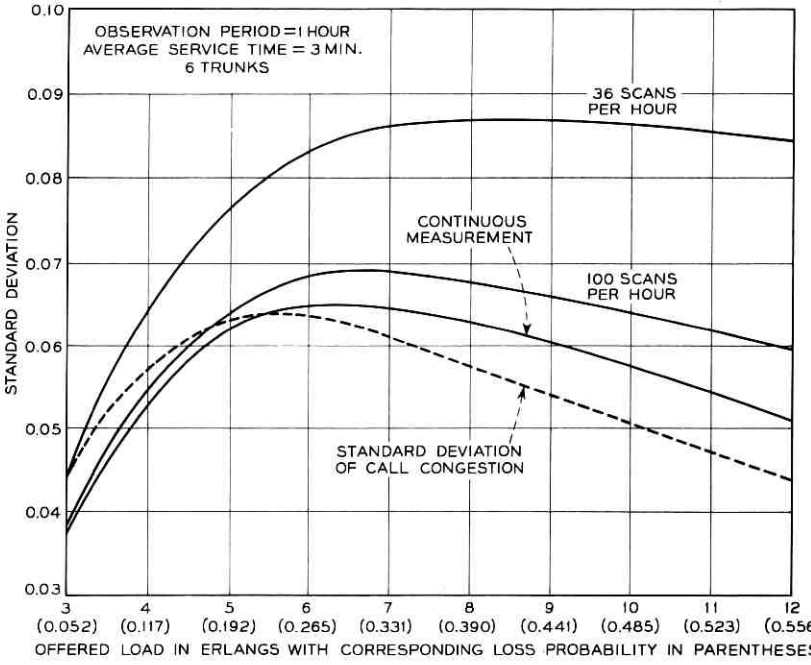


Fig. 11 — Standard deviation of the time congestion estimated by switch-counting.

latter, in turn, is exceeded by the standard deviation of loss estimates based on continuous carried load measurements.

The effect of scanning on the variances of the time congestion and of the loss estimates based on carried load measurements is illustrated in Figs. 11 and 12.

Let us assume now that the length of the observation period is such that (29) closely approximates Var $[S(t)/R(t)]$. Under these conditions, the load beyond which the time congestion is less accurate (in terms of its variance) than the call congestion is approximately equal to the load a determined by the following equation:

$$1 + E_{1,c}(a) = 2E_{1,c}(a)[c + 1 - a + aE_{1,c}(a)]. \qquad (34)$$

This condition is readily seen to be equivalent to the requirement

$$\text{Var}\,[S(t)/R(t)] = \text{Var}\,B(t) \qquad (35)$$

where $B(t)$ stands for the time congestion in an observation period of

length t. Equations (35), (29), and (31) together with the relations (cf. Ref. 3 p. 131)

$$ES(t) = atEB(t)$$

$$\text{Var } S(t) - ES(t) = (at)^2 \text{ Var } B(t)$$

imply (34).

For given c, (34) has a unique positive root, r, which is smaller than c except in the case $c = 1$ where the root is equal to 1. Computations show that this root lies relatively close to c (cf. Fig. 13).

Let $B_n(t)$ be the estimate of the time congestion obtained by switch-

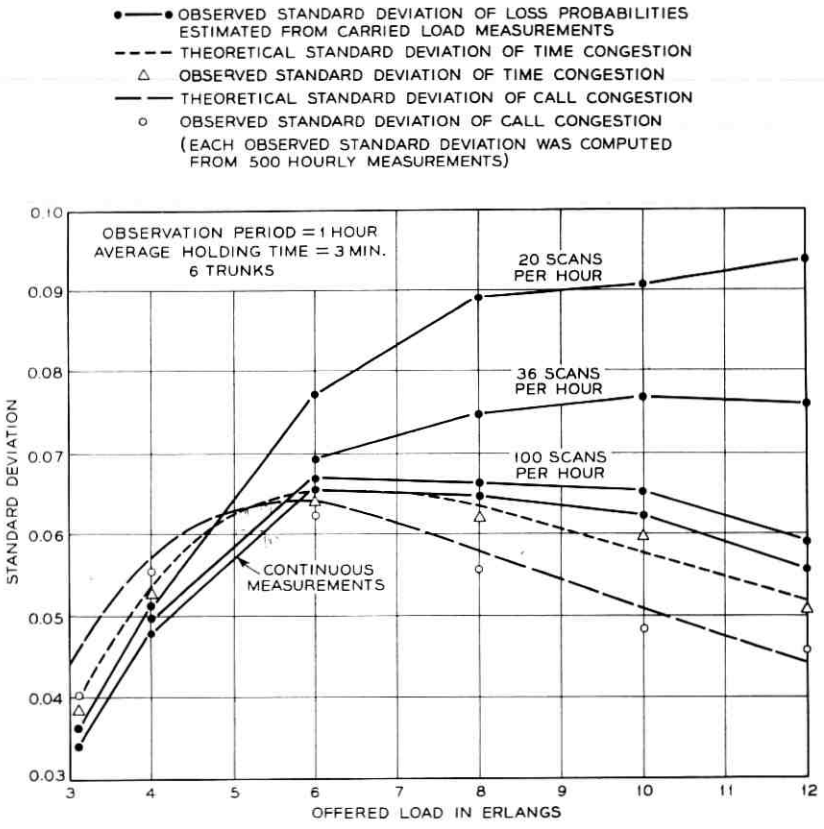


Fig. 12 — Relative accuracy of loss estimates based on call and time congestions and on carried load measurements.

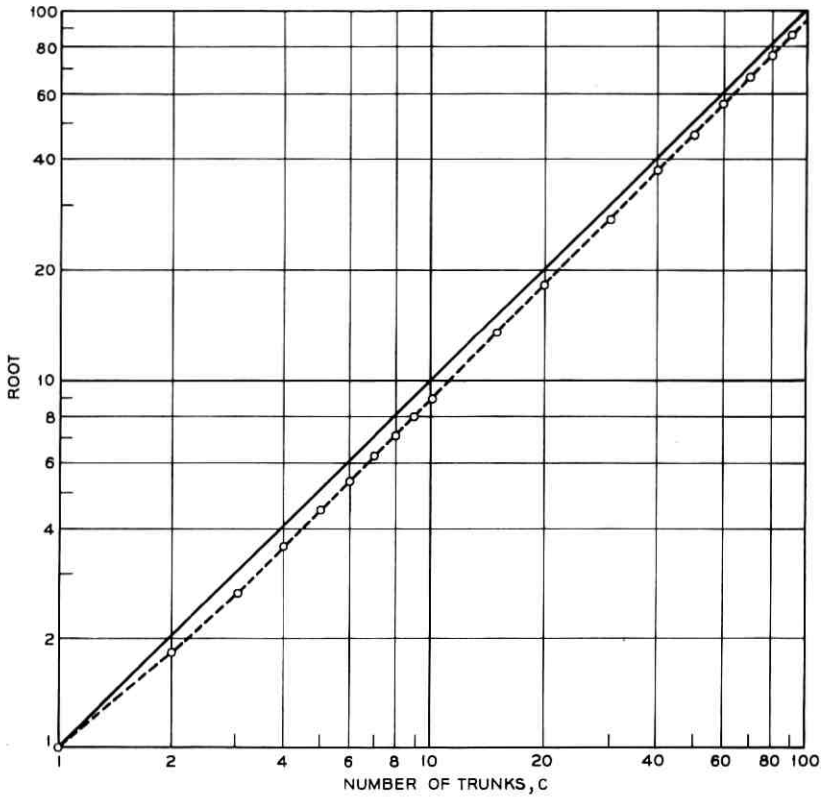


Fig. 13 — Root of equation (34).

counting at the rate of n scans per observation period of length t. We shall now derive an explicit formula for the variance of $B_n(t)$.

Let τ be the interval separating consecutive scans, $N(u)$ be the number of busy trunks at time u, and

$$X(u) = \begin{cases} 1 & \text{if } N(u) = c \\ 0 & \text{if } N(u) < c. \end{cases}$$

Then

$$B_n(t) = n^{-1} \sum_{1}^{n} X(i\tau)$$

and

$$EB_n(t) = E_{1,c}(a)$$

$$\text{Var } B_n(t) = n^{-2} \sum_{i=1}^{n} \sum_{j=1}^{n} \text{Cov } [X(i\tau), X(j\tau)].$$

Now let

$$P(u) = Pr\,[N(u) = c \mid N(0) = c].$$

The function $P(\cdot)$, which is called the recovery function of the process $N(\cdot)$, has the following expression (cf. Ref. 3, p. 85 and Ref. 8, p. 135):

$$P(u) = E_{1,c}(a) - \sum_{j=1}^{c} \frac{e^{w_j|u|}}{w_j} \prod_{i \neq j} \left(1 - \frac{1}{w_j - w_i}\right)$$

where, as before, w_i, $i = 1, \cdots, c$, are the c roots of $\sigma_c(w + 1)$. Since

$$\text{Cov } [X(u + v), X(v)] = E_{1,c}(a)P(u) - E_{1,c}^{\,2}(a)$$

we have (cf. Ref. 8, pp. 136–138)

$$\begin{aligned} \text{Var } B_n(t) &= n^{-2} E_{1,c}(a) \sum_{i=1}^{n} \sum_{j=1}^{n} P(|\, i - j \,|\, \tau) - E_{1,c}^{\,2}(a) \\ &= -n^{-2} E_{1,c}(a) \sum_{-n}^{n} (n - |\, k \,|) \sum_{j=1}^{c} \frac{e^{|k| w_j \tau}}{w_j} \\ &\quad \cdot \prod_{i \neq j} \left(1 - \frac{1}{w_j - w_i}\right) \qquad (36) \\ &= n^{-1} E_{1,c}(a) \sum_{j=1}^{c} \left\{ \frac{1}{w_j} \prod_{i \neq j} \left(1 - \frac{1}{w_j - w_i}\right) \right\} \\ &\quad \cdot \left\{ \text{ctnh}\left(\frac{\tau w_j}{2}\right) + \left(\frac{1 - e^{n\tau w_j}}{2n}\right) \text{csch}^2\left(\frac{\tau w_j}{2}\right) \right\}. \end{aligned}$$

If we let n tend to infinity in this formula, we obtain, in the limit, the variance of the time congestion, $B(t)$, for continuous observation (measurement):

$$\begin{aligned} \text{Var } B(t) = \frac{2E_{1,c}(a)}{t} &\left\{ \sum_{j=1}^{c} w_j^{-2} \prod_{i \neq j} \left(1 - \frac{1}{w_j - w_i}\right) \right. \\ &\left. + \frac{1}{t} \sum_{j=1}^{c} w_j^{-3} (1 - e^{tw_j}) \prod_{i \neq j} \left(1 - \frac{1}{w_j - w_i}\right) \right\}. \end{aligned}$$

This last formula was first obtained, in a slightly different form, by Kosten, Manning and Garwood (cf. Ref. 1).

ACKNOWLEDGMENTS

I wish to thank D. C. Boes for many useful discussions. Many thanks are also due to Miss C. J. Durnan and C. A. Lennon for their skillful programming.

REFERENCES

1. Kosten, L., Manning, J. R., and Garwood, F., 1949, On the Accuracy of Measurements of Probabilities of Loss in Telephone Systems, J. Roy. Statist., Soc. Series B, *11*, pp. 54–67.
2. Szegö, G., Orthogonal Polynomials, American Mathematical Society, Colloquium Publications, *XXIII* (revised edition), 1959.
3. Riordan, J., *Stochastic Service Systems*, Wiley, New York, 1962.
4. Feller, W., *An Introduction to Probability Theory and Its Applications*, *I*, Wiley, New York, 1957.
5. Smith, W. L., 1954, Asymptotic Renewal Theorems, Proc. Roy. Soc. Edinb., A, *64*, pp. 9–48.
6. Descloux, A., Overflow Processes of Trunk Groups with Poisson Inputs and Exponential Service Times, B.S.T.J., *42*, March, 1963, pp. 383–397.
7. Palm, C., 1943, Intensitätsschwankungen im Fernsprechverkehr, Ericsson Technics, *44*, p. 189.
8. Benes, V. E., The Covariance Function of a Simple Trunk Group, with Applications to Traffic Measurement, B.S.T.J., *40*, January, 1961, pp. 117–148.

Spectra of Digital FM

By R. R. ANDERSON and J. SALZ

(Manuscript received March 29, 1965)

Formulas are derived for the spectral density function of an ensemble of continuous-phase, constant-envelope FM waves. The modulation signals are random time series of the form $\sum_n a_n g(t - nT)$, *where* $g(t)$ *is an arbitrary pulse of finite duration* rT, $r \geqq 1$. *The* a_n*'s are independent random variables possessing identical but otherwise arbitrary probability distribution. The derived results are general and are presented in terms of averages of elementary functions. When the* a_n*'s are discrete random variables, both continuous and discrete spectra are treated, and conditions in terms of the modulation parameters are given under which discrete spectral lines are present. Several of our specialized formulas are applicable in the study of multilevel FM data transmission systems as well as in pulse frequency modulation.*

I. INTRODUCTION

Progress in analysis of multilevel frequency shift keying (FSK) has lagged behind that of binary. Inherent difficulties associated with an increase in the number of levels are partly responsible, but activity also has been inhibited by the general impression that multilevel FSK is inferior to differential phase modulation with the same number of levels operating in the same bandwidths.

Recent work, Ref. 1, has evolved design principles showing a possibility of substantial improvements in multilevel FM performance over that formerly thought to be typical. Also there are many existing analog channels, e.g., in microwave radio relay, which operate by FM. Attempts to send digital data efficiently over such channels force consideration of the multilevel FSK problem.

An important item in the statistical description of an information-bearing signal is the spectral density, which defines the average power density of the signal as a function of frequency. In addition to furnishing an estimate of bandwidth requirements, the spectral density is

critically important in optimization procedures for minimizing the effect of channel noise subject to a constraint on mean total transmitting signal power. Evaluation of mutual interference between channels also requires knowledge of spectral distribution.

From the practical point of view the most interesting case of digital FM is that in which the phase is continuous at the transitions, as may be obtained at the output of a keyed oscillator. The memory thus introduced makes the analysis far from trivial. So far as is known, the binary case is the only continuous-phase FSK problem hitherto covered in the literature. The present paper gives a complete analytic solution for a general set of parameters.

An interesting feature is the extent to which sharp spectral peaks occur near the discrete signaling frequencies. These peaks can, under certain conditions, become delta functions indicating steady sine-wave components. Such components make the design of optimum filters difficult because the best results usually demand sharply tuned suppression of the corresponding regions at the transmitter and complementary high gain peaks at the receiver. Furthermore, the interference produced in other channels by untreated peaked spectra can be inordinately severe. One important result of the analysis is an establishment of conditions on the frequency spacing relative to signaling rate such that spectral peaking does not occur.

In this paper we derive compact formulas for the spectral density function of an ensemble of continuous-phase, constant-envelope FM waves. The frequency of the wave is switched every T seconds by a known signal. The phase of the wave is so adjusted as to maintain continuity at the transitions. For example, when the baseband signal is a rectangular pulse of T seconds in duration, the frequency of the wave during each interval T may be one out of many different frequencies picked at random. In general, the baseband signal is not time limited to T seconds. This case will arise when the original time limited signal is passed through a filter.

The random signal whose spectral density we wish to study has the following standard representation:

$$S(t) = A \cos\left\{\omega_c t + \omega_d \sum_{n=0}^{n=\infty} a_n \int_0^t g(t' - nT)\,dt' + \varphi\right\} \qquad (1)$$

$$0 \leqq t \leqq \infty$$

where ω_c and ω_d are arbitrary angular frequencies. The angle φ is a uniformly distributed random variable (r.v.) on $[0,2\pi]$ and the a_n's are in-

dependent r.v.'s with arbitrary but identical probability distribution. The symbol A is an arbitrary real amplitude.

The only restriction on the baseband signal $g(t)$ is that

$$g(t) = \begin{cases} g_r(t), & 0 \leqq t \leqq rT \\ 0, & \text{everywhere else} \end{cases} \tag{2}$$

where r is an arbitrary positive integer. (We naturally require that $g(t)$ be integrable over this interval.) When the a_n's are binary r.v.'s and $g(t)$ is a rectangular pulse of T seconds duration, spectra and correlation functions have recently been derived by Bennett and Rice, Ref. 2. Salz, Ref. 3, extended Bennett's and Rice's results to include arbitrary a_n's possessing arbitrary probability distributions.

In our treatment the distribution of the a_n's as well as $g(t)$ is entirely arbitrary. For instance when $g(t)$ is a rectangular pulse and the a_n's are discrete r.v.'s, the wave (1) represents the ensemble of multilevel FM waves. If the a_n's correspond to the samples of speech taken every T seconds, we have pulse amplitude modulation via frequency shift keying. Many other applications depending on the choice of the a_n's and $g(t)$ may be cited, and are covered in our results.

Our method of attack on the problem is direct. We calculate the segmented Fourier transform of (1), obtain its magnitude squared, average over all random variables, divide by the length of the segment, and then evaluate the limit as the length increases without bound. After considerable amount of bookkeeping, we obtain general formulas. We then specialize the formulas, and investigate some interesting representative cases. The general formula for the continuous spectrum is given in (31). Equation (40) is the general formula for the discrete as well as the continuous spectrum.

II. GENERAL DEVELOPMENT

We found it easier to work with the complex representation of (1). Therefore let

$$S(t) = \frac{A}{2}\left[z(t) + z^*(t)\right] \qquad 0 \leqq t \leqq \infty \tag{3}$$

where

$$z(t) = e^{i\varphi} \exp i\left\{\omega_c t + \omega_d \sum_{n=0}^{n=\infty} a_n \int_{-nT}^{t-nT} g(t')\,dt'\right\}. \tag{4}$$

The symbol * denotes the complex conjugation.

