4212

THE BELL SYSTEM
TECHNICAL JOURNAL

VOLUME XLIII JANUARY 1964 NUMBER 1, PART 1

June 7, 1965

The Ferreed

By A. FEINER

(Manuscript received August 20, 1963)

The advantages of the ferreed as a switching network crosspoint led to an early decision to adopt it for use in electronic switching systems. The prospect of large-scale use of the device gave impetus to a search for an economical, easily fabricated component. This paper describes the considerations which influenced the choices of a suitable magnetic material, magnetic circuit geometry, and coil design that were made for the production model.

I. INTRODUCTION

The concept of the ferreed was presented in an earlier article in this journal.[1] The purpose of this paper is to describe the evolution of this device during its further development.

To recollect, a ferreed is a device born of marriage between miniature sealed reed contacts (see Ref. 2) and an external magnetic circuit containing remanently magnetizable members. Operation or release of the sealed contacts can be controlled by setting the remanent members in one of two magnetic states by means of short current pulses.

Among the several useful properties that can be brought about in the ferreeds by selection of the proper magnetic configurations and coil design is the ability to respond to coordinate excitation — a vital requirement for any device considered for a network crosspoint.

Recognition of the potential advantages of a switching network crosspoint with metallic contacts, absence of holding power and the ability to operate in times much shorter than prior electromechanical devices led to an early decision to adopt it for the network of No. 1 ESS (Electronic Switching System) — the new telephone switching system scheduled for its commercial debut in 1965.

The intended application of the ferreed in the switching network of No. 1 ESS, where it would appear in very large numbers (14–20 crosspoints per line), gave impetus to a search for an economical, easily fabricated embodiment. Several important choices had to be made with regard to the geometry of the magnetic circuit, the winding configuration and the remanent magnetic material. At the same time, the requirements of the sealed reed contact were reexamined, and a modified version of it known as the 237B contact was adopted for ferreed use.

II. THE CROSSPOINT FERREED

2.1 *Choice of Remanent Material*

All original work on the ferreeds was based on the use of a specially developed cobalt ferrite as the remanent material. In time, certain inherent difficulties became apparent: notably, a strong temperature dependence of the magnetic properties and low flux density, leading to structures of large cross section and poor efficiency. Furthermore, as more thought was given to the ferreed as a system component, it was found that the originally postulated microsecond speeds for the actuation of the ferreed were neither required nor practical from the standpoint of driving requirements.

These considerations opened the way to a search for a metallic substitute. Several chromium and tungsten steel compositions were investigated and found wanting due to lack of squareness and fullness of the hysteresis loop — properties whose importance were stressed in Ref. 1.

The attention soon centered on a recent addition to the list of cobalt-iron-vanadium alloys — Remendur. The name of this alloy refers to its primary magnetic characteristic, i.e., a remanence greater than 17,000 gauss. This is coupled with a square hysteresis loop and a coercive force from 1 to 60 oersteds. With a nominal composition of 48 per cent cobalt, 48 per cent iron, 3.5 per cent vanadium and 0.5 per cent manganese, Remendur bridges the gap between the high coercive force of Vicalloy and the low coercive force and high permeability properties of 2V-Permendur and Supermendur. Fig. 1 shows a hysteresis loop obtained on a Remendur strip developed for ferreed use. Of importance to the ferreed application is the squareness B_r/B_s and fullness $\sqrt{H_oB_o/H_cB_r}$

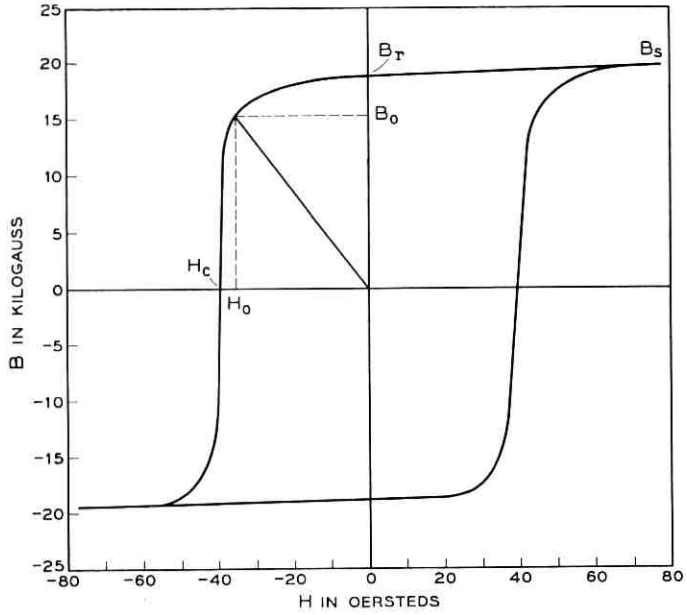


Fig. 1 — Hysteresis loop of Remendur used in ferreeds.

of the hysteresis loop. This property implies that the energy expenditure in establishing a desired end state approaches a minimum, and that the excess flux generated in the same process is small—important in view of the interference problems present in ferreed arrays.

2.2 *Choice of Geometry*

There exist two basic forms of ferreed structures — the parallel and the series ferreeds. These are illustrated in Fig. 2. The choice of Remendur, the need for tight magnetic coupling between the remanent members and the reed contacts, and the relative ease of fabrication led to adoption of the series structure for the crosspoint ferreed.

That structure is shown in Fig. 3 in the form used in the ESS network. Mounted on each side of the reed contacts, which are molded together in plastic to form a single piece part, and extending approximately over the length of the glass envelopes, are two flat plates of Remendur. Notches on the plastic and on the plates permit accurate relative positioning of the two.

The reeds and the remanent plates are inserted into plastic coil forms

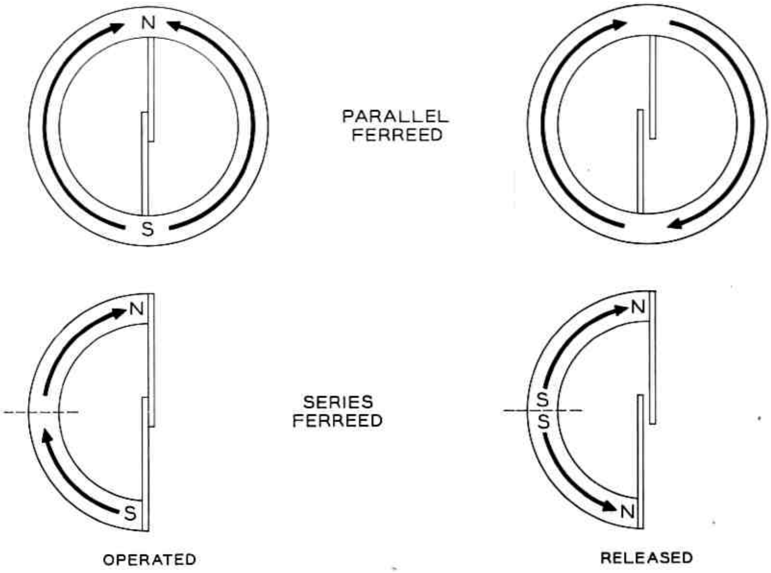


Fig. 2 — Principles of parallel and series ferreeds.

molded into a steel plate. This steel plate acts as a common shunt for the whole array — it divides each crosspoint magnetically into two separately controllable halves, greatly reducing the energy requirement for producing the release state in which, as shown in Fig. 4, the two halves of the remanent members are magnetized in opposing directions. The same steel plate acts as the mechanical backbone of the whole array.

2.3 *Coil Design*

The differential excitation mode was selected to provide coordinate addressing of crosspoints. Fig. 5 reviews this principle as applied to a series ferreed. Each crosspoint has two sets of windings — one for each coordinate. Each set contains a winding of N turns on one side of the shunt plate and one with a larger number, typically $2N$, on the other side. The $2N$-turn winding is connected series opposing the N-turn winding. One pair of windings is in series with the corresponding pairs of all crosspoints in the same row, while the other is in series with the pairs of all crosspoints in the same column of the array. As the paired windings oppose each other, energization produces the release state in every crosspoint energized, except the one where both pairs of windings are energized simultaneously — the crosspoint at the intersection of the energized row and the column.

The logic inherent to differential excitation was found to be well suited to network array operation, in which, in general, only one crosspoint in each row or column need be operated.

No separate release actions are required, as operating a crosspoint automatically releases other crosspoints associated with the same row and column.

The design of the coils has to take in account the energization requirements of a single crosspoint as well as the system requirement

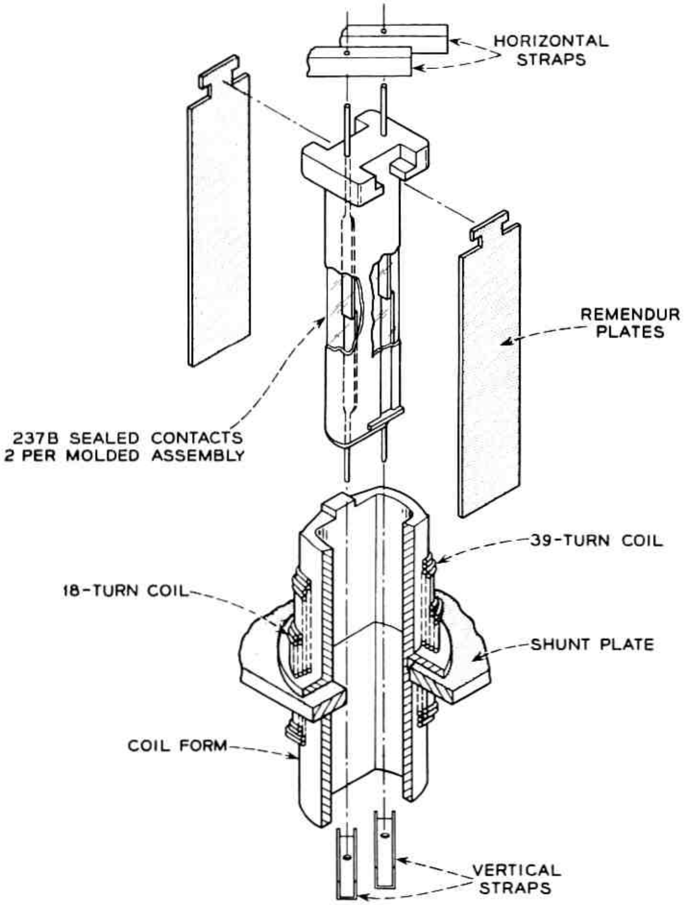


Fig. 3 — Exploded view of the two-wire crosspoint ferreed.

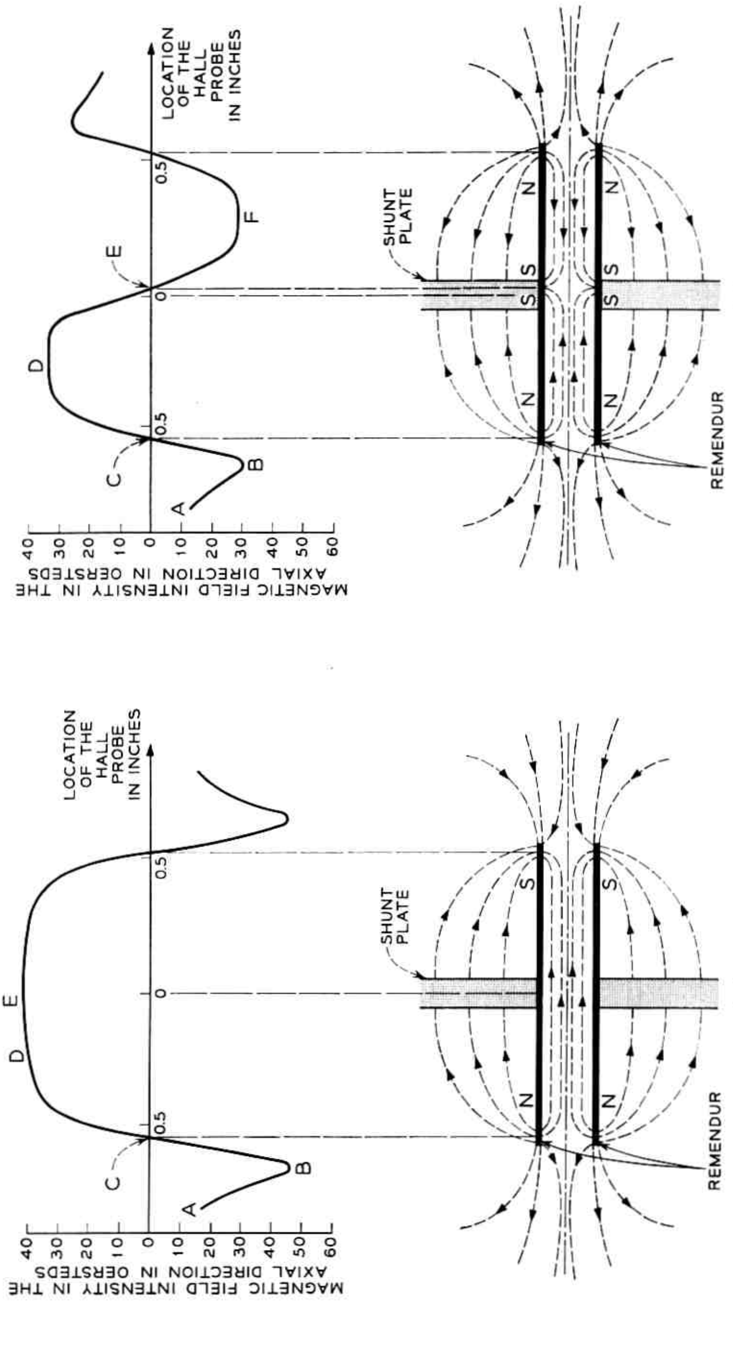

MAGNETIC FIELD INTENSITY IN THE AXIAL DIRECTION IN OERSTEDS
40
30
20
10
0
10
20
30
40
50
60
LOCATION OF THE HALL PROBE IN INCHES
0.5
0
0.5
A
B
C
D
E
SHUNT PLATE
N
S
N
S
REMENDUR
(a)
MAGNETIC FIELD INTENSITY IN THE AXIAL DIRECTION IN OERSTEDS
40
30
20
10
0
10
20
30
40
50
60
LOCATION OF THE HALL PROBE IN INCHES
0.5
0
0.5
A
B
C
D
E
F
SHUNT PLATE
N
S
S
N
N
S
S
N
REMENDUR
(b)

Fig. 4 — Field distribution of the crosspoint ferreed in the operated and released states.

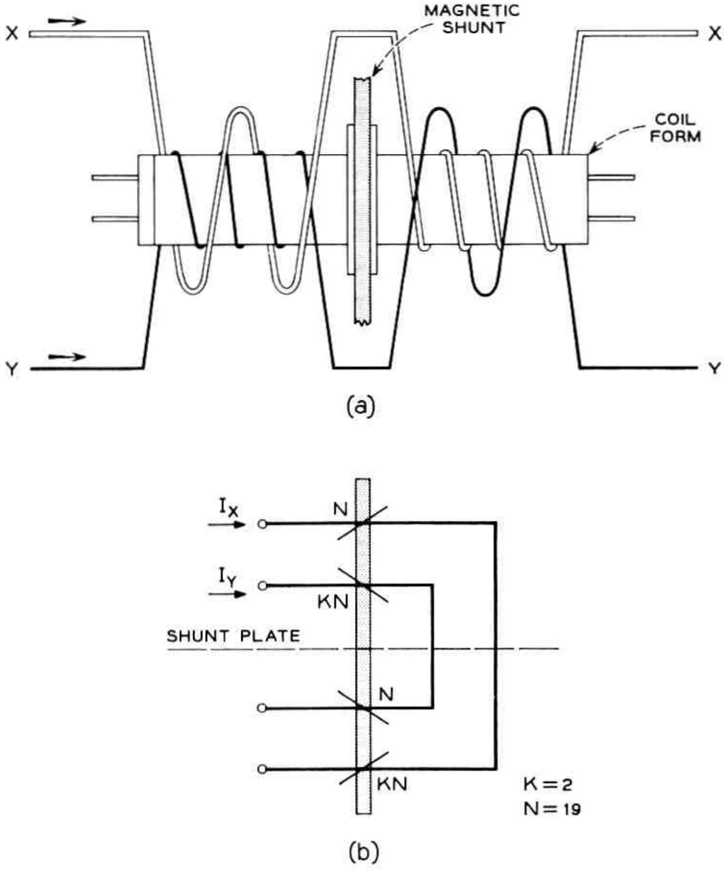


Fig. 5 — Winding configuration for differential excitation of the series ferreed: (a) winding pattern, (b) mirror symbol notation.

calling for simultaneous pulsing of 32 winding pairs in the process of establishing a connection through two stages of ferreed switches.

In ESS, these considerations led to the adoption of coils with windings of 18 and 39 turns wound with 25-gauge copper wire. With these coils, the nominal operating current pulse of 10 amperes peak amplitude and 250 microseconds duration insures adequate margins for both operation and release of the crosspoint.

The coils are wound directly on the coil forms by a machine that winds eight rows (or columns) of crosspoints simultaneously in a continuous succession, each with a single length of wire. This eliminates soldered connections between coils, thus reducing the winding cost and improving the reliability of the assembly.

The winding sense is reversed in adjacent crosspoints. This magnetic "checkerboarding" was found to be an effective means for reducing magnetic interaction phenomena as well as the noise pickup in the transmission pairs due to ferreed energizing pulses.

2.4 *Crosspoint Arrays*

Switching network considerations led to selection of an 8 × 8 crosspoint array as a basic network building block. In Fig. 6, such an array is shown. In addition, specifically for the concentrating stages of the network, several other array types were required: a switch providing each of 16 input terminal pairs with an access to 4 out of 8 available outputs, and 8 × 4 and 4 × 4 switches. It was found that each of these arrays could be derived from the basic 8 × 8 apparatus unit by suitably changing the connections of the control windings and the voice-pair strappings. Fig. 7 shows these connections for all the developed ferreed

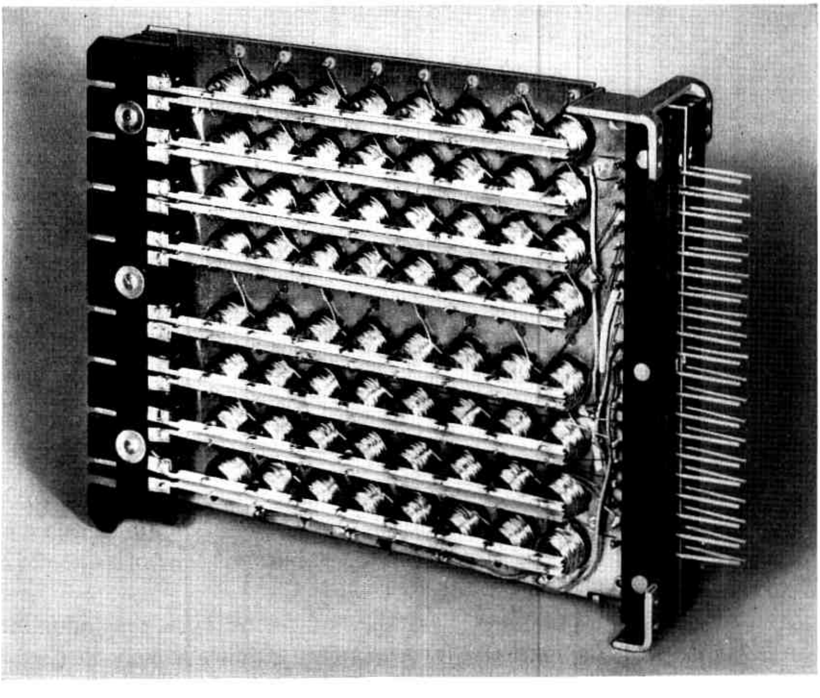


Fig. 6 — An 8 × 8 ferreed switch with covers removed.

switch types. As can be expected, this standardization of the physical size and component parts of the switches has eased the manufacturing and the network equipment design problems.

The connections shown between the ends of the row and column control winding chains stem from the access scheme adopted in the network design. In this scheme, identical current is applied to both coordinates by connecting them effectively in series when energizing a crosspoint at their intersection.

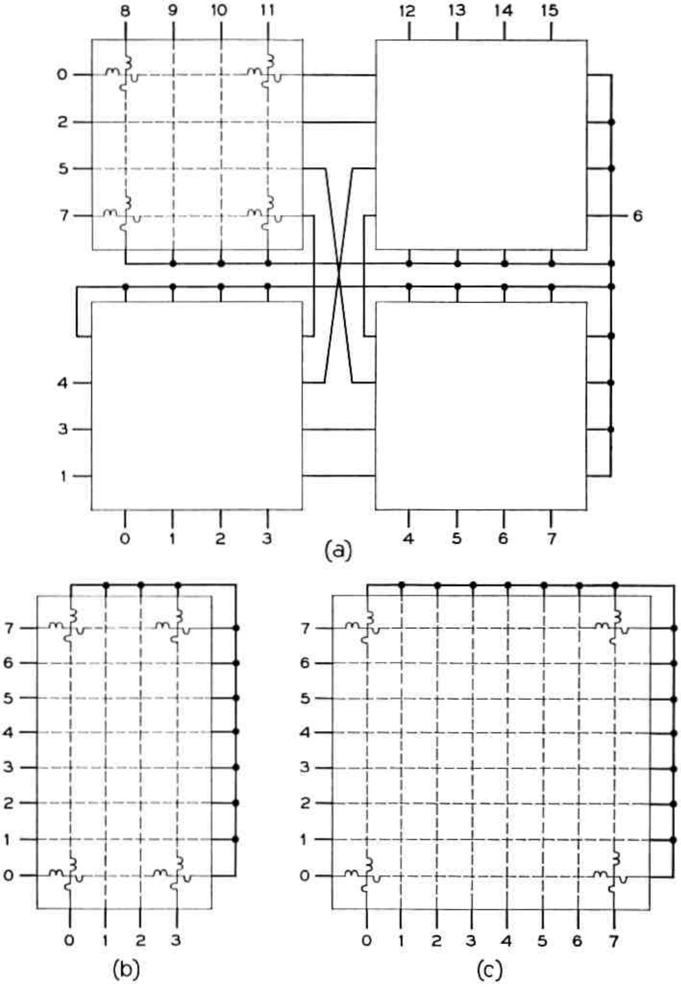


Fig. 7 — Control winding interconnection for three types of two-wire switches: (a) 16 × 4/8, (b) 8 × 4, and (c) 8 × 8.

III. DESIGN TECHNIQUE

When the problem of designing the ferreed was first approached, it was found that the usual lumped-constant, linear magnetic circuit approach, while sufficient to yield a workable device, did not provide the means for its optimization; neither did it give an assurance of margins in face of tolerance allowances that have to be made for the whole structure, and variations in reed contact properties and in the magnetic properties of Remendur. Several attempts were made to refine the analytical tools toward this end. While providing qualitative insight into the operation of the device, they were frustrated from attaining the ultimate goal of a quantitative, explicit solution by the complexity of the problem caused by the rather difficult geometry and the essential nonlinearity of the magnetic materials.

As a result, the refinements in the ferreed design had to be based largely on experimental techniques. Over the years, numerous experimental ferreed study techniques have been devised. These include the use of search coils with integrators, hysteresis measurements of reeds and the remanent magnetic members, Hall probes in the crosspoint structure and the reed gap, and reversible permeability measurements of the reeds. Supplemented by experiments in which the component parts of the structure, their positioning and the driving conditions underwent systematic variations, these techniques were instrumental in arriving at the present structure.

The use of Hall probes provided two study techniques. First, Hall probes were employed to measure longitudinal magnetic field intensity along the ferreed axis, after applying varying operate and release pulses. Second, via the use of specially constructed sealed reeds with Hall probes mounted in the gap of the reed, it was possible to measure the resultant magnetic flux density in the reed gap under varying operating conditions. The drawback of the techniques lies in the upsetting of the ferreed magnetic circuit by the absence of the reed or introduction of a permanently open reed structure.

Reversible permeability measurements of the sealed reeds, accomplished via inductance measurements of small sense coils at about 100 kc, provided a convenient means of determining the instantaneous applied mmf to the sealed reeds under varying operating and interference conditions. The technique was especially useful because it permitted the use of ordinary sealed reeds under actual operating conditions, and it was free of drift problems since no integrator circuits were involved. On the other hand, the nature of the reversible permeability characteristic of the sealed reed is so insensitive in the released state of the sealed reeds as to make its use not suitable in that region.

IV. OTHER FERREED TYPES

4.1 *The Bipolar Ferreed*

In the process of designing a ferreed switching network, the need arose for a device containing a pair of contacts that would be individually controllable. A typical use for this device is disconnection of the line current sensing element at the line circuit whenever a connection is established in the switching network (cutoff relay function). A postulated property of this device — to respond to control current pulse polarity to open or close its contacts — was found to permit integrating the control access with the one for the crosspoints.

An adaptation of the parallel ferreed principle, shown in Fig. 8, provided a suitable embodiment meeting this need. Of the two parallel remanent members, one consists of a permanent magnet material, Cunife I; the other, surrounded by a single coil, of Remendur. Contact closure or release depends on the polarity of the current pulse applied to the coil. Eight such devices packaged together form a single apparatus unit compatible in its length with the crosspoint units.

4.2 *The Four-Wire Crosspoint Array*

For use in switching networks requiring two separate directions of transmission, the two-wire crosspoint design has been extended to permit the operation of four contacts at every crosspoint location. The four contacts are arranged in a square pattern and are surrounded by an open-ended box formed by four remanent plates. The windings are similar to those of the two-wire array and again an eight-by-eight size has been chosen; Fig. 9 shows an individual crosspoint and an overall view of the unit.

V. SUMMARY

Out of the original concept of the ferreed originated a whole class of useful switching devices. Characterized by small size, high speed of operation and absence of holding power, they permit retaining the desirable aspects of metallic contacts in the environment of electronic switching machines without creating undue time compatibility problems.

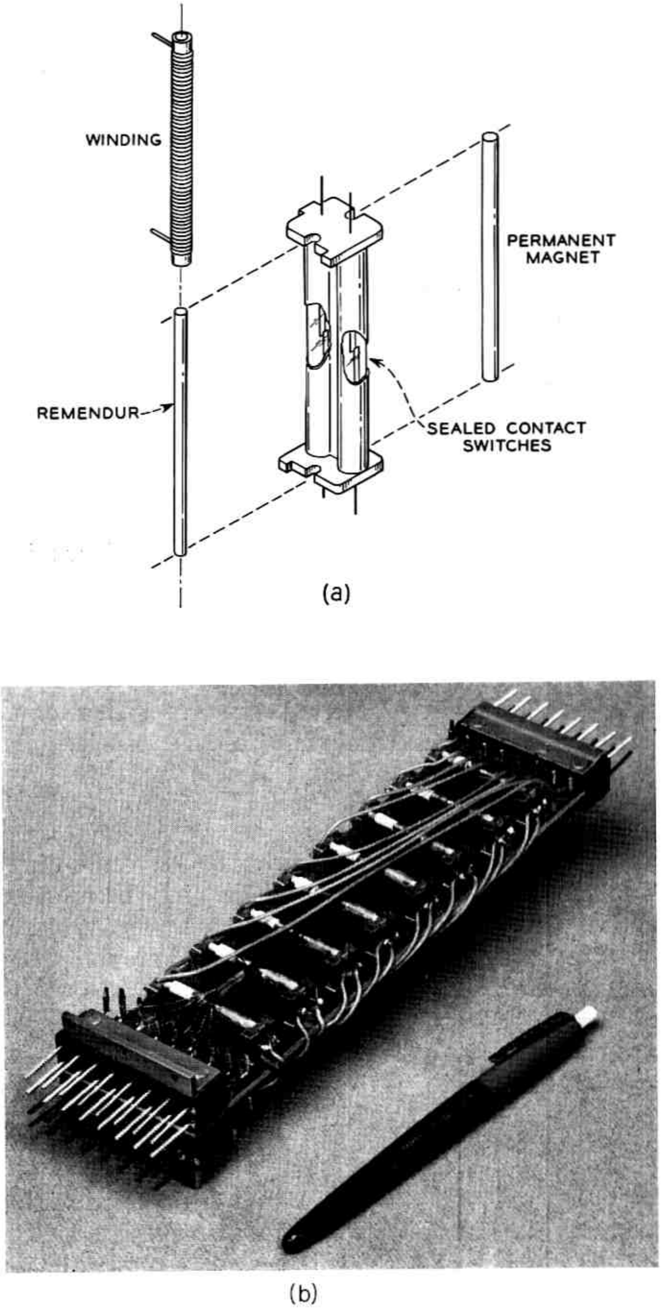


Fig. 8 — (a) The bipolar ferreed; (b) a 1 × 8 apparatus unit.

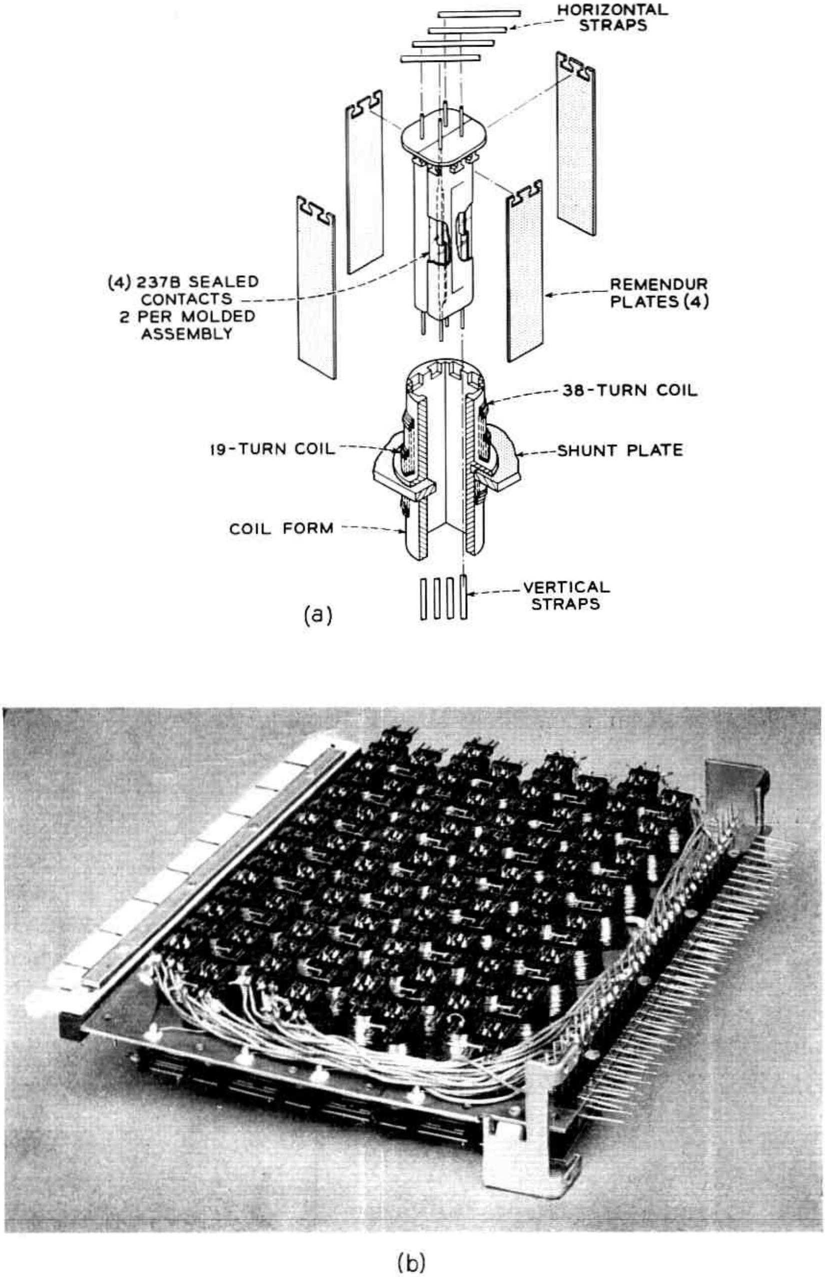


Fig. 9 — (a) Exploded view of a single four-wire crosspoint; (b) over-all view of an 8 × 8 switch with protective covers removed.

Table I — Summary of Ferreed Characteristics

Switch		Dimensions (Inches)			Operate and Release Pulse		Contact Characteristics			
Code	Type	Height	Width	Length	Peak Ampl. (A)	Width (μs)	Max. Res. (ohms)	Max. Operate Time (ms.)	Max. Surge Current	Life
242 A	2-wire 8 × 8	6¾	2⅛	9¼	9	200 to 500	0.2	3	3A*	2×10^6†
242 B	2-wire (2) 8 × 4									
242 C	2-wire 16 × 4/8									
252 A	4-wire 8 × 8	9¾	2⅛	9¼	9	200 to 300				
241 B	2-wire 1 × 8	1⅝	2⅛	9¼	6	200 to 500	5‡	3	3A	2×10^6

* To protect the contacts, crosspoints are operated and released in a dry circuit — maximum surge current refers to current value applied to closed contacts.

† Minimum life of 2×10^6 operations with contact resistance below 0.2 ohm.

‡ This contact breaks a maximum of 40 ma in its operation.

Table I gives a summary of the characteristics of the ferreed codes now in existence.

VI. ACKNOWLEDGMENTS

Many people have contributed important ideas and skills to make the ferreed a success; the author would like to offer his particular appreciation to Messrs. H. L. B. Gould and D. H. Wenny for their work on the Remendur, Messrs. R. L. Peek, F. H. Myers, and H. Raag for their work in magnetic design of the ferreed, and Messrs. H. J. Wirth and R. A. Billhardt for the mechanical design.

The credit for solving the manufacturing problems should go to Mr. G. A. Mitchell of the Western Electric Company at Columbus.

REFERENCES

1. Feiner, A., Lovell, C. A., Lowry, T. N., and Ridinger, P. G., The Ferreed — A New Switching Device, B.S.T.J., **39**, January, 1960, p. 1.
2. Keller, A. C., Recent Developments in Bell System Relays — Particularly Sealed-Contact and Miniature Relays, B.S.T.J., this issue, p. 15.

Recent Developments in Bell System Relays — Particularly Sealed Contact and Miniature Relays

By A. C. KELLER

(Manuscript received February 25, 1963)

Relays are among the most important electromechanical devices. They have been in use for many years and continue, in modern form, to be essential elements in modern Bell System and military applications, including electronic switching systems.

The most important recent developments are miniaturization, sealed contact relays using glass-enclosed contacts, and "remanent" type devices.

Ferreed and bipolar ferreed coordinate arrays and individual units are new and important switching elements. These devices make use of miniature glass-enclosed contacts in combination with "square loop" magnetic material such as ferrite or certain iron alloys. They are magnetic "latching" units and are operated or released by short pulses.*

I. INTRODUCTION

An important article entitled "Relays in the Bell System" was published[1] in the B.S.T.J. in 1924. This was a comprehensive article on relays which were then in use in the Bell System, and it gave some information on typical applications. Since that time, a few articles have appeared in the B.S.T.J. covering relays, particularly the article[2] in 1952 describing the general purpose wire spring relay. This is the most widely used relay in Bell System equipment at the present time. In addition there have been several comprehensive publications on the design of relays[3,4] and several new forms of the wire spring relay, namely the "two-in-one" relay[5] and a magnetic latching form of this device. Miniature wire spring relays have been and are being studied.