Choose a finite interval $[0,NT]$. Over this interval the Fourier transform of $z(t)$ is

$$\begin{aligned} Z(\omega,NT) &= \sum_{k=0}^{k=N-1} \int_{kT}^{(k+1)T} z(t)e^{-\omega t}dt \\ &= e^{i\varphi} \sum_{k=0}^{k=N-1} Q_k \end{aligned} \tag{5}$$

where

$$Q_k = \int_{kT}^{(k+1)T} dt \exp i(\omega_c - \omega)t \prod_{n=0}^{n=\infty} \exp i\left\{\omega_d a_n \int_{-nT}^{t-nT} g(t')dt'\right\}. \tag{6}$$

Set $t - kT = y$ above to obtain

$$Q_k = e^{ikT(\omega_c-\omega)} \int_0^T dy \exp i(\omega_c - \omega)y \prod_{n=0}^{n=\infty} P_{n,k}(y) \tag{7}$$

where

$$P_{n,k}(y) = \exp i\left\{\omega_d a_n \int_{-nT}^{y-(n-k)T} g(t')dt'\right\}. \tag{8}$$

Since $g(t)$ is time limited to rT, where r is a positive integer, it follows that we can write $P_{n,k}(y)$ for $0 \leqq y \leqq T$, as;

$$P_{n,k}(y) = \begin{cases} \exp\left[i\omega_d a_n \int_0^{rT} g_r(t')dt'\right], & 0 \leqq n \leqq k-r \\ \exp\left[i\omega_d a_n \int_0^{y-(n-k)T} g(t')dt'\right], & k-r+1 \leqq n \leqq k \\ 1 & n > k. \end{cases} \tag{9}$$

With this representation, we can take the product in (7) running from $n = 0$ to $n = k - r$ outside the integral sign since $P_{n,k}(y)$ in this range of n does not depend on y. $P_{n,k}(y)$ depends on y only in the range $k - r + 1 \leqq n \leqq k$. Making use of these facts, we write for Q_k

$$Q_k = e^{ikT\nu} \exp i\left\{\alpha_r \sum_{n=0}^{n=k-r} a_n\right\} F(\nu, a_{k-r+1} \cdots a_k) \tag{10}$$

where

$$\begin{aligned} &F(\nu, a_{k-r+1} \cdots a_k) \\ &\quad = \int_0^T e^{i\nu y} \prod_{n=k-r+1}^{n=k} \exp i\left\{\omega_d a_n \int_0^{y-(n-k)T} g_r(t')dt'\right\} dy \\ &\quad = \int_0^T e^{i\nu y} \exp i\left\{\sum_{m=1}^{m=r} a_{k-r+m} V[y - (m-r)T]\right\} dy, \end{aligned} \tag{11}$$

and

$$\begin{aligned} \alpha_r &= \omega_d \int_0^{rT} g_r(t)\,dt \\ \nu &= \omega_c - \omega \\ V(\xi) &= \omega_d \int_0^{\xi} g(t)\,dt. \end{aligned} \tag{12}$$

The segmented Fourier transform of the original signal (3) is

$$\begin{aligned} S(\omega,NT) &= \frac{A}{2}\left[Z(\omega,NT) + Z_c(\omega,NT)\right] \\ &= \frac{A}{2}\left[e^{i\varphi}\sum_{k=0}^{k=N-1} Q_k + e^{-i\varphi}\sum_{k=0}^{k=N-1} Q_{ck}\right] \end{aligned} \tag{13}$$

where $Z_c(\omega,NT)$ is the Fourier Transform of $Z^*(\omega,N)$ given as

$$\begin{aligned} Z_c(\omega,NT) &= \sum_{k=0}^{k=N-1}\int_{kT}^{(k+1)T} z^*(t)\,e^{-i\omega t}\,dt \\ &= Ae^{-i\varphi}\sum_{k=0}^{k=N-1} Q_{ck}, \end{aligned} \tag{14}$$

and

$$\begin{aligned} Q_{ck} = \int_{kT}^{(k+1)T} dt \exp -i(\omega_c + \omega)t \\ \cdot \prod_{n=0}^{n=\infty} \exp -i\left\{\omega_d a_n \int_{-nT}^{t-nT} g(t')\,dt'\right\}. \end{aligned} \tag{15}$$

The magnitude squared of $S(\omega,NT)$ averaged with respect to the r.v. φ is

$$\langle |S(\omega,NT)|^2\rangle_\varphi = \frac{A^2}{4}\left[\sum_{k,s=0}^{k,s=N-1} Q_k Q_s^* + \sum_{k,s=0}^{k,s=N-1} Q_{cs}Q_{ck}^*\right]. \tag{16}$$

The symbol $\langle\cdot\rangle$ denotes the averaging operator.

From the definition of Q_k in (10) we obtain

$$\begin{aligned} Q_k Q_s^* = e^{iT\nu(k-s)}F(\nu,a_{k-r+1}\cdots a_k)F^*(\nu,a_{s-r+1}\cdots a_s) \\ \cdot \exp i\left\{\alpha_r\left[\sum_{n=0}^{n=k-r} a_n - \sum_{n=0}^{n=s-r} a_n\right]\right\}. \end{aligned} \tag{17}$$

We observe that $Q_{cs}Q_{ck}^*$ equals $Q_kQ_s^*$ with $\omega_c - \omega$ replaced by $\omega_c + \omega$.

It is thus sufficient to continue our calculation using only the first sum in (16).

If we let the first term in (16) be $W_+(\omega)$ and the second term $W_-(\omega)$, the desired power spectrum $G(\omega)$ is by definition

$$G(\omega) = \lim_{N\to\infty} (2/TN)\{\langle W_+(\omega)\rangle_{\mathbf{a}} + \langle W_-(\omega)\rangle_{\mathbf{a}}\} \tag{18}$$

where the ensemble average is taken over the collection of r.v.'s $\mathbf{a} = (a_0, a_1 \cdots a_N)$.

We now proceed to calculate the respective averages. From (16)

$$\begin{aligned}\langle W_+(\omega)\rangle_{\mathbf{a}} &= \frac{A^2}{4}\sum_{k,s=0}^{k,s=N-1}\langle Q_k Q_s^*\rangle_{\mathbf{a}} \\ &= \frac{A^2}{2}\,\mathrm{Re}\left[\sum_{\substack{k,s\\ \{k<s\}}}\langle Q_k Q_s^*\rangle_{\mathbf{a}}\right] \\ &\quad + \frac{A^2}{4}\sum_{k=0}^{k=N-1}\langle |\,Q_k\,|^2\rangle_{\mathbf{a}}.\end{aligned} \tag{19}$$

The symbol "Re[·]" denotes the real part.

To facilitate the evaluation of the averages, we rearrange the double sums above in the following manner:

$$\begin{aligned}\sum_{\substack{k,s=0\\ \{k>s\}}}^{k,s=N-1}\langle Q_k Q_s^*\rangle_{\mathbf{a}} &= \sum_{s=0}^{s=N-2}\sum_{k=s+1}^{k=N-1}\langle Q_k Q_s^*\rangle_{\mathbf{a}} \\ &= \sum_{s=0}^{s=N-2}\Big\{\langle Q_{s+1}Q_s^*\rangle_{\mathbf{a}} + \langle Q_{s+2}Q_s^*\rangle_{\mathbf{a}} \\ &\quad + \cdots + \langle Q_{s+r}Q_s^*\rangle_{\mathbf{a}} + \sum_{k=s+r+1}^{k=N-1}\langle Q_k Q_s^*\rangle_{\mathbf{a}}\Big\} \\ &= \sum_{s=0}^{s=N-2}\langle Q_{s+1}Q_s^*\rangle_{\mathbf{a}} + \sum_{s=0}^{s=N-3}\langle Q_{s+2}Q_s^*\rangle_{\mathbf{a}} \\ &\quad + \cdots + \sum_{s=0}^{s=N-r-1}\langle Q_{s+r}Q_s^*\rangle_{\mathbf{a}} \\ &\quad + \sum_{s=0}^{s=N-r-2}\sum_{k=s+r+1}^{k=N-1}\langle Q_k Q_s^*\rangle_{\mathbf{a}}.\end{aligned} \tag{20}$$

Using the explicit representation of F and F^* in 11, we obtain from 17

$$\begin{aligned}\langle Q_{s+j}Q_s^*\rangle_a &= f_{jr}(\nu), \qquad 1 \leqq j \leqq r \\ &= e^{iT\nu j}\langle F(\nu,a_{s+j-r+1} \cdots a_{s+j}) \\ &\quad \cdot F^*(\nu,a_{s-r+1} \cdots a_s) \exp i\left\{a_r \sum_{n=s-r+1}^{n=s-r+j} a_n\right\}\Big\rangle_a \\ &= e^{iT\nu j}\int_0^T\int_0^T dydy'\, e^{i\nu(y-y')} \\ &\quad \cdot \langle \exp i\Big\{\sum_{m=1}^{m=r} a_{s+j-r+m}V[y - (m-r)T] \\ &\quad - \Big\{\sum_{m=1}^{m=r} a_{s-r+m}V[y' - (m-r)T] + a_r\sum_{n=1}^{n=j} a_{s-r+n}\Big\}\Big\rangle_a\end{aligned} \tag{21}$$

Using elementry manipulations we obtain

$$\begin{aligned}f_{jr}(\nu) &= e^{iT\nu j}\int_0^T\int_0^T dydy' e^{i\nu(y-y')}\prod_{m=1}^{m=r-j} C_a\{V[y - (m-r)T] \\ &\quad - [y' - (m-r+j)T]\} \\ &\cdot \prod_{m=1+r-j}^{m=r} C\{V[y - (m-r)T]\}\prod_{m=1}^{m=j} C^*\{V[y' - (m-r)T] - \alpha_r\}.\end{aligned} \tag{22}$$

The function

$$C_a(s) = \langle e^{ias}\rangle_a = \int e^{ias}dF(a) \tag{23}$$

is the characteristic function of the r.v. a, and $F(a)$ its probability distribution.

We next calculate in the same manner as above

$$\begin{aligned}\langle Q_kQ_s^*\rangle_a &= e^{iT\nu(k-s)}\Big\langle F(\nu,a_{k-r+1} \cdots a_k)F^*(\nu,a_{s-r+1} \cdots a_s) \\ &\quad \cdot \exp i\left\{\alpha_r \sum_{n=s-r+1}^{n=k-r} a_n\right\}\Big\rangle_a \\ &= e^{iT\nu(k-s)}\langle F(\nu,a_{k-r+1} \cdots a_k)\rangle_a\Big\langle F^*(\nu,a_{s-r+1} \cdots a_s) \\ &\quad \cdot \exp i\left\{\alpha_r \sum_{n=s-r+1}^{n=s} a_n\right\}\Big\rangle_a C_a^{k-s-r}(\alpha_r)\end{aligned} \tag{24}$$

when

$$k > s + r.$$

When $k = s$ we have

$$\langle | Q_k |^2 \rangle_{\mathbf{a}} = \langle | F(\nu, a_{k-r+1} \cdots a_k) |^2 \rangle_{\mathbf{a}} . \tag{25}$$

From the definition of F or F^* in (11) we obtain explicitly

$$\langle | F(\nu, a_{k-r+1} \cdots a_k) |^2 \rangle_{\mathbf{a}} = \int_0^T \int_0^T dy dy' e^{i\nu(y-y')} \cdot \prod_{m=1}^{m=r} C_a\{V[y - (m-r)T] - V[y' - (m-r)T]\}, \tag{26}$$

$$\langle F(\nu, a_{k-r+1} \cdots a_k) \rangle_{\mathbf{a}} = \int_0^T e^{i\nu y} \prod_{m=1}^{m=r} C_a\{V[y - (m-r)T]\} dy \tag{27}$$

and

$$\langle F^*(\nu, a_{s-r+1} \cdots a_s) \exp i\left\{\alpha_r \sum_{n=s-r+1}^{n=s} a_n\right\} \rangle_{\mathbf{a}} = \int_0^T e^{-i\nu y} \prod_{m=1}^{m=r} C_a^*\{V[y - (m-r)T] - \alpha_r\} dy. \tag{28}$$

We now observe that the various averages in (22)–(25) are independent of the indices k and s and therefore when we divide (19) or (20) by N and take the limit as N approaches infinity we obtain

$$\lim_{N\to\infty} \frac{1}{N} \langle W_+(\omega) \rangle_{\mathbf{a}} = \frac{A^2}{4} \langle | F(\mathbf{a}) |^2 \rangle_{\mathbf{a}} + \frac{A^2}{2} \operatorname{Re}\left\{ \sum_{j=1}^{j=r} f_{jr}(\nu) + \langle F_k(\mathbf{a}) \rangle_{\mathbf{a}} \langle F_s(\mathbf{a}) \exp i\left\{\alpha_r \sum_{n=s-r+1}^{n=s} a_n\right\} \rangle_{\mathbf{a}} \cdot \lim_{N\to\infty} \frac{1}{N} \sum_{s=0}^{s=N-r-2} \sum_{k=s+r+1}^{k=N-1} e^{iT\nu(k-s)} C_a^{k-s-r}(\alpha_r) \right\} \tag{29}$$

where we set $F_k(\mathbf{a}) = F(\nu, a_{k-r+1} \cdots a_k)$; $f_{jr}(\nu)$ is defined in (22).

The limit in (26) can readily be evaluated provided $| C_a(\alpha_r) | < 1$. This is the case when we have only continuous spectra. When $| C_a(\alpha_r) | = 1$, the evaluation of the limit is more involved. But this latter is precisely the case when discrete spectra appear, which we shall study in a forthcoming section. For the moment we proceed to evaluate the limit when the modulus of the characteristic function evaluated at α_r is less than unity.

In this case, from (29) we obtain

$$\lim_{N\to\infty}\frac{1}{N}\sum_{s=0}^{s=N-r-2}\sum_{k=s+r+1}^{k=N-1} e^{iT\nu(k-s)}C_a{}^{k-s-r}(\alpha_r)$$

$$= e^{iT\nu r}\lim_{N\to\infty}\sum_{l=1}^{l=N-r-1}[C_a(\alpha_r)\exp(iT\nu)]^l \tag{30}$$

$$= \frac{e^{iT\nu(r+1)}C_a(\alpha_r)}{1-e^{iT\nu}C_a(\alpha_r)}.$$

Using the definition of the spectrum in (18) and applying the explicit representation of the averages computed in (22)–(30), we obtain finally the positive image spectrum

$$\begin{aligned} G_+(\nu) = {} & \frac{A^2}{2T}\int_0^T\int_0^T dy\,dy'\,e^{i\nu(y-y')} \\ & \cdot\prod_{m=1}^{m=r} C_a\{V[y-(m-r)T]-V[y'-(m-r)T]\} \\ & + \frac{A^2}{T}\,\mathrm{Re}\Bigg\{\sum_{j=1}^{j=r}(\text{Equation 22}) \\ & + \frac{e^{iT\nu(r+1)}C_a(\alpha_r)}{1-e^{iT\nu}C_a(\alpha_r)}\int_0^T dy e^{i\nu y} \\ & \cdot\prod_{m=1}^{m=r} C_a\{V[y-(m-r)T]\}\int_0^T dy e^{-i\nu y} \\ & \cdot\prod_{m=1}^{m=r} C_a{}^*\{V[y-(m-r)T]-\alpha_r\}\Bigg\}. \end{aligned} \tag{31}$$

Although the final formula may appear rather complicated at first, under close scrutiny it will be observed that the formula is in a convenient form for numerical calculation by a digital computer. At most a double integral on a finite dimensional plane needs to be evaluated. We will later demonstrate, by using a few interesting examples, how the numerical work can be carried out, and the results will be exhibited graphically.

III. SINGULAR CASES

So far we have considered only continuous spectra. In order to arrive at the result of (31) we had to sum the series in (30), and that series converges only when the magnitude of the characteristic function evalu-

ated at α_r is less than unity. This turns out to be the requirement for the spectrum to contain no lines.

Whenever the magnitude of the characteristic function, evaluated at $\alpha_r \neq 0$, is unity we are no longer justified in using the results in (30) since the series diverges. This behavior suggests the presence of discrete spectral lines associated with periodicities in the original random process. Mathematically, this result can only occur if the r.v.'s **a** are discrete and have a definite relationship. The characteristic function of a continuous r.v. must satisfy $|C_a(s)| < |C_a(0)|$ when $s \neq 0$. We proceed to identify the conditions on the a_n which give rise to a characteristic function with unit modulus and therefore spectral lines.

Loève, Ref. 4, shows that if $|C_a(s)| = 1$ for $s \neq 0$, the form of $C_a(s)$ must be

$$C_a(s) = \sum_{k=0}^{k=\infty} P_{a_k} \exp(isa_k) \tag{32}$$

where $P_{a_k} \geqq 0$, $\sum_{k=0}^{k=\infty} P_{a_k} = 1$ and the random variables a_k must satisfy

$$a_k = (2\pi/s)k + b \tag{33}$$

where b is an arbitrary constant.

Thus if $|C_a(\alpha_r)| = 1$ the r.v.'s must be integral multiples of one another plus an arbitrary constant common to each of them.

Using (33) in (32) we see that the characteristic function becomes $\exp(ibs)$. We set $C_a = \exp(ibs)$ and evaluate the following limit:

$$\begin{aligned} \Lambda &= e^{iT\nu r} \lim_{N\to\infty} \frac{1}{N} \sum_{s=0}^{s=N-r-2} \sum_{k=s+r+1}^{k=N-1} e^{-i\gamma(k-s-r)} \\ &= e^{-iT\nu r} \lim_{N\to\infty} \left\{ \sum_{l=1}^{l=N} e^{i\gamma l} + \frac{1}{N} \sum_{l=1}^{l=N} (l+r) e^{i\gamma l} \right\} \end{aligned} \tag{34}$$

where

$$\gamma = T\nu + b\alpha_r .$$

Let

$$\begin{aligned} \Lambda_1(A) &= \sum_{l=1}^{l=\infty} [A \exp i\gamma]^l = \frac{A \exp i\gamma}{1 - A \exp i\gamma} \\ &= i \frac{d}{d\gamma} \ln (1 - A \exp i\gamma), \qquad A < 1. \end{aligned} \tag{35}$$

The limit of Λ_1 as $A \to 1$ is the first sum in (34). Obviously this limit

does not exist in the ordinary sense. However the "distribution" limit, denoted by $\lim^{(D)}$, Ref. 5, does exist. Barnard, Ref. 6, has shown that

$$\lim_{A\to 1^-}{}^{(D)} \ln (1 - A \exp i\gamma) = \ln |\sin (\gamma/2)| + \ln 2 + i[(\gamma/2) - \pi R_{2\pi}(\gamma) - 2\pi M] \tag{36}$$

where M is an arbitrary integer, and

$$R_{2\pi}(\gamma) = \sum_{n=0}^{\infty} \mu(\gamma - 2\pi n) - \sum_{n=1}^{\infty} \mu(-\gamma - 2\pi n)$$

$$R_{\pi}(\gamma) = R_{2\pi}(\gamma - \pi)$$

$$u(\gamma) = \begin{cases} 1, & \gamma \geqq 0 \\ 0, & \gamma < 0. \end{cases}$$

He also proved that the right side of (36) constitutes a properly defined generalized function or a distribution.

When (36) is differentiated with respect to γ we obtain

$$i\frac{d}{d\gamma}\left\{\lim_{A\to 1^-} \ln (1 - A \exp i\gamma)\right\} = \frac{i}{2}\cot\frac{\gamma}{2} - \frac{1}{2} + \pi \sum_{n} \delta(\gamma - 2\pi n) \tag{37}$$

where $\delta(\cdot)$ is the well known dirac delta-function.

The limit of the right hand sum in (34) approaches zero since this sum is proportional to the derivative of the first sum divided by N. Since the first sum is a generalized function or a distribution so is its derivative. Consequently in the distribution sense the limit is zero.