* In this paper, this is a remanent material of suitable coercive force range, generally intermediate between the common permanent magnets and the materials used for memory, such as cores, thin films, etc.

It is the purpose of this paper, in part, to bring together in one article some of the newer relays of importance to the Bell System, including a few which are experimental at this time. In this survey, the most important recent developments are miniaturization, sealed contact relays using glass-enclosed contacts, and magnetic latching devices. Frequency sensitive relays[6] are included, as are ferreed[7] and bipolar coordinate arrays. Such arrays consist of individual units of miniature glass-enclosed contacts (typically 2 or 4 at each crosspoint) in combination with a suitable "square loop" magnetic material such as certain ferrites or certain iron alloys which have controllable magnetic remanence. These devices are magnetic latching devices and can be operated or released by pulses as short as 5 microseconds. Arrays of this type are important units in Bell System electronic switching systems such as No. 1 ESS.[8]

Relays are made in larger quantities by the Bell System than ever before, and also more relays are made by more manufacturers outside the Bell System than ever before. The increasing use of relays is of interest in view of the rapid development of solid-state switching devices and systems and their higher switching speeds. In general, solid-state devices operate in microseconds or better compared with milliseconds or longer for electromechanical devices. The reasons[9] for the continued use and expansion of the uses of relay type switching devices are: (*i*) relays, with their large ratio of open to closed contact impedance, often result in equipment designs which are simple and inexpensive yet fast enough to make unimportant any increase in switching speed; (*ii*) relays can be used singly and in small numbers without the associated common control equipment often required to take full advantage of the sensational speeds of solid-state switching devices; (*iii*) the rapid expansion of switching of all kinds requires more of many types of switching equipment, including both solid-state and electromechanical types; and (*iv*) relays and solid-state devices are developing a compatibility, and in fact combinations of both have been developed, notably the ferreed. Compatibility has accelerated the miniaturization of new relay designs because they are often used together. Relay size reductions of $\frac{1}{10}$ or more in volume have been achieved.

Reliability is also becoming increasingly important, and lower failure rates are often required under more severe operating conditions. In military applications, this relates particularly to vibration, shock, temperature and humidity. Miniature relays often perform better under vibration and shock conditions than larger types because of the lower inertia of the moving parts and the higher natural frequencies of their smaller parts.

II. MINIATURE SEALED CONTACTS AND RELAYS USING THESE

There are two general classes of sealed contacts of the glass enclosed type. These are the dry reed[10] type and the mercury-wetted[11] type.

Relays using the larger form of dry reed sealed contacts have been described in previous papers.[10] Two new miniature dry reed sealed contacts are shown in Fig. 1, and for comparison the larger 224A type,[10] which has been in Bell System applications for a number of years, particularly in the digit register package in the No. 5 crossbar system. All of these sealed contacts, shown in Fig. 1, consist of two magnetic reeds sealed in a glass tube. Dry reed sealed contacts are free from external influences such as dust, corrosive atmospheres, and ambient pressures, and are relatively free of temperature effects. They do require a high degree of care and control during manufacture if maximum performance and uniformity are needed. In general, the mating contact surfaces are plated with gold, silver, rhodium, etc., or combinations of these, sometimes diffused under a controlled atmosphere. These operations are necessary in order to achieve a low and stable contact resistance and to avoid sticking, which may be the case with certain soft precious metals. The 237A (or G29) was the first of the miniature dry reed sealed contacts to be applied in systems applications. As described in Ref. 10, it is essentially a scaled-down (1 to 3) version of the larger 224A sealed contact.

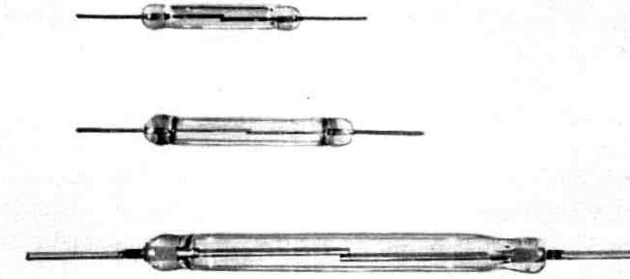


Fig. 1 — Dry reed sealed contacts: top, miniature type 237A (G-29); center, miniature type 237B; lower, standard type 224A.

The 237B miniature dry reed sealed contact was developed specifically for the crosspoint contacts of the switching network in electronic switching systems, although it is now also applied in certain relays in such systems and is suitable for general applications. The new requirements for the crosspoint application are: (*i*) higher breakdown voltage — of the order of 880 volts, (*ii*) closer operate and release values, and (*iii*) contact resistance of less than 0.2 ohm during 1,000,000 operations. These new and more severe requirements made it necessary (*i*) to pressurize the sealed contacts, (*ii*) to control tolerances more closely, and (*iii*) to improve the contact life by combinational plating of gold and silver. In addition, the reeds of the 237B design have been simplified by eliminating the "hinge" sections at a slight sacrifice in size. The increase is from the 237A length of 0.875 inch to 1.00 inch.

Operation of such contacts is by the application of a magnetic field, and several different methods are shown in Fig. 2. Fig. 2(a) shows the operation by passing the current through a winding surrounding the sealed contact. Fig. 2(b) shows one elementary form of ferreed where the operation results from pulse operation and magnetizing a "square loop" ferrite element. In this case the sealed contact remains closed without holding power because it is "magnetically latched." Release is by a pulse smaller in magnitude and of opposite polarity. Figs. 2(c) and 2(d) show other ferreed structures.

Typical values for the operating characteristics of these sealed contacts in air core coils are as shown in Table I. These operate ampere-turn values are minimum values in a simple air core test winding and, in general, faster speeds are obtained by increasing the applied ampere-turns. The minimum operate times as listed result, in general, by applying several times* the minimum operate ampere-turns.

Although sealed contacts can be operated by pulses of sufficient duration in the circuit shown in Fig. 2(a), the contact will remain closed only during an interval approximately the time that the current flows through the winding. Pulse operation of most interest is associated with "magnetic latching." This can be done by using a magnetic bias either by a suitable remanent member — as shown in Fig. 2(b) — or by a biasing winding. The operating time of such devices can be of the order of that obtained with normal neutral operation of sealed contacts. However, the ferreed type of operation can result in "effective" operating times very much faster and in the microsecond region.

There is another form of magnetic latching of sealed contacts which uses remanent reeds for the elements of the sealed contact. In this case,

* Operate time is a function of applied power (EI).

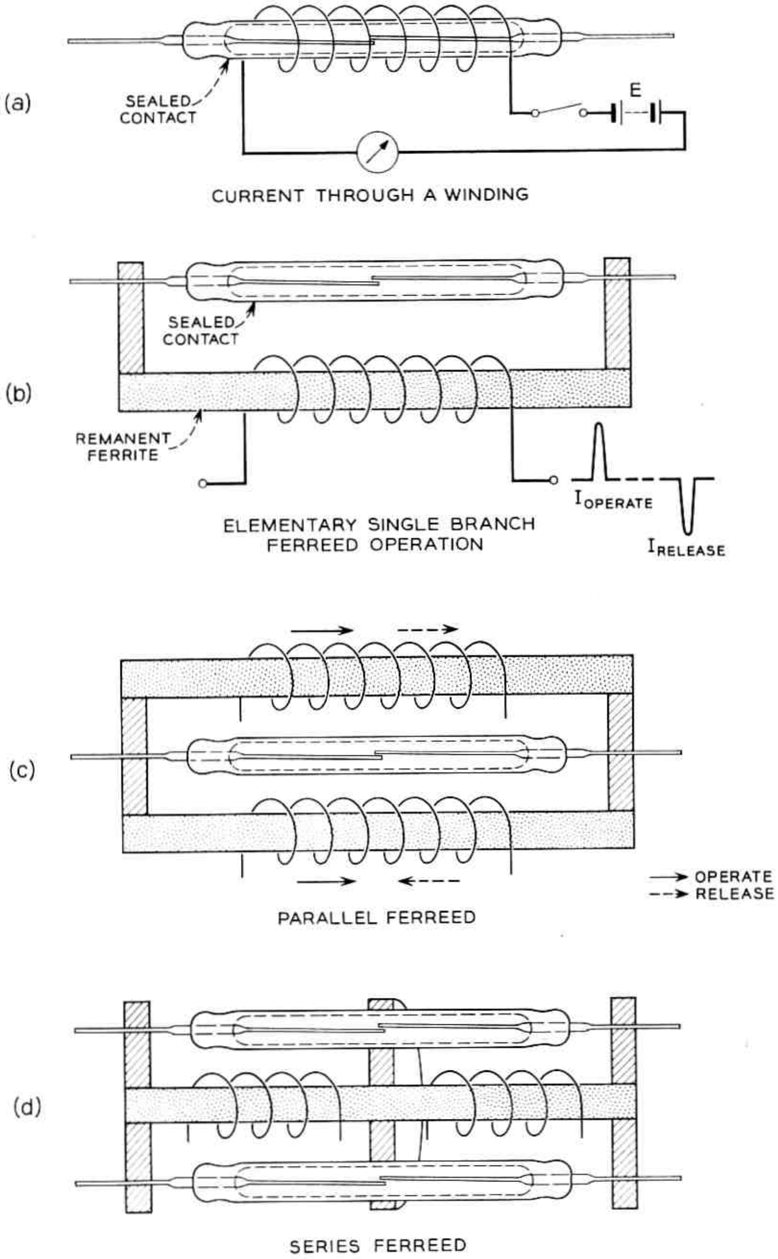


Fig. 2 — Operation of dry reed sealed contacts: (a) current through a winding, (b) elementary single-branch ferreed operation, (c) parallel ferreed, (d) series ferreed.

TABLE I — TYPICAL OPERATING CHARACTERISTICS

Sealed Contact	Operate (Ampere-Turns)	Release (Ampere-Turns)	Approximate Minimum Operate Time
			(milliseconds)
224 A	90 ± 12	34 ± 8	about 1.0
237A (G29)	34 ± 12	18 ± 8	" 0.5
237B	30.5 ± 5.5	15 ± 4	" 0.5

discussed in Refs. 7 and 12, the contacts are also locked by residual magnetism. As is the case with series or parallel ferreeds using nonremanent reed contacts, remanent reed sealed contacts may be operated by pulses shorter than the time of contact closure, but they may also be operated with longer pulses of lower power because the operation is dependent essentially on the input pulse energy. The advantage of remanent reeds is chiefly in the lower energy levels when they are used as crosspoints in a switching network, although these energy levels are somewhat higher than required to operate soft reeds in permanent magnet latching relays of this type. In comparing remanent reed sealed contacts and remanent sleeve crosspoints, the minimum energy in microwatt seconds, EIt* for operate and release, is important. Estimates are shown in Table II.

The energy relations also show how it is possible, in a given ferreed or remanent reed device, to trade time for the magnitude of the pulse current. For example, a 5-microsecond operate time would require a pulse of about 10 times the current value of that required to operate the same device (with a different winding) in 50 microseconds, etc.

Conventional type relays using the miniature 237A and 237B sealed contacts are shown in Fig. 3. Fig. 3(a) shows the 237A (G-29) sealed contact in a 2-make relay (GA 53702) as used in certain missile systems. Fig. 3(b) shows the 311A relay, which is a 3-make switching system relay using the 237B sealed contact. These relays are operated, under nominal conditions, at about 0.2 watt and 0.32 watt, respectively. Other designs with break contacts or transfer contacts have been made of similar size. Such relays make use of permanent magnets to bias the break contacts closed in the unenergized condition. By energizing the coil, these contacts are caused to open. Break and transfer contacts of this type have been made using the larger 224A sealed contact and have been described in a previous article.[13] There are limitations relating to

* E = applied steady-state voltage in volts
I = peak current in amperes
t = time in seconds

Table II — Input Requirements for Operate and Release for Two Sealed Contacts per Crosspoint

	Operation		Release	
	NI_O	$EIt_{\min}$	NI_R	$EIt_{\min}$
Remanent reed contact	32	94	36	80
Remanent sleeve crosspoint	100	1900	70	900

reoperation at high currents through the coil and also to variations with operating current of the break and make sequence in such transfer contacts. In particular, break-before-make contacts cannot always be assured under all operating conditions. For this reason several forms of 3-element transfer sealed contacts have been studied to provide break-before-make action under all conditions. One such experimental dry reed transfer[14] sealed contact is shown in Fig. 4(a). In this particular form, all 3 reeds are made of magnetic material. Fig. 4(b) shows the design relations required for good operation and a sketch of the device. Other dry reed transfer sealed contact forms are also under consideration.

III. FERREEDS AND BISTABLE DEVICES USING MINIATURE SEALED CONTACTS

Ferreeds were first described in an article[7] in the B.S.T.J. in 1960. Figs. 5 to 7 show several ferreed units. Fig. 5(a) shows one of the origi-

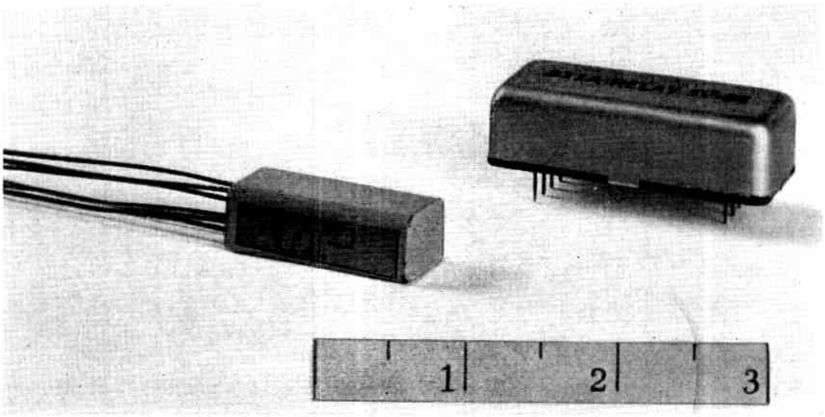


Fig. 3 — Relays using miniature sealed contacts: (a) 2 make contact missile relay GA 53702, (b) 3 make contact relay type 311A.

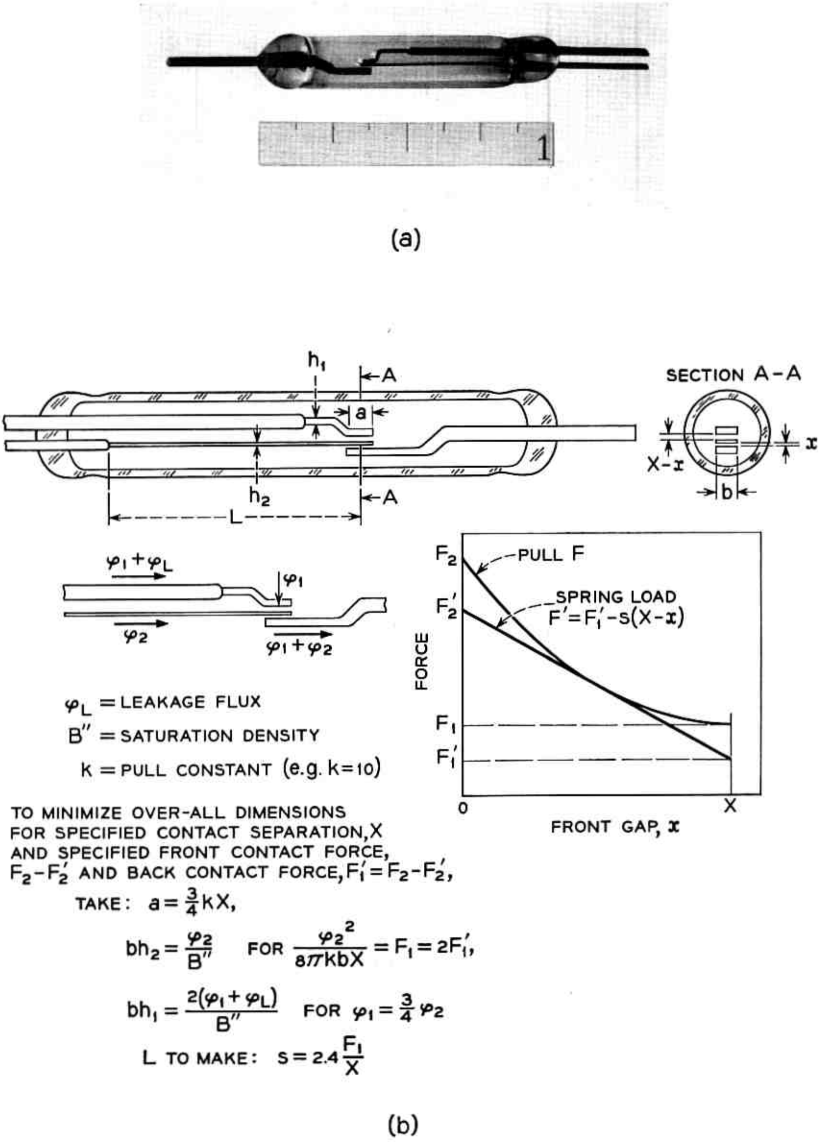

1
(a)
h₁
A
a
SECTION A-A
x
X-x
b
h₂
A
L
φ₁+φ_L
φ₁
φ₂
φ₁+φ₂
φ_L = LEAKAGE FLUX
B″ = SATURATION DENSITY
k = PULL CONSTANT (e.g. k=10)
TO MINIMIZE OVER-ALL DIMENSIONS FOR SPECIFIED CONTACT SEPARATION, X AND SPECIFIED FRONT CONTACT FORCE, F₂-F₂′ AND BACK CONTACT FORCE, F₁′=F₂-F₂′,
TAKE: a = ¾kX,
bh₂ = φ₂/B″ FOR φ₂²/8πkbX = F₁ = 2F₁′,
bh₁ = 2(φ₁+φ_L)/B″ FOR φ₁ = ¾φ₂
L TO MAKE: s = 2.4 F₁/X
F₂
PULL F
F₂′
SPRING LOAD F′=F₁′-s(X-x)
FORCE
F₁
F₁′
0
X
FRONT GAP, x
(b)

Fig. 4 — Miniature dry reed transfer sealed contact: (a) model G-54, (b) optimum design relations.

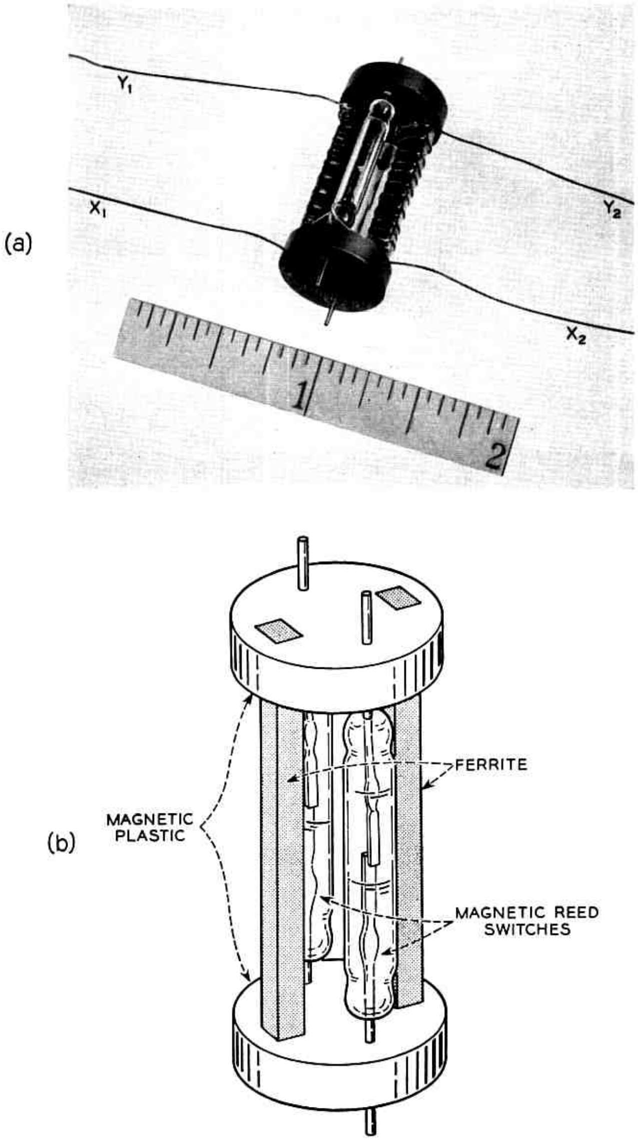


Fig. 5 — Ferreed designs: (a) photograph of 1960 design, (b) drawing of 1960 design.

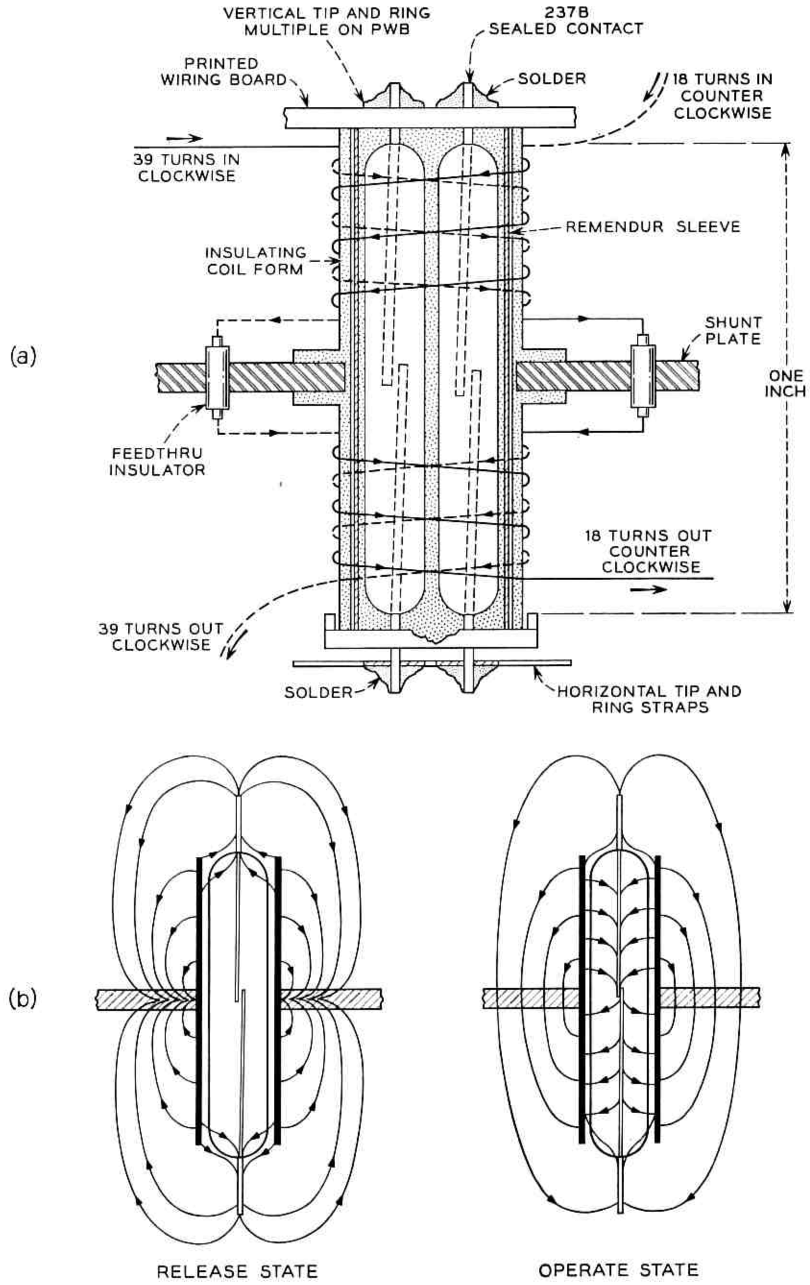


Fig. 6 — Ferreed designs (cont.): (a) crosspoint design with Remendur sleeve, (b) flux patterns with Remendur sleeve.

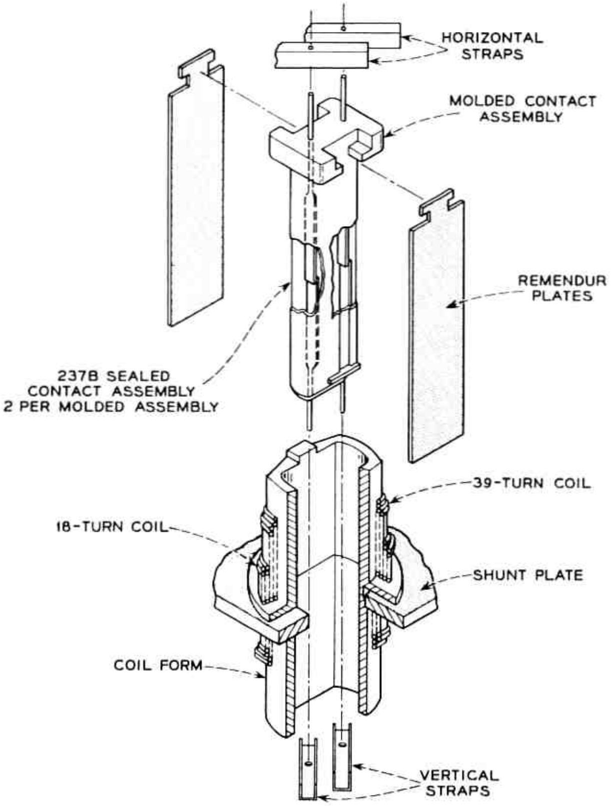


Fig. 7 — Ferreed designs (cont.): crosspoint design with Remendur plates.

nal parallel type ferreeds described in the 1960 article. Fig. 5(b) is a drawing of the same device. Fig. 6(a) shows another later series ferreed in which a sleeve of a "square loop" material (Remendur*) of the iron alloy type is used. Fig. 6(b) shows the flux patterns for the ferreed shown in Fig. 6(a). Fig. 7 shows a crosspoint using Remendur plates. An important characteristic of all of the ferreeds shown in Figs. 5 to 7 is the balanced magnetic release arrangement that eliminates marginal requirements on the release current.

In all cases one remanent member remains magnetized (half the remanent member in the series ferreed) while the field in the other member (or half member) is reversed in changing states. The field

* Remendur is an alloy of vanadium-iron-cobalt.

energy* which must be supplied to the operating coils to reverse magnetization is of the order of 3 to 5 times the remanent field energy of the remanent member and of the order of 10 or more on a pulse energy basis.

There should be no inherent difference in the performance of the parallel and series type ferreeds except (a) due to the energy requirement and (b) due to the dynamic characteristics in the sleeve or plate series ferreed where the flux through the reeds is necessarily reversed during each pulse. In this case the field due to the operating winding is in the opposite direction to the field supplied by the remanent members when the winding is not energized. The energy requirement mentioned in (a) can be less for the parallel type due to somewhat smaller air return reluctance, but on the other hand, the sleeve or plate series type provides better magnetic coupling.

The ferreeds having operate times down to about 5 microseconds use "square loop" ferrite magnetic materials. Somewhat simpler, less expensive and less temperature-sensitive forms of ferreeds use iron alloy metallic remanent materials in sleeve, plate, etc., form at some sacrifice in speed. However, speeds of about 50 microseconds or less are quite feasible. In any of these ferreeds, the magnetic material is set to the magnetized condition in microseconds. As a result of this, the sealed contacts close about 0.2 to 0.5 millisecond later. For almost all practical circuit conditions, this can be taken as operation in microseconds because circuit elements of this type are not usually required to release until other circuit operations are completed. Typical important *ferrite* characteristics for ferreed operation are coercive force, H_c, of 30–35 oersteds at maximum field, H, of 1000 and saturation flux density, B, of 4500 gauss, with corresponding remanence B_R about 2800. Typical magnetic characteristics of an iron alloy (Remendur) used with ferreeds are: H_c, 37–42 oersteds at maximum field, H, 100 and saturation flux density of 21,000 gauss, with corresponding remanence B_R of 17,000.

3.1 *Ferreed and Bistable Arrays*

In switching networks for electronic switching systems,[8] arrays and equipment assemblies of individual ferreed units are needed, for example 8 by 8, 1 by 8, etc. These have been needed in 2-wire and 4-wire forms. Accordingly, in the 8 by 8 array of the 4-wire type, 256 sealed contacts are needed. In one form, such arrays use four flat plates of

* The field energy is proportional to the product of the saturation flux for the reeds and the magnetomotive force required to develop this flux. Better magnetic coupling between the remanent members and the reeds will reduce the field energy required.

Remendur which are rolled in such a direction as to give the maximum magnetic properties in the direction of the reed axes.

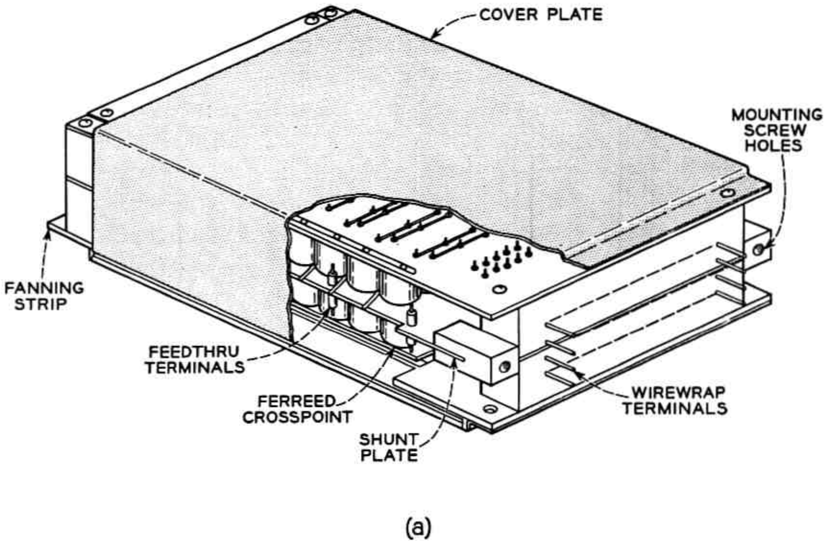
The operation of a ferreed array is somewhat similar to that of a crossbar switch in that a particular crosspoint is operated by the simultaneous operation of particular vertical and horizontal rows. A particular crosspoint is thereby operated and held in this condition without holding power. The winding arrangements of the ferreed elements are such that the other crosspoints remain unoperated. To release the crosspoint, in effect, reverse currents reset the magnetic material to the unmagnetized condition; hence the sealed contacts open. Fig. 8 shows an 8 by 8,2-wire array or switch.

The ferreed shown in Figs. 6(a) or 7 is the basic crosspoint element of the array shown in Fig. 8. This form contains 2 miniature dry reed sealed contacts surrounded by a sleeve (flat plates are more recent) of remanent magnetic material (Remendur). The magnetic shunt plate, positioned at the midpoint of the sleeve, separates the sleeve or plates magnetically into two independent halves. When the two halves are magnetized series-aiding, the flux return is through the reeds, causing the sealed contacts to close. When they are magnetized in series-opposition, the sealed contacts open.

Each end of each crosspoint has two windings. A winding on one end is connected in series-opposition, with the winding of half the number of turns on the other end, as shown in Fig. 6(a). When either of the two sets of windings is energized, the two ends of the sleeve or plate are poled oppositely and the sealed contacts are opened. When the two sets of windings are energized simultaneously with equal currents, the two ends are poled series-aiding and the sealed contacts close.

In a typical switch, 64 ferreed crosspoints are assembled together to form an 8 by 8 switch. Internal to the switch, the windings of rows and columns form a common multiple. To close a crosspoint, current is passed in one column and out one row via a common multiple. The crosspoint at the intersection of the column and row then closes. At the same time, current passes through one of the two windings of all other ferrends in the same row and column, causing any that are operated to release. This is a differential mode of operation, called "destructive mark"; it is characterized by the absence of specific network release operations, i.e. "taking down" connections. Connections are "taken down" as a direct result of, and at the same time as, connections that are set up.

Bipolar ferreeds are also needed in switching systems. Fig. 9 shows the magnetic circuit of one form of an individual bipolar element. A



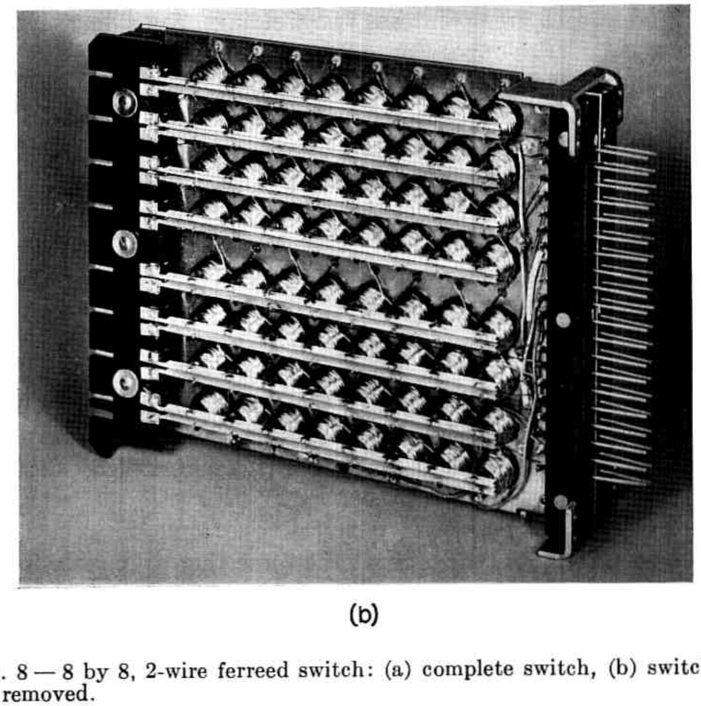


Fig. 8 — 8 by 8, 2-wire ferreed switch: (a) complete switch, (b) switch with cover removed.

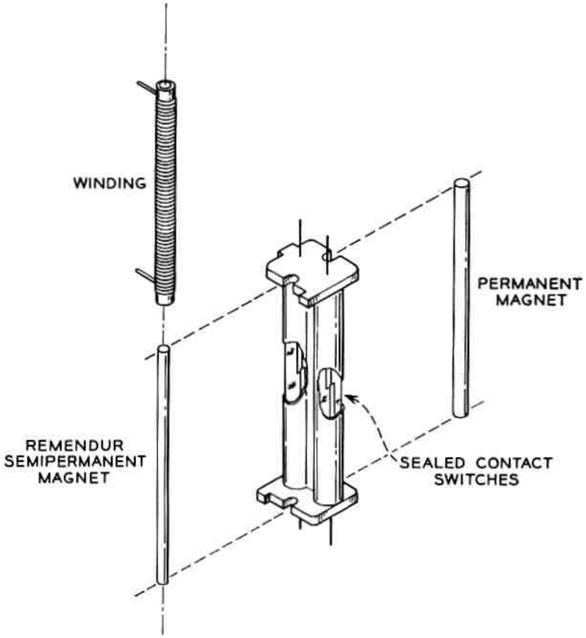


Fig. 9 — 2-contact bipolar ferreed.

combination of a "square loop" material is used together with a permanent magnet arranged as shown in relation to the sealed contacts. In this case more than one sealed contact may be used at each crosspoint. The bipolar unit gives a cutoff relay action. Fig. 10 shows a 1 by 8 unit of the 2-wire type. These open or close the reed contacts in response to the polarity of the current through a single winding.