When the characteristic function is of the form (32), which implies that the r.v.'s satisfy (33), we observe that

$$C_a\{V[y + pT] - \alpha_r\} = e^{-ib\alpha_r} C_a\{V[y + pT]\} \tag{38}$$

$$p = r - m.$$

From this we obtain

$$\prod_{m=1}^{m=j} C_a^*\{V[y + pT] - \alpha_r\} = e^{b\alpha_r j} \prod_{m=1}^{m=j} C^*[y + pT] \tag{39}$$

Applying (38), (39) in (31), and replacing

$$\frac{e^{iT\nu} C_a(\alpha_r)}{1 - e^{iT\nu} C_a(\alpha_r)}$$

by (37), we obtain for the continuous as well as the discrete spectral density

$$\begin{aligned} G_+(\nu) &= \frac{A^2}{2T}\int_0^T\int_0^T dydy' e^{i\nu(y-y')} \\ &\cdot\prod_{p=0}^{p=r-1} C_a\{V[y+pT]-V[y'+pT]\} \\ &+\frac{A^2}{T}\,\mathrm{Re}\left\{\left|\int_0^T dy e^{i\nu y}\prod_{p=0}^{p=r-1} C_a\{V[y+pT]\}\right|^2\right. \\ &\cdot\left[1+e^{i\gamma r}\left(\pi\sum_n \delta(\gamma-2\pi n)-\frac{1}{2}+\frac{i}{2}\cot\frac{\gamma}{2}\right)\right] \\ &\left.+\sum_{j=1}^{j=2-1} f_{jr}(\nu)\right\}. \end{aligned} \tag{40}$$

By recalling the definition of γ in (34) we see that spectral lines occur when

$$T\nu + b\alpha_r = 2\pi n, \tag{41}$$

or

$$\nu = (2\pi n/T) - b(\alpha_r/T)$$

and the minimum spacing $\Delta\nu$ between the lines are given by

$$\Delta\nu = 2\pi\Delta f = \frac{2\pi(n+1)}{T} - \frac{2\pi n}{T} = \frac{1}{T}.$$

IV. SPECIAL CASES

In this section we select several special cases, believed to be of general interest, and exhibit them graphically.

The first case we want to explore is that in which $g(t)$ is a rectangular pulse of unit height and the a_n's are discrete, that is, a frequency shift keying system. For the binary case, the two frequencies are referred to as mark and space frequency, and each is located ω_d from the carrier, ω_c. For the multilevel case the frequency spacing is uniform. In reference to our general formula (31), the following parameters apply:

$$\begin{aligned} r &= 1 \\ V(\xi) &= \omega_d\xi, \qquad \alpha_1 = \omega_d T. \end{aligned} \tag{42}$$

When (42) is applied in (31) and after considerable manipulation we obtain the specialized formula for the one sided continuous density

$$G(\omega) = \frac{2A^2}{T}\left\{\langle |F(\omega - \omega_c - a\omega_d)|^2\rangle_a + 2\,\mathrm{Re}\left[\frac{\exp\,(iT(\omega_c - \omega))\langle F(\omega - \omega_c - \omega_d)\rangle_a}{1 - C_a(\omega_d T)\exp\,(iT(\omega_c - \omega))}\right.\right. \tag{43}$$
$$\left.\left.\cdot\langle F^*(\omega - \omega_c - a\omega_d)\exp\,(i\omega_d aT)\rangle_a\right]\right\},$$

where

$$F(\omega) = \frac{T}{2}\exp\,(-i\omega T/2)\,\frac{\sin\,\omega T/2}{\omega T/2}, \tag{44}$$

and

$$|C_a(\omega_d T)| < 1.$$

Let

$$\begin{aligned}\beta &= (\omega - \omega_c)T/2\pi \\ \gamma &= (\omega - \omega_c - a\omega_d)T/2,\end{aligned} \tag{45}$$

and write (43) as

$$G(\beta) = \frac{2A^2}{T}\left\{\left\langle\left|\frac{T}{2}e^{-i\gamma}\frac{\sin\,\gamma}{\gamma}\right|^2\right\rangle_a + 2\,\mathrm{Re}\left[\frac{e^{-2\pi\beta}\left\langle\frac{T}{2}e^{-i\gamma}\frac{\sin\,\gamma}{\gamma}\right\rangle_a\left\langle\frac{T}{2}e^{i\gamma}\frac{\sin\,\gamma}{\gamma}e^{ik\pi a}\right\rangle_a}{1 - e^{-i2\pi\beta}C_a(\omega_d T)}\right]\right\}.$$

$$\frac{G(\beta)}{A^2T} = \frac{I_1}{2} + \mathrm{Re}\left[\frac{I_2^2}{1 - C_a(\omega_d T)e^{-i2\pi\beta}}\right] \tag{46}$$

where

$$\begin{aligned}I_1 &= \left\langle\frac{\sin^2\,\gamma}{\gamma^2}\right\rangle_a \\ I_2 &= \left\langle\frac{\sin\,\gamma}{\gamma}e^{-i\gamma}\right\rangle_a.\end{aligned} \tag{47}$$

For binary frequency shift keying, the frequency deviation parameter, k, may be defined as the ratio of frequency shift (the difference between mark and space frequency) to the signaling frequency (the sum of the number of marks and the number of spaces in one second). That is,

$$k = \frac{\omega_m - \omega_s}{2\pi/T} = \frac{\omega_d T}{\pi}. \tag{48}$$

The same definition holds for multilevels with frequency assignments which make the frequency spacing uniform and equal to that for the binary case. The frequencies nearest to the carrier are located at $\omega_c \pm \omega_d$, the intermediate frequencies are at $\omega_c \pm (2n - 1)\omega_d$, and the ones furthest from the carrier are at $\omega_c \pm (N - 1)\omega_d$, where N is the number of levels. Thus, the frequency band of the power spectrum will increase approximately with N for constant k. The frequency band occupied can be kept approximately the same by letting k decrease with N.

In this example, the random variables a_n are discrete and may be represented as

$$a_n = 2n - (N + 1), \qquad n = 1, 2, \cdots, N. \tag{49}$$

The argument of F in (43) is

$$\gamma_n = (\omega - \omega_c - a_n\omega_d)T/2 = (\beta - a_n k/2)\pi, \tag{50}$$

and the equations in (47) become

$$\begin{aligned} I_1 &= \frac{1}{N}\sum_{n=1}^{N}\left[\frac{\sin\gamma_n}{\gamma_n}\right]^2 \\ I_2^2 &= \frac{1}{N^2}\sum_{n=1}^{N}\sum_{m=1}^{N}\exp -i(\gamma_n + \gamma_m)\,\frac{\sin\gamma_n}{\gamma_n}\,\frac{\sin\gamma_m}{\gamma_m}. \end{aligned} \tag{51}$$

Since we can alternatively write

$$a_n = \pm(2n - 1), \qquad n = 1, 2, \cdots, N/2$$

we have that

$$\begin{aligned} C_a(\omega_d T) = C_a(k\pi) &= \sum_n P_r(a_n)\exp\,(i\omega_d a_n T) \\ &= (1/N)\sum_{n=1}^{N/2}[e^{ik\pi(2n-1)} + e^{-ik\pi(2n-1)}] \\ &= (2/N)\sum_{n=1}^{N/2}\cos k\pi(2n - 1). \end{aligned} \tag{52}$$

The complex terms from (46) are

$$\begin{aligned} B &= \mathrm{Re}\left[\frac{e^{-i(\gamma_n+\gamma_m)}}{1 - C_a e^{-2\pi\beta}}\right] \\ &= \frac{\cos(\gamma_n + \gamma_m) - C_a\cos(\gamma_n + \gamma_m - 2\pi\beta)}{1 + C_a^2 - 2C_a\cos 2\pi\beta}. \end{aligned} \tag{53}$$

We can now write the N-level normalized spectral density as

$$\frac{G(\beta)}{A^2T} = \frac{1}{N}\sum_{n=1}^{N}\left[\frac{1}{2}\frac{\sin^2\gamma_n}{\gamma_n^2} + \frac{1}{N}\sum_{m=1}^{N} B\frac{\sin\gamma_n}{\gamma_n}\frac{\sin\gamma_m}{\gamma_m}\right], \qquad (54)$$

where γ_n and B are given by (50) and (53).

Using several values of the two parameters — N, the number of levels and k, the deviation — we have calculated numerically the spectral densities from the relations given above, and plotted them against the normalized frequency $\beta = (\omega - \omega_c)T/2\pi$. On this scale, ω_d occurs at $k/2$ for all values of N.

A large number of spectra are presented to indicate the way the shape changes as the frequency deviation varies. For binary FM, these are given in Fig. 1. We point out that the spectra for the binary cases

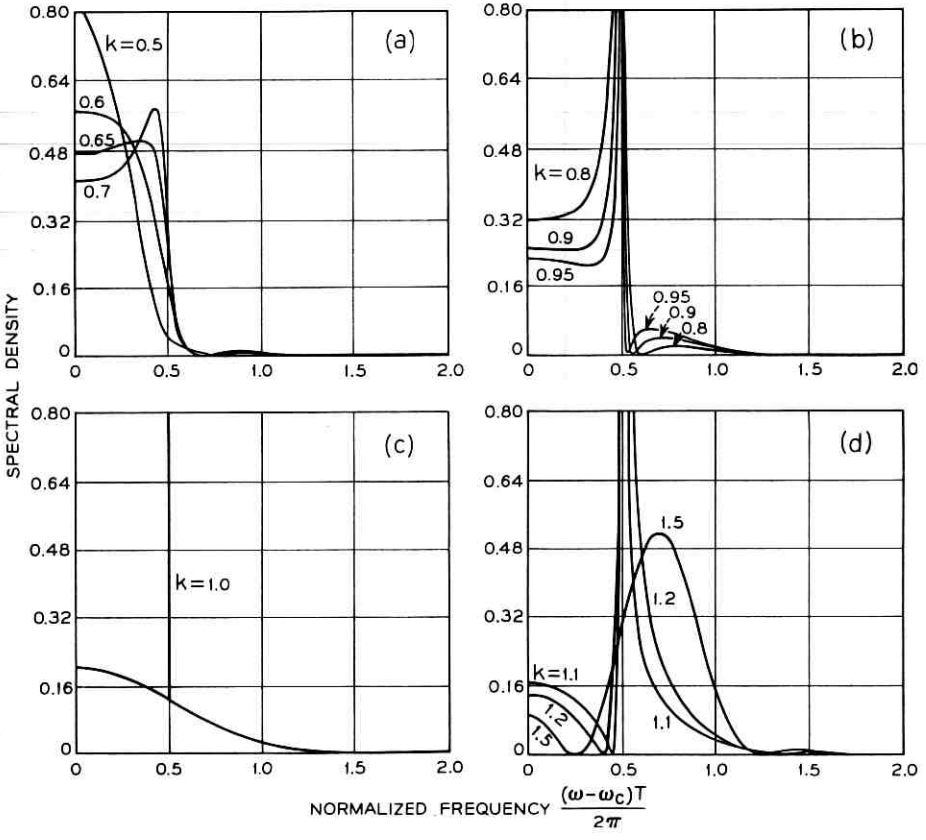


Fig. 1 — Spectral density for 2-level FM.

are the same as those given by Bennett and Rice, *op cit*, except that the origin of the frequency scale has been shifted from the lower (space) frequency to midband.

The spectra for 4- and 8-level FM are given in Figs. 2 and 3, respectively. The multilevel cases show considerable similarity to the binary ones. For small values of k, the spectra are narrow and decrease smoothly towards zero. In particular, Fig. 4 shows the spectra for $k = 1/N$, and these three are nearly identical. As k increases towards unity the spectrum widens, and as predicted, there tends to be concentration of power about the *a priori* chosen frequencies. This concentration is especially marked in the range $1 - 1/2N < k < 1 + 1/2N$. At $k = 1$, there is a spectral line at the frequency $\frac{1}{2}$, and its odd multiples. As k increases from unity the concentration at ω_d is again broadened, and reduced in intensity. We attempt to show these features in several plots as a function of k.

Fig. 5 shows the decrease in spectral density at zero (mid-band) frequency with increasing k. For higher level systems the zero-frequency level is less for any value of k, but the decrease with k is slower.

The position of the spectral peaks, as a function of k for the 8-level system is shown in Fig. 6. Other level systems show similar behavior. For $k = 1$, the *a priori* chosen frequencies are (measured from the carrier) at $\pm(2n - 1)/2$, and the delta functions in the spectral density occur at these same frequencies. For $k < 1$, the peaks of the spectral density are no longer delta functions, and they occur nearer the carrier than the chosen frequencies. They are further from the carrier for $k > 1$.

An interesting phenomenon is observed for the cases where k is the reciprocal of the number of levels. For these relations the principal portion of the spectrum is confined to a relatively narrow band. These curves have approximately the same shape as seen from the curves in Fig. 4 and the following table:

No. of Levels	k	Spectral Density at Freq = 0	.25	.5	.75
2	0.500	0.810	0.500	0.090	0.00
4	0.250	0.750	0.470	0.117	0.011
8	0.125	0.735	0.430	0.124	0.013

The program was extended to calculate the power in the continuous portion of the spectrum. For all values of N, and for k away from unity this power is $\frac{1}{2}$, and for $k = 1$, this power is $\frac{1}{2} - 1/2N$. Thus the total power in the spectral lines is $1/2N$. Clearly the power in each line is $1/2N^2$ since they are assumed to have uniform likelihood of occurrence. It is

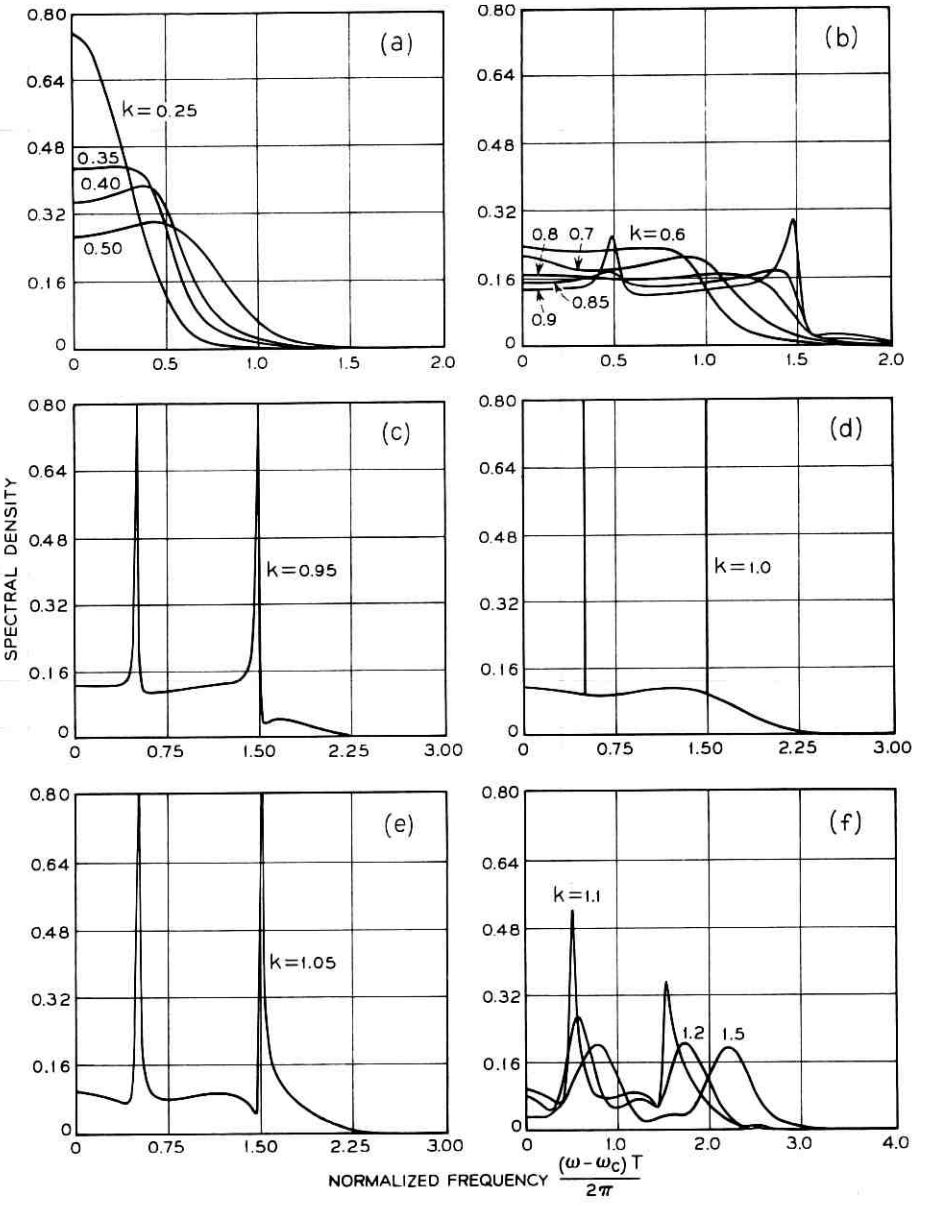


Fig. 2 — Spectral density for 4-level FM.

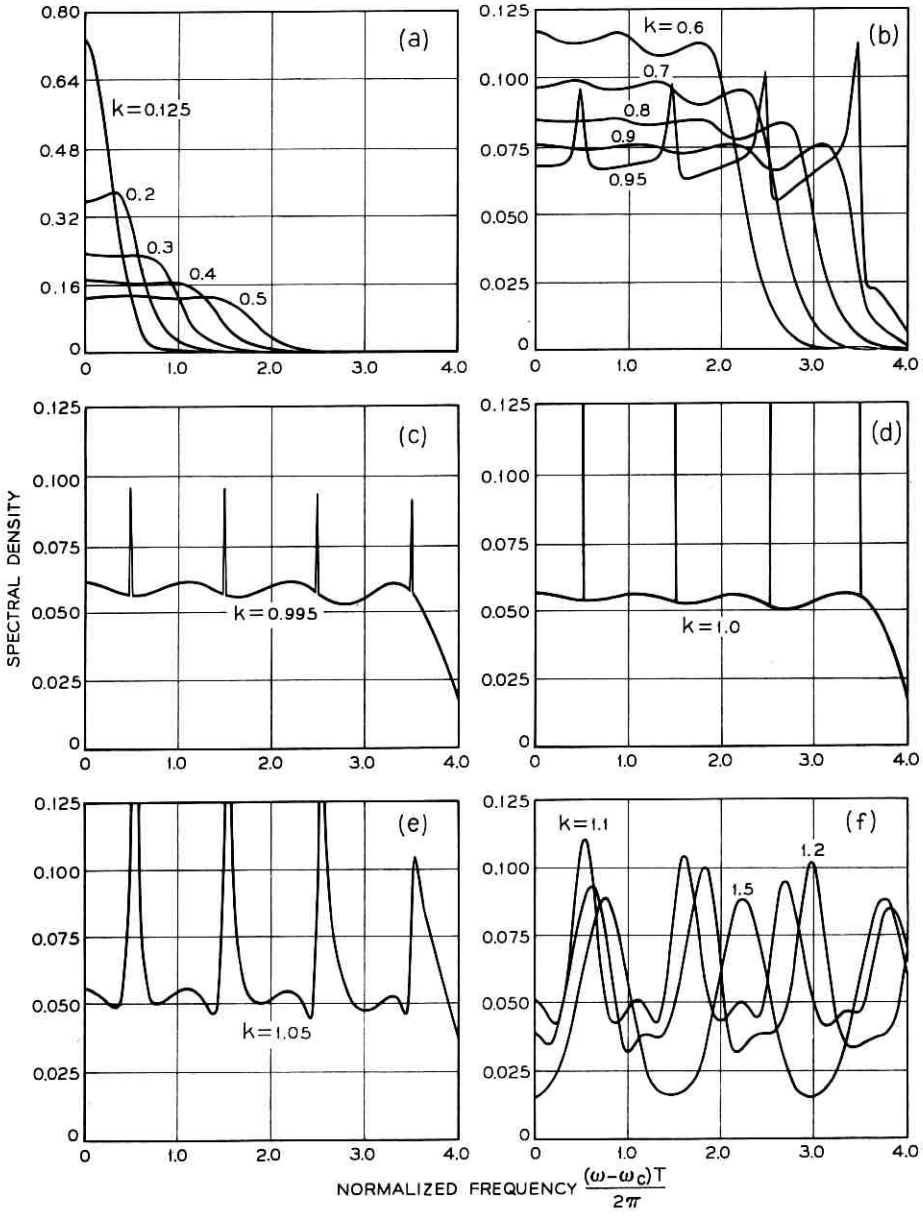


Fig. 3 — Spectral density for 8-level FM.

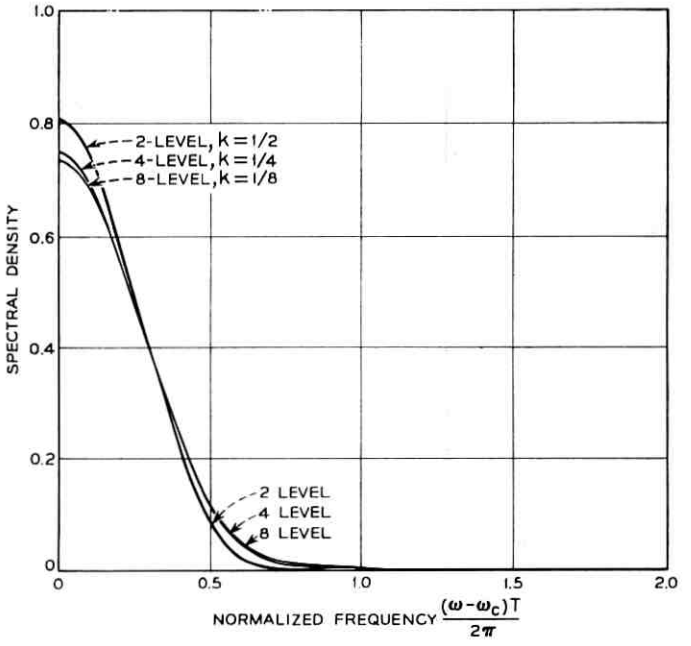


Fig. 4 — Spectral density for k = 1/No. of levels.

also very easy to show from (40) that this is the expected division of power between the continuous spectrum and the discrete spectral lines.

We also thought it interesting to exhibit spectral shapes when the a_n's have a gaussian probability distribution. This situation may arise in pulse frequency modulation with baseband amplitude samples possessing gaussian probability densities. In this case, the probability density of the a_n's is

$$P(a_n) = \frac{1}{\sqrt{2\pi}\,\sigma} \exp\left(-\frac{a_n^2}{2\sigma^2}\right) = \frac{1}{\sqrt{2\pi}\,\sigma} \exp\left[-\frac{(\gamma - \beta\pi)^2}{2\mu^2}\right], \quad (55)$$

and the characteristic function is

$$C_a(\omega_d T) = \exp\left[-\frac{(\omega_d T\sigma)^2}{2}\right] = \exp\,(-2\mu^2),$$

where

$$\mu = \omega_d T\sigma/2, \quad (56)$$

and

$$\gamma = \beta\pi - a\mu/\sigma.$$

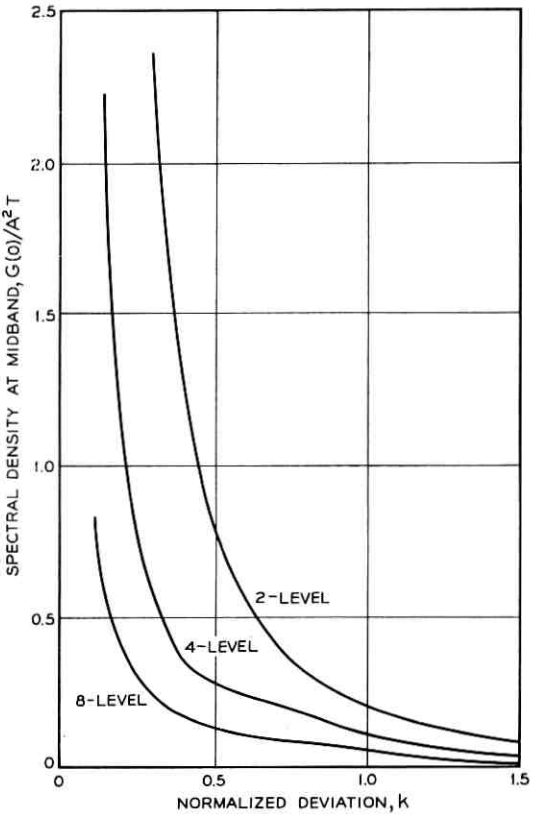


Fig. 5 — Spectral density at midband — Discrete multilevel case.

Equation (51) can be written as

$$I_1 = \frac{1}{\sqrt{2\pi}\,\mu}\int_{-\infty}^{\infty}\frac{\sin^2\gamma}{\gamma^2}\exp\left[-\frac{(\gamma-\beta\pi)^2}{2\mu^2}\right]d\gamma, \tag{57}$$

$$I_2 = \frac{1}{\sqrt{2\pi}\,\mu}\int_{-\infty}^{\infty}\frac{\sin\gamma}{\gamma}\,e^{-i\gamma}\exp\left[-\frac{(\gamma-\beta\pi)^2}{2\mu^2}\right]d\gamma.$$

Using elementary reductions I_1 is written as

$$I_1 = \frac{\exp\left[-\frac{1}{2}\left(\frac{\beta\pi}{\mu}\right)^2\right]}{2}\int_{-2}^{2}\left(1-\frac{|z|}{2}\right)\exp\left[-\frac{1}{2}\left(\mu z - i\frac{\beta\pi}{\mu}\right)^2\right]dz. \tag{58}$$

Let

$$\begin{aligned} t &= (xz/2) - iy \\ t_1 &= iy \\ t_2 &= x + iy \end{aligned} \tag{59}$$

where

$$x = \sqrt{2}\,\mu, \qquad y = \pi\beta/\sqrt{2}\,\mu. \tag{60}$$

Then

$$I_1 = \frac{2(Ax - By)}{x^2} + \frac{e^{-x^2}\cos 2\pi\beta - 1}{x^2}, \tag{61}$$

where

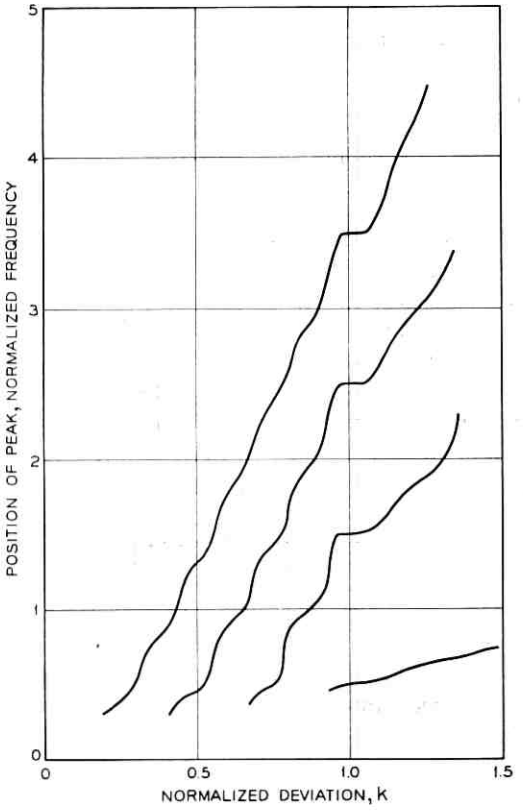


Fig. 6 — Position of spectral peaks — Discrete multilevel case.

$$A = e^{-y^2} \operatorname{Re}\left(\int_0^{t_2} e^{-t^2}\,dt - \int_0^{t_1} e^{-t^2}\,dt\right)$$
$$B = e^{-y^2} \operatorname{Im}\left(\int_0^{t_2} e^{-t^2}\,dt - \int_0^{t_1} e^{-t^2}\,dt\right). \tag{62}$$

In this same manner,

$$I_2 = \frac{1}{2\pi}\int_{-\infty}^{\tau} dz \exp\left(-i\beta\pi z - \frac{\mu^2 z^2}{2}\right)\int_{-\infty}^{\infty} d\gamma\,\frac{\sin\gamma}{\gamma}\,e^{-i\gamma z - i\gamma z}$$
$$= \frac{A - iB}{x} \tag{63}$$

and

$$I_2^2 = \frac{A^2 - B^2 - 2iAB}{x^2}. \tag{64}$$

Substituting (61) and (64) into (46) we obtain

$$\frac{G(\beta)}{A^2T} = \frac{1}{2x^2}\left[2(Ax - \beta y) + e^{-x^2}\cos 2\pi\beta - 1 + \frac{(A^2 - B^2)(e^{-x^2} - \cos 2\pi\beta) + 2AB\cos 2\pi\beta}{\cosh x^2 - \cos 2\pi\beta}\right]. \tag{65}$$

In this case the deviation is controlled by the parameter $\mu = \omega_d T\sigma/2$. In Fig. 7 we display the spectra for several values of this parameter. When $\mu = 0$, the spectrum is a delta function at $\beta = 0$ (midband). As μ increases the spectrum widens, approximately as μ and the midband value decreases approximately as μ^{-2}, for small μ, and as μ^{-1} for larger values. We show these two trends in Figs. 8 and 9.

The values of spectral density at $\beta = 0$, together with the asymptotes, are shown in Fig. 8. Two estimates of the width are shown in Fig. 9. From the definition of the Gabor bandwidth, Ref. 7,

$$\sigma_G = \left[\frac{\int_0^{\infty} G(\beta)\beta^2 d\beta}{\int_0^{\infty} G(\beta)\,d\beta}\right]^{\frac{1}{2}}. \tag{66}$$

We note that σ_G is very nearly equal to μ/π. That is, the standard deviation of the power spectrum is the same as the standard deviation of the input times the normalized deviation frequency:

$$\sigma_G \doteq \sigma\ (\omega_d T/2\pi). \tag{67}$$

Another estimate of the width of the power spectrum comes from the value of β at which the density has fallen by e, namely

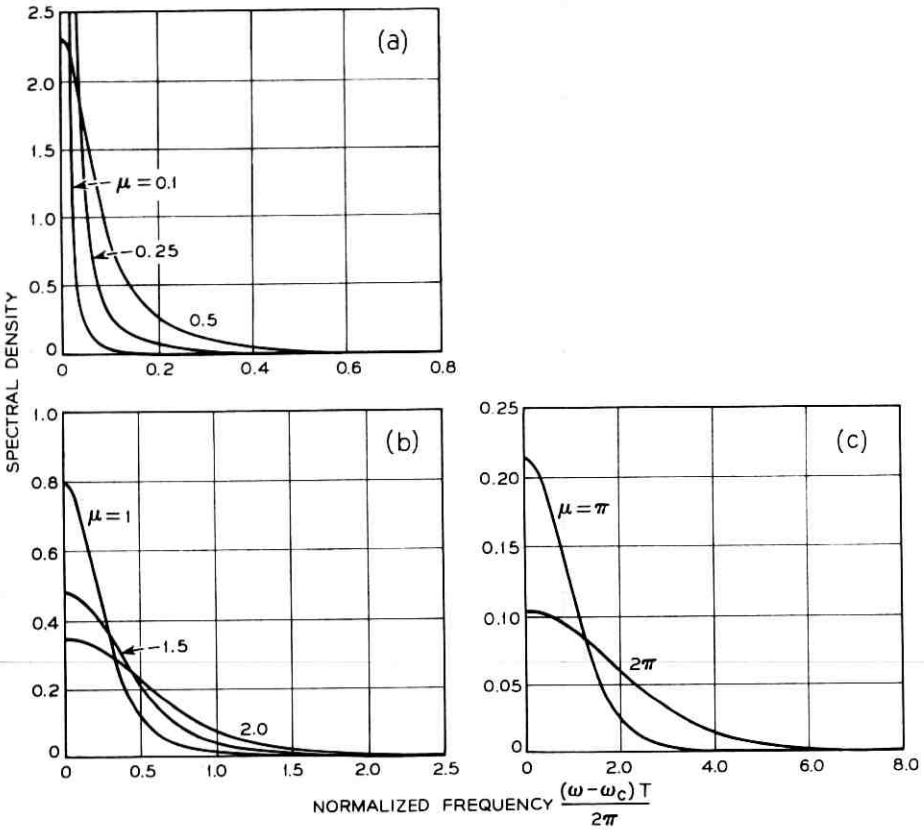


Fig. 7 — Spectral density for gaussian distribution.

$$\sigma_E = \frac{\beta}{\sqrt{\pi}} \quad \text{at} \quad \frac{G(0)}{e}. \tag{68}$$

At high values of σ, where the spectral density curves appear more nearly gaussian, σ_E approaches σ_G.

The gaussian case spectral density curves were also integrated to obtain the power. We obtain $\frac{1}{2}$ in all cases, thus providing a check on our work. It is interesting to note that even for μ as low as 0.5, 98 per cent of the power lies within $3\sigma_G$ of midband.

V. SUMMARY OF CURVES

Spectra are presented for 2, 4, and 8 equally-spaced uniformly distributed frequencies and for normally distributed frequencies. The gen-

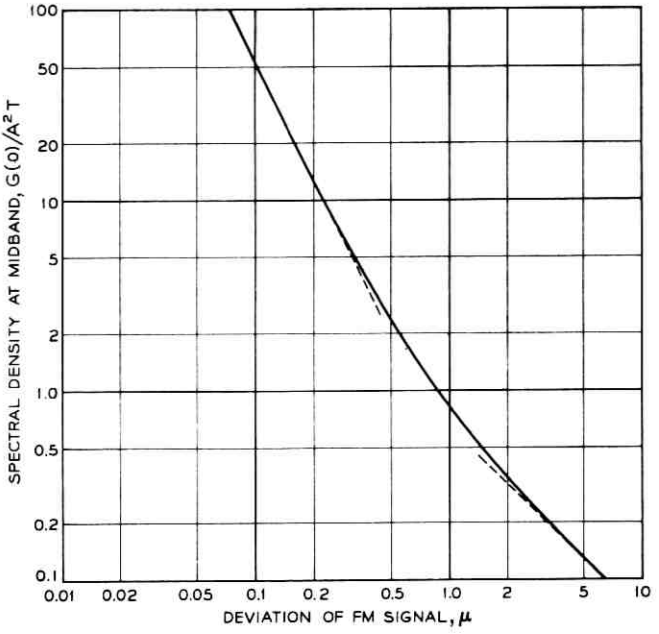


Fig. 8 — Spectral density at midband — Gaussian case.

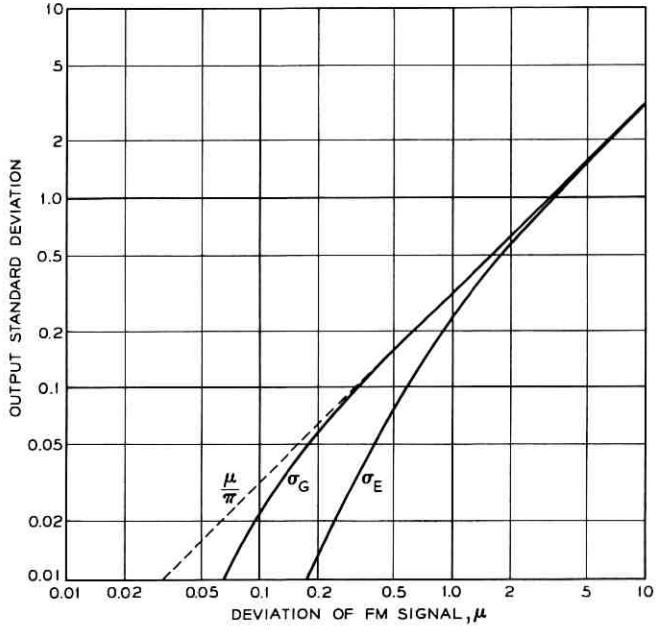


Fig. 9 — Standard deviation of spectral density function — Gaussian case.

eral trend of the curves, as a function of the frequency deviation, is shown in Figs. 5, 6, 8, and 9. As expected, the band occupied by a significant portion of the spectrum increases with the deviation.

For the discrete multilevel case, the frequency deviation parameter is $k = \omega_d T/\pi$. For $k = 1/N$ the spectral density functions for different N are nearly identical. They are relatively narrow and decrease smoothly to zero. Line spectra occur at the *a priori* chosen frequencies when k is an integer.

For the gaussian case the deviation parameter is $\mu = \sigma\omega_d T/2$. For large μ, the shape of the spectral density approaches a gaussian curve with a standard deviation of $\sigma\omega_d T/2\pi$. For lower μ, the curves are slightly narrower with correspondingly longer tails. The maximum value for each μ, which occurs at $\beta = 0$, approaches $\frac{1}{2}\mu^2$ for small μ and $2/\pi\mu$ for large values.

REFERENCES

1. Salz, J., Performance of Multilevel Narrow-Band FM Digital Communication Systems, to be published.
2. Bennett, W. R., and Rice, S. O., Spectral Density and Autocorrelation Functions Associated with Binary Frequency Shift Keying, B.S.T.J., *42*, Sept. 1963.
3. Salz, J., Spectral Density Function of Multilevel Continuous Phase FM, IEEE Trans. Information Theory, July, 1965.
4. Loève, M., *Probability Theory*, D. Van Nostrand Company, pp. 201–202, 1960.
5. Lighthill, M. J., *An Introduction to Fourier Analysis and Generalized Functions*, Cambridge University Press, 1959.
6. Barnard, R. D., A Note on a Special Class of One-Sided Distribution Sums, B.S.T.J., *44*, Nov., 1965.
7. Gabor, D., Theory of Communication, J. Inst. Elect. Engrs., *93* III, p. 429.

On the Decomposition of Lattice-Periodic Functions

By R. L. GRAHAM

(Manuscript received March 31, 1965)

The problem of decomposing an arbitrary periodic function defined on an n-dimensional cubic lattice into finite linear combinations of certain primitive functions is considered. Generally, a primitive function is one which periodically assumes only the values ± 1 and 0. Rather simple necessary and sufficient conditions are derived for such a decomposition and when a decomposition is possible, an algorithm is given which accomplishes it. These results have been used in recent generalizations of the Ewald method.

I. INTRODUCTION

In a study of the classical problem of the calculation of the potential due to an ionic crystal lattice, and in particular, generalizations of the Ewald method (Ref. 1) along the lines of Nijboer and Dewette (Refs. 3, 4), W. J. C. Grant (Ref. 2) proposed the following problem: Suppose we say that an ionic crystal lattice is primitive if for a suitable choice of origin there exist three vectors $\bar{x}_1$, $\bar{x}_2$, $\bar{x}_3$ such that the charge at the point $n_1\bar{x}_1 + n_2\bar{x}_2 + n_3\bar{x}_3$ is just $q_0(-1)^{n_1+n_2+n_3}$ for some fixed q_0 and for all triples of integers (n_1 , n_2 , n_3) and that the charge at all other points is zero. (For example, the ordinary NaCl lattice is primitive with the $\bar{x}_i$ taken to be the unit coordinate vectors and $q_0 = 1$.) The question is then: Which crystal lattices can be decomposed into finite sums of primitive lattices? Different primitive lattices in the decomposition may have different origins and by the sum of two lattices we mean, of course, the component-wise sum.