IV. MERCURY-WETTED SEALED CONTACTS AND RELAYS

Fig. 11 shows a number of mercury-wetted sealed contacts of the transfer contact type. The 226D type is one of the smallest and most recent types. It is different from the others shown in that it is a break-before-make contact. The break-before-make action is the result of design changes, Fig. 11, of the pole-piece contact elements. Sealed contacts with mercury-wetted contacts are important because they have been shown to have the least contact chatter, often none, also have the longest operating life of any relays yet designed, and can exceed one billion operations.

The small size of the 226D mercury sealed contact can be packaged

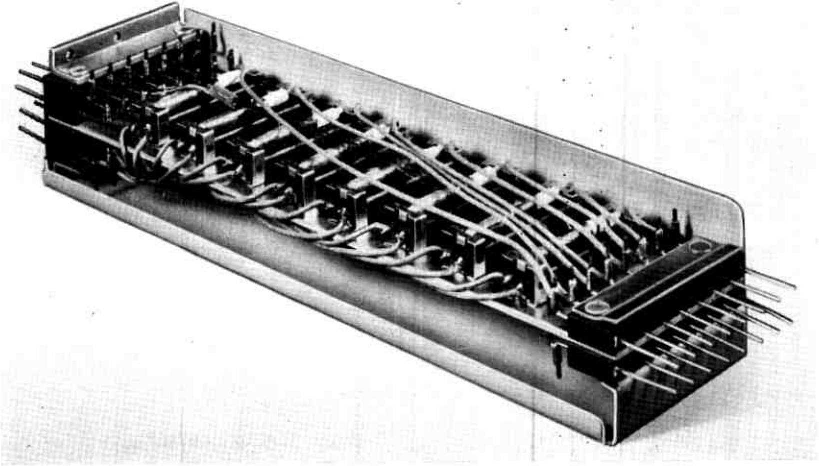


Fig. 10 — 1 by 8 assembly of 2-wire bipolar ferreeds.

in a small-size relay. However, two new relay designs using the new mercury sealed contact, the 314A and the 315A, do not require size reduction because they are chiefly expected to replace larger Bell System relays, namely the 255 and 280 types in certain applications where improved performance is needed.

The 314A is expected to replace the 255 type polar relay in telegraph circuits and to reduce maintenance in these. Fig. 12 shows the 255 relay and the new 314A relay. As can be seen, these are plug-in types and are interchangeable.

The 315A shown in Fig. 13 is a plug-in type and is expected to replace some of the codes of the 280 type polar relay, particularly those used in the No. 5 crossbar system, in order to improve performance and reduce maintenance. This is important in that the 280 type relays used in the No. 5 crossbar system show the highest relay trouble rate in terms of troubles per 1000 relays per year. However, 280 type relays are used in smaller numbers, in such systems, to perform special and exacting functions.

All of the mercury sealed contacts discussed, or used by the Bell System up to the present time, are required to operate in a vertical position within certain limits, usually ± 30 degrees. Military applications, particularly, would be served by an "all-position" mercury sealed contact. Several forms of such contacts have been built and studied. Most of these have been judged to be rather complicated and relatively expensive to control and manufacture. A more recent and simpler experimental design is shown in Fig. 14. Basically, this is a modification of the 226D sealed contact shown in Fig. 11 but modified in two ways: (*i*) excess mercury is removed during manufacture, including the usual pool of mercury, and (*ii*) armature changes have been made to improve the contact performance under shock and vibration conditions. By reducing the amount of available mercury for replenishment at the contact surface, the life of the sealed contact is reduced, but several million op-

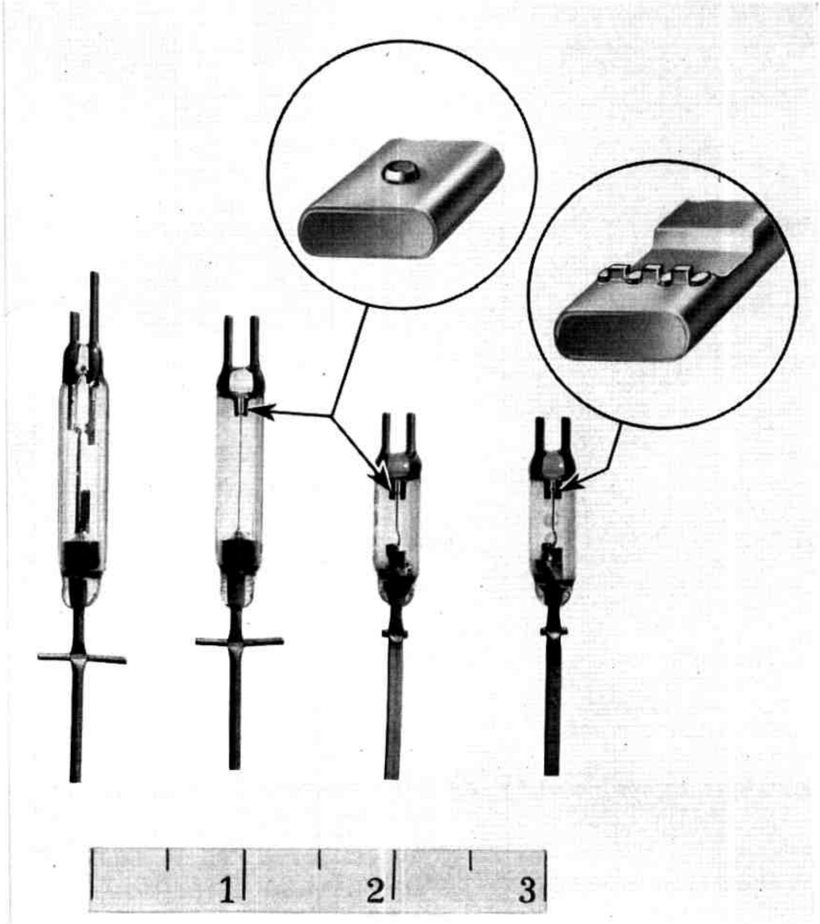


Fig. 11 — Mercury-wetted sealed contacts: left, 218 type; left center, 222 type make-before-break contact; right center, 226B type make-before-break contact; right, 226D type break-before-make contact.

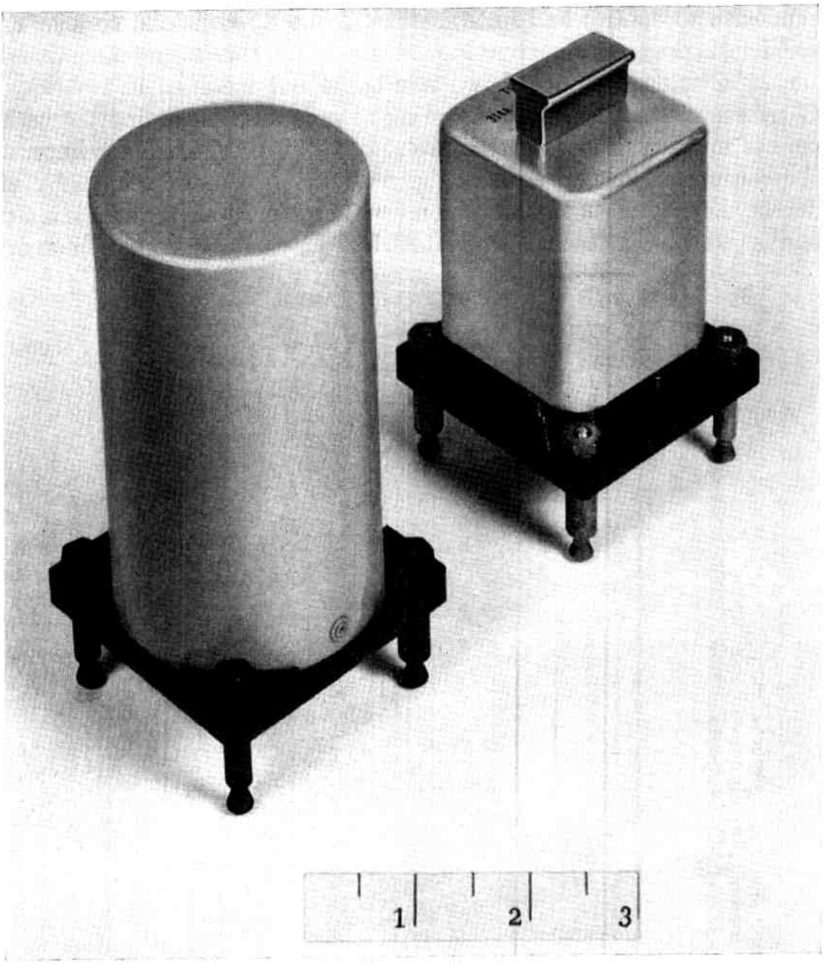


Fig. 12 — Telegraph relays: left, standard 255 type; right, new 314A type using 226D sealed contact.

erations are possible. For many applications this is adequate. This relay is described in detail in an article[15] in the Bell Laboratories Record.

V. MINIATURE ARMATURE TYPE RELAYS

5.1 *Rotary Armature Relays*

A miniature relay of this type was described in a paper[16] in 1959. Fig. 15(a) is a photograph of this relay and Fig. 15(b) is a drawing of its

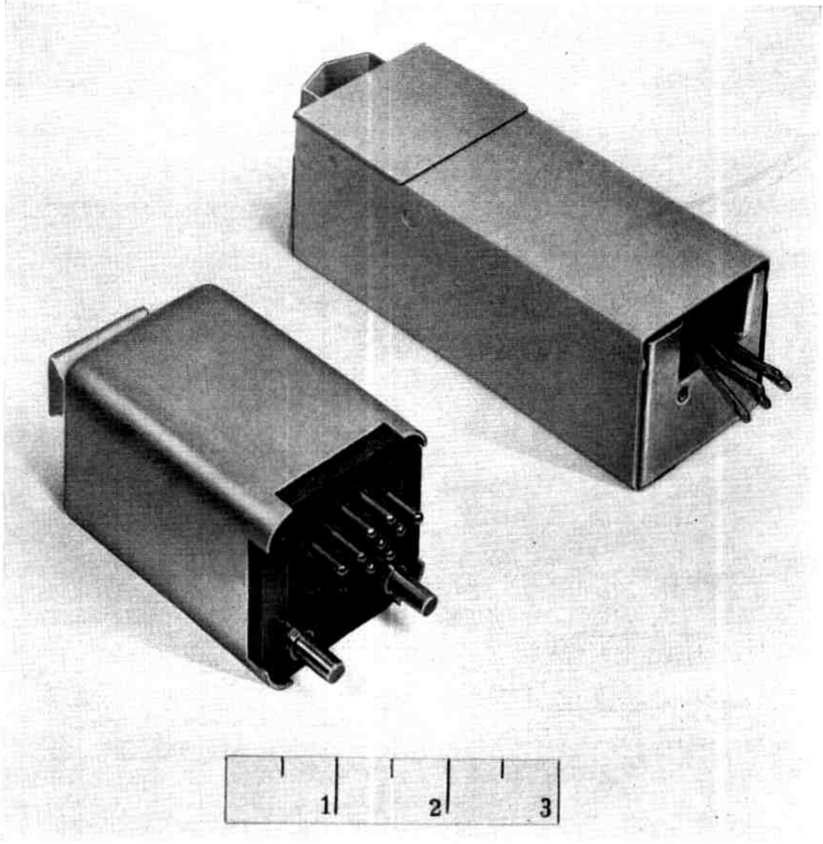


Fig. 13 — Polar relays: right, standard 280 type; left, new 315A type using 226D sealed contact.

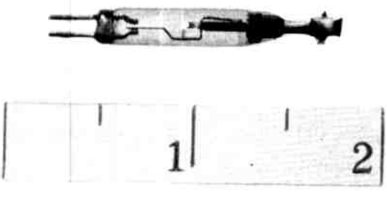


Fig. 14 — Experimental "all position" mercury-wetted sealed contact model T-116.

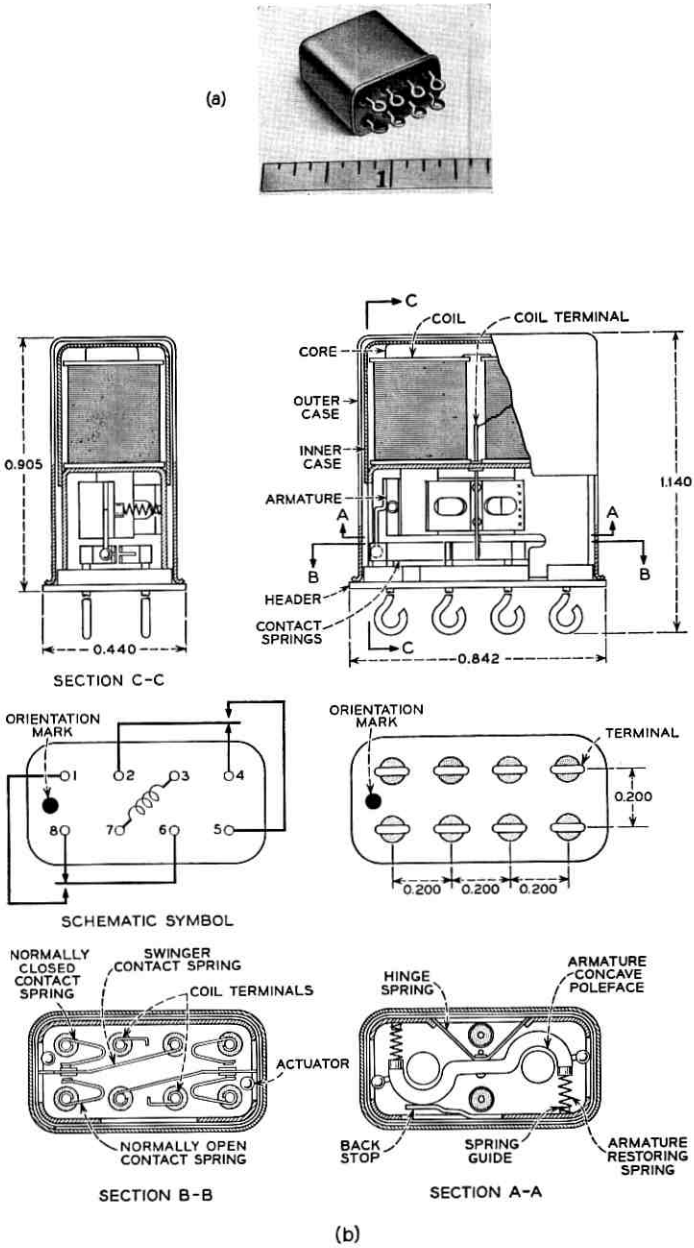


Fig. 15 — Miniature rotary armature relay: (a) photograph of GS 57668 relay, (b) drawing showing relay construction.

major elements. It has been in manufacture for military applications as the GS 57668 relay. It is of the "crystal can" size and has a rotary armature operating two transfer contacts symmetrically arranged. As compared with similar relays it has the following advantages: (*i*) improved contact reliability, particularly in dry circuits, by the use of twin precious metal contacts in a separate sealed contact chamber free of all organic materials; this eliminates the so-called "brown powder" problem in which organic polymers are formed with resulting high-resistance contacts; (*ii*) elimination of bearing friction and the associated erratic performance; this is accomplished by using a reed type spring armature suspension; and (*iii*) a magnetic design of improved sensitivity with corresponding reduced effect due to stray magnetic fields.

5.2 Telstar *Satellite Type Relays*[17]

Fig. 16 shows a relay similar to the "crystal can" relay shown in Fig. 15 except that it operates or releases on pulses. It uses magnetic latching so that no holding power is required. This relay is used in the Bell System Telstar satellites; in fact nine each are used in Telstar I and Telstar II. Fig. 16(a) is a photograph of the relay, and Fig. 16(b) is a drawing of the chief features. It is characterized by the dual armatures in which the two armatures are connected together by a small permanent magnet. Fig. 16(c) shows the control circuit in Telstar I using the relay.

5.3 *MA and MB Miniature Relays*[18]

A new series of relays known as MA and MB types has recently been developed, primarily to save space for equipment installed on the premises of Bell System customers. Manufacture of these was started at the Western Electric Co. plant at Kearny, N. J., in 1962. Fig. 17 shows the MA and MB relays. The MA relay has a maximum contact capacity of 4 transfer contacts and the MB, which uses some of the same piece parts, has a maximum contact capacity of 6 transfer contacts.

These relays have most of the basic features of the standard wire spring relay (Ref. 2), namely: (*i*) code card operation to provide a simple means for a wide variety of contact combinations; (*ii*) low stiffness, pretensioned springs; (*iii*) coplanar spring groups to simplify welding and handling and to standardize assembly in manufacture; (*iv*) contact materials and contact forces identical with the standard wire spring relay; (*v*) essential elimination of locked contacts because of the card operation; (*vi*) twin precious metal contacts; etc. The basic contact springs are shown in Fig. 18 before and after shearing the ends of the

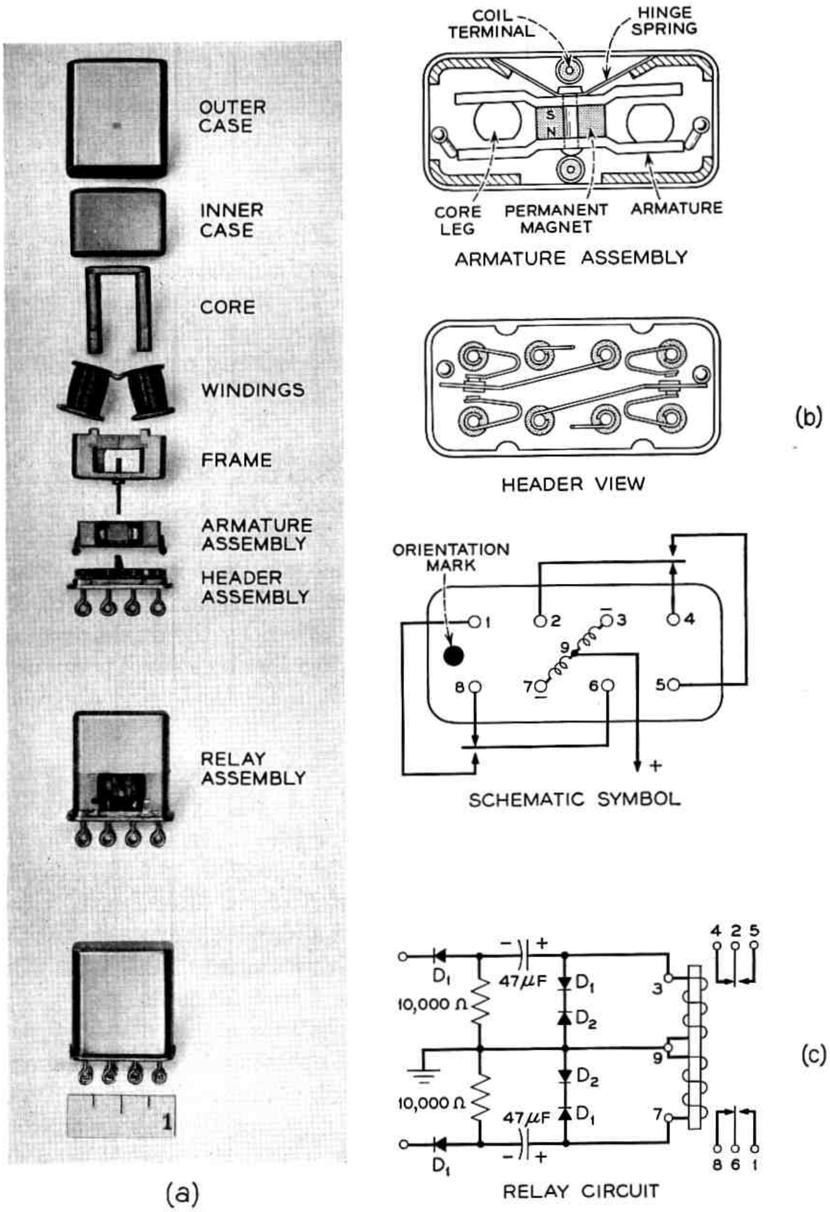


Fig. 16 — Miniature rotary armature latching relay (Telstar): (a) photograph of relay and relay parts, (b) relay armature assembly and circuit used in Telstar satellite, (c) control circuit in Telstar I.

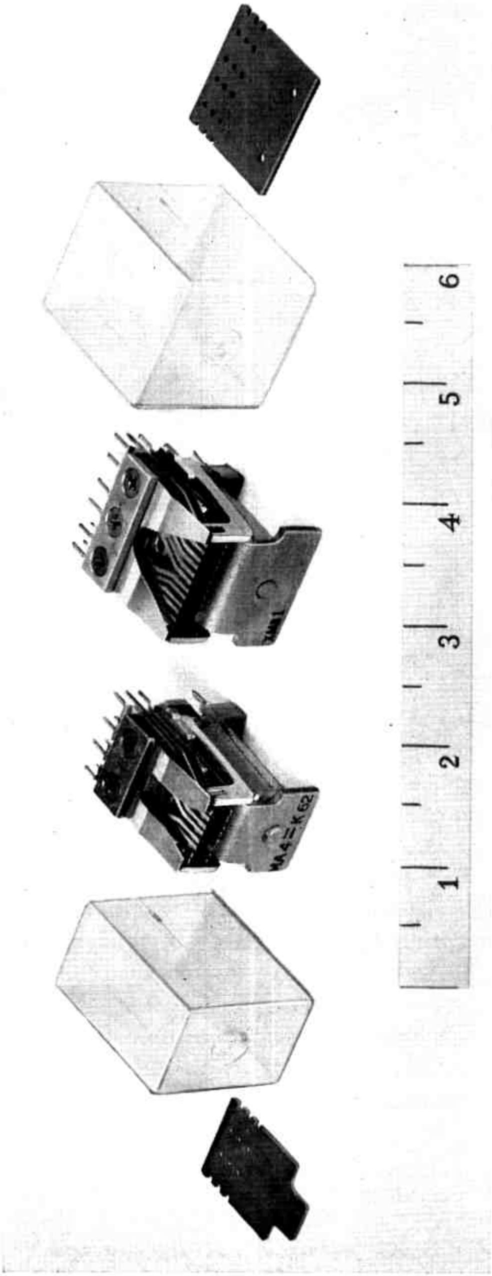


Fig. 17 — Miniature MA and MB type relays: left, MA type; right, MB type.

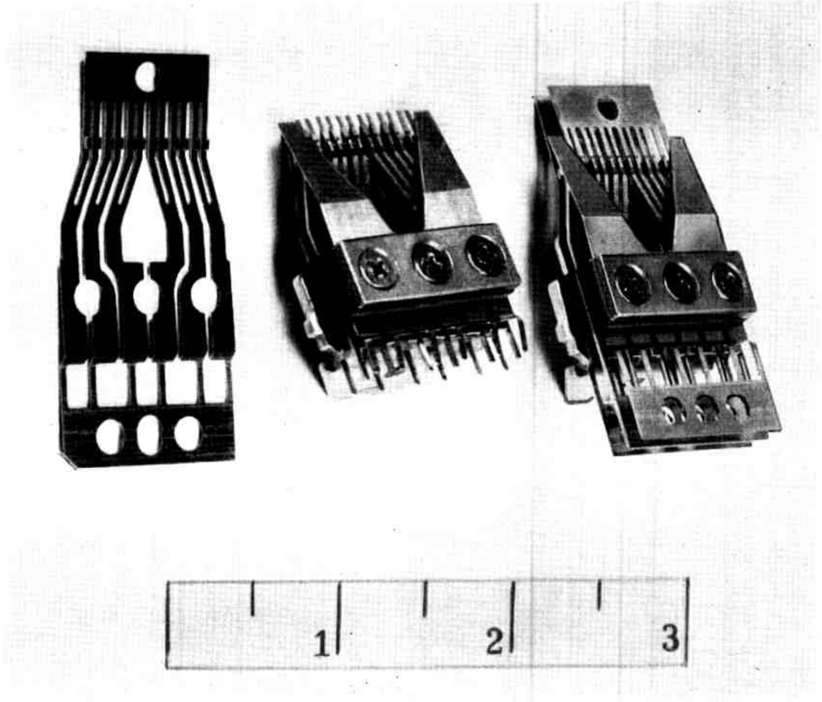


Fig. 18 — Contact springs for MB type relay.

contact spring groups. Typical contact and winding information and operating currents are given in Table III. As is the case for the standard general purpose wire spring relay, a few code cards are sufficient for a large number of contact combinations.

The MA and MB relays do not have the sensitivity* or the contact capacity of the wire spring relays, but they are much smaller, i.e., about $\frac{1}{10}$ the volume, and they are suitable for mounting on printed circuit boards. One such typical plug-in printed circuit package is shown in Figs. 19(a) and 19(b). The same principles used in the MA and MB relays can also be used in crossbar switch designs to reduce the size and weight to about 15 per cent of the present types.

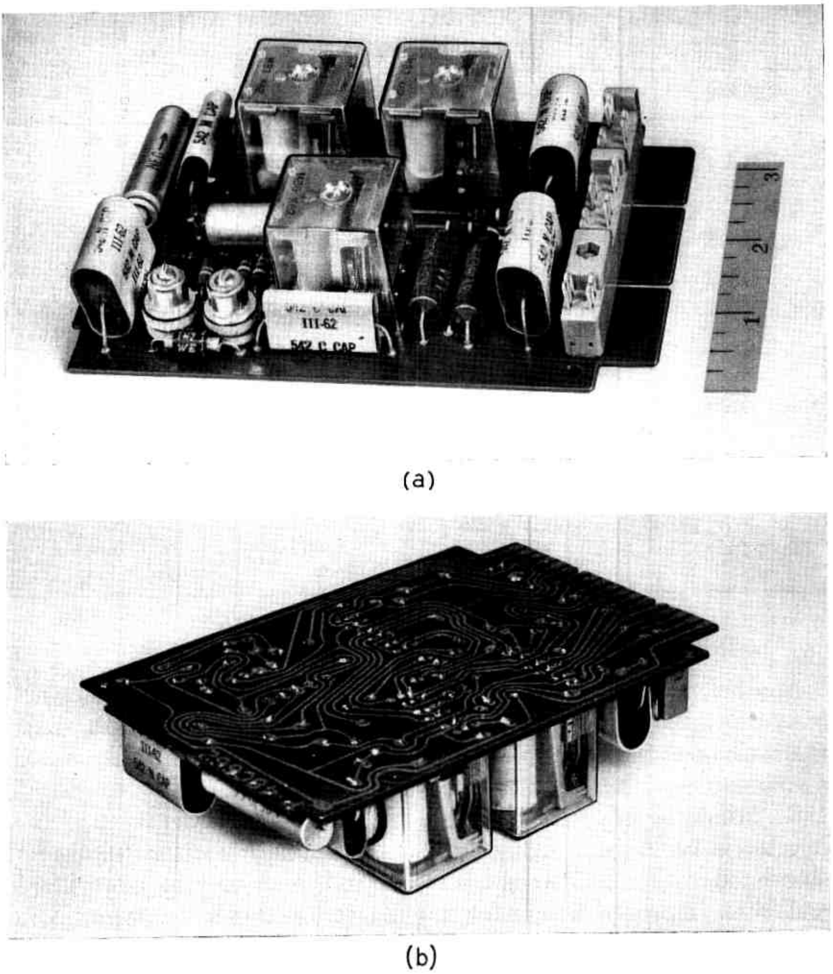
* Ampere-turn sensitivity of the 6 transfer MB relay is about 185, compared with 160 for the AF wire spring relay and 220 for the AK (5 transfer) relay. However, because of the larger coil on wire spring relays, the relative power sensitivities for 6 transfer relays are about: 0.45 watt for the MB, 0.18 for the AF, 0.14 for the AJ, and 0.55 for the AK relay.

Table III — Some Typical MA and MB Relay Code Information

Code	Springs	Winding Resistance	Operate Current
		(ohms)	(amperes)
MA 1	4 transfers	915	0.016
MA 3	2 makes 2 breaks	590	0.013
MA 4	3 transfers 1 continuity	915	0.016
MA 7	3 makes 1 transfer	2100	0.0078
MA 11	2 transfers 2 continuities	590	0.021
MB 1	6 transfers	590	0.024
MB 3	2 transfers 1 continuity 2 makes 1 break	915	0.018
MB 4	6 makes	915	0.016
MB 6	2 continuities 2 makes 1 early break	915	0.0175
MB 7	3 transfers 3 continuities	590	0.024

VI. FREQUENCY SENSITIVE RELAY—THE VIBRATING REED SELECTOR

Another miniature device, shown in Fig. 20, is a frequency sensitive relay called the 215 type tuned reed selector.[6] Fig. 21 shows a drawing of the basic operating principles. The selector shown in Fig. 20 has been in manufacture at the Western Electric Co. in North Carolina, starting in 1962, primarily for the Bell System BELLBOY radio paging service.[19] The selector is basically a highly precise and stable miniature tuning fork associated with a lightweight contact. It is smaller and more stable, and is an improved design for manufacture compared with an earlier similar device known as the type 212 selector.[20] These devices are very sensitive, responsive only to sustained frequencies of the order of 0.5 second, and insensitive to noise interference. Fig. 22 shows the data over a wide temperature range for two of these devices, operating at nominal frequencies of 517.5 and 997.5 cycles per second and at corresponding bandwidths of about 1.1 and 1.3 cycles per second. Sufficient stability has been achieved so that, for the BELLBOY service, 33 different frequencies spaced 15 cycles apart are provided in less than one octave between 517.5 and 997.5 cycles. By using three different frequencies at a time, more than 5000 combinations are possible for selective ringing of a particular customer.



(a)

(b)

Fig. 19 — Plug-in printed wiring board with MB type relays: (a) apparatus side, (b) wiring side.

Stability of materials and design have been measured, and these show the total frequency change from −40°C to +80°C to be less than 0.5 cycle and the bandwidth change to be less than 0.2 cycle. At operating power levels of 100 microwatts, the intermittent contact will close to a low-resistance level over 20 per cent or more of the cycle time. An important factor in this has been the use of a nickel-iron-molybdenum alloy[21] (Vibralloy). This material has controlled elastic and magnetic properties.

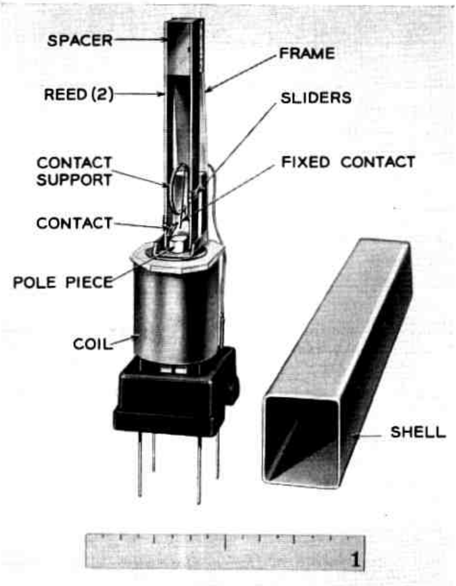


Fig. 20 — Tuned reed selector (BELLBOY) — 215 type.

The lightweight contact is essential so that the selector frequency is unchanged when the intermittent contact is made. The contacts are rhodium against platinum rhodium. Clearly, contact life is important and circuits are used typically to change the potential on an electron tube or transistor to trigger a switching or signaling function without exceeding a contact current of a few milliamperes. In the BELLBOY application a transistor oscillator is triggered to give an audible signal. However, the short contact closures occurring at a rate of hundreds per second may therefore control pulses that have an integrated or average power that is a substantial fraction of a watt. For example, only small

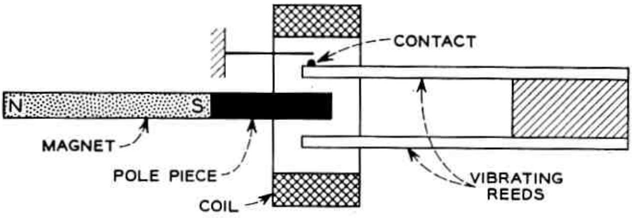


Fig. 21 — Tuned reed selector schematic.

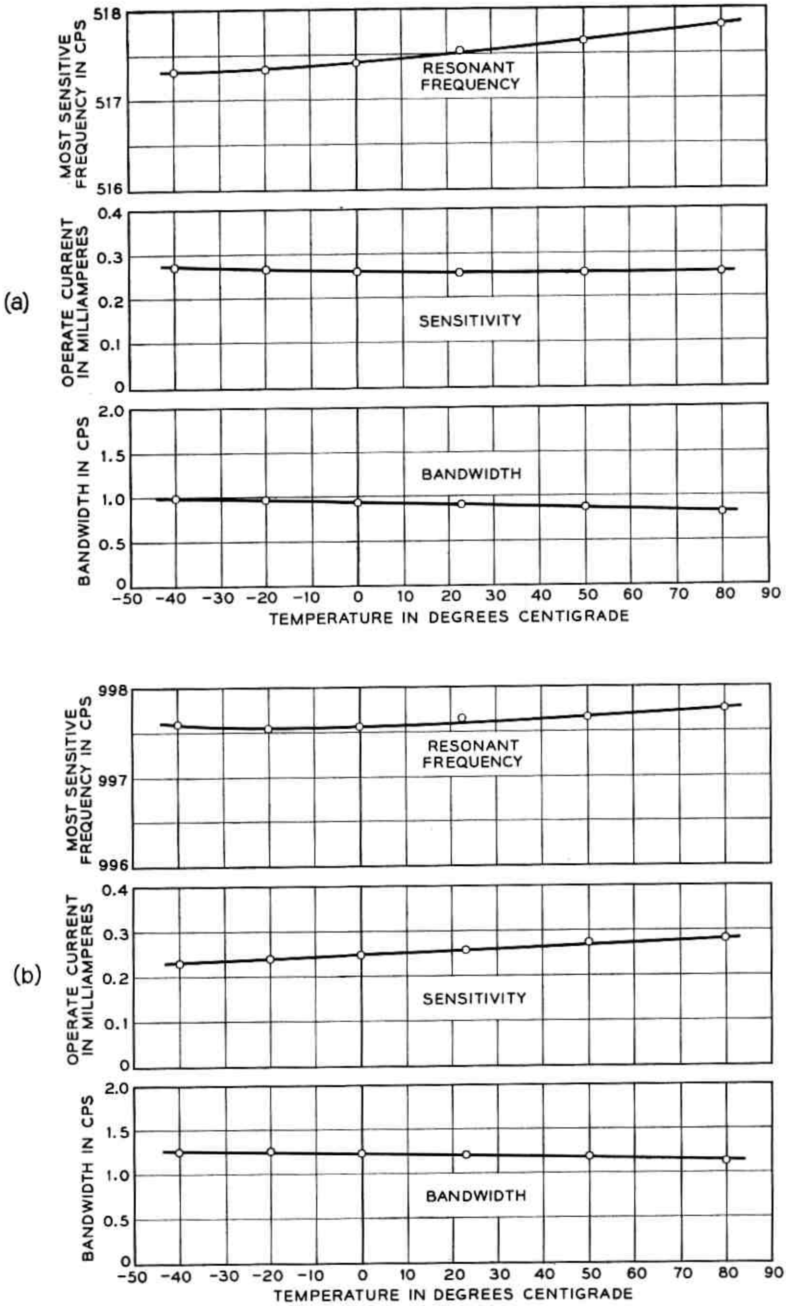


Fig. 22 — 215 type tuned reed selector data: (a) temperature characteristics of 517.5-cycle unit, (b) temperature characteristics of 997.5 cycle unit.

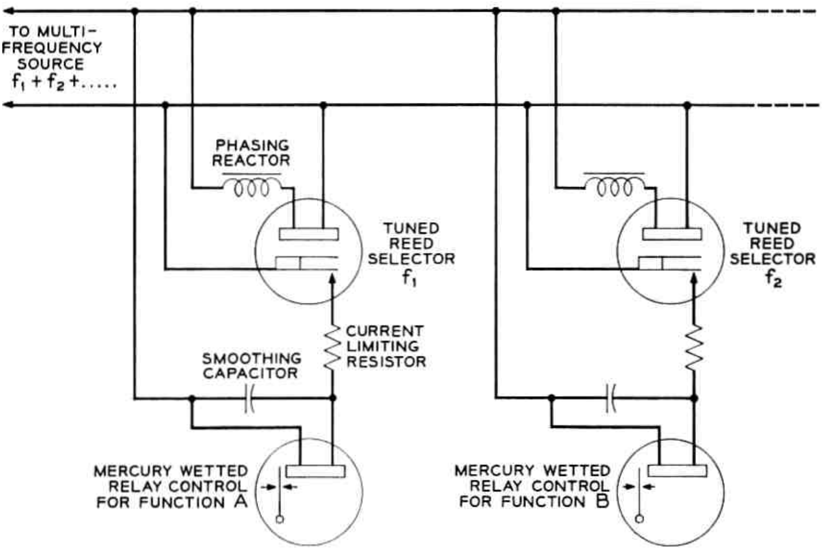

TO MULTI-FREQUENCY SOURCE $f_1 + f_2 + \ldots$
PHASING REACTOR
TUNED REED SELECTOR f_1
TUNED REED SELECTOR f_2
CURRENT LIMITING RESISTOR
SMOOTHING CAPACITOR
MERCURY WETTED RELAY CONTROL FOR FUNCTION A
MERCURY WETTED RELAY CONTROL FOR FUNCTION B

Fig. 23 — Direct operation of mercury-wetted relay from low-level frequency frequency signals via tuned reed selectors.

changes in frequency or sensitivity were measured over a test period of 1500 hours in a 12-volt circuit with a 240-ohm resistor giving a closing current of 50 milliamperes. The power capacity of the contacts can, in fact, be used to operate relays or other devices directly: for example, mercury sealed contact relays with large contact current capacity. One such circuit is shown in Fig. 23. In this circuit the selector contact is used as a synchronous rectifying means to generate dc from the same ac source that operates the selector. When the input frequency corresponds to that of the selector, the contact closes in synchronism once each cycle to send unidirectional pulses to the capacitor and relay in parallel. The capacitor serves to smooth the pulses so that the relay winding has nearly a constant current in it. Combination circuits using reed selectors and mercury-wetted contact relays provide a simple means of selectively controlling substantial powers to perform a multiplicity of functions over a single pair of wires.

VII. REMARKS

In the telecommunications field, rapid advances are being made in many new areas of technology. Devices and systems based on these will naturally be compared and evaluated for Bell System applications with older devices and systems. In such comparisons, care is needed to do this, not only with devices at hand but with the possibilities that presently exist on the basis of general advances made in the older fields. One of the older and important areas is that of electromechanical devices such as the relays discussed in this article. Decisions can then be made and devices chosen, not on the basis of technology, but on the basis of the best performance, cost, and over-all systems requirements. Relays, in modern form, sometimes in miniature form, can be expected to be important devices in the future as they have been in the past.

REFERENCES

1. Shackleton, S. P., and Purcell, H. W., Relays in the Bell System, B.S.T.J., **3**, January, 1924, p. 1.
2. Keller, A. C., A New General Purpose Relay for Telephone Switching Systems, B.S.T.J., **31**, November, 1952, p. 1023.
3. Peek, R. L., Jr., and Wagar, H. N., *Switching Relay Design*, D. Van Nostrand, New York, 1955.
4. Keller, A. C., Wagar, H. N., Peek, R. L., Jr., and Logan, M. A., B.S.T.J., **33**, January, 1954, entire issue; also printed as Bell Telephone System Monograph 2180.
5. Guettich, T. H., The "Two-in-One" Wire Spring Relay, Bell Laboratories Record, **36**, December, 1958, p. 458.
6. Bostwick, L. G., A Miniature Tuned Reed Selector of High Sensitivity and Stability, B.S.T.J., **41**, March, 1962, p. 411.
7. Feiner, A., Lovell, C. A., Lowry, T. N., and Ridinger, P. G., The Ferreed — A New Switching Device, B.S.T.J., **39**, January, 1960, p. 1.
8. Feiner, A., The Ferreed, B.S.T.J., this issue, p. 1.
9. Keller, A. C., Relays and Switches, Proc. I.R.E., **50**, May, 1962, p. 932.
10. Hovgaard, O. M., and Perreault, G. E., Development of Reed Switches and Relays, B.S.T.J., **34**, March, 1955, p. 309; Hovgaard, O. M., and Fontana, W. J., Proc. Electronics Components Conf., Philadelphia, Pa., May, 1959; Peek, R. L., Jr., Magnetization and Pull Characteristics of Mating Magnetic Reeds, B.S.T.J., **40**, March, 1961, p. 523.
11. Ellwood, W. B., Brown, J. T. L., and Pollard, C. E., Recent Developments in Relays, Elec. Eng., **66**, November, 1947, p. 1104.
12. Peek, R. L., Jr., unpublished work; U. S. Patent No. 3,059,075, filed October 22, 1959; issued October 16, 1962.
13. Husta, P., and Perreault, G. E., Magnetic Latching Relays Using Glass Sealed Contacts, B.S.T.J., **39**, November, 1960, p. 1553.
14. Bradford, K. F., An Experimental Dry Reed Sealed Transfer Contact, Bell Laboratories Record, **41**, May, 1963, p. 200.
15. Pollard, C. E., Position Independent Mercury Contacts, Bell Laboratories Record, **41**, February, 1963, p. 58.
16. Schneider, C., and Spahn, C. F., Miniature Relay for High Reliability, Proc. Electronics Components Conf., Philadelphia, Pa., May, 1959. U. S. Patent No. 3,042,773, filed December 19, 1958; issued July 3, 1962.
17. Schneider, C., A Miniature Latch-In Relay for the *Telstar* Satellite, Bell Laboratories Record, **41**, June, 1963, p. 241.
18. Werring, W. W., Miniature Relays for Key Telephone Systems, Bell Laboratories Record, **40**, December, 1962, p. 414.
19. Mitchell, D., and Van Wynen, K. G., A 150-mc Personal Radio Signaling System, B.S.T.J., **40**, September, 1961, p. 1239.
20. Keller, A. C., and Bostwick, L. G., Vibrating Reed Selectors for Mobile Radio Systems, Trans. AIEE, **68**, 1949, p. 383.
21. Fine, M. E., and Ellis, W. C., Thermal Variations of Young's Modulus in Some Fe-Ni-Mo Alloys, Jour. Metals, **3**, September, 1951, p. 761.

Overflow Traffic from a Trunk Group with Balking*

By PETER LINHART

(Manuscript received April 18, 1963)

A stream of telephone calls is submitted to a group of trunks, the first-choice group, according to a recurrent process. We allow balking on this trunk group; i.e., if a call finds k of the first-choice trunks busy it may be served, with probability p_k, or may fail to be served, with probability q_k. A call which fails to receive immediate service on the first-choice trunk group is submitted to a second-choice trunk group, the overflow group. We also allow balking on the overflow group. Calls which fail to receive immediate service on the overflow group are lost to the system. Holding times have negative-exponential distribution.

We give methods for finding the joint distributions of numbers of busy trunks on the first-choice and overflow groups, at overflow instants (i.e., instants at which calls are submitted to the overflow group), at arrival instants, and at arbitrary instants. We consider the transient as well as the limiting distributions (and demonstrate the existence of the limiting distributions).

The methods developed are illustrated by several examples. Numerical results are given for the blocking in the particular case that the first-choice group constitutes a random slip, while the overflow group is full-access (common).

I. INTRODUCTION

1.1 *Balking and Overflow Traffic*

A telephone call is submitted to a group of m trunks. This call may fail to occupy a trunk, even though not all m trunks are busy. There may be a number of reasons for such a failure, e.g.: the calling line may not have access to any *idle* trunks, some equipment other than the

[*] This paper represents part of a doctoral dissertation submitted to the Subcommittee on Applied Mathematics, Columbia University.

trunk itself may be required to complete a connection and this equipment may be busy, or the m trunks may be merely first-stage links in a connecting network and there may be no free path through this network. Whatever the cause of the failure, we shall say that the submitted call *balks* (although the word is perhaps more appropriate in queueing theory applications). In this paper we shall restrict ourselves to the case in which the probability of balking depends only on the number of busy trunks: if an arriving call finds k trunks busy, it is served, with probability p_k, or balks with probability q_k $(p_k + q_k = 1)$. If all trunks are busy, an arriving call cannot be served, and therefore $q_m = 1$. Thus we subsume blocking under the term balking.

The traffic which overflows from a trunk group with balking has different characteristics from that which overflows from a *full-access* group. [By a full-access trunk group we mean one for which $q_k = 0$ $(k < m)$, $q_m = 1$.] Suppose *recurrent* traffic is submitted to a full-access group (when we refer to recurrent input traffic we mean that the intervals between arriving calls are independent, identically distributed random variables). Suppose further that the holding times of calls have negative-exponential distribution. Then, as Conny Palm[1] has shown, the overflow traffic is also recurrent. This is not the case for traffic overflowing from a trunk group with balking.

The traffic which balks on the first-choice group may be submitted to an overflow group of, say, M trunks. There may also be balking on the overflow group. Now L. Takács[2] has treated in detail the process of numbers of busy trunks in a trunk group with balking to which a recurrent stream of calls of negative-exponential holding times is submitted. Thus, if the first-choice group is full-access, we know how to describe what goes on on the overflow group. However, if the first-choice group is not full-access, the stream of calls submitted to the overflow group is not recurrent, and therefore further analysis is required to describe the process of numbers of busy trunks on the overflow group. We attempt to treat this problem in the present paper; in so doing, we are led to consider the joint distribution of numbers of busy trunks on the first-choice and overflow groups, which is also of interest in itself.

1.2 *Mathematical Description of the Problem, and Some Notation*

Calls are submitted to a group of m trunks, the first-choice group, at successive instants $\tau_1, \tau_2, \cdots, \tau_n, \cdots$. The interarrival times, $\theta_n = \tau_n - \tau_{n-1}$ $(n = 2, 3, 4, \cdots)$, are independent, identically distributed random variables with common distribution function

$$P\{\theta_n \leqq x\} = F(x),$$

and we specify further that $P\{\tau_1 \leqq x\} = F(x)$. We assume that the $\{\theta_n\}$ are not *lattice variables* (i.e., that the interarrival times are not confined to multiples of a constant), that $F(0) = 0$ and that

$$0 < \alpha < \infty,$$

where

$$\alpha = \int_0^\infty x dF(x)$$

is the mean interarrival time.

Note that the class of recurrent inputs just described includes, among others: Poisson arrivals, equally spaced arrivals, and, as previously remarked, arrivals which are themselves overflows from a full-access trunk group to which a Poisson process of calls with negative-exponential holding time is submitted.

If the nth call receives service, then its holding time is a random variable, χ_n. The $\{\chi_n\}$ are independent and identically distributed, with common distribution function

$$P\{\chi_n \leqq x\} = \begin{cases} 1 - e^{-x} & \text{for } x \geqq 0 \\ 0 & \text{for } x < 0 \end{cases}$$

and are independent of the arrival process $\{\tau_n\}$.

Note that we are measuring time in units of the mean holding time; thus $a = 1/\alpha$ is the submitted traffic in erlangs.

An arriving call which finds k trunks of the first-choice group busy is served with probability p_k, or balks with probability q_k. We have

$$p_k + q_k = 1 \qquad (k = 0, 1, \cdots, m)$$
$$q_m = 1.$$

A call which balks on the first-choice group is immediately submitted to a second group of M trunks, the *overflow group* (we allow the case $M = \infty$). We denote the sequence of instants at which calls are submitted to the overflow group by $\{T_N\}$ $(N = 1, 2, 3, \cdots)$. If such a call finds K trunks of the overflow group busy, it is served, with probability G_K, or balks, with probability H_K. We have

$$G_K + H_K = 1 \qquad (K = 0, 1, \cdots, M)$$
$$H_M = 1 \qquad (\text{if } M < \infty).$$

We make the following plausible restriction on the balking probabilities

$$p_k > 0 \quad \text{for} \quad k < m$$
$$G_K > 0 \quad \text{for} \quad K < M.$$

A call which balks on the overflow group is said to be *blocked*. It immediately disappears from the system and is not resubmitted; i.e., lost calls are cleared.

We now define the following random variables:

$$\begin{aligned}
\xi(t) &= \text{number of busy trunks on first-choice group at time } t \\
\xi_n &= \xi(\tau_n -) \\
\xi_n^{\,o} &= \xi(T_N -) \text{ (the superscript ``}o\text{'' means ``overflow''.)} \\
\Xi(t) &= \text{number of busy trunks on overflow group at time } t \\
\Xi_n &= \Xi(\tau_n -) \\
\Xi_N^{\,o} &= \Xi(T_N -).
\end{aligned}$$

We also define the following probabilities, which it will be our object to determine:

$$P\{\xi_N^{\,o} = k,\ \Xi_N^{\,o} = K\} = P^o(k,K,N)$$
$$\lim_{N\to\infty} P^o(k,K,N) = P^o(k,K)$$
$$P\{\xi_n = k,\ \Xi_n = K\} = P(k,K,n)$$
$$\lim_{n\to\infty} P(k,K,n) = P(k,K)$$
$$P\{\xi(t) = k,\ \Xi(t) = K\} = P(k,K,t)$$
$$\lim_{t\to\infty} P(k,K,t) = P^*(k,K).$$

When one of the variables k or K in one of these probabilities is not written, it is understood to be summed over, e.g.

$$P(k,t) = \sum_{K=0}^{M} P(k,K,t).$$

A quantity of particular interest in applications is the blocking

$$B = \sum_{k=0}^{m} \sum_{K=0}^{M} q_k H_K P(k,K).$$

We shall also have occasion to refer to the blocking on the first-choice group

$$b = \sum_{k=0}^{m} q_k P(k).$$

Further notation will be introduced as it is needed. The notation will as far as possible conform to that of Takács.[2] We shall, when possible, use lower-case letters to refer to the first-choice group and the corresponding capital letters for the overflow group. Equations of Ref. 2 will be denoted by a T: e.g., "(T44)." We note here only the following definitions:

$$\varphi(s) = \int_0^\infty e^{-sx}\, dF(x)$$

$$C_r = \prod_{j=1}^{r} \frac{\varphi(j)}{1-\varphi(j)} \qquad (C_0 = 1)$$

$$C_r(s) = \prod_{j=0}^{r} \frac{\varphi(s+j)}{1-\varphi(s+j)} \qquad (C_{-1}(s) \equiv 1).$$

1.3 *Previous Results*

Let us denote the interoverflow times by $\Theta_N = T_N - T_{N-1}$. As we have mentioned, if the first-choice group is full-access, the $\{\Theta_N\}$ are independent and identically distributed. In this case let us denote their common distribution function by

$$G(x) = P\{\Theta_N \leqq x\}$$

with Laplace-Stieltjes transform

$$\gamma(s) = \int_0^\infty e^{-sx}\, dG(x).$$

Takács[3] solves a recurrence of Palm[1] to obtain

$$\gamma(s) = \frac{\displaystyle\sum_{r=0}^{m} \binom{m}{r} \frac{1}{C_{r-1}(s)}}{\displaystyle\sum_{r=0}^{m+1} \binom{m+1}{r} \frac{1}{C_{r-1}(s)}}. \tag{1}$$

A. Descloux[4] gives convenient recurrence formulas for calculating $\gamma(s)$ and the moments of $G(x)$ in the case of Poisson input, i.e., when

$$F(x) = \begin{cases} 1 - e^{-ax} & (x \geqq 0) \\ 0 & (x < 0) \end{cases}.$$

Some results exist for $P(k,K)$ in the case of Poisson input [for which, and only for which, as we shall see, $P^*(k,K) = P(k,K)$]. The first of these is due to L. Kosten.[5] He considers a full-access first-choice group and an infinite full-access overflow group. Let us denote binomial moments with respect to the overflow group by

$$U(k,R) = \sum_{K=R}^{M} \binom{K}{R} P(k,K).$$

Then Kosten finds

$$U(k,R) = C_0^R(a)\,\frac{C_0^m(a)C_R^k(a)}{C_R^m(a)C_{R+1}^m(a)}. \tag{2}$$

(See also the appendix by J. Riordan to a paper of R. I. Wilkinson.[6]) The polynomials in (2) are defined by

$$C_R^k(a) = \sum_{j=0}^{k} \binom{j+R-1}{j} \frac{a^{k-j}}{(k-j)!} \tag{3}$$

so that $C_0^k(a) = a^k/k!$, if we agree that $\binom{-1}{0} = 1$. J. Riordan (Ref. 7, p. 120) remarks that these polynomials are closely related to the Poisson-Charlier polynomials $C_n(x,a)$; in fact

$$C_R^k(a) = C_k(-R,a).$$

E. Brockmeyer,[8] N. Bech,[9] and K. Lundkvist[10] consider a problem which differs from Kosten's only in that M is finite ($G_M = 0$). Brockmeyer finds

$$P(k,K) = \sum_{S=0}^{M-K} (-1)^S Y_{S+K} \binom{S+K}{K} C_{K+S}^{k-S}(a) \tag{4}$$

where

$$Y_S = \sum_{J=S}^{M} (-1)^{J-S} \binom{J-1}{S-1} a_J \qquad (S = 1, 2, \cdots, M)$$

$$Y_0 = \frac{1}{C_1^{m+M}(a)}$$

$$a_J = \frac{1}{C_1^{m+M}(a)} \cdot \frac{1}{C_J^m(a)} \sum_{L=J}^{M} \binom{L-1}{J-1} C_0^{m+L}(a).$$

We do not consider here more complicated trunking situations (graded multiples, alternate routing arrangements in which the overflow group is at the same time the first-choice group for other sources of traffic). See, however, Wilkinson,[6] and R. Syski (Ref. 11, chapters 7, 8, 10).

Takács[2] gives, for arbitrary q_k, methods of finding $P(k,n)$, $P(k)$, $P(k,t)$, and $P^*(k)$. Thus in what follows we shall take the attitude that everything we need concerning the first-choice group only is, in principle, known.

1.4 *An Example*

This paper grew out of the following problem, in which both balking and overflow are involved. Subscriber lines are connected to the m trunks of the first-choice group in such a way that each line has access to only γ of them. We refer to a particular set of γ trunks as the access pattern for a particular line or group of lines. Equal traffic is submitted to each of the $\binom{m}{\gamma}$ possible access patterns. When a connection is made, any idle trunk in the subscriber's access pattern is equally likely to be selected. This arrangement is referred to as a *random slip*, or *Erlang's ideal grade*. It is easy to see that the balking probabilities are

$$q_k = 0, \quad \text{for} \quad 0 \leqq k < \gamma, \quad \text{and}$$

$$q_k = \frac{\binom{k}{\gamma}}{\binom{m}{\gamma}}, \qquad \text{for} \quad \gamma \leqq k \leqq m.$$

Traffic which balks on the first-choice group is submitted to a full-access overflow group of M trunks. If a call is blocked on the overflow group, it is lost.

Such an arrangement may be economically desirable. The average traffic carried per trunk (for a given blocking probability, B) is less than for a full-access group of $m + M$ trunks, but the traffic per crosspoint is greater. Knowing the costs of trunks and of crosspoints, and given $m + M$ and the desired value of B, one wishes to select γ and m so as to minimize the cost per unit of carried traffic. We shall give some numerical results for this arrangment.

II. THE STATE OF THE SYSTEM AT OVERFLOW INSTANTS

2.1 *Transient Behaviour*

Unless the first-choice group is full-access, the overflow process $\{T_N\}$ is not recurrent and the sequence $\{\Xi_N{}^o\}$ is not a Markov chain. However, the sequence of pairs of random variables $\{\xi_N{}^o,\ \Xi_N{}^o\}$ is a homogeneous Markov chain. This may be seen as follows. Suppose we know that $\xi(T_N-) = k$ and $\Xi(T_N-) = K$. T_N is an arrival instant; because the arrival process is recurrent and independent of the holding times, the history of the system before T_N has no effect on the epochs of future arrivals. T_N is an overflow instant; whether or not the overflowing call is accepted by the overflow group depends only on the value of K. Because of the exponential distribution of holding times, the stochastic behaviour of the system after T_N is independent of the ages of calls in progress at T_N. Thus the values of $\xi(T_N-)$ and $\Xi(T_N-)$ determine the whole future stochastic behaviour of the system. Therefore we are led first to a consideration of the probabilities $P^o(k,K,N)$.

If $\xi(t) = k$, $\Xi(t) = K$, then we say that at time t the system is in the state (k,K). The values of $\xi_N{}^o$ are limited to those k for which $q_k > 0$. We denote the set of such integers k by $\mathcal{A}$. As initial conditions we take $\xi(0+) = i$, $\Xi(0+) = I < \infty$. (It is not required that $i \in \mathcal{A}$.) Under these initial conditions, we seek $P^o(k,K,N)$ for $k \in \mathcal{A}$; $K = 0, 1, 2, \cdots$; $N = 1, 2, 3, \cdots$.

Let us now define the following quantities:

$$\begin{aligned} G_{jk}(x) &= P\{\xi_{N+1}{}^o = k,\ \Theta_{N+1} \leqq x \mid \xi(T_N+) = j\} \\ &= P\{\xi_{N+1}{}^o = k,\ \Theta_{N+1} \leqq x \mid \xi_N = j\} \\ &= P\{\xi_1{}^o = k,\ T_1 \leqq x \mid \xi(0+) = j\} \end{aligned}$$

with Laplace-Stieltjes transform

$$\gamma_{jk}(s) = \int_0^\infty e^{-sx}\, dG_{jk}\ (x)$$

$$U^o(k,R,N) = \sum_{K=R}^{M} \binom{K}{R} P^o(k,K,N) \qquad (R = 0, 1, \cdots, M)$$

$$V^o(k,R,N) = \sum_{K=R}^{M} \binom{K}{R} G_K P^o(k,K,N) \qquad (R = 0, 1, \cdots, M)$$

$$V^o(k,-1,N) = 0.$$

We may now state:

Theorem 1: The distribution $P^o(k,K,N)$ is uniquely determined by the binomial moments $U^o(k,R,N)$; the latter are determined by

$$U^o(k,R,1) = \binom{I}{R} \gamma_{ik}(R) \tag{5}$$

$$U^o(k,R,N+1) = \sum_{j \in \mathcal{A}} \gamma_{jk}(R)[U^o(j,R,N) + V^o(j,R-1,N)]. \tag{6}$$

Proof: The transition probabilities for the homogeneous Markov chain $\{\xi_N{}^o, \Xi_N{}^o\}$ are given by

$$\begin{aligned} p^o(j,J;k,K) &= P\{\xi_{N+1}{}^o = k,\ \Xi_{N+1}{}^o = K \mid \xi_N{}^o = j,\ \Xi_N{}^o = J\} \\ &= \int_0^\infty P\ \{\Xi_{N+1}{}^o = K \mid \Xi_N{}^o = J,\, \Theta_{N+1} = x\}\ dG_{jk}(x). \end{aligned}$$

It is easy to see that

$$\begin{aligned} P\{\,\Xi_{N+1}{}^o &= K \mid \Xi_N{}^o = J,\ \Theta_{N+1} = x\} \\ &= G_J \binom{J+1}{K} e^{-xK}(1 - e^{-x})^{J+1-K} \\ &\qquad + H_J \binom{J}{K} e^{-xK}(1 - e^{-x})^{J-K}. \end{aligned}$$

Thus

$$\begin{aligned} p^o(j, J; k, K) = \int_0^\infty dG_{jk}(x) \Big[G_J \binom{J+1}{K} e^{-xK}(1 - e^{-x})^{J+1-K} \\ + H_J \binom{J}{K} e^{-xK}(1 - e^{-x})^{J-K} \Big]. \end{aligned} \tag{7}$$

Now

$$P^o(k,K,N+1) = \sum_{j=0}^{m} \sum_{J=0}^{M} p^o(j,J;k,K) P^o(j,J,N). \tag{8}$$

Substituting (7) in (8), and taking the Rth binomial moment with respect to the overflow group, we obtain

$$\begin{aligned} U^o(k, R, N+1) &= \sum_{j \in \mathfrak{a}} \sum_{J=0}^{M} \int_0^\infty dG_{jk}(x) \Big[G_J \binom{J+1}{R} \\ &\qquad + H_J \binom{J}{R} \Big] e^{-xR}\ P^o(j, J, N) \\ &= \sum_{j \in \mathfrak{a}} \sum_{J=0}^{M} \gamma_{jk}(R) \left[\binom{J}{R} + G_J \binom{J}{R-1} \right] P^o(j, J, N) \\ &= \sum_{j \in \mathfrak{a}} \gamma_{jk}(R)\ [U^o(j, R, N) + V^o(j, R-1, N)], \end{aligned}$$

which is (6).

For $N = 1$, we have

$$P^o(k, K, 1) = \int_0^\infty dG_{ik}(x) \binom{I}{K} e^{-xK}(1 - e^{-x})^{I-K}$$

so that

$$U^o(k, R, 1) = \int_0^\infty dG_{ik}(x) \binom{I}{R} e^{-Rx} = \binom{I}{R} \gamma_{ik}(R),$$

which is (5).

From the definition of $U^o(k,R,N)$, we have

$$\begin{aligned}\sum_{R=K}^{M} (-1)^{R-K} \binom{R}{K} U^o(k, R, N) \\ = \sum_{R=K}^{M} (-1)^{R-K} \binom{R}{K} \sum_{J=R}^{M} \binom{J}{R} P^o(k, J, N).\end{aligned} \tag{9}$$

Now, for any finite N the double series on the right contains a finite number of terms, even if $M = \infty$. This is so because

$$P^o(k,J,N) = 0 \quad \text{for} \quad k + J \geqq i + I + N,$$

and we have assumed $I < \infty$.

Thus the double series can be rearranged, and one obtains readily that the binomial moments determine the probabilities according to

$$P^o(k, K, N) = \sum_{R=K}^{M} (-1)^{R-K} \binom{R}{K} U^o(k, R, N). \tag{10}$$

In (5) and (6), the quantities $\gamma_{jk}(R)$ occur as coefficients. We regard these coefficients as known because they can be expressed in terms of certain quantities determined by Takács.[2] Let

$$M_{ik}(x) = \mathbf{E}\,\{\text{number of } \tau_n \text{ in } (o,x] \text{ for which } \xi_n = k \mid \xi(0+) = i\},$$

with Laplace-Stieltjes transform

$$\mu_{ik}(s) = \int_0^\infty e^{-sx}\, dM_{ik}(x).$$

Takács gives a method for finding the $\mu_{ik}(s)$ [(T70), in which, however, the index i is implicit]. The way in which the quantities $\mu_{jk}(R)$ determine the $\gamma_{jk}(R)$ is expressed in the following lemma (in which, it is to be noted, values of the indices j,k, etc. are no longer restricted to the set $\mathcal{A}$).

Lemma 1: Define $M_{ik}^o(x) = \mathbf{E}$ {*number of* T_N *in* $(0,x]$ *for which* $\xi_N^o =$

$k \mid \xi(0+) = i\}$, with Laplace-Stieltjes transform

$$\mu_{ik}^{\circ}(s) = \int_0^{\infty} e^{-sx}\, dM_{ik}^{\circ}(x).$$

Let $\mu^{\circ,R}$ be the square matrix with elements $\mu_{jk}^{\circ}(R)$; $j,k = 0, 1, \cdots, m$.
Let γ^R be the square matrix with elements $\gamma_{jk}(R)$; $j,k = 0, 1, \cdots, m$.
Then, for $R = 1, 2, \cdots$,

$$\gamma^R = \mu^{\circ,R}(E + \mu^{\circ,R})^{-1} \tag{11}$$

where E is the $(m + 1)$ by $(m + 1)$ unit matrix.

Since, obviously

$$\mu_{jk}^{\circ}(R) = q_k \mu_{jk}(R), \tag{12}$$

(11) provides the desired relation between the $\gamma_{jk}(R)$ and the $\mu_{jk}(R)$.

Proof: We shall first show that

$$\mu_{jk}^{\circ}(R) = \gamma_{jk}(R) + \sum_{l=0}^{m} \gamma_{jl}(R)\mu_{lk}^{\circ}(R) \tag{13}$$

for $R = 1, 2, \cdots$.

Suppose $\xi(0+) = j$, and consider a given R-tuple of trunks on the overflow group which are all busy at $t = 0+$. If $T_1 = x$, the probability that the overflow at T_1 will find this R-tuple still busy is e^{-Rx}.

Thus

$$\gamma_{jk}(R) = \int_0^{\infty} e^{-Rx}\, dG_{jk}(x)$$

is the probability that this R-tuple is still busy at T_1 *and* that $\xi(T_1-) = k$.

Again, if this R-tuple remains busy just until $t = x$, the expected number of overflows *from* k to find it busy is $M_{jk}^{\circ}(x)$. Therefore the unconditional expectation of the number of overflows from k to find it busy is

$$\int_0^{\infty} M_{jk}^{\circ}(x)\, d(1 - e^{-Rx}) = \int_0^{\infty} e^{-Rx}\, dM_{jk}^{\circ}(x) = \mu_{jk}^{\circ}(R).$$

Denote (temporarily) by $[\mu_{jk}^{\circ}(R) \mid l]$ the expected number of overflows from k to find this R-tuple still busy, on the condition that $\xi(T_1-) = l$ and the R-tuple is still busy at $t = T_1-$.

Then, by the principle of total expectation,

$$\mu_{jk}^{\circ}(R) = \sum_{l=0}^{m} [\mu_{jk}^{\circ}(R) \mid l]\gamma_{jl}(R). \tag{14}$$

Now because of the exponential holding-time distribution

$$[\mu_{jk}{}^{o}(R) \mid l] = \mu_{lk}{}^{o}(R) \quad \text{for} \quad l \neq k \tag{15}$$

and

$$[\mu_{jk}{}^{o}(R) \mid k] = 1 + \mu_{kk}{}^{o}(R). \tag{16}$$

Substituting (15) and (16) into (14), we obtain (13). Equation (13) may be written

$$\mu^{o,R} = \gamma^{R} + \gamma^{R}\mu^{o,R}. \tag{17}$$

Thus, to prove the lemma, it remains to show that $(E + \mu^{o,R})$ is nonsingular.

From (17)

$$(E - \gamma^{R})\mu^{o,R} = \gamma^{R}.$$

Therefore

$$(E - \gamma^{R})\cdot(E + \mu^{o,R}) = E$$

$$\det\,(E - \gamma^{R}) \cdot \det\,(E + \mu^{o,R}) = 1.$$

Since clearly both $\det\,(E - \gamma^{R})$ and $\det\,(E + \mu^{o,R})$ are finite (for $R > 0$), it follows that $\det\,(E - \gamma^{R}) \neq 0$ and $\det\,(E + \mu^{o,R}) \neq 0$, which completes the proof of the lemma.

We note, for later use, that we have also shown that

$$\mu^{o,R} = (E - \gamma^{R})^{-1}\gamma^{R}. \tag{18}$$

We need a separate method for finding $\gamma_{jk}(0)$, the above argument being invalid because $\mu_{jk}{}^{o}(0) = \infty$ for all $k \in \mathfrak{A}$.

We notice that $\gamma_{jk}(0) = G_{jk}(\infty) = P\{\xi(T_1-) = k \mid \xi(0+) = j\}$.

The quantities $\gamma_{jk}(0)$ are determined by the following system of equations:

$$\begin{aligned}\gamma_{jk}(0) = q_k \int_0^\infty dF(x) \binom{j}{k} e^{-kx}(1 - e^{-x})^{j-k} + \sum_{l=0}^{m} p_l\, \gamma_{l+1,k}(0)\cdot \\ \cdot \int_0^\infty dF(x) \binom{j}{l} e^{-lx}(1 - e^{-x})^{j-l} \qquad (j, k = 0, 1, \cdots, m).\end{aligned} \tag{19}$$

This may be seen as follows:

The event $\{\xi(T_1-) = k\}$ can occur in these mutually exclusive ways:

(i) the first arrival after $t = 0$ encounters k busy trunks on the first-choice group, with probability

$$\int_0^\infty dF(x) \binom{j}{k} e^{-kx}(1 - e^{-x})^{j-k},$$

and overflows, with probability q_k;

(*ii*) the first arrival after $t = 0$ encounters l busy trunks and does not overflow [so that $\xi(T_1+) = l + 1$]; the next overflow following this occurrence is from k [probability $\gamma_{l+1,k}(0)$].

For each k, (19) is a set of linear equations in the $\gamma_{jk}(0)$. These equations determine the $\gamma_{jk}(0)$ uniquely if the coefficient matrix is nonsingular (for each k). Call this matrix $A^{(k)}$. If we can show that $| A_{jj}^{(k)} | > \sum_{l \neq j} A_{jl}^{(k)}$ for each j, it will follow from the theorem of Lévy-Hadamard-Gerschgorin (Ref. 12, p. 79) that det $A^{(k)} \neq 0$. That is, we want to show that

$$\sum_{l=0}^{m} p_l \int_0^\infty dF(x) \binom{j}{l} e^{-lx}(1 - e^{-x})^{j-l} < 1. \tag{20}$$

The left side of (20) is evidently strictly less than

$$\sum_{l=0}^{m} \int_0^\infty dF(x) \binom{j}{l} e^{-lx}(1 - e^{-x})^{j-l} = 1, \quad \text{for each } j, \text{ Q.E.D.}$$

Equations (5) and (6) may be solved, in some cases, by means of generating functions.

Let

$$U^o(k,R,w) = \sum_{N=1}^{\infty} U^o(k,R,N)w^N$$

$$V^o(k,R,w) = \sum_{N=1}^{\infty} V^o(k,R,N)w^N$$

Note that it follows from (10) that

$$\sum_{N=1}^{\infty} P^o(k,K,N)w^N = \sum_{R=K}^{M} (-1)^{R-K} \binom{R}{K} U^o(k,R,w). \tag{21}$$

From (5) and (6) we obtain

$$U^o(k,R,w) = w\left\{\binom{I}{R}\gamma_{ik}(R) + \sum_{j \in \alpha} \gamma_{jk}(R)[U^o(j,R,w) + V^o(j,R-1,w)]\right\}. \tag{22}$$

We illustrate the use of (22) by a simple example.