The object of this paper is threefold:

(*i*) The problem is extended to its natural n-dimensional analogue.

(*ii*) Rather simple necessary and sufficient conditions are given for the existence of the desired decomposition.

(*iii*) When such a decomposition is possible, an algorithm is given which accomplishes it.

II. PRELIMINARY IDEAS

In order to illustrate the basic ideas which will be used in the proofs of the general (n-dimensional) theorem (see p. 1200), we begin by considering the following one-dimensional version.

Suppose we call a real-valued function f defined on the integers *primitive* if for some integers x and c it is true that

$$f(z) = \begin{cases} (-1)^a & \text{if} \quad z = ax + c \\ 0 & \text{otherwise} \end{cases}$$

for all integers a. The question then becomes: What is the set of all those functions which can be represented as real finite linear combinations* of primitive functions? For example, the function g defined by:

$$g(z) = \begin{cases} 0 & \text{if} \quad z \equiv 0 \pmod 4 \\ 1 & \text{if} \quad z \equiv 1 \pmod 4 \\ 2 & \text{if} \quad z \equiv 2 \pmod 4 \\ -3 & \text{if} \quad z \equiv 3 \pmod 4 \end{cases}$$

may be decomposed into a linear combination of primitive functions by:

$$g(z) = -g_1(z) + 2g_2(z) + g_3(z)$$

where

$$g_1(z) = \begin{cases} (-1)^a & \text{if} \quad z = 2a \\ 0 & \text{otherwise} \end{cases},$$

$$g_2(z) = \begin{cases} (-1)^a & \text{if} \quad z = 2a + 1 \\ 0 & \text{otherwise} \end{cases},$$

$$g_3(z) = (-1)^z.$$

We can write this more graphically if we use the notation

$$f\colon \cdots, a_0, a_1, a_2, \cdots$$

to denote the fact that $f(0) = a_0, f(1) = a_1$, etc. We then have

* In this paper, linear combination will always mean *finite* linear combination.

$$\begin{aligned} -g_1 &: \cdots, -1, \quad 0,1, \quad 0, \cdots \\ 2g_2 &: \cdots, \quad 0, \quad 2,0,-2, \cdots \\ g_3 &: \cdots, \quad 1,-1,1,-1, \cdots \\ \hline g &: \cdots, \quad 0, \quad 1,2,-3, \cdots . \end{aligned}$$

Similarly, if we start with

$$h: \cdots, 1,3,-2,-1,-3,2, \cdots$$

then the desired decomposition is easily found to be:

$$\begin{aligned} h &: \cdots, 1,0, \quad 0,-1, \quad 0,0, \cdots \\ 3h_2 &: \cdot\ \cdot, 0,3, \quad 0, \quad 0,-3,0, \cdots \\ -2h_3 &: \cdots, 0,0,-2, \quad 0, \quad 0,2, \cdots \\ \hline h_1 &: \cdots, 1,3,-2,-1,-3,2, \cdots . \end{aligned}$$

In general, it is clear that any linear combination f of primitive functions is *periodic* and that *within a period the sum of the function values of f must be zero*. If a function has these latter two properties, we say that the function has *mean zero*. It might at first be surmised that any function with mean zero could be written as a linear combination of primitive functions. However, attempts to decompose the periodic function

$$g: \cdot\cdot\ \overline{,1,-1,0,1,-1,0,} \cdots .$$

(the bar indicating a complete period) soon lead one to suspect that this initial guess is incorrect. (In fact, g *cannot* be decomposed into primitive functions.)

One question which arises immediately is exactly which periods the primitive components of a function f might have, if f itself has some period p (where we say that f has period p if $f(z + p) = f(z)$ for all z). In the preceding example, while g has period 3, perhaps there is a decomposition of g for which the primitive components have much larger periods. (It will turn out, however, that this is not possible.)

To answer these questions, we first introduce some notation. If g is a function defined on the integers,* then by $g(z/r)$ we mean the function defined by:

$$g\left(\frac{z}{r}\right) = \begin{cases} g\left(\frac{z}{r}\right) & \text{if } \frac{z}{r} \text{ is an integer} \\ 0 & \text{otherwise} \end{cases}$$

* In general, in this paper all functions assume the value 0 on points with non-integral coordinates.

Let $i(z)$ denote the function which assumes the value 1 on all integers. Thus, if the function f which we wish to decompose has period

$$p = 2^a(2m + 1),$$

then by forming the functions $i[(z - k)/(2m + 1)]f(z), 0 \leqq k < 2m + 1$, we have functions which "sample" f at points separated by a distance of $2m + 1$. For example, if f is given by

$$f\colon \cdots, \overline{1,3,-6,3,-5,4,1,3,-6,3,-5,4,} \cdots$$

so that the period of f is $6 = 2 \cdot 3$ (where we will assume that $f(0) = 1$) then we have

$$i\left(\frac{z}{3}\right)f(z)\colon \cdots, \overline{1, 0, \quad 0, 3, \quad 0, 0,} \cdots$$

$$i\left(\frac{z-1}{3}\right)f(z)\colon \cdots, \overline{0, 3, \quad 0, 0, -5, 0,} \cdots$$

$$i\left(\frac{z-2}{3}\right)f(z)\colon \cdots, \overline{0, 0, -6, 0, \quad 0, 4,} \cdots$$

Note that $i[(z - k)/(2m + 1)]f(z)$ also has period p and, in general,

$$f(z) = \sum_{k=0}^{2m} i\left(\frac{z-k}{2m+1}\right)f(z).$$

The result toward which the remainder of this section will be devoted can now be expressed simply in the following way: If f has period

$$p = 2^a(2m + 1)$$

then f *can be expressed as a linear combination of primitive functions* if and only if *for each* k *the function* $i[(z - k)/(2m + 1)]f(z)$ *has mean zero.*

It follows from this, for example, that if $p = 2^a$ then f can be decomposed into primitive functions if f has mean zero. On the other hand, if f has an *odd* period $p = 2m + 1$, then each function $i[(z - k)/(2m + 1)]$ has just one nonzero value per period so that f can be decomposed into primitive function if it is identically zero.

We now give a series of lemmas, informal proofs and examples which will indicate the ideas needed for the proof of the general theorem. An outline of our plan of attack is to establish the following results:

> If f is a linear combination of primitive functions then for any k and for any $r \neq 0, f[(z - k)/r]$ also is a linear combination of primitive functions. (1)

If f is a linear combination of primitive functions and f has period $p = 2^a(2m + 1)$ then for all k, $i[(z - k)/(2m + 1)]f(z)$ has mean zero. (2)

If f has period 2^a and mean zero then f is a linear combination of primitive functions. (3)

Assuming we have established (1), (2) and (3), the proof of the original assertion follows directly. One direction follows immediately from (2). To show the other direction assume that for each k, $i[(z - k)/(2m + 1)]f(z)$ has mean zero. Notice that each function $i[(z - k)/(2m + 1)]f(z)$ is just an "expanded" copy of a function $f_k(z)$ which has period 2^a and mean zero (i.e., $i[(z - k)/(2m + 1)]f(z) = f_k[(z - k)/(2m + 1)]$). Hence, by (3), $f_k(z)$ is a linear combination of primitive functions and it then follows by (1) that this is also true of $f_k[(z - k)/(2m + 1)]$. Consequently

$$f(z) = \sum_{k=0}^{2m} i\left(\frac{z - k}{2m + 1}\right) f(z) = \sum_{k=0}^{2m} f_k\left(\frac{z - k}{2m + 1}\right)$$

is a linear combination of primitive functions and the proof is completed.

It remains to prove (1), (2) and (3).

The proof of (1) is straightforward. We first note that if $f(z)$ is primitive then $f[(z - k)/r]$ is also primitive for any k and for any $r \neq 0$. For by hypothesis there exist x and c such that

$$f(z) = \begin{cases} (-1)^a & \text{if } z = ax + c \\ 0 & \text{otherwise} \end{cases}.$$

On the other hand, by definition we have

$$f\left(\frac{z - k}{r}\right) = \begin{cases} f(y) & \text{if } z = ry + k \\ 0 & \text{otherwise} \end{cases}.$$

Hence

$$f\left(\frac{z - k}{r}\right) = \begin{cases} (-1)^a & \text{if } z = r(ax + c) + k \\ 0 & \text{otherwise} \end{cases}$$

$$= \begin{cases} (-1)^a & \text{if } z = a(rx) + (rc + k) \\ 0 & \text{otherwise} \end{cases}$$

and so $f[(z - k)/r]$ is primitive. The extension to linear combinations of primitive functions follows at once and (1) is proved.

In order to prove (2) we first need an auxiliary result (a simplified version of Lemma 2). This is: Suppose f has periods $p = 2^a(2m + 1)$

and $p' = 2^{a'}(2m' + 1)$. Then for any k, $i[(z - k)/(2m + 1)]f(z)$ has mean zero iff $i[(z - k)/(2m' + 1)]f(z)$ has mean zero. To prove this, let us first assume that $a' = a$ and $2m + 1$ divides $2m' + 1$. The sum $\sum_{z=0}^{p-1} i[(z - k)/(2m + 1)]f(z)$ is exactly the sum of the $f(z)$ for which $z - k \equiv 0 \pmod{2m + 1}$, and $0 \leqq z - k \leqq p - 1$ (since $i[(z - k)/(2m + 1)]f(z)$ has period p). There are 2^a such values of $z - k$, namely,

$$z - k \,\varepsilon\, A = \{0, 2m + 1, 2(2m + 1), \cdots, (2^a - 1)(2m + 1)\}.$$

Similarly the sum $\sum_{z=0}^{p'-1} i[(z - k)/(2m' + 1)]f(z)$ is exactly the sum of the $f(z)$ for which $z - k \equiv 0 \pmod{2m' + 1}$ and $0 \leqq z - k \leqq p' - 1$. Again there are 2^a such values of $z - k$, namely,

$$z - k \,\varepsilon\, B = \{0, 2m' + 1, 2(2m' + 1), \cdots, (2^a - 1)(2m' + 1)\}.$$

All the elements of A and B are congruent to zero modulo $2m + 1$ (since $2m + 1$ divides $2m' + 1$). Also since $2m + 1$ and $2m' + 1$ are *odd* then both sets A and B contain a *complete residue system modulo* 2^a. Hence modulo p, A and B are identical. Since f has period p then

$$\sum_{z=0}^{p-1} i\left(\frac{z - k}{2m + 1}\right) f(z) = \sum_{z \varepsilon A} f(z) = \sum_{z \varepsilon B} f(z) = \sum_{z=0}^{p'-1} i\left(\frac{z - k}{2m' + 1}\right) f(z).$$

If we now assume that $a' \geqq a$ (instead of $a' = a$) then it is not difficult to see that

$$\sum_{z=0}^{p'-1} i\left(\frac{z - k}{2m' + 1}\right) f(z) = 2^{a'-a} \sum_{z=0}^{p-1} i\left(\frac{z - k}{2m + 1}\right) f(z).$$

Thus, what we have shown is that if f has periods $p = 2^a(2m + 1)$ and $p' = 2^{a'}(2m' + 1)$ where p divides p' then $i[(z - k)/(2m + 1)]f(z)$ has mean zero iff $i[(z - k)/(2m' + 1)]f(z)$ has mean zero. Since in general a function which has periods q and q' also has period (q,q') (the greatest common divisor of q and q'), then the initial assertion follows at once. As a simple example consider the function f given by

$$f\colon \cdots, \overline{1,3,-2,-1,4,2}, 1,3,-2,-1,4,2, \cdots.$$

This function has $6 = 2 \cdot 3$ as a period and $i(z/3)f(z)$ has mean zero since

$$\sum_{z=0}^{5} i(z/3)f(z) = f(0) + f(3) = 1 - 1 = 0.$$

However we may also consider f as having a period of $12 = 2^2 \cdot 3$ in which case $i(z/3)f(z)$ has also mean zero since

$$\sum_{z=0}^{11} i(z/3)f(z) = f(0) + f(3) + f(6) + f(9) = 0.$$

Finally, f has a period of $18 = 2\cdot 3^2$ and $i(z/9)f(z)$ still has mean zero since

$$\sum_{z=0}^{17} i(z/9)f(z) = f(0) + f(9) = 0.$$

Our next step will be to prove (2) using the result just established. We first show that if f is *primitive* and has period $p = 2^a(2m + 1)$ then

$$\sum_{z=0}^{p-1} i\left(\frac{z - k}{2m + 1}\right) f(z) = 0 \quad \text{for all } k. \tag{4}$$

To see this, we partition the integers into two-element subsets $\{u_i, v_i\}$ such that $v_i = u_i + 2m + 1$ for each i. Since f is primitive there exist integers x and c such that

$$f(z) = \begin{cases} (-1)^a & \text{if } z = ax + c \\ 0 & \text{otherwise} \end{cases}.$$

Since

$$u_i x \equiv v_i x \pmod{2m + 1}$$

and $v_i - u_i = 2m + 1$ is odd then it follows that

$$f(u_i x + c) = -f(v_i x + c) \quad \text{for all } i.$$

But

$$i\left(\frac{u_i x - k}{2m + 1}\right) = i\left(\frac{v_i x - k}{2m + 1}\right) \quad \text{for all } i \text{ and } k.$$

Consequently it follows from the fact that f has period p that

$$\sum_{z=0}^{p-1} i\left(\frac{z - k}{2m + 1}\right) f(z) = 0 \quad \text{for all } k$$

and (4) is established.

To establish (2) assume that f has period $p = 2^a(2m + 1)$ and is a linear combination of primitive functions f_i, $1 \leqq i \leqq t$. If f_i has period $p_i = 2^{a_i}(2m_i + 1)$ then by (4) we know that

$$\sum_{z=0}^{p_i-1} i\left(\frac{z - k}{2m_i + 1}\right) f_i(z) = 0 \quad \text{for all } k.$$

Hence, if we choose $q = p_1 p_2 \cdots p_t p = 2^{a'}(2m' + 1)$ then by the simplified version of Lemma 2, we have

$$\sum_{z=0}^{q-1} i\left(\frac{z-k}{2m'+1}\right) f_i(z) = 0 \quad \text{for} \quad 1 \leqq i \leqq t \quad \text{and all } k.$$

Consequently

$$\sum_{z=0}^{q-1} i\left(\frac{z-k}{2m'+1}\right) f(z) = 0$$

since by hypothesis f is a linear combination of the f_i. But f has period p so applying the Lemma 2 result again we find

$$\sum_{z=0}^{p-1} i\left(\frac{z-k}{2m+1}\right) f(z) = 0$$

and (2) is proved.

We are left with (3) to prove. To do this we first establish the following result: If $h(z)$ is defined by $h(z) = (-1)^z$ then for a fixed n, the $2^n - 1$ functions $h[(z-k)/2^r]$, $0 \leqq k < 2^r$, $0 \leqq r < n$, are linearly independent over the reals. This is easy to see since for a fixed r, $h[(z-k)/2^r]$ assigns a nonzero value only to those z such that $z \equiv k \pmod{2^r}$. Hence for $k = 0, 1, \cdots, 2^r - 1$, the $h[(z-k)/2^r]$ assume nonzero values on disjoint sets. On the other hand, $h[(z-k)/2^r]$ assigns *different* values to the points k and $k + 2^r$ while any $h[(z-k')/2^{r'}]$ assigns the *same* value to these points for $r' < r$. Thus, $h[(z-k)/2^r]$ is not a linear combination of other $h[(z-k)/2^s]$ for $s \leqq r$. This establishes the independence of the h's. Note that for $0 \leqq k < 2^r$ and $0 \leqq r < n$, the function $h[(z-k)/2^r]$ has period 2^n and mean zero. By taking suitable linear combinations of the $2^n - 1$ independent $h[(z-k)/2^r]$, we can form functions f which assume any desired values on the points 0, 1, 2, $\cdots$, $2^n - 2$. Of course, we must have

$$f(2^n - 1) = -\sum_{z=0}^{2^n-2} f(z).$$

Consequently the $2^n - 1$ functions $h[(z-k)/2^r]$ form a *basis* for the set of all periodic functions with period 2^n and mean zero. That is, any function f with period 2^n and mean zero can be written as a linear combination of primitive functions with period 2^n. This completes the proof of (3).

To conclude this section we give an example which illustrates the ease with which the primitive components of a function may be found. Consider the function g given by:

$$g\colon \cdots, \overline{1,2,-5,\pi,\tfrac{1}{2},1-\pi,\tfrac{1}{2},0,} \cdots.$$

g has period 8 and mean zero. The only component $h[(z-k)/2^r]$ which can cause a difference in $g(0)$ and $g(4)$ is

$$h(z/4)\colon \cdots, 1,0,0,0,-1,0,0,0, \cdots.$$

Since $\alpha\, h(z/4)$ assigns the points 0 and 4 values which differ by 2α and

$$g(0) - g(4) = \tfrac{1}{2}$$

then by choosing $\alpha = \frac{1}{4}$ we obtain the $h(z/4)$ component of g. Performing similar calculations for $h[(z-k)/4]$, $k = 1, 2, 3$, we obtain

$$g_1(z) = g(z) - \frac{1}{4}h\left(\frac{z}{4}\right) - \left(\frac{\pi+1}{2}\right)h\left(\frac{z-1}{4}\right) + \frac{11}{4}h\left(\frac{z-2}{4}\right) - \frac{\pi}{2}h\left(\frac{z-3}{4}\right)$$

given by

$$g_1(z)\colon \cdots, \overline{\frac{3}{4}, \frac{3-\pi}{2}, -\frac{9}{4}, \frac{\pi}{2}, \frac{3}{4}, \frac{3-\pi}{2}, -\frac{9}{4}, \frac{\pi}{2},} \cdots$$

(which has period 4). We apply the same arguments to the decomposition of $g_1(z)$ into the $h[(z-k)/2]$, $k = 0,1$, and find

$$g_2(z) = g_1(z) - \frac{3}{2}h\left(\frac{z}{2}\right) - \left(\frac{3}{4} - \frac{\pi}{2}\right)h\left(\frac{z-1}{2}\right)$$

given by

$$g_2(z)\colon \cdots, \overline{-\frac{3}{4}, \frac{3}{4},} -\frac{3}{4}, \frac{3}{4}, -\frac{3}{4}, \frac{3}{4}, -\frac{3}{4}, \frac{3}{4}, \cdots$$

so that $g_2(z) = -\frac{3}{4}h(z)$. Consequently g has been decomposed into primitive functions. Graphically we have:

$$\begin{array}{rl}
\frac{1}{4}h\left(\frac{z}{4}\right): & \cdots,\ \frac{1}{4},\ 0,\ 0,\ 0,-\frac{1}{4},\ 0,\ 0,\ 0,\ \cdots \\
\left(\frac{\pi+1}{2}\right)h\left(\frac{z-1}{4}\right): & \cdots,\ 0,\frac{\pi+1}{2},\ 0,\ 0,\ 0,-\left(\frac{\pi+1}{2}\right),\ 0,\ 0,\ \cdots \\
-\frac{11}{4}h\left(\frac{z-2}{4}\right): & \cdots,\ 0,\ 0,-\frac{11}{4},\ 0,\ 0,\ 0,\frac{11}{4},\ 0,\ \cdots \\
\frac{\pi}{2}h\left(\frac{z-3}{4}\right): & \cdots,\ 0,\ 0,\ 0,\ \frac{\pi}{2},\ 0,\ 0,\ 0,-\frac{\pi}{2},\ \cdots \\
\frac{3}{2}h\left(\frac{z}{2}\right): & \cdots,\ \frac{3}{2},\ 0,-\frac{3}{2},\ 0,\ \frac{3}{2},\ 0,-\frac{3}{2},\ 0,\ \cdots \\
\left(\frac{3}{4}-\frac{\pi}{2}\right)h\left(\frac{z-1}{2}\right): & \cdots,\ 0,\frac{3}{4}-\frac{\pi}{2},\ 0,\frac{\pi}{2}-\frac{3}{4},\ 0,\ \frac{3}{4}-\frac{\pi}{2},\ 0,\frac{\pi}{2}-\frac{3}{4},\ \cdots \\
-\frac{3}{4}h(z): & \cdots,-\frac{3}{4},\ \frac{3}{4},-\frac{3}{4},\ \frac{3}{4},-\frac{3}{4},\ \frac{3}{4},-\frac{3}{4},\ \frac{3}{4},\ \cdots \\
\hline
g(z): & \cdots,\ 1,\ 2,\ -5,\ \pi,\ \frac{1}{2},\ 1-\pi,\ \frac{1}{2},\ 0,\ \cdots
\end{array}$$

III. THE GENERAL THEOREM

We are ready to proceed to the n-dimensional generalizations of the results of Section II. The proofs given will use basically the same ideas as before although the technical details become somewhat more formal and involved. We begin with some definitions.