Example 1:

If the first-choice group is full-access (the only element of $\mathfrak{A}$ is m), then $U^o(k,R,N)$ and $V^o(k,R,N)$ vanish except for $k = m$. For simplicity, we assume that $i = m$; then the only relevant element of the matrix γ^R is $\gamma_{mm}(R)$, and (22) becomes:

$$U^o(m,R,w) = w\,\gamma_{mm}(R)\left[\binom{I}{R} + U^o(m,R,w) + V^o(m,R-1,w)\right],$$

whence

$$U^o(m, R, w) = \frac{w\,\gamma_{mm}(R)}{1 - w\,\gamma_{mm}(R)}\left[\binom{I}{R} + V^o(m, R - 1, w)\right]. \qquad (23)$$

$\gamma_{mm}(s)$ is the Laplace-Stieltjes transform of the interoverflow-time distribution, i.e., it is just the function $\gamma(s)$ given by (1). Thus (23) is exactly equivalent to (T32), and merely serves to illustrate our remark (Section 1.1) that if the first-choice group is full-access, we can use the methods of Ref. 2 to describe the behaviour of the sequence $\{\Xi_N{}^o\}$.

2.2 *The Limiting Distribution* $P^o(k,K)$

Theorem 2: The quantities $P^o(k,K) = \lim_{N\to\infty} P^o(k,K,N)$ *exist, are strictly positive, form a probability distribution independent of the initial state, and are uniquely determined by the binomial moments* $U^o(k,R) = \sum_{K=R}^{M}\binom{K}{R}P^o(k,K)$*; the latter are determined by*

$$U^o(k,R) = q_k \sum_{j\in\mathfrak{A}} \mu_{jk}(R)V^o(j,R-1) \qquad (R = 1, 2, \cdots, M) \quad (24)$$

and

$$U^o(k, 0) = \frac{q_kP(k)}{b} \qquad (25)$$

where

$$V^o(k,R) = \sum_{K=R}^{M}\binom{K}{R} G_KP^o(k,K).$$

Proof: We first show the existence of the limiting distribution.

In this section, we use theorems given in Feller,[13] chapter 15, sections 5 and 6.

The Markov chain $\{\xi_N{}^o, \Xi_N{}^o\}$ is evidently irreducible (since $p_k > 0$ for $k < m$) and aperiodic. Therefore $\lim_{N\to\infty} P^o(k,K,N)$ exists. Since it is irreducible, the chain has either all transient, all recurrent null, or all recurrent non-null states.

If a state (k,K) is transient or recurrent null, then $\lim_{N\to\infty} P(k,K,N) = 0$.

Therefore, to show that all states are recurrent non-null it will suffice to show that for *some* state (k,K), $\lim_{N\to\infty} P^o(k,K,N) > 0$. It will then follow that this is so for all states, and that $\sum_{k\in\mathfrak{a}} P^o(k,K) = 1$. We look at the state $(0,0)$:

To see that $\lim_{N\to\infty} P^o(0,0,N) > 0$, we compare our system (with arbitrary balking probabilities) to the special system for which $m = 0$, $M = \infty$, $H_K = 0$ (always assuming the same input process). For this special system, write $P\{\Xi_N{}^o = K\} = \tilde{P}^o(K,N)$, and take as initial condition: $\Xi(0+) = i + I$.

It is clear that for any system with $M = \infty$, and with the same initial condition,

$$P^o(0,0,N) \geqq \tilde{P}^o(0,N),$$

for each N, whence

$$\lim_{N\to\infty} P^o(0,0,N) \geqq \lim_{N\to\infty} \tilde{P}^o(0,N).$$

But it is known[3] that $\lim_{N\to\infty} \tilde{P}^o(0,N) > 0$; thus

$$\lim_{N\to\infty} P^o(0,0,N) = P^o(0,0) > 0$$

and all states are recurrent non-null. Hence, since the chain is also irreducible and aperiodic, it is ergodic.

We now know also that a unique stationary distribution exists and that it coincides with the limiting distribution. From (6), we must have

$$U^o(k,R) = \sum_{j\in\mathfrak{a}} \gamma_{jk}(R)[U^o(j,R) + V^o(j,R-1)]. \tag{26}$$

Denote by $U^{o,R}$ the row-vector with components $U^o(k,R)$, $0 \leqq k \leqq m$.

Then (26) may be written

$$U^{o,R} = (U^{o,R} + V^{o,R-1})\gamma^R.$$

Thus, from (18),

$$U^{o,R} = V^{o,R-1}\mu^{o,R}. \tag{27}$$

Writing out (27) in components, and using (12), we obtain (24).

We now prove (25). Denote by $C^{(n)}$ the event that the nth arrival overflows. Thus,

$$b = \lim_{n\to\infty} P\{C^{(n)}\}.$$

Now,

$$P^o(k,K) = \lim_{N\to\infty} P\{\xi_N{}^o = k,\ \Xi_N{}^o = K\} = \lim_{n\to\infty} P\{\xi_n = k,\ \Xi_n = K \mid C^{(n)}\}$$

$$= \lim_{n\to\infty} \frac{P\{\xi_n = k, \Xi_n = K\}\ P\ \{C^{(n)} \mid \xi_n = k, \Xi_n = K\}}{P\{C^{(n)}\}}.$$

But

$$P\{C^{(n)} \mid \xi_n = k,\ \Xi_n = K\} = P\{C^{(n)} \mid \xi_n = k\} = q_k .$$

Therefore

$$P^o(k, K) = \frac{q_k\, P(k, K)}{b} \tag{28}$$

and

$$U^o(k, 0) = \sum_{K=0}^{M} P^o(k, K) = \frac{q_k \sum_{K=0}^{M} P(k, K)}{b}$$

$$= \frac{q_k\ P(k)}{b}, \text{ Q.E.D.}$$

To complete the proof of Theorem 2, it remains to show that the binomial moments $U^o(k,R)$ uniquely determine the probabilities $P^o(k,K)$. This proof will be easier after we have discussed the stationary distribution at arrival moments, $P(k,K)$, and we therefore defer it until then.

It is sometimes convenient to work with the double binomial moments

$$B^o(r, R) = \sum_{k=r}^{m} \binom{k}{r} U^o(k, R)$$

$$C^o(r, R) = \sum_{k=r}^{m} \binom{k}{r} V^o(k,R).$$

In terms of these, (24) and (25) of Theorem 2 become

$$B^o(r,R) = \sum_{j=0}^{m} [f_{jr}(R) - g_{jr}(R)]C^o(j,R-1) \qquad (29)$$

$$(R = 1, 2, \cdots, M)$$

$$B^o(r, 0) = \frac{1}{b}\sum_{k=r}^{m} \binom{k}{r} q_k\, P(k). \qquad (30)$$

Here we have used the following definitions: $f_{lr}(s)$ and $g_{lr}(s)$ are the lth differences of $\Phi_{0r}(s)$ and $\Psi_{0r}(s)$:

$$f_{lr}(s) = \sum_{j=0}^{l} (-1)^{l-j} \binom{l}{j} \Phi_{jr}(s) \qquad (31)$$

$$g_{lr}(s) = \sum_{j=0}^{l} (-1)^{l-j} \binom{l}{j} \Psi_{jr}(s) \qquad (32)$$

where $\Phi_{jr}(s)$ and $\Psi_{jr}(s)$ are defined, following Takács [(T59), (T60)], by

$$\Phi_{jr}(s) = \sum_{k=r}^{m} \binom{k}{r} \mu_{jk}(s) \qquad (33)$$

$$\Psi_{jr}(s) = \sum_{k=r}^{m} \binom{k}{r} p_k\mu_{jk}(s) \qquad (34)$$

and must satisfy [(T61) and (T62)]

$$\Phi_{j0}(s) = \frac{\varphi(s)}{1 - \varphi(s)} \qquad (35)$$

and

$$\frac{\Phi_{jr}(s)}{C_r(s)} = \frac{1}{C_{r-1}(s)}\left[\binom{j}{r} + \Psi_{j,r-1}(s)\right] \qquad (36)$$

as well as the relations in r implied by their definitions [see (T25)],

$$\Psi_{jr}(s) = \sum_{l=r}^{m} \binom{l}{r} (\Delta^{l-r} p_r)\Phi_{jl}(s). \qquad (37)$$

Examples of the application of the methods of this section will be found in Section V.

III. THE STATE OF THE SYSTEM AT ARRIVAL INSTANTS

3.1 *Transient Behaviour*

The sequence $\{\xi_n, \Xi_n\}$ is clearly a homogeneous Markov chain. We assume initial conditions $\xi(0+) = i$, $\Xi(0+) = I$, and seek the distribution $P(k,K,n)$. We no longer restrict our attention to states (k,K) for which $q_k > 0$, but consider all states (k,K), $0 \leqq k \leqq m \leqq \infty$, $0 \leqq K \leqq M \leqq \infty$.

We shall prove the following:

Theorem 3: The distribution $P(k,K,n)$ is uniquely determined by the double binomial moments

$$B(r,R,n) = \sum_{k=r}^{m} \sum_{K=R}^{M} \binom{k}{r} \binom{K}{R} P(k,K,n);$$

the latter are determined by

$$B(r,R,1) = \varphi_{r+R} \binom{i}{r} \binom{I}{R} \tag{38}$$

$$(r = 0, 1, \cdots, m; R = 0, 1, \cdots, M)$$

$$\begin{aligned} B(r,R,n+1) = \varphi_{r+R}[&B(r,R,n) + D(r-1,R,n) \\ &+ C(r,R-1,n) - E(r,R-1,n)] \end{aligned} \tag{39}$$

$$(r = 0, 1, \cdots, m; R = 0, 1, \cdots, M; n = 1, 2, \cdots).$$

Here

$$C(r,R,n) = \sum_{k=r}^{m} \sum_{K=R}^{M} \binom{k}{r} \binom{K}{R} G_K P(k,K,n)$$

$$D(r,R,n) = \sum_{k=r}^{m} \sum_{K=R}^{M} \binom{k}{r} \binom{K}{R} p_k P(k,K,n)$$

$$E(r,R,n) = \sum_{k=r}^{m} \sum_{K=R}^{M} \binom{k}{r} \binom{K}{R} p_k G_K P(k,K,n)$$

and all these quantities are defined to be zero if $r < 0$ or $R < 0$.

Proof: If the arrival at τ_n finds the system in the state (j,J), it may either get on the first-choice group, with probability p_j, or balk on the first-choice group with probability q_j; in the latter case, it may get on the overflow group, with probability G_J, or balk there too, with probability H_J. Thus the transition probabilities are given by

$$\begin{aligned} p(j,J;k,K) &= P\{\xi_{n+1} = k,\ \Xi_{n+1} = K \mid \xi_n = j,\ \Xi_n = J\} \\ = \int_0^\infty dF(x) \Bigg\{ & p_j \binom{j+1}{k} e^{-xk}(1-e^{-x})^{j+1-k} \binom{J}{K} e^{-xK}(1-e^{-x})^{J-K} \\ & + q_j \binom{j}{k} e^{-xk}(1-e^{-x})^{j-k} \Bigg[G_J \binom{J+1}{K} e^{-xK}(1-e^{-x})^{J+1-K} \\ & + H_J \binom{J}{K} e^{-xK}(1-e^{-x})^{J-K} \Bigg] \Bigg\}. \end{aligned} \tag{40}$$

Now

$$P(k,K,n+1) = \sum_{j=0}^{m} \sum_{J=0}^{M} p(j,J;k,K)\, P(j,J,n). \tag{41}$$

Substituting (40) in (41) and taking binomial moments with respect to both the first-choice and overflow groups, we obtain:

$$\begin{aligned} B(r,R,n+1) &= \varphi_{r+R} \sum_{j=0}^{m} \sum_{J=0}^{M} \left\{ p_j \binom{j+1}{r} \binom{J}{R} \right. \\ &\quad \left. + q_j \binom{j}{r} \left[G_J \binom{J+1}{R} + H_J \binom{J}{R} \right] \right\} P(j,J,n). \end{aligned} \tag{42}$$

Note that the quantity in braces in (42) is

$$\left\{ \binom{j}{r}\binom{J}{R} + p_j \binom{j}{r-1}\binom{J}{R} + q_j G_J \binom{j}{r}\binom{J}{R-1} \right\}. \tag{43}$$

Substituting (43) in (42), we obtain (39).
For $n = 1$, we have

$$P(k,K,1) = \int_0^\infty dF(x) \binom{i}{k} e^{-xk}(1-e^{-x})^{i-k} \binom{I}{K} e^{-xK}(1-e^{-x})^{I-K};$$

taking binomial moments with respect to both trunk groups, we obtain (38).

From the double binomial moments, one obtains the probabilities $P(k,K,n)$ by using:

$$U(k,R,n) = \sum_{r=k}^{m} (-1)^{r-k} \binom{r}{k} B(r,R,n) \tag{44}$$

and

$$P(k,K,n) = \sum_{R=K}^{M} (-1)^{R-K} \binom{R}{K} U(k,R,n). \tag{45}$$

Clearly $P(k,K,n) = 0$ for $k + K \geqq i + I + n$; it follows that the sums in (44) and (45) contain a finite number of terms for finite n, even if $M = \infty$, and there are no problems about convergence.

Equations (38) and (39) may be solved, in some cases, by means of generating functions; we give an example.

Example 2:

We consider the simplest possible case, in which

$$q_k = 0 \qquad (k = 0, 1, \cdots, m-1)$$

$$q_m = 1$$

$$M = \infty$$

$$H_K = 0 \qquad (K = 0, 1, 2, \cdots).$$

In this case,

$$C(r,R,n) = B(r,R,n), \tag{46}$$

$$E(r,R,n) = D(r,R,n), \tag{47}$$

and

$$D(r,R,n) = B(r,R,n) - \binom{m}{r} B(m,R,n). \tag{48}$$

Substituting (46), (47), and (48) in (39), we get

$$\begin{aligned} B(r,R,n+1) = \varphi_{r+R}[&B(r,R,n) + B(r-1,R,n) \\ &-\binom{m}{r-1} B(m,R,n) + \binom{m}{r} B(m,R-1,n)]. \end{aligned} \tag{49}$$

Let

$$B(r,R,w) = \sum_{n=1}^{\infty} B(r,R,n)w^n.$$

From (38) and (49):

$$\begin{aligned} B(r,R,w) = \frac{w\varphi_{r+R}}{1 - w\varphi_{r+R}}\Bigg[&\binom{i}{r}\binom{I}{R} + B(r-1,R,w) \\ &-\binom{m}{r-1} B(m,R,w) + \binom{m}{r} B(m,R-1,w)\Bigg]. \end{aligned} \tag{50}$$

The solution of (50) is

$$\begin{aligned} B(r,R,w) = \Gamma_{r+R}(w)\Bigg\{&\frac{\sum_{j=r}^{m}\binom{m}{j}\frac{1}{\Gamma_{j+R}(w)}}{\sum_{j=0}^{m}\binom{m}{j}\frac{1}{\Gamma_{j+R}(w)}}\cdot\sum_{S=0}^{R}\binom{I}{S}\sum_{j=0}^{i}\binom{i}{j}\frac{1}{\Gamma_{j+S-1}(w)} \\ &-\frac{\sum_{j=r+1}^{m}\binom{m}{j}\frac{1}{\Gamma_{j+R-1}(w)}}{\sum_{j=0}^{m}\binom{m}{j}\frac{1}{\Gamma_{j+R-1}(w)}}\cdot\sum_{S=0}^{R-1}\binom{I}{S}\sum_{j=0}^{i}\binom{i}{j}\frac{1}{\Gamma_{j+S-1}(w)} \\ &-\binom{I}{R}\sum_{j=r+1}^{m}\binom{i}{j}\frac{1}{\Gamma_{j+R-1}(w)}\Bigg\} \end{aligned}$$

where we have defined

$$\Gamma_r(w) = \prod_{j=0}^{r} \frac{w\varphi_j}{1 - w\varphi_j}, \qquad (r = 0, 1, 2, \cdots)$$

$$\Gamma_{-1}(w) \equiv 1.$$

3.2 *The Limiting Distribution $P(k,K)$*

Theorem 4: The quantities $P(k,K) = \lim_{n\to\infty} P(k,K,n)$ exist, are strictly positive, form a probability distribution independent of the initial state, and are uniquely determined by the double binomial moments $B(r,R) = \sum_{k=r}^{m} \binom{k}{r} U(k,R)$, where $U(k,R) = \sum_{K=R}^{M} \binom{K}{R} P(k,K)$; the $B(r,R)$ are given by

$$B(r,R) = bC_{r+R}\left[\sum_{j=r}^{m} \frac{B^{\circ}(j,R)}{C_{j+R}} - \sum_{j=r+1}^{m} \frac{C^{\circ}(j,R-1)}{C_{j+R-1}}\right] \tag{51}$$

$$(r = 0, 1, \cdots, m;\ R = 0, 1, \cdots, M).$$

Here

$$C^{\circ}(r,R) = \sum_{k=r}^{m} \binom{k}{r}\binom{K}{R} G_K P^{\circ}(k,K).$$

Proof: That the limits $P(k,K)$ exist and are independent of the initial state again follows from the fact that the Markov chain $\{\xi_n, \Xi_n\}$ $(n = 1, 2, \cdots)$ is irreducible and aperiodic. To show that the $P(k,K)$ are strictly positive and form a probability distribution, we must show that there exists some state (k,K) such that $P(k,K) > 0$. This can be done by a method similar to that used in the proof of Theorem 2; we omit the argument. It follows that a unique stationary distribution exists and that it coincides with the limiting distribution. We express this stationary distribution in terms of the stationary distribution $P^{\circ}(k,K)$ in the following way:

Consider the arrival which occurs at τ_n (under equilibrium conditions).

It either overflows, with probability b, or does not, with probability $(1 - b)$.

If it overflows, the probability that it encountered the state (j,J) is $P^{\circ}(j,J)$.

If it does not overflow, let us denote the probability that it encountered the state (j,J) by $P^{\emptyset}(j,J)$.

We note that

$$P(j,J) = bP^{\circ}(j,J) + (1 - b)P^{\emptyset}(j,J). \tag{52}$$

Suppose that $\theta_{n+1} = x$.

If the arrival at τ_n encountered the state (j,J) and overflowed, the probability that the arrival at τ_{n+1} encounters the state (k,K) is:

$$\binom{j}{k} e^{-xk}(1 - e^{-x})^{j-k}\Big[G_J\binom{J+1}{K} e^{-xK}(1 - e^{-x})^{J+1-K} + H_J\binom{J}{K} e^{-xK}(1 - e^{-x})^{J-K}\Big] = \alpha(x), \text{ say.} \tag{53}$$

If the arrival at τ_n encountered the state (j,J) and did not overflow, the probability that the arrival at τ_{n+1} encounters the state (k,K) is:

$$\binom{j+1}{k} e^{-xk}(1 - e^{-x})^{j+1-k}\binom{J}{K} e^{-xK}(1 - e^{-x})^{J-K} = \beta(x), \text{ say.} \tag{54}$$

Taking account of both these possibilities, and removing the condition on θ_{n+1},

$$P(k,K) = \sum_{j=0}^{m}\sum_{J=0}^{M}\int_0^\infty dF(x)[bP^{o}(j,J)\alpha(x) + (1 - b)P^{\emptyset}(j,J)\beta(x)].$$

Using (52),

$$P(k,K) = \sum_{j=0}^{m}\sum_{J=0}^{M}\int_0^\infty dF(x)\{bP^{o}(j,J)[\alpha(x) - \beta(x)] + P(j,J)\beta(x)\}.$$

Taking binomial moments with respect to both trunk groups, and using (53) and (54),

$$\begin{aligned} B(r,R) &= \varphi_{r+R}\sum_{j=0}^{m}\sum_{J=0}^{M}\left\{bP^{o}(j,J)\left[\binom{j}{r}\left(G_J\binom{J+1}{R} + H_J\binom{J}{R}\right)\right.\right. \\ &\qquad \left.\left. - \binom{j+1}{r}\binom{J}{R}\right] + P(j,J)\binom{j+1}{r}\binom{J}{R}\right\} \\ &= \varphi_{r+R}\{B(r,R) + B(r - 1,R) \\ &\qquad + b[C^{o}(r,R - 1) - B^{o}(r - 1,R)]\}. \end{aligned} \tag{55}$$

The solution of (55) is

$$\frac{B(r,R)}{C_{r+R}} = \frac{B(m,R)}{C_{m+R}} + b\left[\sum_{j=r}^{m-1}\frac{B^{o}(j,R)}{C_{j+R}} - \sum_{j=r+1}^{m}\frac{C^{o}(j,R - 1)}{C_{j+R-1}}\right]. \tag{56}$$

Now note that, from (28),

$$bB^{o}(m,R) = B(m,R). \tag{57}$$

Substituting (57) in (56), we obtain (51).

To complete the proof of Theorem 4, it remains to show that the double binomial moments $B(r,R)$ uniquely determine the probabilities $P(k,K)$. It is clear that the $B(r,R)$ uniquely determine the $U(k,R)$ through the equation

$$U(k,R) = \sum_{r=k}^{m} (-1)^{r-k} \binom{r}{k} B(r,R) \tag{58}$$

because m is finite. Thus we must show that

$$P(k,K) = \sum_{R=K}^{M} (-1)^{R-K} \binom{R}{K} U(k,R) \tag{59}$$

when M is infinite; it will suffice to show that the series on the right converges absolutely.

From (39) we have

$$B(0,R) = \frac{\varphi_R}{1 - \varphi_R} [C(0,R-1) - E(0, R-1)]. \tag{60}$$

Now,

$$C(0,R) - E(0,R) = \sum_{k=r}^{m} \sum_{K=R}^{M} \binom{k}{r} \binom{K}{R} q_k G_K P(k,K) \leqq B(0,R). \tag{61}$$

Therefore,

$$B(0,R) \leqq \frac{\varphi_R}{1 - \varphi_R} B(0,R-1). \tag{62}$$

Now

$$\lim_{R\to\infty} \varphi_R = \lim_{s\to\infty} \varphi(s) = F(0+) = 0$$

whence

$$\lim_{R\to\infty} \frac{\varphi_R}{1 - \varphi_R} = 0.$$

Thus

$$\lim_{R\to\infty} \frac{B(0,R)}{B(0,R-1)} = 0. \tag{63}$$

Equation (63) is sufficient to insure that

$$\sum_{R=K}^{M} \binom{R}{K} B(0,R)$$

converges.

Consider for simplicity the case $m = 1$. Then we have

$$B(0,R) = U(0,R) + U(1,R). \tag{64}$$

At least one of the statements

$$\lim_{R\to\infty} \frac{U(0,R)}{U(0,R-1)} = 0 \tag{65}$$

$$\lim_{R\to\infty} \frac{U(1,R)}{U(1,R-1)} = 0 \tag{66}$$

must be true, for if both failed to be true, then for some $\epsilon > 0$ there would be terms for which

$$\frac{U(0,R)}{U(0,R-1)} > \epsilon$$

$$\frac{U(1,R)}{U(1,R-1)} > \epsilon$$

for arbitrarily large R; it would follow that for arbitrarily large R

$$\frac{B(0,R)}{B(0,R-1)} = \frac{U(0,R) + U(1,R)}{U(0,R-1) + U(1,R-1)} > \epsilon$$

which contradicts (63).

Say (65) is true. Then the series

$$\sum_{R=K}^{M} \binom{R}{K} U(0,R)$$

converges; thus

$$\sum_{R=K}^{M} \binom{R}{K} U(1,R) = \sum_{R=K}^{M} \binom{R}{K} B(0,R) - \sum_{R=K}^{M} \binom{R}{K} U(0,R)$$

converges, and this proves (59) for $m = 1$. The generalization to arbitrary m is straightforward.

Corollary: We can now easily complete the proof of Theorem 2 by remarking that [using (28)]

$$\begin{aligned} bU^{\circ}(k,R) &= b\sum_{J=R}^{M} \binom{J}{R} P^{\circ}(k,J) \\ &= \sum_{J=R}^{M} \binom{J}{R} q_k P(k,J) \leqq \sum_{J=R}^{M} \binom{J}{R} P(k,J) = U(k,R) \end{aligned}$$

so that the series

$$P^{\circ}(k,K) = \sum_{R=K}^{M} (-1)^{R-K} \binom{R}{K} U^{\circ}(k,R)$$

converges absolutely, Q.E.D.

We again defer examples to Section V.

IV. THE STATE OF THE SYSTEM AT ANY TIME

4.1 *Transient Behaviour*

Let

$$B(r,R,t) = \sum_{k=r}^{m} \sum_{K=R}^{M} \binom{k}{r} \binom{K}{R} P(k,K,t)$$

with *Laplace* transform

$$\beta(r,R,s) = \int_0^{\infty} e^{-st} B(r,R,t)dt.$$

Let $M_{ik}{}^{IK}(t)$ be the expected number of arrivals in $(0,t]$ to encounter k trunks busy on the first-choice group and K on the overflow group, on the condition that $\xi(0+) = i$, $\Xi(0+) = I$, with Laplace-Stieltjes transform

$$\mu_{ik}{}^{IK}(s) = \int_0^{\infty} e^{-sx}\, dM_{ik}{}^{IK}(x).$$

We also define several kinds of double binomial moments:

$$\Phi_{ir}{}^{IR}(s) = \sum_{k=r}^{m} \sum_{K=R}^{M} \binom{k}{r} \binom{K}{R} \mu_{ik}{}^{IK}(s)$$

$$X_{ir}{}^{IR}(s) = \sum_{k=r}^{m} \sum_{K=R}^{M} \binom{k}{r} \binom{K}{R} G_K \mu_{ik}{}^{IK}(s)$$

$$\Psi_{ir}{}^{IR}(s) = \sum_{k=r}^{m} \sum_{K=R}^{M} \binom{k}{r} \binom{K}{R} p_k \mu_{ik}{}^{IK}(s)$$

$$Y_{ir}{}^{IR}(s) = \sum_{k=r}^{m} \sum_{K=R}^{M} \binom{k}{r} \binom{K}{R} p_k G_K \mu_{ik}{}^{IK}(s).$$

Theorem 5:

$$\Phi_{ir}{}^{IR}(s) = \frac{\varphi(s+r+R)}{1-\varphi(s+r+R)} \cdot\left[\binom{i}{r}\binom{I}{R} + \Psi_{i,r-1}{}^{IR}(s) + X_{ir}{}^{I,R-1}(s) - Y_{ir}{}^{I,R-1}(s)\right]. \qquad (67)$$

Proof: Consider a certain set of r first-choice trunks and a certain set of R overflow trunks. We shall call the union of these two sets an (r,R)-tuple of trunks, and if the r first-choice trunks and the R overflow trunks are all busy at time t, we shall say that this particular (r,R)-tuple of trunks is busy at time t. Thus, when the system is in the state (k,K), the number of busy (r,R)-tuples is $\binom{k}{r}\binom{K}{R}$. Let us make the convention that there is always one busy $(0,0)$-tuple. The expected number of busy (r,R)-tuples at time t is evidently $B(r,R,t)$.

Let us now calculate the expected total number of encounters between arriving calls and busy (r,R)-tuples in the interval $(0,t]$. Denote this expectation by $E_{ir}^{IR}(t)$.

If the nth arrival occurs in $(0,t]$, and if $(\xi_n = k,\ \Xi_n = K)$, then the nth arrival encounters $\binom{k}{r}\binom{K}{R}$ busy (r,R)-tuples. Thus

$$E_{ir}^{IR}(t) = \sum_{n=1}^{\infty}\sum_{k=r}^{m}\sum_{K=R}^{M}\binom{k}{r}\binom{K}{R}\int_0^{\infty} dP\{\tau_n \leqq u,\ \xi_n = k,\ \Xi_n = K\}.$$

But

$$\sum_{n=1}^{\infty} P\{\tau_n \leqq u,\ \xi_n = k,\ \Xi_n = K\} = M_{ik}^{IK}(u). \tag{68}$$

Therefore

$$E_{ir}^{IR}(t) = \sum_{k=r}^{m}\sum_{K=R}^{M}\binom{k}{r}\binom{K}{R} M_{ik}^{IK}(t)$$

with Laplace-Stieltjes transform

$$\epsilon_{ir}^{IR}(s) = \Phi_{ir}^{IR}(s). \tag{69}$$

But $\epsilon_{ir}^{IR}(s)$ can be found in another way. If $(\xi_n = k,\ \Xi_n = K)$, then at time τ_n+, the system is in the state $(k+1,K)$ with probability p_k, the state $(k,K+1)$ with probability q_kG_K, or the state (k,K) with probability q_kH_K. Thus the expected number of busy (r,R)-tuples at time τ_n+, under the stated condition, is

$$p_k\binom{k+1}{r}\binom{K}{R} + q_k\binom{k}{r}\left[G_K\binom{K+1}{R} + H_K\binom{K}{R}\right]$$
$$= \binom{k}{r}\binom{K}{R} + p_k\binom{k}{r-1}\binom{K}{R} + q_kG_K\binom{k}{r}\binom{K}{R-1},$$

and the expected number of busy (r,R)-tuples *created* by the nth arrival, under the stated condition, is

$$p_k\binom{k}{r-1}\binom{K}{R} + (1 - p_k)G_K\binom{k}{r}\binom{K}{R-1}.$$

Now the probability that the life of a busy (r,R)-tuple will be longer than x is exp $(-(r + R)x)$. Thus the expected number of encounters between arriving calls and created (r,R)-tuples in the interval $(0,t]$ is:

$$\begin{aligned}\sum_{n=1}^{\infty}\sum_{k=r-1}^{m}\sum_{K=R-1}^{M}\int_0^t dP\,\{\tau_n \leqq u,\, \xi_n = k,\, \Xi_n = K\}&\\ \cdot\left[p_k\binom{k}{r-1}\binom{K}{R} + (1 - p_k)G_K\binom{k}{r}\binom{K}{R-1}\right]&\qquad(70)\\ \cdot\int_0^{t-u} e^{-(r+R)x}\,dM(x)&\end{aligned}$$

where $M(x)$ is the expected number of arrivals in an interval of length x, when there was an arrival at the start of the interval. $M(x)$ has Laplace-Stieltjes transform

$$\mu(s) = \frac{\varphi(s)}{1 - \varphi(s)}.$$

Equation (70) is a convolution. Recalling (68), we see that (70) has Laplace-Stieltjes transform,

$$\begin{aligned}\sum_{k=r-1}^{m}\sum_{K=R-1}^{M}\Bigg[p_k\binom{k}{r-1}\binom{K}{R}&\\ + (1 - p_k)G_K\binom{k}{r}\binom{K}{R-1}\Bigg]\mu_{ik}{}^{IK}(s)\mu(s + r + R).&\qquad(71)\end{aligned}$$

We must not forget the (r,R)-tuples which were busy initially; the expected number of encounters between arriving calls and these is

$$\binom{i}{r}\binom{I}{R}\sum_{n=1}^{\infty}\int_0^t dP\{\tau_n \leqq u\}e^{-(r+R)u} = \binom{i}{r}\binom{I}{R}\int_0^t dM(u)e^{-(r+R)u}$$

with Laplace-Stieltjes transform

$$\binom{i}{r}\binom{I}{R}\mu(s + r + R). \qquad (72)$$

Adding (71) and (72) we get

$$\begin{aligned}\epsilon_{ir}{}^{IR}(s) = \frac{\varphi(s + r + R)}{1 - \varphi(s + r + R)}&\\ \cdot\left[\binom{i}{r}\binom{I}{R} + \Psi_{i,r-1}{}^{IR}(s) + X_{ir}{}^{I,R-1}(s) - Y_{ir}{}^{I,R-1}(s)\right].&\qquad(73)\end{aligned}$$

Now comparing (69) and (73), we obtain (67).

Theorem 6: The distribution $P(k,K,t)$ $(t > 0)$ *is determined by*

$$\beta(r,R,s) = \frac{1 - \varphi(s + r + R)}{\varphi(s + r + R)} \cdot \frac{1}{s + r + R} \Phi_{ir}{}^{IR}(s). \tag{74}$$

Proof: We have

$$\begin{aligned} P(k,K,t) = {} & \binom{i}{k} e^{-tk}(1 - e^{-t})^{i-k} \binom{I}{K} e^{-tK}(1 - e^{-t})^{I-K}[1 - F(t)] \\ & + \sum_{n=1}^{\infty} \sum_{j=0}^{m} \sum_{J=0}^{M} \int_0^t dP\{\xi_n = j, \Xi_n = J, \tau_n \leqq u\} \\ & \cdot \left\{ p_j \binom{j+1}{k} e^{-(t-u)k}(1 - e^{-(t-u)})^{j+1-k} \binom{J}{K} \right. \\ & \cdot e^{-(t-u)K}(1 - e^{-(t-u)})^{J-K} + q_j \binom{j}{k} e^{-(t-u)k} \\ & \cdot (1 - e^{-(t-u)})^{j-k} \left[G_J \binom{J+1}{K} e^{-(t-u)K} \right. \\ & \cdot (1 - e^{-(t-u)})^{J+1-K} + H_J \binom{J}{K} e^{-(t-u)K} \\ & \left. \left. \cdot (1 - e^{-(t-u)})^{J-K} \right] \right\} [1 - F(t - u)]. \end{aligned} \tag{75}$$

This may be seen as follows: either no calls arrive in the interval $(0,t]$, or the last call to arrive in that interval is the nth $(n = 1, 2, \cdots)$, i.e. the nth call arrives at time u and no calls arrive in the interval $(u,t]$. If this call encounters the state (j,J) it may get on the first-choice group (probability p_j), the overflow group (probability $q_j G_J$), or neither (probability $q_j H_J$). Then enough calls must end in the interval $(u,t]$ so that the state at time t is (k,K).

From (75), and keeping in mind (68),

$$\begin{aligned} B(r,R,t) = {} & \binom{i}{r}\binom{I}{R} e^{-t(r+R)}[1 - F(t)] + \sum_{j=0}^{m} \sum_{J=0}^{M} \int_0^t dM_{ij}{}^{IJ}(u) \\ & \cdot e^{-(t-u)(r+R)} \left\{ \binom{j}{r}\binom{J}{R} + p_j \binom{j}{r-1}\binom{J}{R} + q_j G_J \binom{j}{r} \right. \\ & \left. \cdot \binom{J}{R-1} \right\} [1 - F(t - u)], \end{aligned}$$

and taking the Laplace transform,

$$\beta(r,R,s) = \frac{1 - \varphi(s + r + R)}{s + r + R}\left[\binom{i}{r}\binom{I}{R} + \Phi_{ir}{}^{IR}(s) + \Psi_{i,r-1}{}^{IR}(s) + X_{ir}{}^{I,R-1}(s) - Y_{ir}{}^{I,R-1}(s)\right]. \tag{76}$$

From (76) and (67) we obtain (74).

It remains to show that the double binomial moments uniquely determine the probabilities $P(k,K,t)$. As in the proof of Theorem 4, it will suffice to show that for all $t > 0$

$$\lim_{R\to\infty} \frac{B(0,R,t)}{B(0,R-1,t)} = 0. \tag{77}$$

From (67), for $R > I$,

$$\Phi_{i0}{}^{IR}(s) \leqq \frac{\varphi(s + R)}{1 - \varphi(s + R)}\,\Phi_{i0}{}^{I,R-1}(s). \tag{78}$$

But, *for all* $s > 0$,

$$\lim_{R\to\infty} \frac{\varphi(s + R)}{1 - \varphi(s + R)} = 0.$$

Therefore

$$\lim_{R\to\infty} \frac{\Phi_{i0}{}^{IR}(s)}{\Phi_{i0}{}^{I,R-1}(s)} = 0. \tag{79}$$

Now from (74),

$$\frac{\beta(0,R,s)}{\beta(0,R-1,s)} = \frac{1 - \varphi(s + R)}{\varphi(s + R)}\,\frac{\varphi(s + R - 1)}{1 - \varphi(s + R - 1)} \cdot \frac{s + R - 1}{s + R}\,\frac{\Phi_{i0}{}^{IR}(s)}{\Phi_{i0}{}^{I,R-1}(s)}$$

and so

$$\lim_{R\to\infty} \frac{\beta(0,R,s)}{\beta(0,R-1,s)} = \lim_{R\to\infty} \frac{\Phi_{i0}{}^{IR}(s)}{\Phi_{i0}{}^{I,R-1}(s)} = 0, \tag{80}$$

since

$$\lim_{s\to\infty} \frac{\varphi(s)}{\varphi(s-1)} = 1.$$

From (80), and the inversion formula for the Laplace transform, the result (77) follows.

Example 3:

Consider the case

$$\begin{aligned} q_k &= 0 && (k = 0, 1, \cdots, m-1)\\ q_m &= 1 &&\\ M &= \infty &&\\ H_K &= H,\ G_K = G\ (G + H = 1) && (K = 0, 1, 2, \cdots). \end{aligned}$$

This example may be of some practical interest. It represents a situation in which some equipment, other than a free trunk, is needed to set up a connection on the overflow group. If this equipment is serving a large number of trunk groups, the chance of its being idle may be substantially independent of the situation on the particular overflow group being considered here, and may be represented by a constant, G.

In this case we have

$$\begin{aligned} X_{ir}{}^{IR}(s) &= G\Phi_{ir}{}^{IR}(s)\\ Y_{ir}{}^{IR}(s) &= G\Psi_{ir}{}^{IR}(s) \end{aligned}$$

and

$$\Psi_{ir}{}^{IR}(s) = \Phi_{ir}{}^{IR}(s) - \binom{m}{r}\Phi_{im}{}^{IR}(s).$$

Equation (67) becomes

$$\begin{aligned} \Phi_{ir}{}^{IR}(s) = \frac{\varphi(s + r + R)}{1 - \varphi(s + r + R)} \Bigg\{ &\binom{i}{r}\binom{I}{R} + \Phi_{i,r-1}{}^{IR}(s)\\ &- \binom{m}{r-1}\Phi_{im}{}^{IR}(s) + G\binom{m}{r}\Phi_{im}{}^{I,R-1}(s)\Bigg\}. \end{aligned} \tag{81}$$

The solution of (81) is:

$$\begin{aligned} \Phi_{ir}{}^{IR}(s) = C_{r+R}(s)\Bigg\{ &\frac{\sum_{j=r}^{m}\binom{m}{j}\frac{1}{C_{j+R}(s)}}{\sum_{j=0}^{m}\binom{m}{j}\frac{1}{G^R C_{j+R}(s)}} \cdot \sum_{S=0}^{R}\binom{I}{S}\sum_{j=0}^{i}\binom{i}{j}\frac{1}{G^S C_{j+S-1}(s)}\\ &- \frac{\sum_{j=r+1}^{m}\binom{m}{j}\frac{1}{C_{j+R-1}(s)}}{\sum_{j=0}^{m}\binom{m}{j}\frac{1}{G^R C_{j+R-1}(s)}} \cdot \sum_{S=0}^{R-1}\binom{I}{S}\sum_{j=0}^{i}\binom{i}{j}\frac{1}{G^S C_{j+S-1}(s)}\\ &- \binom{I}{R}\sum_{j=r+1}^{m}\binom{i}{j}\frac{1}{C_{j+R-1}(s)}\Bigg\}. \end{aligned} \tag{82}$$

The expression for $\beta(r,R,s)$ can now be obtained from (82), using (74).

4.2 *The Limiting Distribution $P^*(k,K)$*

Theorem 7: The quantities $P^*(k,K)$ exist, are strictly positive, form a probability distribution, are independent of the initial state, and are uniquely determined by the double binomial moments

$$B^*(r,R) = \sum_{k=r}^{m} \sum_{K=R}^{M} \binom{k}{r} \binom{K}{R} P^*(k,K);$$

the latter satisfy

$$\begin{aligned} B^*(r,R) &= \frac{a}{r+R} \frac{1-\varphi_{r+R}}{\varphi_{r+R}} B(r,R), \qquad \text{for} \quad r+R>0 \qquad (83) \\ B^*(0,0) &= 1. \end{aligned}$$

Proof: To prove the existence, we consider the limit of (75) as $t \to \infty$. Clearly the first term goes to zero, and we have

$$\begin{aligned} P^*(k,K) = \lim_{t\to\infty} \sum_{J=0}^{M} \int_0^t \sum_{j=0}^{m} dM_{ij}{}^{IJ}(u) \\ \cdot \biggl\{ p_j \binom{j+1}{k} e^{-(t-u)k} (1-e^{-(t-u)})^{j+1-k} \binom{J}{K} \\ \cdot e^{-(t-u)K} (1-e^{-(t-u)})^{J-K} + q_j \binom{j}{k} \\ \cdot e^{-(t-u)k} (1-e^{-(t-u)})^{j-k} \biggl[G_J \binom{J+1}{K} \\ \cdot e^{-(t-u)K} (1-e^{-(t-u)})^{J+1-K} + H_J \binom{J}{K} \\ \cdot e^{-(t-u)K} (1-e^{-(t-u)})^{J-K} \biggr] \biggr\} [1-F(t-u)]. \end{aligned} \qquad (84)$$

It follows from Smith's "fundamental theorem,"[14] the assumption that $F(x)$ is not a lattice distribution, and the fact that $P(j,J) > 0$ for all j and J, that the limit in (84) exists and is given by

$$\begin{aligned}P^*(k,K) = \sum_{j=0}^{m}\sum_{J=0}^{M} \frac{P(j,J)}{\alpha} \int_0^\infty du[1 - F(u)] \\ \cdot\left\{p_j \binom{j+1}{k} e^{-uk}(1 - e^{-u})^{j+1-k}\right. \\ \cdot\binom{J}{K} e^{-uK}(1 - e^{-u})^{J-K} + q_j \binom{j}{k} e^{-uK}(1 - e^{-u})^{j-k} \qquad (85) \\ \cdot\left[G_J \binom{J+1}{K} \cdot e^{-uK}(1 - e^{-u})^{J+1-K}\right. \\ \left.\left. + H_J \binom{J}{K} e^{-uk}(1 - e^{-u})^{J-K}\right]\right\}.\end{aligned}$$

It is clear from (85) that $P^*(k,K) > 0$ for all (k,K), since the integrands are all strictly positive. (Note also that we have assumed $\alpha > 0$.) The dependence on (i,I) has disappeared, and it is easy to show from (85) that

$$\sum_{k=0}^{m}\sum_{K=0}^{M} P^*(k,K) = 1.$$

Thus $B^*(0,0) = 1$. To show (83), we take a different tack:

Consider any state (k,K). Transitions into the state (k,K) are of four types:

$$\begin{aligned} &(k - 1,K) \rightarrow (k,K) \qquad &&(\text{type } a) \\ &(k,K - 1) \rightarrow (k,K) &&(\text{type } b) \\ &(k + 1,K) \rightarrow (k,K) &&(\text{type } c) \\ &(k,K + 1) \rightarrow (k,K) &&(\text{type } d). \end{aligned}$$

Transitions out of the state (k,K) are also of four types:

$$\begin{aligned} &(k,K) \rightarrow (k - 1,K) \qquad &&(\text{type } a') \\ &(k,K) \rightarrow (k,K - 1) &&(\text{type } b') \\ &(k,K) \rightarrow (k + 1,K) &&(\text{type } c') \\ &(k,K) \rightarrow (k,K + 1) &&(\text{type } d'). \end{aligned}$$

Denote by $N_y(t)$ the expected number of transitions of type y in the interval $(0,t]$.

If we consider the process only at times when the state (k,K) exists, transitions of type (a') form a Poisson process of density k, and transitions of type (b') form a Poisson process of density K. Thus,

$$N_{a'}(t) = k \int_0^t P(k, K, t)dt \qquad (86a')$$

$$N_{b'}(t) = K \int_0^t P(k, K, t)dt. \qquad (86b')$$

Similarly,

$$N_c(t) = (k + 1) \int_0^t P(k + 1, K, t)dt \qquad (86c)$$

$$N_d(t) = (K + 1) \int_0^t P(k, K + 1, t)dt. \qquad (86d)$$

Now $\{\xi_n = k,\ \Xi_n = K\}$ is a recurrent event, with mean recurrence time $[\alpha/P(k,K)] > 0$. Thus, from the "elementary renewal theorem,"[15]

$$\lim_{t\to\infty} \frac{M_{ik}{}^{IK}(t)}{t} = \frac{P(k,K)}{\alpha}\, .$$

But clearly,

$$N_{d'}(t) = q_k G_K M_{ik}{}^{IK}(t),$$

so that

$$\lim_{t\to\infty} \frac{N_{d'}(t)}{t} = \frac{q_k G_K P(k,K)}{\alpha} = \frac{bG_K P^o(k,K)}{\alpha}\, . \qquad (86d')$$

Similarly,

$$\lim_{t\to\infty} \frac{N_b(t)}{t} = \frac{G_{K-1} b P^o(k,\, K - 1)}{\alpha} \qquad (86b)$$

$$\lim_{t\to\infty} \frac{N_{c'}(t)}{t} = \frac{p_k P(k,K)}{\alpha} = \frac{P(k,K) - bP^o(k,K)}{\alpha} \qquad (86c')$$

$$\lim_{t\to\infty} \frac{N_a(t)}{t} = \frac{P(k - 1,\, K) - bP^o(k - 1,\, K)}{\alpha}\, . \qquad (86a)$$

We now notice that in any interval $(0,t]$, the number of transitions out of the state (k,K) can differ from the number of transitions into the state (k,K) by at most 1. From this remark, and all the equations (86), it follows that

$$\begin{aligned}(k + K)P^*(k,K) &+ aP(k,K) - abH_K P^o(k,K) \\ &= ab[G_{K-1}P^o(k,K - 1) - P^o(k - 1,K)] + aP(k - 1,K) \qquad (87) \\ &\qquad + (k + 1)P^*(k + 1,K) + (K + 1)P^*(k,K + 1).\end{aligned}$$

Taking the double binomial moment of (87), one obtains

$$\begin{aligned}(r+R)B^*(r,R) = a\Big\{B(r-1,R) - \binom{m+1}{r}B(m,R) \\ + b\Big[C^o(r,R-1) - B^o(r-1,R) + \binom{m+1}{r}B^o(m,R)\Big]\Big\}.\end{aligned} \tag{88}$$

We now note that, according to (51),

$$\begin{aligned}a\Big[B(r-1,R) - \binom{m+1}{r}B(m,R)\Big] \\ = abC_{r+R-1}\Big[\sum_{j=r-1}^{m}\frac{B^o(j,R)}{C_{j+R}} - \sum_{j=r}^{m}\frac{C^o(j,R-1)}{C_{j+R-1}}\Big] \\ - ab\binom{m+1}{r}B^o(m,R).\end{aligned} \tag{89}$$

Putting (89) into (88), we obtain (83).

It is now easy to see that the $B^*(r,R)$ determine the $P^*(k,K)$. For from (83)

$$\begin{aligned}\lim_{R\to\infty}\frac{B^*(0,R)}{B^*(0,R-1)} = \lim_{R\to\infty}\frac{r+R-1}{r+R}\frac{\varphi_{R-1}}{\varphi_R}\frac{B(0,R)}{B(0,R-1)} \\ = \lim_{R\to\infty}\frac{B(0,R)}{B(0,R-1)} = 0.\end{aligned}$$

Corollary: For Poisson input, $P^*(k,K) = P(k,K)$.

Proof: For Poisson input, $F(x) = 1 - e^{-ax}$, $0 < a < \infty$; $a = 1/\alpha$. Thus

$$\varphi(s) = \frac{a}{a+s}, \qquad \varphi_r = \frac{a}{a+r}$$

$$B^*(r,R) = \frac{a}{r+R}\frac{r+R}{a}B(r,R) = B(r,R),$$

and since the double binomial moments determine the probabilities uniquely, the result follows.

Examples will be found in the next section.

V. EXAMPLES FOR THE STATIONARY PROCESS

5.1 *Categories of Examples*

In this section we will try to find the stationary binomial moments $B^o(r,R)$, $B(r,R)$, and $B^*(r,R)$ for certain special cases, or categories of cases. In the easiest cases we will succeed in finding explicit expressions for all these moments; in a harder case we will find explicit expressions only when $R = 1$ or $R = 2$; in the most complicated example (the random slip with overflow group, mentioned in Section I), the treatment is numerical, and only the results for the over-all blocking, B, are reported.

If the first-choice group is full-access, the situation is particularly simple, since overflow can only occur if $\xi_n = m$; the vector equations (24) for $U^o(k,R)$ then become scalar, and $B^o(r,R) = \binom{m}{r} U^o(m,R)$.

If the balking on the first-choice group is arbitrary, but the overflow group is infinite with no balking, or with constant balking probability, as in Example 3 above, some simplification occurs. For then,

$$V^o(k,R) = GU^o(k,R)$$

and hence (24) becomes a recurrence relation, although the quantities it relates are vectors. In such a case it is straightforward to find the first few moments of the distribution on the overflow group.

In cases in which neither of the above simplifications occur, the form of the balking probabilities may still be such as to facilitate calculation; an example of this is the random slip with overflow group.

5.2 *Full-Access First-Choice Group*

We suppose

$$q_k = 0 \qquad (k = 0, 1, \cdots, m - 1)$$

$$q_m = 1.$$

Equations (24) reduce to the single equation

$$U^o(m,R) = \mu_{mm}(R) V^o(m, R - 1) \tag{90}$$

and from (13),

$$\mu_{mm}(R) = \frac{\gamma(R)}{1 - \gamma(R)} \qquad (R = 1,2, \cdots).$$

$\gamma(R)$ is given by (1); it easily follows that

$$\mu_{mm}(R) = \frac{\sum_{j=0}^{m} \binom{m}{j} \frac{1}{C_{j-1}(R)}}{\sum_{j=0}^{m} \binom{m}{j} \frac{1}{C_j(R)}}. \tag{91}$$

Noting that, from the definitions,

$$C_j(R) = \frac{C_{j+R}}{C_{R-1}}, \tag{92}$$

(91) becomes

$$\mu_{mm}(R) = \frac{\sum_{j=0}^{m} \binom{m}{j} \frac{1}{C_{j+R-1}}}{\sum_{j=0}^{m} \binom{m}{j} \frac{1}{C_{j+R}}}. \tag{93}$$

We also know [from (25)] that

$$U^{o}(m,0) = \frac{P(m)}{b} = 1. \tag{94}$$

Example 4:

We now consider a slight generalization of the system considered by Brockmeyer (see Section I). Namely, let

$$\begin{aligned} q_k &= 0 \qquad (k = 0, 1, \cdots, m-1) \\ q_m &= 1 \\ H_K &= H \qquad (K = 0, 1, \cdots, M-1) \\ H_M &= 1. \end{aligned}$$

In this case we have

$$V^{o}(m,R) = G\left[U^{o}(m,R) - \binom{M}{R} U^{o}(m,M)\right].$$

Thus, from (90),

$$U^{o}(m,R) = \mu_{mm}(R) G\left[U^{o}(m,R-1) - \binom{M}{R-1} U^{o}(m,M)\right] \qquad (R = 1, 2, \cdots, M). \tag{95}$$

The solution of (95) is

$$U^{o}(m,R) = \left[G^{R} \prod_{Q=1}^{R} \mu_{mm}(Q)\right] \frac{\sum_{J=R}^{M} \binom{M}{J} \left[G^{J} \prod_{Q=1}^{J} \mu_{mm}(Q)\right]^{-1}}{\sum_{J=0}^{M} \binom{M}{J} \left[G^{J} \prod_{Q=1}^{J} \mu_{mm}(Q)\right]^{-1}}. \tag{96}$$

Now, from (93),

$$\prod_{Q=1}^{R} \mu_{mm}(Q) = \frac{\sum_{j=0}^{m} \binom{m}{j} \frac{1}{C_j}}{\sum_{j=0}^{m} \binom{m}{j} \frac{1}{C_{j+R}}} \qquad (R = 1, 2, \cdots). \tag{97}$$

Thus,

$$B^{\circ}(r,R) = \binom{m}{r} G^R \frac{\sum_{j=0}^{m} \binom{m}{j} \frac{1}{C_j}}{\sum_{j=0}^{m} \binom{m}{j} \frac{1}{C_{j+R}}} \cdot \frac{\sum_{J=R}^{M} \frac{\binom{M}{J}}{G^J} \sum_{j=0}^{m} \binom{m}{j} \frac{1}{C_{j+J}}}{\sum_{J=0}^{M} \frac{\binom{M}{J}}{G^J} \sum_{j=0}^{m} \binom{m}{j} \frac{1}{C_{j+J}}}. \tag{98}$$

We notice [see (T54)] that

$$\sum_{j=0}^{m} \binom{m}{j} \frac{1}{C_j} = \frac{1}{b} = \frac{1}{P(m)}.$$

Thus, from (51),

$$\begin{aligned} B(r,R) = {} & G^R C_{r+R} \frac{\sum_{J=R}^{M} \frac{\binom{M}{J}}{G^J} \sum_{j=0}^{m} \binom{m}{j} \frac{1}{C_{j+J}}}{\sum_{J=0}^{M} \frac{\binom{M}{J}}{G^J} \sum_{j=0}^{m} \binom{m}{j} \frac{1}{C_{j+J}}} \\ & \cdot \left\{ \frac{\sum_{j=r}^{m} \binom{m}{j} \frac{1}{C_{j+R}}}{\sum_{j=0}^{m} \binom{m}{j} \frac{1}{C_{j+R}}} - \frac{\sum_{j=r+1}^{m} \binom{m}{j} \frac{1}{C_{j+R-1}}}{\sum_{j=0}^{m} \binom{m}{j} \frac{1}{C_{j+R-1}}} \right\} \qquad (R = 1, 2, \cdots, M). \end{aligned} \tag{99}$$

$B^*(r,R)$ follows from (83).

When $G = 1$, (99) is the generalization to recurrent input of Brockmeyer's result, (4). It can indeed be verified that (99), for Poisson input and for $G = 1$, agrees with (4).

For infinite full-access overflow group ($M = \infty, G = 1$), (99) becomes

$$B(r,R) = C_{r+R} \left\{ \frac{\sum_{j=r}^{m} \binom{m}{j} \frac{1}{C_{j+R}}}{\sum_{j=0}^{m} \binom{m}{j} \frac{1}{C_{j+R}}} - \frac{\sum_{j=r+1}^{m} \binom{m}{j} \frac{1}{C_{j+R-1}}}{\sum_{j=0}^{m} \binom{m}{j} \frac{1}{C_{j+R-1}}} \right\}. \tag{100}$$

Equation (100) is the generalization to recurrent input of Kosten's result, (2). Again it can be verified that (100), for Poisson input, agrees with (2).

5.3 *Constant-Balking Overflow Group*

We suppose that $M = \infty$

$$G_K = G \qquad (K = 0, 1, 2, \cdots).$$

Then (24) becomes

$$U^{o}(k,R) = q_k G \sum_{j \in \mathfrak{A}} \mu_{jk}(R) U^{o}(j,R-1) \qquad (R = 1, 2, 3, \cdots). \quad (101)$$

Example 5

Suppose further that

$$q_k = q \qquad (k = 0, 1, \cdots, m-1)$$

$$q_m = 1.$$

This might describe a system in which some auxiliary equipment is needed to set up a connection on the first-choice group, some other auxiliary equipment is needed to set up a connection on the overflow group, and the probability that the auxiliary equipment is idle is constant, but this probability is different for the two groups. This is a rather plausible system, except that the overflow group is infinite.

We note that the blocking for such a system is

$$B = \sum_{k=0}^{m} \sum_{K=0}^{\infty} q_k H_K P(k,K) = H[q + pP(m)].$$

It is easy to show by the methods of Ref. 2 that in this example

$$B(r,0) = p^r C_r \frac{\displaystyle\sum_{j=r}^{m} \binom{m}{j} \frac{1}{p^j C_j}}{\displaystyle\sum_{j=0}^{m} \binom{m}{j} \frac{1}{p^j C_j}} \qquad (102)$$

so that in particular

$$P(m) = B(m,0) = \frac{1}{\displaystyle\sum_{j=0}^{m} \binom{m}{j} \frac{1}{p^j C_j}}.$$

Thus,

$$B = H\left[q + \frac{p}{\displaystyle\sum_{j=0}^{m} \binom{m}{j} \frac{1}{p^j C_j}} \right].$$

Instead of (101), we use (29), which in our case becomes

$$B^{o}(r,R) = G \sum_{j=0}^{m} [f_{jr}(R) - g_{jr}(R)]B^{o}(j,R-1) \qquad (R = 1, 2, \cdots). \tag{103}$$

In this case we have, from (37),

$$\Psi_{jr}(s) = p\left[\Phi_{jr}(s) - \binom{m}{r}\Phi_{jm}(s)\right]. \tag{104}$$

We can solve (35), (36), and (104) to obtain

$$\Phi_{jr}(s) = \frac{p^{r}C_{r}(s)}{\sum_{l=0}^{m}\binom{m}{l}\frac{1}{p^{l}C_{l}(s)}}\left\{\left[\sum_{l=0}^{j}\binom{j}{l}\frac{1}{p^{l}C_{l-1}(s)}\right]\left[\sum_{l=r}^{m}\binom{m}{l}\frac{1}{p^{l}C_{l}(s)}\right] - \left[\sum_{l=0}^{m}\binom{m}{l}\frac{1}{p^{l}C_{l}(s)}\right]\cdot\left[\sum_{l=r+1}^{m}\binom{j}{l}\frac{1}{p^{l}C_{l-1}(s)}\right]\right\}.$$

It follows from (31) that

$$f_{lr}(s) = p^{r}C_{r}(s)\left[\frac{\sum_{k=r}^{m}\binom{m}{k}\frac{1}{p^{k}C_{k}(s)}}{\sum_{k=0}^{m}\binom{m}{k}\frac{1}{p^{k}C_{k}(s)}}\cdot\frac{1}{p^{l}C_{l-1}(s)} - \begin{cases}\frac{1}{p^{l}C_{l-1}(s)} & \text{if } l > r\\ 0 & \text{if } l \leqq r\end{cases}\right]. \tag{105}$$

From (105), $f_{lr}(s) - g_{lr}(s)$ can easily be calculated by observing that in this example

$$f_{lr}(s) - g_{lr}(s) = qf_{lr}(s) + p\binom{m}{r}f_{lm}(s).$$

Then, from (103) one obtains

$$B^{o}(r,R) = G\left\{\frac{p\binom{m}{r} + qp^{r}C_{r}(R)\sum_{k=r}^{m}\binom{m}{k}\frac{1}{p^{k}C_{k}(R)}}{\sum_{k=0}^{m}\binom{m}{k}\frac{1}{p^{k}C_{k}(R)}}\cdot\sum_{j=0}^{m}\frac{B^{o}(j,R-1)}{p^{j}C_{j-1}(R)} - qp^{r}C_{r}(R)\sum_{j=r+1}^{m}\frac{B^{o}(j,R-1)}{p^{j}C_{j-1}(R)}\right\}. \tag{106}$$

Noting that, from (30) and (102),

$$B^o(r,0) = \frac{qB(r,0) + p\binom{m}{r}B(m,0)}{q + pB(m,0)} \tag{107}$$

$$= \left[qp^rC_r \sum_{k=r}^{m} \binom{m}{k} \frac{1}{p^kC_k} + p\binom{m}{r}\right]\left[q\sum_{k=0}^{m}\binom{m}{k}\frac{1}{p^kC_k} + p\right]^{-1},$$

we can use (106) to find $B^o(r,1)$, $B^o(r,2)$, etc., and in particular, the first and second moments of the distribution on the overflow group only, at overflow instants, $B^o(0,1)$, $B^o(0,2)$. The formulas are long; we quote only:

$$B^o(0,1) = G\left\{qC_1 + \frac{p}{\sum_{k=0}^{m}\binom{m}{k}\frac{1}{p^kC_{k+1}}} \cdot \frac{\sum_{k=0}^{m}(1+kq)\binom{m}{k}\frac{1}{p^kC_k}}{p + q\sum_{k=0}^{m}\binom{m}{k}\frac{1}{p^kC_k}}\right\}. \tag{108}$$

5.4 *Other Cases*

Once $B^o(r,R)$ is known, it is straightforward to determine $B(r,R)$ and $B^*(r,R)$, using (51) and (83) respectively. [If $B^o(r,R)$ is known, $C^o(r,R)$ can be determined, for use in (51), from the relation, which follows from their definitions:

$$C^o(r,R) = \sum_{J=R}^{M}\binom{J}{R}(\Delta^{J-R}G_R)B^o(r,J); \tag{109}$$

see (T45).] The problem is thus to determine $B^o(r,R)$, from (29) and (30), or equivalently to determine $U^o(k,R)$ from (24) and (25). We consider the latter method.

To use (24) and (25), one must first of all determine $\mu_{jk}(R)$ for all relevant j, k, and R [say, from (T70)], as well as $P(k)$ [say, from (T44) and (T45)]. Then the $V^o(k,R)$ must be expressed in terms of the $U^o(k,R)$; in general $V^o(k,R)$ can be expressed in terms of the $U^o(k,J)$, with $J \geqq R$, by a relation analogous to (109):

$$V^o(k,R) = \sum_{J=R}^{M}\binom{J}{R}(\Delta^{J-R}G_R)U^o(k,J). \tag{110}$$

When (110) is substituted in (24), one obtains a set of simultaneous equations for the $U^o(k,R)$. Equation (25) serves as a boundary condition. If M is finite, (24) can be used to express $U^o(k,M-1)$, $U^o(k,M-2)$, $\cdots$, $U^o(k,0)$ successively in terms of $U^o(k,M)$, and (25) can then be used to determine $U^o(k,M)$.

When the $U^{o}(k,R)$ are known, one finds the $B^{o}(r,R)$ by taking binomial moments, and then the $B(r,R)$ from (51). The probabilities $P(k,K)$ then follow by inverting the binomial moments, and the over-all blocking is determined by

$$B = \sum_{k=0}^{m} \sum_{K=0}^{M} q_k H_K P(k,K).$$

Example 6

We consider the system described in Section I

$$q_k = \binom{k}{\gamma} \Big/ \binom{m}{\gamma} \qquad (k = 0, 1, \cdots, m)$$

$$H_K = 0 \qquad (K = 0, 1, \cdots, M - 1)$$

$$H_M = 1.$$

The IBM 7090 computer at Murray Hill was programmed to find the blocking probability B for certain values of the parameters, namely:

$$m + M = 10$$

$$\gamma + M = 6.$$

The calculations were carried out for two kinds of input:

(*i*) Poisson

(*ii*) That sort of recurrent input which is itself the overflow from a group of m_0 trunks to which a Poisson stream of calls (with negative-exponential holding times) of mean intensity a_0 is submitted. Note that, since Poisson traffic is completely characterized by one parameter (its mean, in our case a_0), this sort of recurrent input is completely characterized by two parameters (a_0 and m_0).

Note also that this program allows one to calculate B for certain more complicated trunking arrangements, in the case of Poisson input, e.g., 2 common trunks overflowing to a random slip of 3 on 7 which in turn overflows to 1 common trunk. (This arrangement also involves a total of 10 trunks and 6 crosspoints per line.)

The results (blocking probability B as a function of input traffic a) are shown in Tables I and II and Fig. 1. The cases treated were $m_0 = 0$ (Poisson input, $a = a_0$) and $m_0 = 2$, in which case, of course,

$$a = \frac{a_0^3}{2} \Big/ \left(1 + a_0 + \frac{a_0^2}{2}\right);$$

γ was given the values 2,3,4,5,6. (Note that if $\gamma = 6$, then $M = 0$; there is no overflow group.)

TABLE I — RANDOM SLIP. BLOCKING AS A FUNCTION OF SUBMITTED TRAFFIC, FOR RECURRENT INPUT ($m_0 = 2$)

a_0 (call-hours)	a (call-hours)	Blocking, for the Configurations				
		2/6 + 4	3/7 + 3	4/8 + 2	5/9 + 1	6/10
1.0	0.2000	4.111×10^{-8}	2.711×10^{-8}	2.928×10^{-8}	5.785×10^{-8}	4.619×10^{-7}
1.5	0.4655	1.272×10^{-6}	8.990×10^{-7}	9.132×10^{-7}	1.425×10^{-6}	6.594×10^{-6}
2.0	0.8000	1.398×10^{-5}	1.039×10^{-5}	1.024×10^{-5}	1.290×10^{-5}	4.402×10^{-5}
2.5	1.179	8.454×10^{-5}	6.534×10^{-5}	6.338×10^{-5}	7.870×10^{-5}	1.891×10^{-4}
3.0	1.588	3.451×10^{-4}	2.757×10^{-4}	2.654×10^{-4}	3.101×10^{-4}	6.064×10^{-4}
3.5	2.018	1.066×10^{-3}	8.762×10^{-4}	8.407×10^{-4}	9.417×10^{-4}	1.575×10^{-3}
4.0	2.461	2.675×10^{-3}	2.253×10^{-3}	2.161×10^{-3}	2.348×10^{-3}	3.484×10^{-3}
4.5	2.916	5.713×10^{-3}	4.917×10^{-3}	4.727×10^{-3}	5.020×10^{-3}	6.792×10^{-3}
5.0	3.378	1.074×10^{-2}	9.421×10^{-3}	9.073×10^{-3}	9.486×10^{-3}	1.196×10^{-2}
5.5	3.847	1.823×10^{-2}	1.625×10^{-2}	1.570×10^{-2}	1.621×10^{-2}	1.935×10^{-2}
6.0	4.320	2.846×10^{-2}	2.573×10^{-2}	2.492×10^{-2}	2.552×10^{-2}	2.921×10^{-2}
6.5	4.797	4.149×10^{-2}	3.797×10^{-2}	3.688×10^{-2}	3.751×10^{-2}	4.158×10^{-2}
7.0	5.277	5.714×10^{-2}	5.286×10^{-2}	5.148×10^{-2}	5.208×10^{-2}	5.635×10^{-2}

TABLE II — RANDOM SLIP. BLOCKING AS A FUNCTION OF SUBMITTED TRAFFIC, FOR POISSON INPUT ($m_0 = 0$).

a_0 (call-hours)	a (call-hours)	Blocking, for the Configurations				
		2/6 + 4	3/7 + 3	4/8 + 2	5/9 + 1	6/10
1.0	1.0	1.010×10^{-6}	6.673×10^{-7}	6.771×10^{-7}	1.141×10^{-6}	6.407×10^{-6}
1.5	1.5	2.203×10^{-4}	1.573×10^{-5}	1.527×10^{-5}	2.099×10^{-5}	6.975×10^{-5}
2.0	2.0	1.750×10^{-4}	1.324×10^{-4}	1.265×10^{-4}	1.554×10^{-4}	3.664×10^{-4}
2.5	2.5	7.890×10^{-4}	6.250×10^{-4}	5.943×10^{-4}	6.815×10^{-4}	1.257×10^{-3}
3.0	3.0	2.474×10^{-3}	2.034×10^{-3}	1.935×10^{-3}	2.124×10^{-3}	3.307×10^{-3}
3.5	3.5	6.032×10^{-3}	5.112×10^{-3}	4.881×10^{-3}	5.204×10^{-3}	7.174×10^{-3}
4.0	4.0	1.225×10^{-2}	1.064×10^{-2}	1.020×10^{-2}	1.067×10^{-2}	1.347×10^{-2}
4.5	4.5	2.167×10^{-2}	1.922×10^{-2}	1.857×10^{-2}	1.909×10^{-2}	2.263×10^{-2}
5.0	5.0	3.449×10^{-2}	3.113×10^{-2}	3.010×10^{-2}	3.073×10^{-2}	3.481×10^{-2}
5.5	5.5	5.054×10^{-2}	4.627×10^{-2}	4.490×10^{-2}	4.553×10^{-2}	4.989×10^{-2}
6.0	6.0	6.938×10^{-2}	6.429×10^{-2}	6.259×10^{-2}	6.316×10^{-2}	6.755×10^{-2}
6.5	6.5	9.044×10^{-2}	8.464×10^{-2}	8.264×10^{-2}	8.309×10^{-2}	8.734×10^{-2}
7.0	7.0	1.131×10^{-1}	1.067×10^{-1}	1.045×10^{-1}	1.048×10^{-1}	1.087×10^{-1}

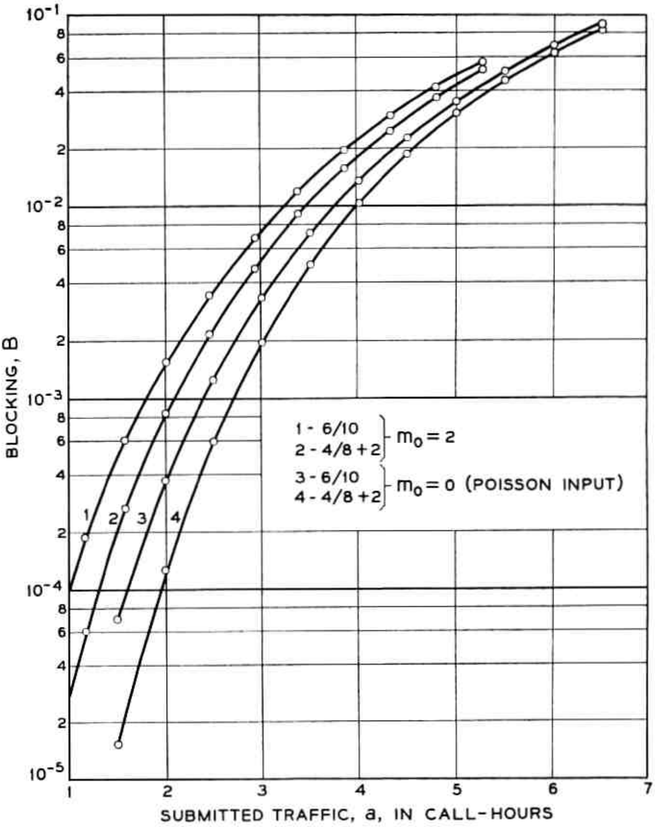


Fig. 1 — Blocking, B, vs submitted traffic, a.