Let Z^n denote the ring of n-tuples of integers with component-wise addition and multiplication. That is, if $a = (a_1, \cdots, a_n)$ and $b = (b_1, \cdots, b_n)$ are elements of Z^n then

$$a + b = (a_1 + b_1, \cdots, a_n + b_n)$$

and

$$a \cdot b = (a_1 \cdot b_1, \cdots, a_n \cdot b_n).$$

In general, unless otherwise noted, lower case Latin letters without subscripts will denote elements of Z^n; lower case letters with subscripts will denote elements of Z, i.e., integers. If $\alpha \in Z$ and $q = (q_1, q_2, \cdots, q_n) \in Z^n$ then we define αq to be $(\alpha q_1, \alpha q_2, \cdots, \alpha q_n)$. The n-tuple $(1,1, \cdots, 1)$ will be denoted by e. By $a < b$ we mean $a_i < b_i$ for $1 \leqq i \leqq n$.

A function $f: Z^n \to R$ (the real numbers) is said to be *primitive* if there exist $a, x^{(1)}, \cdots, x^{(n)} \in Z^n$ such that

$$f(z) = \begin{cases} (-1)^{c_1+\cdots+c_n} & \text{if } z = c_1x^{(1)} + \cdots + c_nx^{(n)} + a \\ 0 & \text{otherwise} \end{cases}$$

for all $z \in Z^n$.

Let $\mathcal{G}$ denote the real vector space generated by the set of all primitive functions. Z^{n+} will denote the subset of Z^n consisting of those n-tuples which have all *positive* coordinates. If $m \in Z^{n+}$ then P_m is defined by

$$P_m = \{x \in Z^n: 0 \leqq x < m\}$$

(i.e., $0 \leqq x_i < m_i$ for $1 \leqq i \leqq n$, where 0 will be used to designate both an element of Z and also the n-tuple $(0, 0, \cdots, 0)$.)

A function $f: Z^n \to R$ is said to have *period* m if

$$f(z + km) = f(z) \quad \text{for all } z, \quad k \in Z^n.$$

If f has period m and

$$\sum_{z \in P_m} f(z) = 0$$

then f is said to have *mean zero*. Let $\mathfrak{F}_m$ denote the real vector space of all functions of period m which have mean zero. Next, we define

$$f[(z - a)/b], \qquad b \neq 0,$$

by

$$f\left(\frac{z-a}{b}\right) = \begin{cases} f(y) & \text{if } z = by + a \\ 0 & \text{otherwise} \end{cases}.$$

For $\alpha \in Z$, $\alpha \neq 0$, let $E(\alpha)$ and $O(\alpha)$ denote the "even part" and "odd part" of α respectively. In other words, if $\alpha = 2^\beta(2\mu + 1)$ for $\beta, \mu \in Z$ then $E(\alpha) = 2^\beta$ and $O(\alpha) = 2\mu + 1$. For $m = (m_1, \cdots, m_n) \in Z^n$, $E(m)$ will denote the n-tuple $(E(m_1), \cdots, E(m_n))$ with $O(m)$ defined similarly.

Finally, for $m \in Z^n$, let $\mathfrak{F}_m{}^*$ denote the real vector space generated by the set of functions

$$\left\{f\left(\frac{z-a}{O(m)}\right): f \in \mathfrak{F}_{E(m)}, a \in Z^n\right\}.$$

We note that if $m = e = (1, 1, \cdots, 1)$ then $\mathfrak{F}_m{}^* = \mathfrak{F}_m$; in general, we always have $\mathfrak{F}_m{}^* \subset \mathfrak{F}_m$.

We come now to the main result of the paper. This is the following:

Theorem.

$$\mathcal{G} \cap \mathfrak{F}_m = \mathfrak{F}_m^*$$

for $m \in Z^{n+}$.

The proof of this theorem will proceed in a series of Lemmas paralleling the steps taken in the preceding section.

Lemma 1. If $g(z) \in \mathcal{G}$ *then* $g[(z - a)/r] \in \mathcal{G}$ *for all* $a \in Z^n$ *and* $r \in Z^{n+}$.

Proof. We first show that if $f(z)$ is primitive then $f[(z - a)/r]$ is primitive. If we assume $f(z)$ is primitive then by definition there exist $b, x^{(1)}, \cdots, x^{(n)} \in Z^n$ such that

$$f(z) = \begin{cases} (-1)^{c_1+\cdots+c_n} & \text{if} \quad z = c_1x^{(1)} + \cdots + c_nx^{(n)} + b. \\ 0 & \text{otherwise} \end{cases}$$

On the other hand

$$f\left(\frac{z-a}{r}\right) = \begin{cases} f(y) & \text{if} \quad z = ry + a \\ 0 & \text{otherwise} \end{cases}.$$

Therefore

$$f\left(\frac{z-a}{r}\right) = \begin{cases} (-1)^{c_1+\cdots+c_n} & \text{if} \quad z = r(c_1x^{(1)} + \cdots + c_nx^{(n)} + b) + a \\ 0 & \text{otherwise} \end{cases}$$

$$= \begin{cases} (-1)^{c_1+\cdots+c_n} & \text{if} \quad z = c_1(rx^{(1)}) + \cdots + c_n(rx^{(n)}) + rb + a \\ 0 & \text{otherwise} \end{cases}$$

and hence, $f[(z - a)/r]$ is primitive. By applying this result to a linear combination of primitive functions, i.e., an element of $\mathcal{G}$, the lemma follows.

Let $i, \pi: Z^n \to R$ be defined by

$$i(z) = 1, \qquad \pi(z) = z_1z_2 \cdots z_n \quad \text{for all} \quad z = (z_1, z_2, \cdots, z_n) \in Z^n.$$

Lemma 2. Let $c, m \in Z^{n+}$, $r = cm$ *and suppose* $f: Z^n \to R$ *has period* m. *Then for any* $a \in Z^n$

$$\sum_{z\in P_r} i\left(\frac{z-a}{O(r)}\right) f(z) = \pi(E(c)) \sum_{z\in P_m} i\left(\frac{z-a}{O(m)}\right) f(z).$$

Proof. We first note that since

$$\begin{aligned} r = cm &= E(c)O(c)E(m)O(m) \\ &= E(c)E(m)O(c)O(m) = E(r)O(r) \end{aligned}$$

then

$$E(r) = E(c)E(m) \quad \text{and} \quad O(r) = O(c)O(m).$$

By definition

$$\sum_{z\in P_{O(c)m}} i\left(\frac{z-a}{O(O(c)m)}\right) f(z) = \sum_{z\in A} f(z)$$

and

$$\sum_{z\in P_m} i\left(\frac{z-a}{O(m)}\right) f(z) = \sum_{z\in B} f(z)$$

where

$$A = \{z: 0 \leqq z = kO(c)O(m) + a < O(c)m \quad \text{for some } k\}$$

and

$$B = \{z: 0 \leqq z = kO(m) + a < m \quad \text{for some } k\}.$$

Hence, for each set, the values which k may assume are just a translation of $P_{E(m)}$, there being $\pi(E(m))$ values in all. Since $O(m)$ and $O(c)O(m)$ are odd (i.e., each component is odd), then A and B both contain a *complete residue system modulo* $E(m)$. Consequently, since all the elements of A and B are congruent to a modulo $O(m)$ then *modulo* $E(m)O(m)$, A *and* B *are identical.* Since $m = E(m)O(m)$ and f has period m then the sums $\sum_{z\in A} f(z)$ and $\sum_{z\in B} f(z)$ are equal.

We also note in general that for any $s \in Z^{n+}$

$$\sum_{P_{E(c)s}} i\left(\frac{z-a}{O(s)}\right) f(z) = \pi(E(c)) \sum_{P_s} i\left(\frac{z-a}{O(s)}\right) f(z)$$

since $P_{E(c)s}$ is the disjoint union of $\pi(E(c))$ copies of P_s. Therefore we have

$$\sum_{P_r} i\left(\frac{z-a}{O(r)}\right) f(z) = \sum_{P_{cm}} i\left(\frac{z-a}{O(cm)}\right) f(z)$$

$$= \sum_{P_{E(c)O(c)m}} i\left(\frac{z-a}{O(c)O(m)}\right) f(z)$$

$$= \pi(E(c)) \sum_{P_{O(c)m}} i\left(\frac{z-a}{O(c)O(m)}\right) f(z)$$

$$= \pi(E(c)) \sum_{P_m} i\left(\frac{z-a}{O(m)}\right) f(z)$$

and the lemma is proved.

We should note that as a corollary to this lemma we obtain:

$$i\left(\frac{z-a}{cm}\right) f(z) \quad \text{has mean zero iff} \quad i\left(\frac{z-a}{m}\right) f(z) \quad \text{has mean zero.} \tag{5}$$

We are now in a position to prove the important

Lemma 3. Suppose $g \in \mathcal{G}$ *has period* $p \in Z^{n+}$. *Then for all* $a \in Z^n$,

$$\sum_{P_p} i\left(\frac{z-a}{O(p)}\right) g(z) = 0.$$

Proof. We first show that the above conclusion holds if we assume that $g = f$ is primitive with period $q = (\alpha, \alpha, \cdots, \alpha)$. In this case there exists $c, x^{(1)}, \cdots, x^{(n)} \in Z^n$ such that

$$f(z) = \begin{cases} (-1)^{a_1+\cdots+a_n} & \text{if } z = a_1x^{(1)} + \cdots + a_nx^{(n)} + c \\ 0 & \text{otherwise} \end{cases}.$$

To each $u = (u_1, \cdots, u_n) \in Z^n$ we can associate the *unique* point $v = (v_1, \cdots, v_n) \in Z^n$ such that $v_1 = u_1 \pm O(q)$, $v_i = u_i$ for $i > 1$, where the $\pm$ sign is chosen so that Z^n is decomposed into the union of disjoint pairs $\{u,v\}$. It follows at once that

$$u_1x^{(1)} + \cdots + u_nx^{(n)} \equiv v_1x^{(1)} + \cdots + v_nx^{(n)} \pmod{O(q)}.$$

Since

$$\sum_{i=1}^{n} v_i - \sum_{i=1}^{n} u_i = \pm O(q)$$

is odd then

$$f(u_1x^{(1)} + \cdots + u_nx^{(n)} + c) = -f(v_1x^{(1)} + \cdots + v_nx^{(n)} + c).$$

Also, note that for all $a \in Z^n$

$$i\left(\frac{u_1x^{(1)} + \cdots + u_nx^{(n)} - a}{O(q)}\right) = i\left(\frac{v_1x^{(1)} + \cdots + v_nx^{(n)} - a}{O(q)}\right).$$

Since f has period q then we must have

$$\sum_{P_q} i\left(\frac{z - a}{O(q)}\right) f(z) = 0$$

as asserted.

We may now remove the restriction that f has period of the form $q = (\alpha, \alpha, \cdots, \alpha) = \alpha e$. If we assume f has an arbitrary period $p \in Z^{n+}$ then it is certainly true that f also has period $\pi(p)e$. By above we have

$$\sum_{P_{\pi(p)e}} i\left(\frac{z - a}{O(\pi(p)e)}\right) f(z) = 0.$$

Since p divides $\pi(p)e$ then by (5) we see that

$$\sum_{P_p} i\left(\frac{z - a}{O(p)}\right) f(z) = 0. \tag{6}$$

Finally, to prove the lemma assume that

$$g = \sum_{j=1}^{t} \alpha_j f_j$$

where the α_j are real and the f_j are primitive. If f_j has period $p^{(j)}$ then by (6) we have

$$\sum_{P_{p^{(j)}}} i\left(\frac{z - a}{O(p^{(j)})}\right) f_j(z) = 0 \quad \text{for} \quad 1 \leqq j \leqq t.$$

If g has period p and q denotes $p^{(1)}p^{(2)} \cdots p^{(t)}p$ then by Lemma 2

$$\sum_{P_q} i\left(\frac{z - a}{O(q)}\right) f_j(z) = 0 \quad \text{for} \quad 1 \leqq j \leqq t.$$

Therefore

$$\sum_{P_q} i\left(\frac{z - a}{O(q)}\right) g(z) = \sum_{j=1}^{t} \alpha_j \sum_{P_q} i\left(\frac{z - a}{O(q)}\right) f_j(z) = 0$$

so that applying Lemma 2 again we obtain

$$\sum_{P_p} i\left(\frac{z - a}{O(p)}\right) g(z) = 0$$

since g has period p. This proves the lemma.

The final lemma is an n-dimensional generalization of (3). Its proof, however, is considerably more complicated.

Lemma 4. If $m = 2^{\alpha}e$ *for some* $\alpha \in Z^{+}$ *then* $\mathfrak{F}_m \subset \mathcal{G}$.

Proof. It will be sufficient to show that there exist $2^{n\alpha} - 1$ functions in $\mathcal{G}$ which also belong to $\mathfrak{F}_m$ and which are linearly independent over R. Let $C^{(k)}$ denote the $k \times k$ matrix of the form

$$C^{(k)} = \begin{pmatrix} 1 & 1 & 1 & 1 & \cdots & 1 \\ 1 & -1 & 1 & 1 & \cdots & 1 \\ 1 & 1 & -1 & 1 & \cdots & 1 \\ 1 & 1 & 1 & -1 & \cdots & 1 \\ . & . & . & . & \cdots & . \\ 1 & 1 & 1 & 1 & \cdots & -1 \end{pmatrix}$$

and let $D^{(k)}$ denote the $k \times k$ matrix of the form

$$D^{(k)} = \begin{pmatrix} 0 & 0 & 0 & \cdots & 0 & 0 & 1 \\ 0 & 0 & 0 & \cdots & 0 & 1 & 1 \\ 0 & 0 & 0 & \cdots & 1 & 1 & 1 \\ . & . & . & \cdots & . & . & . \\ . & . & . & \cdots & . & . & . \\ 1 & 1 & 1 & \cdots & 1 & 1 & 1 \end{pmatrix}$$

In other words,

$$c_{ij} = \begin{cases} 1 & \text{for } i = 1 \\ 1 - 2\delta_{ij} & \text{for } i > 1 \end{cases}$$

and

$$d_{ij} = \begin{cases} 0 & \text{if } i + j \leqq k \\ 1 & \text{otherwise} \end{cases}$$

where δ_{ij} is the Kronecker δ-function. Define $B_\kappa^{(\nu)}$ to be the $\nu \times \nu$ matrix of the form

$$B_\kappa^{(\nu)} = \begin{pmatrix} 0 & C^{(\nu-\kappa+1)} \\ D^{(\kappa-1)} & 0 \end{pmatrix} \quad \text{for} \quad 1 \leqq \kappa \leqq \nu$$

where 0 denotes the appropriate zero matrix. Let $r_{\kappa,\lambda}^{(\nu)}$ denote the point of Z^ν formed from the λth row of $B_\kappa^{(\nu)}$. Finally, let $f_\kappa^{(\nu)}$ denote the function in $\mathcal{G}$ defined by

$$f_\kappa^{(\nu)}(z) = \begin{cases} (-1)^{a_1+\cdots+a_n} & \text{if} \quad z = a_1 r_{\kappa,1}^{(\nu)} + \cdots + a_\nu r_{\kappa,\nu}^{(\nu)} \\ 0 & \text{otherwise} \end{cases}.$$

We show first that the functions $f_\kappa^{(\nu)}$, $1 \leqq \kappa \leqq \nu$, are linearly independent over R. To accomplish this it suffices to show that for any κ, $1 \leqq \kappa \leqq \nu$, there are two points p and q in Z^ν such that

$$f_\kappa^{(\nu)}(p) \neq f_\kappa^{(\nu)}(q)$$

while

$$f_\tau^{(\nu)}(p) = f_\tau^{(\nu)}(q) \quad \text{for} \quad \kappa < \tau \leqq \nu.$$

What we show in fact is that if $|f_\kappa^{(\nu)}(p)| = 1$ then $f_{\kappa+1}^{(\nu)}(p) = 1$ for $1 \leqq \kappa < \nu$. This may be proved by showing that if $s \in Z^\nu$ is any Z-linear combination of the $r_{\kappa,\lambda}^{(\nu)}$, $1 \leqq \lambda \leqq \nu$, then s can be written as a Z-linear combination of the $r_{\kappa+1,\lambda}^{(\nu)}$, $1 \leqq \lambda \leqq \nu$, such that the sum of the coefficients is divisible by 2. We proceed by induction on ν. For $\nu = 2$ we have

$$B_1^{(2)} = \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix}, \qquad B_2^{(2)} = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}.$$

Since

$$r_{1,1}^{(2)} = r_{2,1}^{(2)} + r_{2,2}^{(2)}$$

and

$$r_{1,2}^{(2)} = r_{2,2}^{(2)} - r_{2,1}^{(2)}$$

then any Z-linear combination of the $r_{1,\lambda}^{(2)}$ can be written as a Z-linear combination of the $r_{2,\lambda}^{(2)}$ with the sum of the coefficients divisible by 2 and the assertion is true for this case. Now assume the hypothesis for ν and let

$$s = \sum_{\lambda=1}^{\nu+1} a_\lambda r_{\kappa,\lambda}^{(\nu+1)}$$

be a Z-linear combination of the $r_{\kappa,\lambda}^{(\nu+1)}$. There are two cases.