Before commenting on the results, we mention parenthetically several special features introduced into the calculation by the special form of the balking probabilities and by the kind of input process considered in this example. First, as to finding the $P(k)$: (T44) and (T45) read, in our notation

$$B(r,0) = \frac{\varphi_r}{1 - \varphi_r} D(r - 1,0) \tag{111}$$

$$D(r,0) = \sum_{j=r}^{m} \binom{j}{r} (\Delta^{j-r} p_r) B(j,0). \tag{112}$$

In the present example,

$$\frac{\varphi_r}{1 - \varphi_r} = \frac{a_0}{r} \frac{C_r^{m_0}(a_0)}{C_{r+1}^{m_0}(a_0)} \qquad (r = 1, 2, \cdots) \tag{113}$$

and

$$\Delta^{j-r}p_r = -\frac{\binom{r}{j-\gamma}}{\binom{m}{\gamma}} \qquad (j = r+1, r+2, \cdots). \tag{114}$$

Also, since the overflow group is full-access (although finite), the relation (110) becomes

$$V^o(k,R) = U^o(k,R) - \binom{M}{R} U^o(k,M). \tag{115}$$

In Tables I and II and Fig. 1, we have used the notation $\gamma/m + M$ to describe a random-slip configuration in which each line has access to γ out of the m first-choice trunks and all the overflow trunks, except that the case $\gamma = 6$, $m = 10$, $M = 0$ is referred to as 6/10. The curves in Fig. 1 have been drawn, to avoid crowding, only for 4/8 + 2 and 6/10.

The following conclusions can be drawn from these results:

(*i*) The blocking is higher, for the same mean traffic, when $m_0 = 2$ than when $m_0 = 0$. This is consistent with the intuitive notion that overflow traffic is "peaky".

(*ii*) In a practical range of blocking ($B = 0.001$ or 0.01), 4/8 + 2 is the "best" arrangement and 6/10 is the "worst" of those considered, from the point of view of the traffic capacity of the system for a fixed blocking probability. It can be seen from the curves that if one wanted an arrangement using 6 crosspoints per line and 10 trunks, one would gain about 8 per cent (for $m_0 = 2$) or 6 per cent (for $m_0 = 0$) in traffic capacity at $B = 0.01$, by using the arrangement 4/8 + 2 instead of 6/10. At a blocking probability $B = 0.001$, these gains would be about 16 and 11 per cent respectively. Such increases in traffic capacity are not negligible; they seem to be larger for peaky traffic than for Poisson traffic.

(*iii*) For higher blockings ("overload" conditions), the advantage of 4/8 + 2 relative to 6/10 diminishes.

A study for a practical case would involve calculations of the blocking for other values of $\gamma + M$, a knowledge of the relative costs of trunks and crosspoints, and of course many other considerations, such as the relative costs of building and controlling 4/8 + 2 and 6/10 switches. Also, in such a study, one would want to keep in mind the approximations implicit in the model used in this paper. For example:

(*i*) In reality, blocked calls may wait or be resubmitted.

(*ii*) In reality, the number of traffic sources (lines) is finite, so that the arrival process after any instant is dependent on the number of trunks busy at that instant; thus the input is not, in reality, recurrent.

(*iii*) As a further result of the finiteness of the number of lines, the complete set of $\binom{m}{\gamma}$ access patterns required for a perfect random slip probably could not be used, and even if it could, equal traffic would not be submitted to each access pattern (so that the blocking experienced by different subscribers would be different).

VI. ACKNOWLEDGMENTS

I wish to express my thanks to Professor Lajos Takács for his help and encouragement on this problem, and to Miss G. E. Gioumousis, who worked with me on the computer program.

REFERENCES

1. Palm, C., Intensitätsschwankungen in Fernsprechverkehr, Eric. Tech., **44,** 1943, pp. 1–189.
2. Takács, L., Stochastic Processes with Balking in the Theory of Telephone Traffic, B.S.T.J., **40,** May, 1961, pp. 795–820.
3. Takács, L., On the Limiting Distribution of the Number of Coincidences Concerning Telephone Traffic, Ann. Math. Stat., **30,** 1959, pp. 134–141.
4. Descloux, A., On Overflow Processes of Trunk Groups with Poisson Inputs and Exponential Service Times, B.S.T.J., **42,** March, 1963, pp. 383–397.
5. Kosten, L., Uber Sperrungswahrscheinlichkeit bei Staffelschaltungen, Elect. Nach. Tech., **14,** 1937, pp. 5–12.
6. Wilkinson, R. I., Theories for Toll Traffic Engineering in the U.S.A., B.S.T.J., **35,** March, 1956, pp. 421–514.
7. Riordan, J., *Stochastic Service Systems*, Wiley, New York, 1962.
8. Brockmeyer, E., The Simple Overflow Problem in the Theory of Telephone Traffic, Teleteknik, **5,** 1954, pp. 361–374.
9. Bech, N., Method for Computing the Loss in Alternative Trunking and Grading Systems, Teleteknik, **5,** 1954, pp. 435–448.
10. Lundkvist, K., Analysis of General Theory for Telephone Traffic, Eric. Tech., **11,** 1955, pp. 3–32.
11. Syski, R., *Introduction to Congestion Theory in Telephone Systems*, Oliver and Boyd, Edinburgh, 1960.
12. Bodewig, E., *Matrix Calculus*, Interscience, New York, 1959.
13. Feller, W., *An Introduction to Probability Theory and its Applications*, Vol. I, Wiley, New York, 1957 (2nd ed.).
14. Smith, W. L., Asymptotic Renewal Theorems, Proc. Roy. Soc. Edinb., **64,** 1954, pp. 9–48.
15. Smith, W. L., Renewal Theory and its Ramifications, J. Roy. Stat. Soc., Ser. B, **20,** 1958, pp. 243–302.

On the Properties of Some Systems that Distort Signals — II

By I. W. SANDBERG

(Manuscript received April 19, 1963)

In this paper we study the recoverability of square-integrable bandlimited signals (with arbitrary frequency bands) that are distorted by a frequency-selective time-variable nonlinear operator and subsequently are bandlimited to the original bands. The distortion operator characterizes a very general class of systems containing linear time-invariant elements and a single time-variable nonlinear element. The subsequent bandlimiting of the system's output signals can be thought of as being due to transmission through a channel that performs filtering.

Our principal result asserts that, under certain conditions that are satisfied by many realistic systems, it is possible to uniquely determine the bandlimited input to the system from a knowledge of the bandlimited version of the output, in spite of the intermediate distortion which generally produces signals that are not bandlimited to the original frequency bands. We show that the input signal can be determined by a stable iteration procedure in which the approximating functions converge to their limit at a rate that is at least geometric.

I. INTRODUCTION

In this paper we study the recoverability of square-integrable bandlimited signals (with arbitrary frequency bands) that are distorted by a frequency-selective time-variable nonlinear operator and subsequently are bandlimited to the original bands. The distortion operator characterizes a very general class of systems containing linear time-invariant elements and a single time-variable nonlinear element. The subsequent bandlimiting of the system's output signals can be thought of as being due to transmission through a channel that performs filtering.

Our principal result asserts that, under certain conditions that are satisfied by many realistic systems, it is possible to uniquely determine the bandlimited input to the system from a knowledge of the bandlimited version of the output, in spite of the intermediate distortion which generally produces signals that are not bandlimited to the original frequency bands. Of course the distortion operator is assumed to be known. We show that the input signal can be determined by a stable iteration procedure in which the approximating functions converge to their limit at a rate that is at least geometric. When the physical system consists of only a single nonlinear element, our result reduces to that of Landau and Miranker,[1] and Zames.[2]

In the electronic circuitry of a communication system, it is often the case that an ideally linear amplifier is supplied with an approximately bandlimited input signal and that the circuitry subsequent to the amplifier introduces approximate bandlimiting. Under the assumption that the bandlimiting is ideal, our results imply that in many cases it is possible to completely reverse the effect of nonlinear distortion that may be introduced by such an amplifier due to the malfunctioning of, for example, a transistor or its bias supply, even though, as is typically the case, the transistor may be in a feedback loop. Of course it is necessary to know the properties of the distorting circuit. Results of this type may be useful in situations in which received signals are recorded and the time delay introduced by the recovery scheme is not important. For example, it is conceivable that this type of result may be useful in improving the quality of distorted signals obtained from a transmitter in a space vehicle containing a television camera, in which the distortion is due to a faulty video amplifier.

Section II considers some mathematical preliminaries. In Section III we state our principal results after discussing in detail a mathematical model of the physical system to be considered which focuses attention on the influence of the time-variable nonlinear element. Sections IV and V are concerned with the proof of the results. In particular, Section V considers the rate of convergence and stability of the recovery procedure. Section VI is concerned with some results that relate to the necessity of the conditions introduced earlier.

II. PRELIMINARIES

It is assumed that the reader is familiar with the contraction-mapping fixed-point theorem stated in Part I.[3,4]

As in Part I, $\mathfrak{L}_2$ denotes the Hilbert space of complex-valued square-integrable functions with inner product

$$(f,g) = \int_{-\infty}^{\infty} f\underline{g}\, dt$$

in which $\underset{\sim}{g}$ is the complex conjugate of g. The norm of f [i.e., $(f,f)^{\frac{1}{2}}$] is denoted by $\| f \|$. The intersection of the space $\mathfrak{L}_2$ with the set of real-valued functions is denoted by $\mathfrak{L}_{2R}$.

We take as the definition of the Fourier transform of $f(t)$ in $\mathfrak{L}_2$:

$$F(\omega) = \int_{-\infty}^{\infty} f(t)\, e^{-i\omega t}\, dt$$

and consequently

$$f(t) = \frac{1}{2\pi} \int_{-\infty}^{\infty} F(\omega)\, e^{i\omega t}\, d\omega.$$

With this definition, the Plancherel identity reads:

$$2\pi \int_{-\infty}^{\infty} f(t)\underset{\sim}{g}(t)\, dt = \int_{-\infty}^{\infty} F(\omega)\underset{\sim}{G}(\omega)\, d\omega.$$

As the notation above suggests, lower and upper case versions of a letter are used to denote, respectively, a function and its Fourier transform.

We shall be concerned with the following subspace of $\mathfrak{L}_{2R}$:

$$\mathfrak{B}(\Omega) = \{f(t) \mid f(t) \in \mathfrak{L}_{2R} ;\quad F(\omega) = 0,\, \omega \notin \Omega\}$$

where Ω is a union of disjoint intervals. The measure of Ω is denoted by $\mu(\Omega)$, which, unless stated otherwise, is not assumed to be finite. In particular, Ω may be the entire real line.

The operator that projects an arbitrary element of $\mathfrak{L}_{2R}$ onto $\mathfrak{B}(\Omega)$ is denoted by **P**. In electrical engineering terms, **P** is an ideal filtering operation.

The symbols **I** and **O** denote, respectively, the identity operator and the null operator (i.e., $\mathbf{O}f = 0$ for all $f \in \mathfrak{L}_2$).

III. MATHEMATICAL DESCRIPTION OF THE PHYSICAL SYSTEM AND STATEMENT OF PRINCIPAL RESULTS

Consider a nonlinear time-variable element imbedded in a linear physical system. Let s_1 and s_2 , respectively, denote the system's input and output signals, and let v and w, respectively denote the input and output signals associated with the nonlinear device, which is assumed to be characterized by the equation

$$w = \varphi(v,t) = \varphi[v], \tag{1}$$

where $\varphi(v,t)$ is a real-valued function of the real variables v and t.

It is assumed that v, w, $s_2 \in \mathfrak{L}_{2R}$, $s_1 \in \mathfrak{B}(\Omega)$, and that there exist well-defined linear operators $\mathbf{\Gamma}$ and $\mathbf{\Lambda}$, with domain $\mathcal{B}(\Omega) \times \mathcal{L}_{2R}$, such that† $v = \mathbf{\Gamma}[s_1, w]$ and $s_2 = \mathbf{\Lambda}[s_1, w]$.

We shall be concerned throughout with the four linear operators **A**, **B**, **C**, and **D** derived from $\mathbf{\Gamma}$ and $\mathbf{\Lambda}$ in the following manner:

$$\begin{aligned} v = \mathbf{\Gamma}[s_1, w] &= \mathbf{\Gamma}[s_1, 0] + \mathbf{\Gamma}[0, w] \\ &= \mathbf{A}s_1 + \mathbf{C}w \end{aligned} \tag{2}$$

$$\begin{aligned} s_2 = \mathbf{\Lambda}[s_1, w] &= \mathbf{\Lambda}[s_1, 0] + \mathbf{\Lambda}[0, w] \\ &= \mathbf{D}s_1 + \mathbf{B}w. \end{aligned} \tag{3}$$

3.1 *Representation of the Operators* A, B, C *and* D

We assume throughout that

$$\mathbf{A}f = \int_{-\infty}^{\infty} a(t - \tau) f(\tau) d\tau, \qquad \mathbf{B}f = \int_{-\infty}^{\infty} b(t - \tau) f(\tau) d\tau$$

$$\mathbf{C}f = \int_{-\infty}^{\infty} c(t - \tau) f(\tau) d\tau, \qquad \mathbf{D}f = \int_{-\infty}^{\infty} d(t - \tau) f(\tau) d\tau$$

where each of the real symbolic functions $a(t)$, $b(t)$, $c(t)$, and $d(t)$ is most generally the sum of an element of $\mathcal{L}_{2R}$ and a delta function. It is assumed throughout that $|C(\omega)|$ and $|B(\omega)|$ are uniformly bounded for all ω and that $|A(\omega)|$ and $|D(\omega)|$ are uniformly bounded for all $\omega \varepsilon \Omega$. It follows that **C** and **B** are bounded mappings of $\mathcal{L}_{2R}$ into itself and that **A** and **D** are bounded mappings of $\mathcal{B}(\Omega)$ into itself.

3.2 *The Projection Operation and the Basic Flow Graph*

We shall suppose that s_2, the system's output signal, is the input to a device that projects signals in $\mathcal{L}_{2R}$ onto the subspace $\mathcal{B}(\Omega)$. This device may be thought of as representing an ideal transmission channel of the low-pass, bandpass, or multiband type. If the output of the device is denoted by s_3, then clearly

$$s_3 = \mathbf{P}s_2 = \mathbf{T}^{-1} P \mathbf{T} s_2 \tag{4}$$

where

$$\begin{aligned} P = P(\omega) &= 1, \qquad \omega \varepsilon \Omega \\ &= 0, \qquad \omega \not\varepsilon \Omega \end{aligned}$$

and $\mathbf{T}s_2$ denotes S_2, the Fourier transform of s_2.

† This assumption is almost invariably satisfied in mathematical models of physical systems of interest.

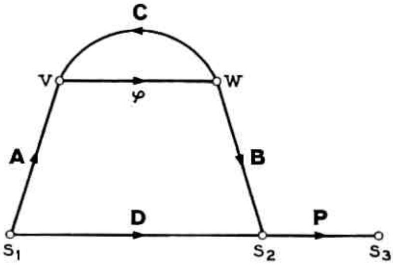


Fig. 1 — Signal-flow graph characterization of the relation between s_1, s_2, s_3, v, and w.

The equations we have introduced give rise to the signal-flow graph shown in Fig. 1 which summarizes the basic situation.

Our primary interest is in (*i*) obtaining conditions under which s_3 uniquely determines s_1, when s_1 is known to lie in the same subspace as s_3 [i.e., in $\mathcal{B}(\Omega)$], and (*ii*) obtaining a technique for recovering s_1.

3.3 *The Time-Variable Nonlinear Element*

We shall denote by $\psi(w,t)$ the inverse nonlinear characteristic; that is, $\psi(\varphi[v],t) = v$ for all v and t. It is assumed throughout that $\psi(0,t) = 0$ for all t, that $\psi\,[w(t)]$ is a measurable function of t whenever w is measurable, and that there exist two positive constants α and β with the properties that $\frac{1}{2}(\alpha + \beta) = 1$ and

$$\alpha(w_1 - w_2) \leqq \psi(w_1, t) - \psi(w_2, t) \leqq \beta(w_1 - w_2) \tag{5}$$

for all t and all $w_1 \geqq w_2$. Of course no loss of generality is introduced by the normalization $\frac{1}{2}(\alpha + \beta) = 1$, which happens to be convenient for our purposes. Observe that $0 < \alpha \leqq 1$.

It follows from (5) that

$$\beta^{-1}(v_1 - v_2) \leqq \varphi(v_1, t) - \varphi(v_2, t) \leqq \alpha^{-1}(v_1 - v_2)$$

for all t and all $v_1 \geqq v_2$. Observe that $w \in \mathcal{L}_{2R}$ if and only if $v \in \mathcal{L}_{2R}$.

3.4 *Assumptions Regarding* $|A(\omega)|$, $|B(\omega)|$, *and* $|D(\omega)|$

In addition to the uniform boundedness of $|A(\omega)|$, $|B(\omega)|$, $|C(\omega)|$, and $|D(\omega)|$ mentioned earlier, it is assumed, unless stated otherwise, that there exists a union of disjoint intervals Ω_D such that $\Omega_D \subseteq \Omega$,

$$\left.\begin{aligned} |D(\omega)| &= 0 \\ |B(\omega)| &\geqq k_1 \\ |A(\omega)| &\geqq k_2 \end{aligned}\right\} \omega \in \Omega_D ,$$

and

$$| D(\omega) | \geqq k_3 , \qquad \omega \, \varepsilon \, (\Omega - \Omega_D)$$

where k_1, k_2, and k_3 are positive constants. In most cases of engineering interest either $\Omega_D = \Omega$ or Ω_D is the null set.†

3.5 *Statement of Principal Results*

Our main result is

Theorem I: Let **A**, **B**, **C**, **D**, α, *and* ψ *be as defined in Sections* 3.1, 3.3, *and* 3.4. *Let*

$$\inf_{\omega \varepsilon (\Omega - \Omega_D)} | C - AD^{-1}B - 1 | > 1 - \alpha$$

$$\inf_{\omega \not\varepsilon \Omega} | C - 1 | > 1 - \alpha.$$

Then to each $s_3 \, \varepsilon \, \mathcal{B}(\Omega)$ *there correspond unique functions* $s_1 \, \varepsilon \, \mathcal{B}(\Omega)$ *and* w, v, $s_2 \, \varepsilon \, \mathfrak{L}_{2R}$ *such that*

$$\begin{aligned} s_3 &= \mathbf{P}s_2 \\ s_2 &= \mathbf{D}s_1 + \mathbf{B}w \\ v &= \mathbf{A}s_1 + \mathbf{C}w \\ v &= \psi[w] \end{aligned}$$

[*i.e., such that* (1), (2), (3), *and* (4) *are satisfied*]. *Furthermore if*

$$\begin{aligned} \bar{s}_3 &= \mathbf{P}\bar{s}_2 \\ \bar{s}_2 &= \mathbf{D}\bar{s}_1 + \mathbf{B}\bar{w} \\ \bar{v} &= \mathbf{A}\bar{s}_1 + \mathbf{C}\bar{w} \\ \bar{v} &= \psi[\bar{w}] \end{aligned}$$

where $\bar{w}$, $\bar{v}$, $\bar{s}_2 \, \varepsilon \, \mathfrak{L}_{2R}$ *and* $\bar{s}_1$, $\bar{s}_3 \, \varepsilon \, \mathcal{B}(\Omega)$,

$$\| s_1 - \bar{s}_1 \| \leqq k_4 \| s_3 - \bar{s}_3 \|$$

where k_4 *is a positive constant that depends only on* **A**, **B**, **C**, **D** *and* ψ.

Suppose that $\psi[w] = \mathbf{C}w + \mathbf{A}s_1$ {i.e., (2) with $v = \psi[w]$} possesses a unique solution $w \, \varepsilon \, \mathfrak{L}_{2R}$ for any $s_1 \, \varepsilon \, \mathcal{B}(\Omega)$ and that if $\psi[\bar{w}] = \mathbf{C}\bar{w} + \mathbf{A}\bar{s}_1$

† The assumptions in this section facilitate a common treatment of these two important cases. Observe that, with the exception of these cases, it is assumed here that $| D(\omega) |$ is discontinuous on Ω. However, as indicated in the Appendix this is by no means a necessary condition for the recoverability of s_1.

in which $\bar{s}_1 \varepsilon \mathcal{B}(\Omega)$ and $\bar{w} \varepsilon \mathcal{L}_{2R}$, $\| w - \bar{w} \| \leqq k_5 \| s_1 - \bar{s}_1 \|$, where k_5 is a constant that does not depend on s_1 or $\bar{s}_1$. [A direct application of Theorem II (in Section IV) shows that this is the case if $\inf_\omega | C - 1 | > (1 - \alpha)$.] It follows directly from the properties of ψ and the assumptions regarding **A**, **B**, **C**, and **D** that if $s_1 \varepsilon \mathcal{B}(\Omega)$, there exist unique functions v, s_2, $s_3 \varepsilon \mathcal{L}_{2R}$ such that (1) (2), (3), and (4) are satisfied. Let $\mathbf{\Phi}$ denote the operator that associates with each $s_1 \varepsilon \mathcal{B}(\Omega)$ the corresponding s_3. The assumptions regarding $\psi[w] = \mathbf{C}w + \mathbf{A}s_1$ together with the boundedness of **B** and **D** imply that $\mathbf{\Phi}$ is a bounded mapping of $\mathcal{B}(\Omega)$ into itself. Under the conditions stated in Theorem I, $\mathbf{\Phi}$ possesses a bounded inverse.

The invertibility conditions are established in Section IV and the boundedness of $\mathbf{\Phi}^{-1}$ is considered in Section V.

The method used to establish the invertibility conditions is constructive. In particular, $\mathbf{\Phi}^{-1}s_3$ can be computed in accordance with a stable iteration procedure for which the successive approximations converge in the $\mathcal{L}_{2R}$ norm at a rate that is at least geometric. The approximations converge also in the supremum norm at a rate that is geometric or greater if $\mu(\Omega)$ is finite.

As indicated earlier, in most cases of engineering interest either $\Omega_D = \Omega$ (the single-loop feedback system case), or Ω_D is the null set (i.e., the magnitude of the "direct transmission" $D(\omega)$ is uniformly bounded away from zero on Ω). The invertibility conditions stated above are satisfied in many cases of practical interest.

The situation considered by Landau and Miranker,[1] and Zames[2] corresponds to one in which $\mathbf{A} = \mathbf{B} = \mathbf{I}$, $\mathbf{D} = \mathbf{C} = \mathbf{O}$, and $\Omega_D = \Omega$. The inequalities are obviously satisfied in this case. In fact they are satisfied when $\Omega_D = \Omega$ and $C(\omega) = 0$, $\omega \not\varepsilon \Omega$. More generally, observe that the inequalities are met if and only if $(C - AD^{-1}B)$, for all $\omega \varepsilon (\Omega - \Omega_D)$, and C, for all $\omega \not\varepsilon \Omega$, are bounded away from the disk centered in the complex plane at [1,0] and having radius $1 - \alpha$ where $0 < \alpha \leqq 1$.

IV. DERIVATION OF INVERTIBILITY CONDITIONS

In the following discussion we shall denote by $\mathbf{P}_D$ the operator that projects elements of $\mathcal{L}_{2R}$ onto $\mathcal{B}(\Omega_D)$. That is,

$$\mathbf{P}_D f = \mathbf{T}^{-1} P_D \mathbf{T} f, \qquad f \varepsilon \mathcal{L}_{2R} \tag{6}$$

where

$$\begin{aligned} P_D = P_D(\omega) &= 1, \qquad \omega \varepsilon \Omega_D \\ &= 0, \qquad \omega \not\varepsilon \Omega_D \end{aligned}$$

and, as before, $\mathbf{T}f$ denotes the Fourier transform of f. Recall that $\mathbf{D}$ is an invertible mapping of $\mathcal{B}(\Omega - \Omega_D)$ into itself, that $\mathbf{A}$ and $\mathbf{B}$ are invertible mappings of $\mathcal{B}(\Omega_D)$ into itself, and that $\mathbf{D}$ annihilates $\mathcal{B}(\Omega_D)$. We shall denote by $\tilde{\mathbf{D}}^{-1}$ the inverse of the restriction of $\mathbf{D}$ to $\mathcal{B}(\Omega - \Omega_D)$, and by $\tilde{\mathbf{A}}^{-1}$ and $\tilde{\mathbf{B}}^{-1}$, respectively, the inverses of the restrictions of $\mathbf{A}$ and $\mathbf{B}$ to $\mathcal{B}(\Omega_D)$.

From (3) and (4)

$$s_3 = \mathbf{D}s_1 + \mathbf{PB}w, \qquad s_1 \varepsilon \mathcal{B}(\Omega) \tag{7}$$

and from (2) and $\psi[w] = v$

$$\psi[w] = \mathbf{C}w + \mathbf{A}s_1 . \tag{8}$$

Our objective is to determine w in order to find s_1 from (7) and (8). The corresponding functions s_2 and v can of course be computed from (3) and $v = \psi[w]$.

Since $\mathbf{D}$ annihilates $\mathcal{B}(\Omega_D)$, $\mathbf{P}_D s_3 = \mathbf{P}_D\mathbf{B}w$ and, since $\mathbf{P}_D$ and $\mathbf{B}$ commute,

$$\mathbf{P}_D w = \tilde{\mathbf{B}}^{-1}\mathbf{P}_D s_3 . \tag{9}$$

The problem therefore reduces to the determination of $(\mathbf{I} - \mathbf{P}_D)w$. Before proceeding it is convenient to set $w_a = \mathbf{P}_D w$ and $w_b = (\mathbf{I} - \mathbf{P}_D)w$, and to introduce

Definition I: Let

$$\begin{aligned} \eta(x) &= \beta - x, \qquad & x \leqq 1 \\ &= x - \alpha, \qquad & x \geqq 1. \end{aligned}$$

From (8),

$$(\mathbf{I} - \mathbf{P}_D)\psi[w_a + w_b] = \mathbf{C}w_b + \mathbf{A}(\mathbf{P} - \mathbf{P}_D)s_1 , \tag{10}$$

since $\mathbf{C}$ and $\mathbf{A}$ commute with $(\mathbf{I} - \mathbf{P}_D)$. From (7),

$$(\mathbf{P} - \mathbf{P}_D)s_3 = \mathbf{D}(\mathbf{P} - \mathbf{P}_D)s_1 + (\mathbf{P} - \mathbf{P}_D)\mathbf{B}w,$$

and

$$(\mathbf{P} - \mathbf{P}_D)s_1 = \tilde{\mathbf{D}}^{-1}(\mathbf{P} - \mathbf{P}_D)s_3 - \tilde{\mathbf{D}}^{-1}(\mathbf{P} - \mathbf{P}_D)\mathbf{B}w. \tag{11}$$

Thus,

$$(\mathbf{I} - \mathbf{P}_D)\psi[w_a + w_b] = \mathbf{C}w_b - \mathbf{A}\tilde{\mathbf{D}}^{-1}(\mathbf{P} - \mathbf{P}_D)\mathbf{B}w_b + \mathbf{A}\tilde{\mathbf{D}}^{-1}(\mathbf{P} - \mathbf{P}_D)s_3$$

from which

$$
\begin{aligned}
(\mathbf{I} - \mathbf{P}_D)\,\{\psi[w_a + w_b] - \psi_0 w_b\} \\
&= [\mathbf{C} - \mathbf{A}\tilde{\mathbf{D}}^{-1}(\mathbf{P} - \mathbf{P}_D)\mathbf{B} - \psi_0\mathbf{I}]w_b + \mathbf{A}\tilde{\mathbf{D}}^{-1}(\mathbf{P} - \mathbf{P}_D)s_3
\end{aligned}
$$

where ψ_0 is a real constant to be chosen subsequently.

Thus, regarding $[\mathbf{C} - \mathbf{A}\tilde{\mathbf{D}}^{-1}(\mathbf{P} - \mathbf{P}_D)\mathbf{B} - \psi_0\mathbf{I}]$ as a mapping of the orthogonal complement of $\mathcal{B}(\Omega_D)$ into itself, and assuming that it possesses a bounded inverse $[\mathbf{C} - \mathbf{A}\tilde{\mathbf{D}}^{-1}(\mathbf{P} - \mathbf{P}_D)\mathbf{B} - \psi_0\mathbf{I}]^{-1}$,

$$\mathrm{R}w_b = w_b$$

where

$$
\begin{aligned}
\mathrm{R}w_b = [\mathbf{C} - \mathbf{A}\tilde{\mathbf{D}}^{-1}(\mathbf{P} - \mathbf{P}_D)\mathbf{B} - \psi_0\mathbf{I}]^{-1}(\mathbf{I} - \mathbf{P}_D)\{\psi[w_a + w_b] - \psi_0 w_t\} \\
- [\mathbf{C} - \mathbf{A}\mathbf{D}^{-1}(\mathbf{P} - \mathbf{P}_D)\mathbf{B} - \psi_0\mathbf{I}]^{-1}\mathbf{A}\tilde{\mathbf{D}}^{-1}(\mathbf{P} - \mathbf{P}_D)s_3\,.
\end{aligned}
$$

The operator $\mathbf{R}$ is a mapping of a complete metric space into itself. We next establish a condition under which $\mathbf{R}$ is a contraction. Let $\mathbf{H} = [\mathbf{C} - \mathbf{A}\tilde{\mathbf{D}}^{-1}(\mathbf{P} - \mathbf{P}_D)\mathbf{B} - \psi_0\mathbf{I}]^{-1}$, and let f and g belong to the orthogonal complement of $\mathcal{B}(\Omega_D)$. Then

$$
\begin{aligned}
\| \mathbf{R}f - \mathbf{R}g \| &\leqq \| \mathbf{H}(\mathbf{I} - \mathbf{P}_D) \| \cdot \| \psi[w_a + f] - \psi[w_a + g] - \psi_0(f - g) \| \\
&\leqq \| \mathbf{H}(\mathbf{I} - \mathbf{P}_D) \| \, \eta(\psi_0) \, \| f - g \|,
\end{aligned}
$$

since

$$\left| \frac{\psi[w_a + f] - \psi[w_a + g]}{f - g} - \psi_0 \right| \leqq \eta(\psi_0).$$

Thus $\mathbf{R}$ is a contraction for some ψ_0 if

$$r = \inf_{\psi_0} \| \mathbf{H}(\mathbf{I} - \mathbf{P}_D) \| \, \eta(\psi_0) < 1. \tag{12}$$

It turns out that the optimal choice of ψ_0 is unity, the median of α and β. Consequently we could have simply set $\psi_0 = 1$ at the outset. However, we prefer to establish the significance of this choice.

4.1 *Evaluation of* $\| \mathbf{H}(\mathbf{I} - \mathbf{P}_D) \|$

Let $H = [C - AD^{-1}(P - P_D)B - \psi_0]^{-1}$ with the understanding that $D^{-1}(P - P_D) = 0$, $\omega \not\in (\Omega - \Omega_D)$. Our result is†

Lemma I:

$$\| \mathbf{H}(\mathbf{I} - \mathbf{P}_D) \| = \underset{\omega \not\in \Omega_D}{ess\ sup} \mid H(\omega) \mid.$$

† The notation $\underset{\omega}{\text{ess sup}}\, Q(\omega)$ denotes $\inf_{\mathfrak{N}} \sup_{\omega \not\in \mathfrak{N}} Q(\omega)$ where $\mathfrak{N}$ is an arbitrary zero-measure subset of the real line.

Proof:

The norm of $\mathbf{H}(\mathbf{I} - \mathbf{P}_D)$ is $\sup\{\| z \|; \| f \| = 1\}$ where $z = \mathbf{H}(\mathbf{I} - \mathbf{P}_D)f$ and $f \,\varepsilon\, \mathfrak{L}_{2R}$. An application of the Plancherel identity yields, in terms of the frequency domain representation of $\mathbf{H}$,

$$\| z \|^2 = \frac{1}{2\pi} \int_{\omega \notin \Omega_D} | H(\omega) |^2 \cdot | F(\omega) |^2 \, d\omega.$$

Hence

$$\sup\{\| z \|; \quad \| f \| = 1\} \leqq \operatorname*{ess\,sup}_{\omega \notin \Omega_D} | H(\omega) |.$$

However if $\operatorname*{ess\,sup}_{\omega \notin \Omega_D} | H(\omega) | < \infty$, for any $\epsilon > 0$ there exists a set of nonzero measure $\mathcal{E}$ which is disjoint from Ω_D and such that $| H(\omega) | \geqq \operatorname*{ess\,sup}_{\omega \notin \Omega_D} | H(\omega) | - \epsilon$, $\omega \,\varepsilon\, \mathcal{E}$. Since $| F(\omega) |$ is permitted to be nonzero only on $\mathcal{E}$, it follows that

$$\sup\{\| z \|; \| f \| = 1\} \geqq \operatorname*{ess\,sup}_{\omega \notin \Omega_D} | H(\omega) | - \epsilon.$$

Thus if $\operatorname*{ess\,sup}_{\omega \notin \Omega_D} | H(\omega) | < \infty$,

$$\| \mathbf{H}(\mathbf{I} - \mathbf{P}_D) \| = \operatorname*{ess\,sup}_{\omega \notin \Omega_D} | H(\omega) |. \tag{13}$$

It is clear that (13) remains valid if $\operatorname*{ess\,sup}_{\omega \notin \Omega_D} | H(\omega) | = \infty$. This proves the lemma.

It follows from (12) and Lemma I that

$$r = \inf_{\psi_0} \operatorname*{ess\,sup}_{\omega \notin \Omega_D} | [C - AD^{-1}(P - P_D)B - \psi_0]^{-1} | \, \eta(\psi_0). \tag{14}$$

4.2 *Determination of ψ_0 and Statement of Theorem II*

The following lemma indicates that the optimal choice of ψ_0 is independent of $[C - AD^{-1}(P - P_D)B]$.

Lemma II: Let ξ be a complex number and suppose that

$$| \xi - \psi_0 |^{-1} \eta(\psi_0) < 1.$$

Then

$$| \xi - \psi_0 |^{-1} \eta(\psi_0) \geqq | \xi - 1 |^{-1} \eta(1).$$

Proof:

Suppose first that $\psi_0 \leqq 1$ and that

$$| \xi - \psi_0 | > k(\beta - \psi_0), \qquad k > 1.$$

Then, since $| \xi - \psi_0 | \leqq | \xi - 1 | + | 1 - \psi_0 |$,

$$| \xi - 1 | + | 1 - \psi_0 | - k(1 - \psi_0) > k(\beta - 1),$$

and hence $| \xi - 1 | > k(\beta - 1)$. Suppose now that $\psi_0 \geqq 1$ and that

$$| \xi - \psi_0 | > k(\psi_0 - \alpha), \qquad k > 1.$$

Then,

$$| \xi - 1 | + | \psi_0 - 1 | - k(\psi_0 - 1) > k(1 - \alpha),$$

and hence $| \xi - 1 | > k(1 - \alpha)$.