(i) Suppse $\kappa \geqq 2$. Note that $B_\kappa^{(\nu+1)}$ and $B_{\kappa+1}^{(\nu+1)}$ have the following forms:

$$B_\kappa^{(\nu+1)} = \begin{pmatrix} 0 & & \\ 0 & & \\ \vdots & & B_{\kappa-1}^{(\nu)} \\ 0 & & \\ \underbrace{1\ 1\ \cdots\ 1}_{\kappa-1}\ 0\ \cdots\ 0 & & \end{pmatrix}$$

$$B_{\kappa+1}^{(\nu+1)} = \begin{pmatrix} 0 & & \\ 0 & & \\ \vdots & & B_\kappa^{(\nu)} \\ 0 & & \\ \underbrace{1\ 1\ \cdots\ 1}_{\kappa}\ 0\ \cdots\ 0 & & \end{pmatrix}$$

We also note the important fact that $r_{\kappa+1,\nu-\kappa}^{(\nu+1)}$ has all zero components except for the κth component which is 1. Thus we have

$$r_{\kappa,\nu+1}^{(\nu+1)} = r_{\kappa+1,\nu+1}^{(\nu+1)} - r_{\kappa+1,\nu-\kappa}^{(\nu+1)}.$$

Now s can be written as

$$s = \sum_{\lambda=1}^{\nu} a_\lambda r_{\kappa,\lambda}^{(\nu+1)} + a_{\nu+1} r_{\kappa,\nu+1}^{(\nu+1)}.$$

By the induction hypothesis (since the $r_{\kappa,\lambda}^{(\nu+1)}$ differ from the $r_{\kappa,\lambda}^{(\nu)}$, $1 \leqq \lambda \leqq \nu$, only in an extra zero component) we can write

$$\sum_{\lambda=1}^{\nu} a_\lambda r_{\kappa,\lambda}^{(\nu+1)} = \sum_{\lambda=1} b_\lambda r_{\kappa+1,\lambda}^{(\nu+1)}$$

where $\sum_{\lambda=1}^{\nu} b_\lambda$ is an even integer. Hence

$$s = \sum_{\lambda=1}^{\nu} b_\lambda r_{\kappa+1,\lambda}{}^{(\nu+1)} + a_{\nu+1}(r_{\kappa+1,\nu+1}{}^{(\nu+1)} - r_{\kappa+1,\nu-\kappa}{}^{(\nu+1)}).$$

Here, s has been written as a Z-linear combination of the $r_{\kappa+1,\lambda}{}^{(\nu+1)}$ with the sum of the coefficients *even* and consequently this case is finished.

(ii) Suppose $\kappa = 1$. We write $B_1{}^{(\nu+1)}$ and $B_2{}^{(\nu+1)}$ as

$$B_1{}^{(\nu+1)} = \begin{pmatrix} & & & & 1 \\ & B_1{}^{(\nu)} & & & 1 \\ & & & & \vdots \\ & & & & 1 \\ & & & & 1 \\ 1 & 1 & \cdots & 1 & 1 & -1 \end{pmatrix}$$

$$B_2{}^{(\nu+1)} = \begin{pmatrix} & & & & 1 \\ & B_2{}^{(\nu)} & & & 1 \\ & & & & \vdots \\ & & & & 1 \\ & & & & 0 \\ 0 & 1 & \cdots & 1 & 1 & -1 \end{pmatrix}$$

We note that

$$r_{1,\nu+1}{}^{(\nu+1)} = r_{2,\nu}{}^{(\nu+1)} + r_{2,\nu+1}{}^{(\nu+1)}.$$

As before, by the induction hypothesis we can write

$$\sum_{\lambda=1}^{\nu} a_\lambda r_{1,\lambda}{}^{(\nu+1)} = \sum_{\lambda=1}^{\nu} b_\lambda r_{2,\lambda}{}^{(\nu+1)}$$

where $\sum_{\lambda=1}^{\nu} b_\lambda$ is even. Consequently we have

$$s = \sum_{\lambda=1}^{\nu} a_\lambda r_{1,\lambda}{}^{(\nu+1)} + a_{\nu+1} r_{1,\nu+1}{}^{(\nu+1)}$$

$$= \sum_{\lambda=1}^{\nu} b_\lambda r_{2,\lambda}{}^{(\nu+1)} + a_{\nu+1}(r_{2,\nu}{}^{(\nu+1)} + r_{2,\nu+1}{}^{(\nu+1)})$$

which expresses s as a Z-linear combination of the $r_{2,\lambda}{}^{(\nu+1)}$ with an even coefficient sum. This completes the induction step and the proof of the assertion that the $f_\kappa{}^{(\nu)}$, $1 \leqq \kappa \leqq \nu$, are linearly independent over R.

A more careful examination of $B_\kappa{}^{(\nu)}$ reveals the following:

(a) $r_{\kappa,\lambda+1}{}^{(\nu)} - r_{\kappa,\lambda}{}^{(\nu)} = (\delta_{1,\nu-\lambda}, \delta_{2,\nu-\lambda}, \cdots, \delta_{\nu,\nu-\lambda})$
for $\nu - \kappa + 1 \leqq \lambda \leqq \nu - 1$.

(b) $r_{\kappa,1}{}^{(\nu)} - r_{\kappa,\lambda}{}^{(\nu)} = (2\delta_{1,\kappa+\lambda-1}, 2\delta_{2,\kappa+\lambda-1}, \cdots, 2\delta_{\nu,\kappa+\lambda-1})$
for $2 \leqq \lambda \leqq \nu - k + 1$.

(c) $\sum_{\lambda=2}^{\nu-\kappa} r_{\kappa,\lambda}{}^{(\nu)} - (\nu - \kappa - 2) r_{\kappa,1}{}^{(\nu)} = (2\delta_{1,\kappa}, 2\delta_{2,\kappa}, \cdots, 2\delta_{\nu,\kappa})$.

Since the linear combinations of the $r_{\kappa,\lambda}{}^{(\nu)}$ in (a), (b) and (c) all have the sum of coefficients an even integer then

$$f_\kappa{}^{(\nu)}((2\delta_{1,\lambda}, 2\delta_{2,\lambda}, \cdots, 2\delta_{\nu,\lambda})) = f_\kappa{}^{(\nu)}(0)$$

for $1 \leqq \lambda, \kappa \leqq \nu$. Hence $f_\kappa{}^{(\nu)}$ has period $2e = (2, 2, \cdots, 2)$. Also we note that the only points in $P_{2e} \subset Z^\nu$ at which $f_\kappa{}^{(\nu)}$ is nonzero are just those Z-linear combinations of the $r_{\kappa,\lambda}{}^{(\nu)}$ which have all coordinates 0 or 1. It is not difficult to see that the only points of this type which may be generated are the 2^κ points of the form $(c_1, c_2, \cdots, c_{\kappa-1}, c_0, c_0, \cdots, c_0)$ where $c_j = 0$ or 1. By a *translation of* $f_\kappa{}^{(\nu)}$ *by* a we mean the function $f_{\kappa,a}{}^{(\nu)}$ defined by

$$f_{\kappa,a}{}^{(\nu)}(z) = f_\kappa{}^{(\nu)}(z - a).$$

By letting the a range over the set of points

$$A_\kappa = \{(\underbrace{0,0,\cdots,0}_{\kappa}, d_1, d_2, \cdots, d_{\nu-\kappa}) : d_\lambda = 0 \text{ or } 1\},$$

the $2^{\nu-\kappa}$ translations $f_{\kappa,a}{}^{(a)}$, $a \in A_\kappa$, have the property that for each $p \in P_{2e}$, *exactly one* of the $f_{\kappa,a}{}^{(\nu)}$ assumes a nonzero value at p. In fact, if we define an inner product $(f_{\kappa,a}{}^{(\nu)}, f_{\lambda,b}{}^{(\nu)})$ for $f_{\kappa,a}{}^{(\nu)}$ and $f_{\lambda,b}{}^{(\nu)}$ by

$$(f_{\kappa,a}{}^{(\nu)}, f_{\lambda,b}{}^{(\nu)}) = \sum_{p \in P_{2e}} f_{\kappa,a}{}^{(\nu)}(p) f_{\lambda,b}{}^{(\nu)}(p)$$

then the inner product of any two distinct functions $f_{\kappa,a}{}^{(\nu)}, f_{\lambda,b}{}^{(\nu)}$, $a \in A_\kappa$, $b \in A_\lambda$, $1 \leqq \kappa, \lambda \leqq \nu$, is zero. Since

$$(f_{\kappa,a}{}^{(\nu)}, f_{\kappa,a}{}^{(\nu)}) = 2^\kappa$$

then by introducing the "normalized" basis functions

$$\hat{f}_{\kappa,a}^{(\nu)} = 2^{-\kappa/2} f_{\kappa,a}^{(\nu)}$$

we see that the set of $(1 + 2 + \cdots + 2^{\nu-1}) = 2^{\nu} - 1$ functions $\hat{f}_{\kappa,a}^{(\nu)}$, $a \in A_{\kappa}$, $1 \leqq \kappa \leqq \nu$, are orthonormal. Thus, we have shown that $F_{2^e} \subset \mathcal{G}$.

The extension of this technique to show that $F_{2^{\alpha_e}} \subset \mathcal{G}$ is quite similar and will be omitted. The basic idea is simply to introduce the "expansions" $f_{\kappa,a,\lambda}^{(\nu)}$ of $f_{\kappa,a}^{(\nu)}$ defined by

$$f_{\kappa,a,\lambda}^{(\nu)}(z) = f_{\kappa,a}^{(\nu)}(z/\lambda) \quad \text{for} \quad \lambda = 2^0, 2^1, \cdots, 2^{\alpha-1},$$

and then by taking suitable normalized translations of these functions, obtain an orthonormal basis (in $\mathcal{G}$) for $\mathfrak{F}_{2^{\alpha_e}}$. This shows that

$$\mathfrak{F}_m = \mathfrak{F}_{2^{\alpha_e}} \subset \mathcal{G}$$

and the proof of the lemma is completed.

We are now ready to proceed to the proof of the

Theorem.

$$\mathcal{G} \cap \mathfrak{F}_m = \mathfrak{F}_m^*$$

for $m \in Z^{n+}$.

Proof: $\underline{\mathcal{G} \cap \mathfrak{F}_m \subset \mathfrak{F}_m^*}$

Let $f \in \mathcal{G} \cap \mathfrak{F}_m$. Since f has period m then by Lemma 3, we have for all $a \in Z^n$

$$\sum_{z \in P_m} i\left(\frac{z - a}{O(m)}\right) f(z) = 0.$$

Since

$$f(z) = \sum_{a \in P_{O(m)}} i\left(\frac{z - a}{O(m)}\right) f(z)$$

and each of the functions $i[(z - a)/O(m)]f(z)$ can be written as $h[(z - a)/O(m)]$ for some $h \in \mathfrak{F}_{E(m)}$ then $f \in \mathfrak{F}_m^*$ and this direction is established.

$\underline{\mathcal{G} \cap \mathfrak{F}_m \supset \mathfrak{F}_m^*}$

We have already noted that $\mathfrak{F}_m^* \subset \mathfrak{F}_m$. It remains to show that $\mathfrak{F}_m^* \subset \mathcal{G}$. By definition $\mathfrak{F}_m^*$ is the real vector space generated by the set

$$\{h[(z - a)/O(m)]\colon\ h \in \mathfrak{F}_{E(m)},\ a \in Z^n\}.$$

By Lemma 4 we have

$$\mathfrak{F}_{E(m)} \subset \mathfrak{F}_{\pi(E(m))_e} \subset \mathfrak{G}.$$

Thus, if $h \in \mathfrak{F}_{E(m)}$ then $h \in \mathfrak{G}$. But by Lemma 1, $h \in \mathfrak{G}$ implies

$$h[(z - a)/O(m)] \in \mathfrak{G}.$$

Therefore, since $\mathfrak{G}$ contains a set of generators for $\mathfrak{F}_m^*$ then $\mathfrak{F}_m^* \subset \mathfrak{G}$. This completes the proof of the theorem.

IV. CONCLUDING REMARKS

As a concluding example of the results of the preceding section, we consider the decomposition of the function f generated by the charge distribution of the crystal structure of potassium tantalate, $KTaO_3$. This compound forms face-centered cubic crystals with a charge distribution as shown in Fig. 1. That is, a $+1$ is situated at each vertex, a -2 at each face-center and a $+5$ is located in the center of the cube. The periodic function f defined by this distribution has period (2,2,2) and is shown in Fig. 2 (which is the forward upper left octant of Fig. 1). We have

$$B_1^{(3)} = \begin{pmatrix} 1 & 1 & 1 \\ 1 & -1 & 1 \\ 1 & 1 & -1 \end{pmatrix}, \quad B_2^{(3)} = \begin{pmatrix} 0 & 1 & 1 \\ 0 & 1 & -1 \\ 1 & 0 & 0 \end{pmatrix}, \quad B_3^{(3)} = \begin{pmatrix} 0 & 0 & 1 \\ 0 & 1 & 0 \\ 1 & 1 & 0 \end{pmatrix}$$

so that the 7 basis functions into which f will be decomposed are as shown in Fig. 3.

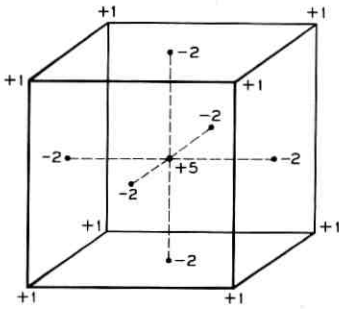


Fig. 1 — Charge distribution of $KTaO_3$.

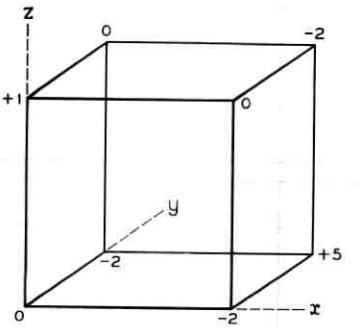


Fig. 2 — A period of the periodic function.

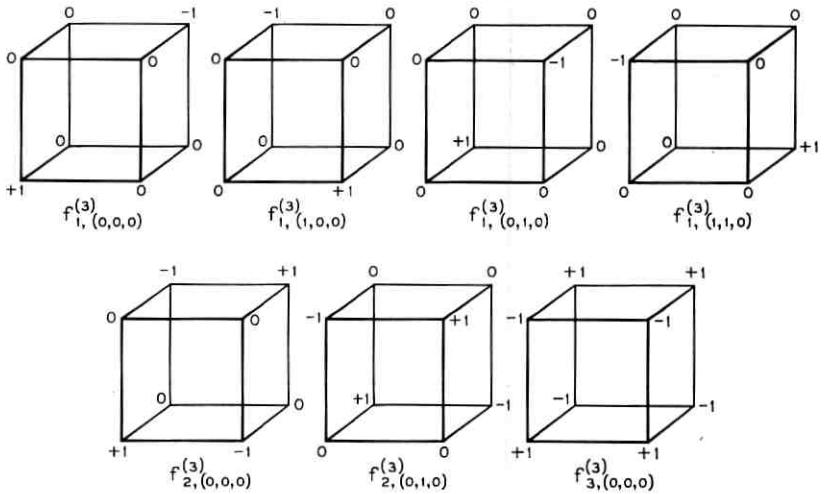


FIG. 3 — The seven basis functions.

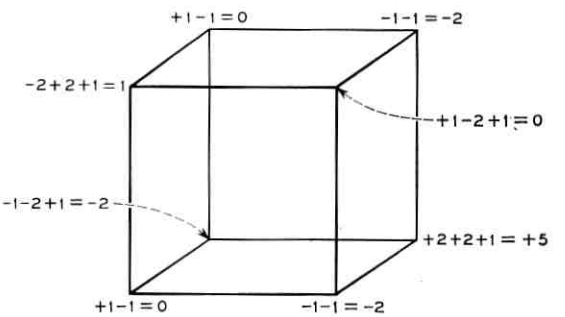


Fig. 4 — Decomposition of the periodic function.

The coefficient of $f_{1,(0,0,0)}^{(3)}$ is obviously $\frac{1}{2}(0 - (-2)) = 1$, etc., so that we obtain

$$f = f_{1,(0,0,0)}^{(3)} - f_{1,(1,0,0)}^{(3)} - f_{1,(0,1,0)}^{(3)} + 2f_{1,(1,1,0)}^{(3)} - 2f_{2,(0,1,0)}^{(3)} - f_{3,(0,0,0)}^{(3)}$$

Graphically, this equality is shown in Fig. 4.

The author gratefully acknowledges many enlightening discussions on this subject with H. O. Pollak (whose ideas along the lines of generating functions led to a short solution of the one-dimensional problem) and W. J. C. Grant (who originated the problem).

REFERENCES

1. Ewald, P. P., Ann. Physik., *64*, 1921, p. 253.
2. W. J. C. Grant, B.S.T.J., *44*, 1965, p. 427.
3. Nijboer, B. R. A. and Dewette, F. W., Physica *23*, 1957, p. 309.
4. Nijboer, B. R. A. and Dewette, F. W., Physica *24*, 1958, p. 422.

Contributors to This Issue

Richard R. Anderson, B.S.M.E., 1949, Northwestern University; M.S.E.E., 1960, Stevens Institute of Technology; Bell Telephone Laboratories, 1949—. Mr. Anderson first engaged in research on electronic switching systems for telephone central offices. In 1956 he joined the data transmission exploratory development department and made several prototype magnetic-tape transports for storing digital data. He has conducted theoretical studies of data transmission systems by computer simulation. Member, AAAS, Sigma Xi and Tau Beta Pi.

A. Descloux, Math. Dipl., 1948, École Polytechnique Fédérale (Swiss Federal Institute of Technology); Ph.D., Mathematical Statistics, University of North Carolina, 1961. After spending 1955–56 on the staff of the University of Washington where he taught mathematics and statistics, Mr. Descloux joined Bell Telephone Laboratories. He has been concerned chiefly with the application of probability theory to traffic problems. Member, Institute of Mathematical Statistics, American Mathematical Society, and Society for Industrial and Applied Mathematics.

Ronald L. Graham, B.S., 1958, University of Alaska; M.A., Ph.D., 1962, University of California (Berkeley); Bell Telephone Laboratories, 1962—. Mr. Graham has been engaged in research in a variety of combinatorial problems arising in coding theory, crystallography, multiprocessing, and fluctuation theory. Member, American Mathematical Society, Mathematical Association of America, Sigma Xi.

T. G. Grau, B.A., 1960, Ohio Wesleyan University; M.S., 1962, Ohio State University; Bell Telephone Laboratories, 1960—. He is associated with the wire spring relay group of the Switching Apparatus Department at the Columbus Laboratory. Mr. Grau is currently engaged in the study of the magnetic characteristics of wire spring relays.

William H. C. Higgins, B.S.E.E., 1929, E.E., 1934, Purdue Universsity; Bell Telephone Laboratories, 1934—. From 1929 to 1934 he was a member of the Development and Research Department of A. T. & T. where he was engaged in systems engineering and field trials of carrier telephone and transcontinental program transmission systems. With the

transfer of that organization to BTL he became associated with development of radio telephone and telegraph systems for point-to-point, ship-to-shore, and ground-to-air applications, and with the development of radio altimeters. He was associated with the development of Army and Navy gun fire control radars, the NIKE guided missile system, Distant Early Warning Line, aircraft bombing and navigation systems, underwater sound detection systems, command guidance system for the TITAN ICBM, and data processing for NIKE ZEUS. Since 1961 he has been Executive Director, Electronic Switching Division and has responsibility for development work on electronic central offices, electronic PBX and military electronic switching systems. Fellow IEEE, Tau Beta Pi, Eta Kappa Nu, Sigma Xi (associate), American Ordnance Association, Armed Forces Communications Electronics Association, Association of the U. S. Army.

D. C. Hogg, B.Sc., 1949, University of Western Ontario; M.Sc., 1950, and Ph.D., McGill University; Bell Telephone Laboratories, 1953—. His work has included studies of artificial dielectrics for microwaves, diffraction of microwaves, and over-the-horizon and millimeter wave propagation. He has been concerned with evaluation of sky noise, analysis of performance characteristics of microwave antennas, and propagation of optical waves. Fellow, IEEE; member, Commission 2, U.R.S.I., Sigma Xi, and AAAS.

J. Salz, B.S.E.E., 1955, M.S.E., 1956, Ph.D., 1961, University of Florida; The Martin Company, 1958–1960; Bell Telephone Laboratories, 1961—. He first worked on the remote line concentrators for the electronic switching system. He has since engaged in theoretical studies of data transmission systems. Member, IEEE; associate member, Sigma Xi.

Morton I. Schwartz, B.E.E., 1956, College of the City of New York; M.E.E., 1959, Eng.Sc.D., 1964, New York University; International Telephone and Telegraph Laboratories, 1956–1961; Bell Telephone Laboratories 1961—. Mr. Schwartz has been concerned with the study of digital FM and the analysis and design of radar systems. He is presently engaged in research in communication theory. Member, Eta Kappa Nu, IEEE and Sigma Xi.

A. K. Spiegler, B.S., 1957, M.S., 1958, University of Illinois; Bell

Telephone Laboratories, 1958—. Mr. Spiegler is a member of the Switching Apparatus Department at the Columbus Branch Laboratory. He is presently involved in a study of the magnetic circuit of the wire spring relays and a new family of relays — the miniature wire spring relays. Member, American Physical Society and the AAAS.

H. G. SUYDERHOUD, Diploma of Ingenieur (EE), Technische Hogeschool Delft, Netherlands, 1955; Bell Telephone Laboratories, 1956 —. Mr. Suyderhoud's work has been primarily in the transmission systems engineering of toll carrier systems. This includes the statistical evaluation and characterization of broadband communication channels. He was also involved in formulating equalization plans permitting wideband data communications. Presently he is engaged in studies of new methods of echo suppression in the exchange plant. Member, IEEE, American Statistical Association, Royal Institute of Dutch Engineers, and Association of Delft Engineers.

ROBERT W. WILSON, B.A., 1957, Rice University; Ph.D., 1962, California Institute of Technology; Bell Telephone Laboratories 1963 —. Since coming to Bell Laboratories, Mr. Wilson has engaged in research in radio astronomy. Member American Astronomical Society, Phi Beta Kappa, Sigma Xi.

AARON D. WYNER, B.S., 1960, Queens College; B.S.E.E., 1960, M.S., 1961, Ph.D., 1963, Columbia University; Bell Telephone Laboratories, 1963—. He has been engaged in research in various aspects of information theory. He is also Adjunct Assistant Professor of Electrical Engineering at Columbia University. Member, IEEE, Tau Beta Pi, Eta Kappa Nu, Sigma Xi.

B.S.T.J. BRIEFS

Axis-Crossing Intervals of Rayleigh Processes

By A. J. RAINAL

(Manuscript received May 12, 1965)

I. INTRODUCTION

Let $R(t,a)$ denote the envelope of a stationary random process consisting of a sinusoidal signal of amplitude $\sqrt{2a}$ and frequency f_0 plus Gaussian noise of unit variance having a narrow-band power spectral density which is symmetrical about f_0. When $a = 0$ Rice[1] presented some theoretical results which are very useful for studying statistical properties of the axis-crossing intervals of $R(t,0)$. The axis-crossing points and the axis-crossing intervals of the Rayleigh process $R(t,a)$ are defined in Fig. 1. Some recent work concerning the axis-crossing intervals of $R(t,a)$ was reported by Levin and Fomin[2], Goryainov,[3] and Rainal.[4,5] The purpose of this brief is to present some theoretical results when $a \geqq 0$. These results stem from a straightforward extension of Rice's analysis. The Rayleigh process $R(t,a)$ occurs at the output of a typical radio or radar receiver during the reception of a sinusoidal signal immersed in Gaussian noise.

II. THEORETICAL RESULTS

Using a notation consistent with Refs. 4 and 5 we define the following probability functions at an arbitrary level R of Fig. 1:

(1) $Q^-(\tau,R,a)d\tau$, the conditional probability that an upward axis-crossing occurs between $t + \tau$ and $t + \tau + d\tau$ given a downward axis-crossing at t.

(2) $Q^+(\tau,R,a)d\tau$, the conditional probability that a downward axis-crossing occurs between $t + \tau$ and $t + \tau + d\tau$ given an upward axis-crossing at t.

(3) $[U(\tau,R,a) - Q(\tau,R,a)]d\tau$, the conditional probability that an upward axis-crossing occurs between $t + \tau$ and $t + \tau + d\tau$ given an upward axis-crossing at t.

The reader should refer to Rice[1] for the definition of all notation which is not defined in this note. When $a \geqq 0$, Rice's (86) becomes:

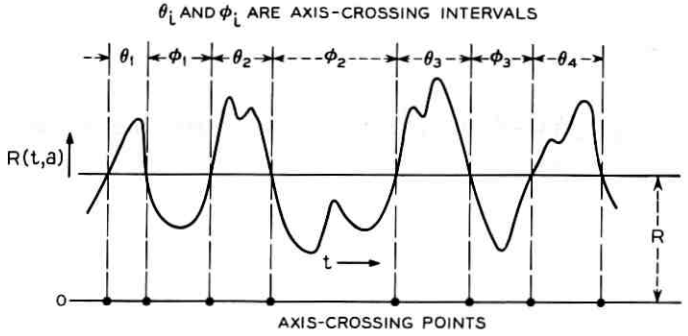


Fig. 1 — The level R defines the axis-crossing points and the axis-crossing intervals of the Rayleigh process $R(t,a)$.

$$Q^-(\tau,R,a) = -\left(\frac{2\pi}{\beta}\right)^{\frac{1}{2}} R^{-1} e^{R^2/2}\, I_0^{-1}\,(RQ)\, e^a \cdot \int_0^{2\pi} d\theta_1 \int_0^{2\pi} d\theta_2 \int_{-\infty}^{0} dR_1' \int_0^{\infty} dR_2' R_1' R_2' p(R,R_1',R_2',R,\theta_1,\theta_2) \tag{1}$$

where: $I_0(\kappa)$ = Bessel function of imaginary argument

$$p(R,R_1',R_2',R,\theta_1,\theta_2) = \frac{R^2[M_{22}^2 - M_{23}^2 c^2]^{-\frac{1}{2}}}{8\pi^3} \cdot \exp\left\{-\frac{1}{2M}\left[A(R_1'^2 + R_2'^2) + 2ArR_1'R_2' + 2DR_1' + 2ER_2' + F\right]\right\}$$

$$A = [M_{22}^2 - M_{23}^2 c^2]^{-1}[MM_{22}(1 - m^2)] \qquad Q = \sqrt{2a}$$

$$c = \cos(\theta_1 - \theta_2) \qquad s = \sin(\theta_1 - \theta_2) \qquad r = \frac{cM_{23}}{M_{22}}$$

$$D = [M_{22}^2 - M_{23}^2 c^2]^{-1}\{R[M_{22} - M_{23}c][Mm'(c - m)] + Q[M_{12} - M_{13}][M_{23}s(M_{22}\sin\theta_2 + M_{23}c\sin\theta_1) - \cos\theta_1(M_{22}^2 - M_{23}^2 c^2)]\}$$

$$E = [M_{22}^2 - M_{23}^2 c^2]^{-1}\{-R[M_{22} - M_{23}c][Mm'(c - m)] + Q[M_{12} - M_{13}][M_{23}s(M_{22}\sin\theta_1 + M_{23}c\sin\theta_2) + \cos\theta_2(M_{22}^2 - M_{23}^2 c^2)]\}$$

$$F = [M_{22}^2 - M_{23}^2 c^2]^{-1}\{2[Q^2 - QR(\cos\theta_1 + \cos\theta_2)] \cdot [M_{11} + M_{14}][M_{22}^2 - M_{23}^2 c^2]$$

$$- 2M_{13}QRs[M_{12} - M_{13}][M_{22} - M_{23}c][\sin\theta_1 - \sin\theta_2]$$
$$- M_{22}Q^2[M_{12} - M_{13}]^2[\sin^2\theta_1 + \sin^2\theta_2]$$
$$- 2M_{23}Q^2c[M_{12} - M_{13}]^2 \sin\theta_1 \sin\theta_2$$
$$+ 2R^2[M_{22} - M_{23}c][(M_{22} + M_{23}c)(M_{11} + M_{14}c) - M_{13}^2s^2]\}.$$

Equation (1) can be put in a form analogous to Rice's (97) and (55):

$$Q^-(\tau,R,a) = \frac{e^aRM_{22}e^{R^2/2}I_0^{-1}(RQ)}{2\pi\sqrt{2\pi\beta}\,(1-m^2)^2}\int_0^{2\pi}\int_0^{2\pi} e^{-(G/2M)}J(r,h,k)\,d\theta_1 d\theta_2 \quad (2)$$

where:

$$J(r,h,k) \equiv \frac{1}{2\pi s_1}\int_h^\infty dx\int_k^\infty dy\,(x-h)\,(y-k)\,e^z$$

$$z = -\frac{x^2+y^2-2rxy}{2(1-r^2)}; \qquad h = -a_1\left[\frac{1-m^2}{M_{22}}\right]^{\frac{1}{2}};$$

$$k = a_2\left[\frac{1-m^2}{M_{22}}\right]^{\frac{1}{2}}$$

$$a_1 = A^{-1}[1-r^2]^{-1}[D-rE] \qquad a_2 = A^{-1}[1-r^2]^{-1}[E-rD]$$

$$G = A^{-1}[1-r^2]^{-1}[2rDE - D^2 - E^2] + F; \qquad s_1 = \sqrt{1-r^2}.$$

We also find that:

$$J(r,h,k) = \frac{s_1}{2\pi}\exp\left[-\frac{(k^2-2rhk+h^2)}{2(1-r^2)}\right] - \frac{he^{-k^2/2}}{2\sqrt{2\pi}}$$
$$\cdot\left[1 - P\left(\frac{h-rk}{s_1}\right)\right] - \frac{ke^{-h^2/2}}{2\sqrt{2\pi}}\left[1 - P\left(\frac{k-rh}{s_1}\right)\right] \quad (3)$$
$$+ (hk+r)K(r,h,k)$$

where:

$$P(x) = \frac{2}{\sqrt{2\pi}}\int_0^x e^{-t^2/2}\,dt$$

$$K(r,h,k) \equiv \text{Karl}^6\text{ Pearson's}\left(\frac{d}{N}\right) = \frac{1}{2\pi s_1}\int_h^\infty dx\int_k^\infty dy e^z.$$

For a recent table of $K(r,h,k)$ see Ref. 7. For a recent discussion of $K(r,h,k)$ see the recent work of Gupta.[8] In these latter two references $K(r,h,k)$ is denoted by $L(h,k,r)$.

$Q^+(\tau,R,a)$ is obtained from (1) by changing the signs of the ∞'s in

the limits of integration. We find that $Q^+(\tau,R,a)$ is equal to the right-hand side of (2) with h,k replaced by $-h,-k$.

$[U(\tau,R,a) - Q(\tau,R,a)]$ is obtained from (1) by changing the lower limit of integration of R_1' to $+\infty$. We find that:

$$U(\tau,R,a) - Q(\tau,R,a) = \frac{e^a R M_{22} e^{R^2/2}\, I_0^{-1}(RQ)}{2\pi\sqrt{2\pi\beta}\,(1-m^2)^2} \cdot \int_0^{2\pi}\int_0^{2\pi} e^{-(G/2M)} J_1(r,h,k)\, d\theta_1 d\theta_2 \tag{4}$$

where:

$$J_1(r,h,k) \equiv \frac{1}{2\pi s_1}\int_h^{-\infty} dx \int_k^{\infty} dy\,(x-h)\,(y-k)\,e^z.$$

We find that $J(r,h,k)$ and $J_1(r,h,k)$ are related by:

$$J_1(r,h,k) = J(r,h,k) + \frac{h}{\sqrt{2\pi}} e^{-k^2/2} - \frac{(hk+r)}{2}[1 - P(k)]. \tag{5}$$

Equations (4) and (5) are the generalizations of (64) and (35) of Ref. 5.

III. STATISTICAL DEPENDENCE OF AXIS-CROSSING INTERVALS

By expanding $m(\tau)$ as:

$$m(\tau) = 1 - \frac{\beta}{2}\tau^2 + \frac{b_3|\tau^3|}{3!} + \frac{b_4\tau^4}{4!} + \frac{b_5|\tau^5|}{5!} + \frac{b_6\tau^6}{6!} + \frac{b_7|\tau^7|}{7!} + o(\tau^7) \tag{6}$$

we find that as $\tau \to 0$ from the right:

$$M_{11} = 2\beta b_3\tau - (b_3^2 - \beta b_4 + \beta^3)\tau^2 + o(\tau^2) \tag{7}$$

$$M_{12} = \beta b_3\tau^2 - \tfrac{1}{2}(b_3^2 - \beta b_4 + \beta^3)\tau^3 + o(\tau^3) \tag{8}$$

$$M_{13} = \beta b_3\tau^2 - \tfrac{1}{2}(b_3^2 - \beta b_4 + \beta^3)\tau^3 + o(\tau^3) \tag{9}$$

$$M_{14} = -2\beta b_3\tau + (b_3^2 - \beta b_4 + \beta^3)\tau^2 + o(\tau^2) \tag{10}$$

$$M_{22} = \tfrac{2}{3}\beta b_3\tau^3 + \tfrac{1}{4}(\beta b_4 - b_3^2 - \beta^3)\tau^4 + o(\tau^4) \tag{11}$$

$$M_{23} = \tfrac{1}{3}\beta b_3\tau^3 + \tfrac{1}{12}(3\beta b_4 - b_3^2 - 3\beta^3)\tau^4 + o(\tau^4). \tag{12}$$

When $b_3 \neq 0$ we find that:

$$M_{12} - M_{13} = -\tfrac{1}{6}\beta^2 b_3\tau^4 + o(\tau^4) \tag{13}$$

$$M_{11} + M_{14} = \tfrac{1}{6}\beta b_3^2\tau^4 + o(\tau^4) \tag{14}$$

$$M_{22} - M_{23} = \tfrac{1}{3}\beta b_3\tau^3 + o(\tau^3) \tag{15}$$

$$M = \tfrac{1}{3}\beta b_3^2\tau^4 + o(\tau^4). \tag{16}$$

When $b_3 = 0$ and $b_5 \neq 0$ we find that:

$$M_{12} - M_{13} = \tfrac{1}{60}\beta^2 b_5\tau^6 + o(\tau^6) \tag{17}$$

$$M_{11} + M_{14} = -\tfrac{1}{120}\beta b_4 b_5\tau^7 + o(\tau^7) \tag{18}$$

$$M_{22} - M_{23} = -\tfrac{1}{30}\beta b_5\tau^5 + o(\tau^5) \tag{19}$$

$$M = \frac{\beta b_5}{60}(\beta^2 - b_4)\tau^7 + o(\tau^7). \tag{20}$$

When $b_3 = b_5 = 0$ we find that:

$$M_{12} - M_{13} = \frac{5\beta}{720}(\beta b_6 + b_4^2)\tau^7 + o(\tau^7) \tag{21}$$

$$M_{11} + M_{14} = \frac{-5b_4}{1440}(\beta b_6 + b_4^2)\tau^8 + o(\tau^8) \tag{22}$$

$$M_{22} - M_{23} = -\tfrac{1}{72}(\beta b_6 + b_4^2)\tau^6 + o(\tau^6) \tag{23}$$

$$M = \tfrac{1}{144}(\beta^2 - b_4)(\beta b_6 + b_4^2)\tau^8 + o(\tau^8). \tag{24}$$

As $\tau \to 0$ we see that the terms of the quantities D, E, and F which involve the sine wave amplitude Q are of higher order in τ than the terms which do not involve Q. This behavior as $\tau \to 0$ is consistent with a result reported by Levin and Fomin[2]. Thus, a theorem presented in Ref. 5 also applies to the Rayleigh process $R(t,a)$. That is: If $R(t,a)$ is a Rayleigh process, defined in paragraph one, having a finite expected number of axis-crossing points per unit time at any level R, then two successive axis-crossing intervals at that level R are statistically dependent.

The theorem implies that the successive axis-crossing points of the Rayleigh process $R(t,a)$ at any level R do not form a Markov point process.

IV. ACKNOWLEDGMENT

It gives me great pleasure to acknowledge stimulating discussions with S. O. Rice.

REFERENCES

1. Rice, S. O., Distribution of the Duration of Fades in Radio Transmission, B.S.T.J., *37*, May 1958, pp. 581–635.

2. Levin, B. R. and Fomin, Ya. A., Approximate Determination of the Distribution Function of Fade Lengths Below Threshold Level for the Envelope of the Sum of a Deterministic Signal and Gaussian Stationary Noise, Telecommunications and Radio Engineering, Part II — Radio Eng., *18*, No. 5, May 1963.
3. Goryainov, V. T., Distribution of Axis-Crossing Intervals for the Smoothed Envelope of Quasi-Sinusoidal Noise, Telecommunications and Radio Engineering, Part II — Radio Eng. *18*, No. 8, August 1963.
4. Rainal, A. J., Zero-Crossing Intervals of Rayleigh Processes, Tech. Report No. AF-108, DDC No. AD-600-393, The Johns Hopkins University, Carlyle Barton Laboratory, Baltimore, Maryland, May 1964. Abstracted in IEEE Trans. Info. Theory, *IT-11*, No. 1, p. 159, Jan. 1965.
5. Rainal, A. J., Zero-Crossing Intervals of Envelopes of Gaussian Processes, Tech. Report, No. AF-110, DDC No. AD-601-231, The Johns Hopkins University, June 1964. Abstracted in IEEE Trans. Info. Theory, *IT-11*, No. 1, p. 159, Jan. 1965.
6. Pearson, Karl, ed., *Tables for Statisticians and Biometricians*, Cambridge University Press, 1931, Part II, Table VIII, Vols. of Normal Bivariate Surface, pp. 78–109.
7. National Bureau of Standards (1959), Tables of the Bivariate Normal Distribution Function and Related Functions, Applied Math. Series 50. U. S. Government Printing Office, Washington 25, D. C.
8. Gupta, S. S., Probability Integrals of Multivariate Normal and Multivariate t, The Ann. of Math. Statistics, *34*, No. 3, Sept. 1963, p. 792.

Errata

A Note on a Signal Recovery Problem, by I. W. Sandberg, B.S.T.J., *43*, November 1964, pp. 3065–3067.

On page 3066, replace $|f_2(t)|^{\frac{1}{2}}$ by $|f_2(t)|^2$, and replace $\tilde{\psi}[w] = \tilde{\psi}[w] - w$ by $\tilde{\psi}[w] = \psi[w] - w$. On page 3067, replace $\max(c_1, c_2)$ by $(c_1^2 + c_2^2)^{\frac{1}{2}}$.

T. R.
1307-4-2