It follows from (14) and Lemma II that if $r < 1$,

$$\begin{aligned} r &= \operatorname*{ess\,sup}_{\omega \notin \Omega_D} | [C - AD^{-1}(P - P_D)B - 1]^{-1} | \eta(1) \\ &= \operatorname*{ess\,sup}_{\omega \notin \Omega_D} | [C - AD^{-1}(P - P_D)B - 1]^{-1} | (1 - \alpha). \end{aligned}$$

At this point we are in a position to state

Theorem II: Let **A**, **B**, **C**, *and* **D** *be the bounded linear operators defined in Section 3.1. Let* **D**, *but not necessarily* **A** *and* **B**, *have the properties stated in Section 3.4. Let* $\tilde{\mathbf{D}}^{-1}$ *denote the inverse of the restriction of* **D** *to* $\mathcal{B}(\Omega_D)$, *and let* $\mathbf{P}_D$ *denote the operator that projects elements of* $\mathcal{L}_{2R}$ *onto* $\mathcal{B}(\Omega_D)$. *Suppose that*

$$r = max\,[r_1 , r_2] < 1,$$

where

$$\begin{aligned} r_1 &= \operatorname*{ess\,sup}_{\omega \epsilon (\Omega - \Omega_D)} | [C - AD^{-1}B - 1]^{-1} | (1 - \alpha) \\ r_2 &= \operatorname*{ess\,sup}_{\omega \notin \Omega} | [C - 1]^{-1} | (1 - \alpha). \end{aligned}$$

Then for any w_a *and* g, *respectively elements of* $\mathcal{B}(\Omega_D)$ *and its orthogonal complement with respect to* $\mathcal{L}_{2R}$, *there exists a unique* w_b *in the orthogonal complement of* $\mathcal{B}(\Omega_D)$ *such that*

$$(\mathbf{I} - \mathbf{P}_D)\psi[w_a + w_b] = [\mathbf{C} - \mathbf{A}\tilde{\mathbf{D}}^{-1}(\mathbf{P} - \mathbf{P}_D)\mathbf{B}]w_b + g.$$

In fact, $w_b = \lim_{i \to \infty} w_{bi}$ *where*

$$\begin{aligned} w_{b(i+1)} = [\mathbf{C} - \mathbf{A}\tilde{\mathbf{D}}^{-1}(\mathbf{P} - \mathbf{P}_D)\mathbf{B} - \mathbf{I}]^{-1}(\mathbf{I} - \mathbf{P}_D)\{\psi[w_a + w_{bi}] - w_{bi}\} \\ - [\mathbf{C} - \mathbf{A}\tilde{\mathbf{D}}^{-1}(\mathbf{P} - \mathbf{P}_D)\mathbf{B} - \mathbf{I}]^{-1}g \end{aligned}$$

and w_{b0} *is an arbitrary element in the orthogonal complement of* $\mathcal{B}(\Omega_D)$.

If $\bar{w}_b$ is a solution corresponding to $\bar{w}_a$ and $\bar{g}$,

$$\| w_b - \bar{w}_b \| \leqq \frac{r}{1-r} \| w_a - \bar{w}_a \| + \frac{r}{(1-r)(1-\alpha)} \| g - \bar{g} \|.$$

Proof:

With the exception of the last inequality, the proof follows from the fact that if $r < 1$, $\mathbf{R}$ (with $\psi_0 = 1$) is a contraction mapping of a complete metric space into itself.† The inequality is obtained as follows. Let $\mathbf{J} = [\mathbf{C} - \mathbf{A}\tilde{\mathbf{D}}^{-1}(\mathbf{P} - \mathbf{P}_D)\mathbf{B} - \mathbf{I}]^{-1}$ (i.e., let $\mathbf{J}$ be $\mathbf{H}$ with $\psi_0 = 1$). Then, since

$$w_b = \mathbf{J}(\mathbf{I} - \mathbf{P}_D)\{\psi[w_a + w_b] - w_b\} - \mathbf{J}g,$$

$$w_b - \bar{w}_b = \mathbf{J}(\mathbf{I} - \mathbf{P}_D)\{\psi[w_a + w_b] - \psi[\bar{w}_a + \bar{w}_b] - (w_a + w_b) + (\bar{w}_a + \bar{w}_b)\} - \mathbf{J}(g - \bar{g}).$$

Therefore

$$\| w_b - \bar{w}_b \| \leqq \| \mathbf{J}(\mathbf{I} - \mathbf{P}_D) \| \, \eta(1) \, \| w_a - \bar{w}_a + w_b - \bar{w}_b \| + \| \mathbf{J}(\mathbf{I} - \mathbf{P}_D) \| \cdot \| g - \bar{g} \|,$$

and since $r = \| \mathbf{J}(\mathbf{I} - \mathbf{P}_D) \| \, \eta(1)$, $\eta(1) = (1 - \alpha)$, and

$$\| w_a - \bar{w}_a + w_b - \bar{w}_b \| \leqq \| w_a - \bar{w}_a \| + \| w_b - \bar{w}_b \|,$$

$$\| w_b - \bar{w}_b \| \leqq \frac{r}{1-r} \| w_a - \bar{w}_a \| + \frac{r}{(1-r)(1-\alpha)} \| g - \bar{g} \|.$$

With regard to the "essential supremum" notation used in the statements of Lemma I and Theorem II, it is of course true that

$$\operatorname*{ess\,sup}_{\omega \notin \Omega_D} | H(\omega) | = \sup_{\omega \notin \Omega_D} | H(\omega) |$$

in at least almost all cases of engineering interest.

4.3 *The Complete Recovery Scheme*

Let us now consider our over-all objective, the recovery of s_1. From (8) and (11), using the definition of $\tilde{\mathbf{A}}^{-1}$,

$$(\mathbf{P} - \mathbf{P}_D)s_1 = \tilde{\mathbf{D}}^{-1}(\mathbf{P} - \mathbf{P}_D)s_3 - \tilde{\mathbf{D}}^{-1}(\mathbf{P} - \mathbf{P}_D)Bw$$

$$\mathbf{P}_D s_1 = \tilde{\mathbf{A}}^{-1}\mathbf{P}_D\{\psi[w] - \mathbf{C}w\}.$$

† In particular, our assumption regarding the inverse of $[\mathbf{C} - \mathbf{A}\tilde{\mathbf{D}}^{-1}(\mathbf{P} - \mathbf{P}_D) - \mathbf{B} - \mathbf{I}]$ is satisfied, since $| C - AD^{-1}(P - P_D) - 1 |$ is bounded away from zero for all ω in the complement of Ω_D.

Therefore,

$$s_1 = (\mathbf{P} - \mathbf{P}_D)s_1 + \mathbf{P}_D s_1 = [\tilde{\mathbf{D}}^{-1}(\mathbf{P} - \mathbf{P}_D) - \tilde{\mathbf{A}}^{-1}\mathbf{C}\tilde{\mathbf{B}}^{-1}\mathbf{P}_D]s_3 + \tilde{\mathbf{A}}^{-1}\mathbf{P}_D\{\psi[\tilde{\mathbf{B}}^{-1}\mathbf{P}_D s_3 + w_b]\} - \tilde{\mathbf{D}}^{-1}(\mathbf{P} - \mathbf{P}_D)\mathbf{B}w_b \tag{15}$$

where we have used (9), the fact that $(\mathbf{P} - \mathbf{P}_D)\mathbf{B}w_a = 0$, and the identity $\tilde{\mathbf{A}}^{-1}\mathbf{P}_D\mathbf{C}\tilde{\mathbf{B}}^{-1}\mathbf{P}_D s_3 = \tilde{\mathbf{A}}^{-1}\mathbf{C}\tilde{\mathbf{B}}^{-1}\mathbf{P}_D s_3$. This proves the first part of Theorem I. The second part, which is concerned with the boundedness of $\mathbf{\Phi}^{-1}$, is considered in Section 5.1.

We define s_{1n}, the nth approximation to s_1, by

$$s_{1n} = [\tilde{\mathbf{D}}^{-1}(\mathbf{P} - \mathbf{P}_D) - \tilde{\mathbf{A}}^{-1}\mathbf{C}\tilde{\mathbf{B}}^{-1}\mathbf{P}_D]s_3 + \tilde{\mathbf{A}}^{-1}\mathbf{P}_D\{\psi[\tilde{\mathbf{B}}^{-1}\mathbf{P}_D s_3 + w_{bn}]\} - \tilde{\mathbf{D}}^{-1}(\mathbf{P} - \mathbf{P}_D)\mathbf{B}w_{bn} \tag{16}$$

where w_{bn} is the nth approximation to w_b as defined in Theorem II. Observe that

$$s_{1n} - s_1 = \tilde{\mathbf{A}}^{-1}\mathbf{P}_D\{\psi[\tilde{\mathbf{B}}^{-1}\mathbf{P}_D s_3 + w_{bn}] - \psi[\tilde{\mathbf{B}}^{-1}\mathbf{P}_D s_3 + w_b]\} - \tilde{\mathbf{D}}^{-1}(\mathbf{P} - \mathbf{P}_D)\mathbf{B}(w_{bn} - w_b),$$

from which, using the right inequality of (5) satisfied by ψ,

$$\| s_{1n} - s_1 \| \leqq \{ \| \tilde{\mathbf{A}}^{-1}\mathbf{P}_D \| \beta + \| \tilde{\mathbf{D}}^{-1}(\mathbf{P} - \mathbf{P}_D)\mathbf{B} \| \} \| w_{bn} - w_b \|. \tag{17}$$

An argument very similar to that used in the proof of Lemma I suffices to show that

$$\| \tilde{\mathbf{A}}^{-1}\mathbf{P}_D \| = \operatorname*{ess\,sup}_{\omega \varepsilon \Omega_D} | A^{-1} | \tag{18}$$

$$\| \tilde{\mathbf{D}}^{-1}(\mathbf{P} - \mathbf{P}_D)\mathbf{B} \| = \operatorname*{ess\,sup}_{\omega \varepsilon (\Omega - \Omega_D)} | D^{-1}B |. \tag{19}$$

Our assumptions regarding **A** and **B** imply that the right-hand side of (18) and the right-hand side of (19) are bounded. Therefore, since $w_b = \lim_{n\to\infty} w_{bn}$, (17) implies that $s_1 = \lim_{n\to\infty} s_{1n}$.

The convergence of s_{1n} to s_1 established in the last paragraph is in the mean-square sense. If $\mu(\Omega) < \infty$, it is also true that s_{1n} converges to s_1 pointwise uniformly in t, that is

$$\lim_{n\to\infty} \sup_t | s_{1n} - s_1 | = 0.$$

This result follows from the inequality:†

† This inequality is proved in Ref. 1 for the case in which Ω is a single interval centered at the origin. The extension to arbitrary sets of finite measure is trivial.

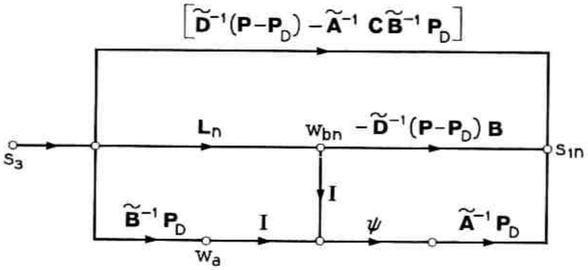

$[\tilde{D}^{-1}(P-P_D) - \tilde{A}^{-1} C \tilde{B}^{-1} P_D]$
L_n
w_{bn}
$-\tilde{D}^{-1}(P-P_D) B$
s_3
s_{1n}
I
$\tilde{B}^{-1} P_D$
I
ψ
$\tilde{A}^{-1} P_D$
w_a

Fig. 2 — Idealized recovery scheme.

$$\sup | f(t) | \leqq \left(\frac{\mu(\Omega)}{2\pi}\right)^{\frac{1}{2}} \| f \|, \qquad f \varepsilon \mathcal{B}(\Omega)$$

and the fact that s_{1n}, $s_1 \varepsilon \mathcal{B}(\Omega)$.

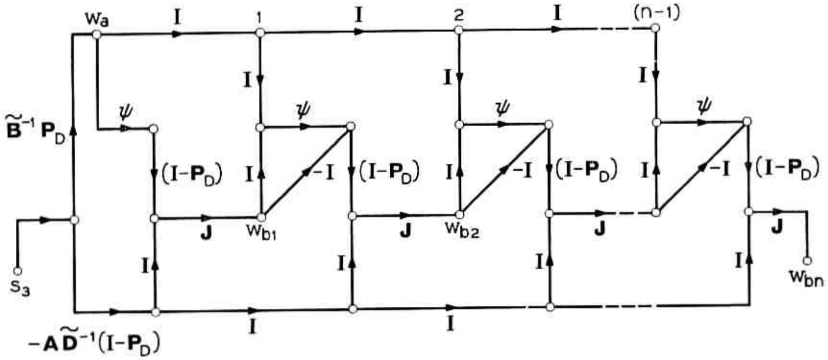


Fig. 3 — The iterative operation $\mathbf{L}_n$.

4.4 *Signal-Flow Graph for a Complete Recovery Scheme*

One complete idealized scheme for obtaining the nth approximation to s_1 , based on (16) and the solution for w_b given in Theorem II with $g = \mathbf{A\tilde{D}}^{-1}(\mathbf{P} - \mathbf{P}_D)s_3$ and $w_{b0} = 0$, is summarized in Fig. 2. The iterative operation† $\mathbf{L}_n$ is shown in detail in Fig. 3 in which, as defined earlier,

† In the special case in which Ω_D is the null set and $C - AD^{-1}PB = 0$ identically in ω, $w = \varphi[\mathbf{A\tilde{D}}^{-1}s_3]$ and hence the iteration stage is not required. The condition that $C - AD^{-1}PB$ vanish identically in ω, under which $\mathbf{\Phi}$ is by no means a trivial mapping of $\mathcal{B}(\Omega)$ into $\mathcal{B}(\Omega)$, is equivalent in engineering terms to requiring that the feedback transmission, for $\omega \not\varepsilon \Omega$, and the null feedback transmission, for $\omega \varepsilon \Omega$, both vanish.

$\mathbf{J} = [\mathbf{C} - \mathbf{A}\tilde{\mathbf{D}}^{-1}(\mathbf{P} - \mathbf{P}_D)\mathbf{B} - \mathbf{I}]^{-1}$. Fig. 4 shows a flow-graph representation of $\mathbf{J}$ in terms of $[\mathbf{C} - \mathbf{A}\tilde{\mathbf{D}}^{-1}(\mathbf{P} - \mathbf{P}_D)\mathbf{B}]$ and elementary operations. The flow graphs in Figs. 2 and 3 simplify in obvious ways in the important special cases in which $\mathbf{D} = \mathbf{O}$ on $\mathcal{B}(\Omega)$ or $\mathbf{D}$ possesses a bounded inverse on $\mathcal{B}(\Omega)$.

The analog implementation of the scheme presented in Fig. 2 requires consideration of the time delay inherent in the approximation of the impulse response functions corresponding to the nonrealizable operators† $\mathbf{P}$ and $\mathbf{P}_D$, as well as the time delay that might be required in the approximation of $\mathbf{J}$. These considerations imply that time delay sections must be inserted at various points in the recovery system and that the time variation of the nonlinear elements must be staggered. Of course the output of the recovery system will be a delayed version of an approximation of $s_1(t)$.

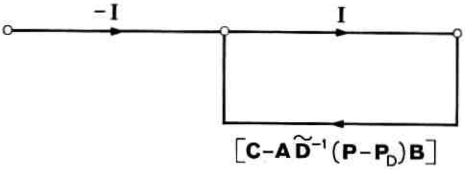


Fig. 4 — Flow-graph representation of the operator $\mathbf{J}$.

There are many variations possible in the implementation of the recovery system. For example, the iteration can be performed with a recording device and a *single* typical stage of the type used in Fig. 3.

V. RATE OF CONVERGENCE AND STABILITY OF THE RECOVERY SCHEME

The key element in the recovery scheme is of course the iteration procedure. We show first that the approximating functions w_{bi} converge to their limit w_b at a rate that is at least geometric. This type of convergence is a direct consequence of the fact that $w_{bi} = \mathbf{R}^i w_{b0}$ where $\mathbf{R}$ is a contraction mapping.

Since

$$\begin{aligned} w_{bi} &= w_{b0} + [w_{b1} - w_{b0}] + [w_{b2} - w_{b1}] + \cdots + [w_{bi} - w_{b(i-1)}], \\ \| w_{bi} - w_b \| &= \| [w_{b(i+1)} - w_{bi}] + [w_{b(i+2)} - w_{b(i+1)}] + \cdots \| \\ &\leqq \| w_{b(i+1)} - w_{bi} \| + \| w_{b(i+2)} - w_{b(i+1)} \| + \cdots . \end{aligned}$$

Repeated applications of the inequality:

† Of course we are ignoring the cases in which $\mathbf{P} = \mathbf{I}$ or $\mathbf{P}_D = \mathbf{O}$.

$$\| w_{bl} - w_{b(l-1)} \| = \| \mathbf{R}w_{b(l-1)} - \mathbf{R}w_{b(l-2)} \|$$
$$\leqq r \| w_{b(l-1)} - w_{b(l-2)} \|, \qquad l \geqq 2$$

lead to

$$\| w_{bi} - w_b \| \leqq \frac{r^i}{1-r} \| w_{b1} - w_{b0} \|. \tag{20}$$

If $w_{b0} = 0$, $w_{b1} = \mathbf{J}(\mathbf{I} - \mathbf{P}_D)\psi[\tilde{\mathbf{B}}^{-1}\mathbf{P}_D s_3] - \mathbf{J}\mathbf{A}\tilde{\mathbf{D}}^{-1}(\mathbf{P} - \mathbf{P}_D)s_3$, and hence

$$\begin{aligned}\| w_{bi} - w_b \| &\leqq \frac{r^i}{1-r} \| \mathbf{J}(\mathbf{I} - \mathbf{P}_D)\{\psi[\tilde{\mathbf{B}}^{-1}\mathbf{P}_D s_3] - \tilde{\mathbf{B}}^{-1}\mathbf{P}_D s_3 \\ &\qquad - \mathbf{A}\tilde{\mathbf{D}}^{-1}(\mathbf{P} - \mathbf{P}_D)s_3\}\| \\ &\leqq \frac{r^i}{1-r} \| \mathbf{J}(\mathbf{I} - \mathbf{P}_D) \| \, \{\eta(1) \, \| \tilde{\mathbf{B}}^{-1}\mathbf{P}_D \| \\ &\qquad + \| \mathbf{A}\tilde{\mathbf{D}}^{-1}(\mathbf{P} - \mathbf{P}_D) \|\} \, \| s_3 \| \\ &\leqq \frac{r^{i+1}}{1-r} \left\{ \| \tilde{\mathbf{B}}^{-1}\mathbf{P}_D \| + \frac{\| \mathbf{A}\tilde{\mathbf{D}}^{-1}(\mathbf{P} - \mathbf{P}_D) \|}{1-\alpha} \right\} \| s_3 \|\end{aligned}$$

where, in accordance with the arguments used in the proof of Lemma I,

$$\| \tilde{\mathbf{B}}^{-1}\mathbf{P}_D \| = \operatorname*{ess\,sup}_{\omega\varepsilon\Omega_D} | B^{-1} |$$
$$\| \mathbf{A}\mathbf{D}^{-1}(\mathbf{P} - \mathbf{P}_D) \| = \operatorname*{ess\,sup}_{\omega\varepsilon(-\Omega_D)} | AD^{-1} |.$$

5.1 *Stability of the Recovery Scheme*

We consider here the degree of immunity of the recovery scheme to two important types of errors.

It is assumed first that the input to the recovery system, which we shall denote by $\bar{s}_3$, differs† from s_3. Let overbarred symbols denote signals due to the input $\bar{s}_3$. We have from (15)

$$\begin{aligned}\| s_1 - \bar{s}_1 \| &= \| [\tilde{\mathbf{D}}^{-1}(\mathbf{P} - \mathbf{P}_D) - \tilde{\mathbf{A}}^{-1}\mathbf{C}\tilde{\mathbf{B}}^{-1}\mathbf{P}_D](s_3 - \bar{s}_3) \\ &\qquad + \tilde{\mathbf{A}}^{-1}\mathbf{P}_D\{\psi[\tilde{\mathbf{B}}^{-1}\mathbf{P}_D s_3 + w_b] - \psi[\tilde{\mathbf{B}}^{-1}\mathbf{P}_D\bar{s}_3 + \bar{w}_b]\} \\ &\qquad - [\tilde{\mathbf{D}}^{-1}(\mathbf{P} - \mathbf{P}_D)\mathbf{B}](w_b - \bar{w}_b) \| \\ &\leqq \| \tilde{\mathbf{D}}^{-1}(\mathbf{P} - \mathbf{P}_D) - \tilde{\mathbf{A}}^{-1}\mathbf{C}\tilde{\mathbf{B}}^{-1}\mathbf{P}_D \| \cdot \| s_3 - \bar{s}_3 \| \\ &\qquad + \| \tilde{\mathbf{A}}^{-1}\mathbf{P}_D \| \, \beta\{ \, \| \tilde{\mathbf{B}}^{-1}\mathbf{P}_D \| \cdot \| s_3 - \bar{s}_3 \| + \| w_b - \bar{w}_b \| \, \} \\ &\qquad + \| \tilde{\mathbf{D}}^{-1}(\mathbf{P} - \mathbf{P}_D)\mathbf{B} \| \cdot \| w_b - \bar{w}_b \|.\end{aligned} \tag{21}$$

† The departure of $\bar{s}_3$ from s_3 might be due to the presence of noise in either the transmission channel or the initial stages of the receiver.

However, from Theorem II with $g = \mathbf{A}\tilde{\mathbf{D}}^{-1}(\mathbf{P} - \mathbf{P}_D)s_3$,

$$\| w_b - \bar{w}_b \| \leqq \frac{r}{1-r} \| \tilde{\mathbf{B}}^{-1}\mathbf{P}_D \| \cdot \| s_3 - \bar{s}_3 \| + \frac{r}{(1-r)(1-\alpha)} \| \mathbf{A}\tilde{\mathbf{D}}^{-1}(\mathbf{P} - \mathbf{P}_D) \| \cdot \| s_3 - \bar{s}_3 \|. \tag{22}$$

In view of our earlier assumptions which imply the boundedness of all of the norms in (21) and (22), it is evident that there exists a positive constant k_4 such that

$$\| s_1 - \bar{s}_1 \| \leqq k_4 \| s_3 - \bar{s}_3 \| \tag{23}$$

for all s_3, $\bar{s}_3 \ \varepsilon \ \mathcal{B}(\Omega)$. In other words, our assumptions imply that $\mathbf{\Phi}^{-1}$ is bounded. This means that the error in the recovered signal is at most proportional to the error in the input to the recovery system. In particular, the recovered signal depends continuously on the input to the recovery system.

We show next that the recovery scheme is not critically dependent upon either an exact knowledge of the operator $\mathbf{J}$ or the projection property of $\mathbf{P}_D$. Specifically, we shall compare the functions w_b and $\hat{w}_b$ defined by

$$w_b = \mathbf{R}w_b, \quad \mathbf{R}w_b = \mathbf{J}(\mathbf{I} - \mathbf{P}_D)\{\psi[w_a + w_b] - w_b\} - \mathbf{J}\mathbf{A}\tilde{\mathbf{D}}^{-1}(\mathbf{P} - \mathbf{P}_D)s_3 \tag{24}$$

$$\hat{w}_b = \hat{\mathbf{R}}\hat{w}_b, \quad \hat{\mathbf{R}}\hat{w}_b = \mathbf{Q}\{\psi[w_a + \hat{w}_b] - \hat{w}_b\} - \mathbf{S}\mathbf{A}\tilde{\mathbf{D}}^{-1}(\mathbf{P} - \mathbf{P}_D)s_3 \tag{25}$$

where $\mathbf{Q}$ and $\mathbf{S}$ are bounded linear mappings of $\mathfrak{L}_{2R}$ into itself. We assume that $r < 1$ and that

$$\hat{r} = \| \mathbf{Q} \| \, \eta(1) < 1. \tag{26}$$

Hence $\hat{\mathbf{R}}$ is assumed to be a contraction mapping of $\mathfrak{L}_{2R}$ into itself. Note that inequality (26) is satisfied if $r = \| \mathbf{J}(\mathbf{I} - \mathbf{P}_D) \| \, \eta(1) < 1$ and $\| \mathbf{J}(\mathbf{I} - \mathbf{P}_D) - \mathbf{Q} \|$ is sufficiently small. A comparison of w_b and $\hat{w}_b$ yields an estimate of the error, due to the departure of $\mathbf{Q}$ from $\mathbf{J}(\mathbf{I} - \mathbf{P}_D)$ and to the departure of $\mathbf{S}$ from $\mathbf{J}$, in the limit function approached by the iteration procedure in the recovery system.

From (24) and (25),

$$\begin{aligned} w_b - \hat{w}_b = {} & (\mathbf{S} - \mathbf{J})\mathbf{A}\tilde{\mathbf{D}}^{-1}(\mathbf{P} - \mathbf{P}_D)s_3 + \mathbf{J}(\mathbf{I} - \mathbf{P}_D)\{\psi[w_a + w_b] - w_b\} \\ & - \mathbf{Q}\{\psi[w_a + w_b] - w_b\} + \mathbf{Q}\{\psi[w_a + w_b] - w_b\} - \mathbf{Q}\{\psi[w_a + \hat{w}_b] - \hat{w}_b\}, \end{aligned}$$

from which

$$\| w_b - \hat{w}_b \| \leqq \| (\mathbf{S} - \mathbf{J})\mathbf{A}\tilde{\mathbf{D}}^{-1}(\mathbf{P} - \mathbf{P}_D)s_3 \| + \| [\mathbf{J}(\mathbf{I} - \mathbf{P}_D) - \mathbf{Q}] \{\psi[w] - w_b\} \| + \| \mathbf{Q} \| \, \eta(1) \, \| w_b - \hat{w}_b \|,$$

and

$$\| w_b - \hat{w}_b \| \leqq \frac{1}{1 - \hat{r}} \| (\mathbf{S} - \mathbf{J})\mathbf{A}\tilde{\mathbf{D}}^{-1}(\mathbf{P} - \mathbf{P}_D)s_3 \| + \frac{1}{1 - \hat{r}} \| [\mathbf{J}(\mathbf{I} - \mathbf{P}_D) - Q]\{\psi[w] - w_b\} \|.$$

Therefore, if the departure of $\mathbf{Q}$ from $\mathbf{J}(\mathbf{I} - \mathbf{P}_D)$ is not too large (i.e., if $\hat{r} < 1$), the error in the limit function approached by the iteration technique is, for fixed s_3 (and hence fixed w), at most a linear combination of two terms, one that approaches zero as $\| \mathbf{S} - \mathbf{J} \|$ approaches zero, and another that approaches zero as $\| \mathbf{J}(\mathbf{I} - \mathbf{P}_D) - \mathbf{Q} \|$ approaches zero.

VI. SOME NEGATIVE RESULTS

In this final section we consider some results that relate to the necessity of the conditions introduced earlier.

The equation $\psi[w] = \mathbf{C}w + \mathbf{A}s_1$, in which $s_1 \,\varepsilon\, \mathcal{B}(\Omega)$, plays a central role in defining the mapping $\mathbf{\Phi}$. As stated in Section 3.5, Theorem II implies that this equation possesses a unique solution $w \,\varepsilon\, \mathcal{L}_{2R}$ if

$$\inf_{\omega} | C - 1 | > 1 - \alpha. \tag{27}$$

It is of interest to note that there exists a function ψ such that the equation $\psi[w] = \mathbf{C}w + \mathbf{A}s_1$ possesses no solution $w \,\varepsilon\, \mathcal{L}_{2R}$ for any non-identically zero $\mathbf{A}s_1$ if (27) is not satisfied, Ω is a bounded set, and $\mathbf{C} = c\mathbf{I}$ where c is a real constant. This follows directly from the fact that if (27) is violated, $\alpha \leqq c \leqq (2 - \alpha) = \beta$. Specifically, throughout a neighborhood of the origin let ψ be independent of t and linear in w with slope c. Then clearly, $\psi[w] - cw = 0$ whenever $| w | < \epsilon$ where ϵ is some positive constant. Since $\mathbf{A}s_1$ is assumed to be nonzero almost everywhere, the validity of our assertion is evident.

Let $\mathbf{U}$ denote the mapping of the orthogonal complement of $\mathcal{B}(\Omega_D)$ into itself defined by $\mathbf{U}w_b = (\mathbf{I} - \mathbf{P}_D)\psi[w_a + w_b] - \mathbf{E}w_b$, where $w_a \,\varepsilon\, \mathcal{B}(\Omega_D)$ and $\mathbf{E} = \mathbf{C} - \mathbf{A}\tilde{\mathbf{D}}^{-1}(\mathbf{P} - \mathbf{P}_D)\mathbf{B}$. Theorem II asserts that $\mathbf{U}$ possesses a bounded inverse if $E(\omega) = C - AD^{-1}(P - P_D)B$, for all ω contained in the complement of Ω_D, is bounded away from the disk in the complex plane centered at [0,1] and having radius $(1 - \alpha)$.

Theorem III: Let γ be a real constant and let Ξ_1 denote an open interval contained in the complement of Ω_D such that $E(\omega)$ is continuous on Ξ_1 and

$$\mathop{inf}_{\omega \varepsilon \Xi_1} | E(\omega) - \gamma | = 0.$$

Let ψ be independent of t and continuously differentiable with respect to x on an interval Ξ_2 where

$$\mathop{inf}_{x \varepsilon \Xi_2} \left| \frac{d\psi(x)}{dx} - \gamma \right| = 0.$$

Then **U** *does not possess a bounded inverse.*

Remark: Note that the hypotheses regarding ψ are satisfied if ψ is independent of t, continuously differentiable with respect to x, and γ is any point on the real-axis diameter of the disk mentioned above. Of course we assume that

$$\inf_x \frac{d\psi(x)}{dx} = \alpha, \qquad \text{and} \qquad \sup_x \frac{d\psi(x)}{dx} = \beta.$$

Proof of Theorem III:

We need the following lemma.

Lemma III: Let Δ_1 denote the real interval $[-T,T]$, let ϵ_1 and ϵ_2 be real positive constants, and let $h(t)$ be a continuous real function defined on Δ_1. Then there exists a function $g(t)$ in the orthogonal complement of $\mathcal{B}(\Omega_D)$ (assuming that Ω_D is a proper subset of the real line) such that

$$| h(t) - g(t) | \leqq \epsilon_1, \qquad t \varepsilon (\Delta_1 - \Delta_2)$$

where Δ_2 is a set of points contained in disjoint intervals of total measure not exceeding ϵ_2.

Proof:

If the complement of Ω_D contains an interval centered at the origin, the result is known and in fact is true with Δ_2 the null set. The following very direct argument makes use of the known result to treat the case in which the complement of Ω_D does not contain an interval centered at the origin.

Let ω_1 and ω_2 be real positive constants such that the interval $[\omega_1 - \omega_2, \omega_1 + \omega_2]$, where $\omega_1 > \omega_2$, is contained in the complement of Ω_D. Let Ω' be an interval of length $2\omega_2$ centered at the origin. Let Ω' be an interval of length $2\omega_2$ centered at the origin. Let $\{t_1, t_2, \cdots, t_n\} = \{t \mid t \varepsilon \Delta_1;\ \cos \omega_1 t = 0\}$. Let I_j denote an interval of length ϵ_2/n centered at t_j. For any $\epsilon_3 > 0$, there exists a function $l(t) \; \varepsilon \; \mathcal{B}(\Omega')$ such that

$$\left| l(t) - \frac{h(t)}{\cos \omega_1 t} \right| \leqq \epsilon_3, \qquad t \, \varepsilon \, (\Delta_1 - \Delta_2)$$

where $\Delta_2 = \bigcup_{j=1}^{n} I_j$. Choose ϵ_3 such that $\epsilon_2 = \epsilon_3 \inf_{t \varepsilon (\Delta_1 - \Delta_2)} \cos \omega_2 t$. It is evident that $l(t) \cos \omega_1 t$ possesses the properties of $g(t)$ stated in the lemma.

To prove Theorem III it suffices to show that for any $\epsilon > 0$, there exist two functions w_{1b} and w_{2b}, belonging to the orthogonal complement of $\mathcal{B}(\Omega_D)$, such that $\| w_{1b} - w_{2b} \| = 1$ and $\| \psi[w_a + w_{1b}] - \psi[w_a + w_{2b}] - \mathbf{E}(w_{1b} - w_{2b}) \| < \epsilon$.

Let ϵ_4, ϵ_5, and ϵ_6 be arbitrary positive constants. Since $\inf_{\omega \varepsilon \Xi_1} | E(\omega) - \gamma | = 0$ and $E(-\omega)$ is equal to the complex conjugate of $E(\omega)$, there exists an $\omega_3 \; \varepsilon \; \Xi_1$ such that $| E(\pm\omega_3) - \gamma | \leqq \frac{1}{2}\epsilon_4$. Let Π_1 and Π_2 denote two finite intervals of equal length $\mu(\Pi_1)$ contained in Ξ_1 and centered, respectively, at $-\omega_3$ and $+\omega_3$. Let $(w_{1b} - w_{2b}) \; \varepsilon \; \mathcal{B}(\Pi_1 \cup \Pi_2)$ with $\| w_{1b} - w_{2b} \| = 1$. Choose $\mu(\Pi_1)$ and T such that

$$\sup_{\omega \varepsilon \Pi_1} | E(\omega) - \gamma | \leqq \epsilon_4, \qquad \| w_{1b} - w_{2b} \|_{\substack{|t| > T \\ t \varepsilon \Delta_3}} \leqq \epsilon_5$$

where Δ_3 is any subset of $\Delta_1 = [-T,T]$ with measure not exceeding k_6, a sufficiently small positive constant. The second inequality can always be satisfied since, in accordance with the inequality stated in Section 4.3, $\sup_t | w_{1b} - w_{2b} | \leqq [\pi^{-1}\mu(\Pi_1)]^{\frac{1}{2}}$.

Since $\inf_{x \varepsilon \Xi_2} | [d\psi(x)/dx] - \gamma | = 0$, there exists a real constant $x_0 \; \varepsilon \; \Xi_2$ such that

$$\left| \frac{\psi[w_a + w_{1b}] - \psi[w_a + w_{2b}]}{w_{1b} - w_{2b}} - \gamma \right| \leqq \epsilon_6 \tag{28}$$

whenever $| w_a + w_{1b} - x_0 |$ and $| w_{1b} - w_{2b} |$ are sufficiently small. We may assume that $\mu(\Pi_1)$ is so small that the condition on $| w_{1b} - w_{2b} |$ is satisfied. Choose w_{1b} in accordance with Lemma III so that (28) is satisfied on $(\Delta_1 - \Delta_2)$ where Δ_2 is a set of measure not exceeding k_6. Let $(\Delta_1 - \Delta_2)^*$ denote the complement of $(\Delta_1 - \Delta_2)$. Observe that

$$\begin{aligned} &\| \psi[w_a + w_{1b}] - \psi[w_a + w_{2b}] - \mathbf{E}(w_{1b} - w_{2b}) \| \\ &\quad \leqq \| \psi[w_a + w_{1b}] - \psi[w_a + w_{2b}] - \gamma(w_{1b} - w_{2b}) \| \\ &\qquad + \| (\mathbf{E} - \gamma\mathbf{I})(w_{1b} - w_{2b}) \| \end{aligned}$$

$$\leqq \epsilon_6 \| w_{1b} - w_{2b} \| + \| \psi[w_a + w_{1b}] - \psi[w_a + w_{2b}] - \gamma(w_{1b} - w_{2b}) \|_{(\Delta_1 - \Delta_2)^*}$$
$$+ \| (\mathbf{E} - \gamma \mathbf{I})(w_{1b} - w_{2b}) \|$$
$$\leqq \epsilon_6 + (\beta + | \gamma |) \epsilon_5 + \| (\mathbf{E} - \gamma \mathbf{I})(w_{1b} - w_{2b}) \|$$
$$\leqq \epsilon_6 + (\beta + | \gamma |) \epsilon_5 + \epsilon_4 .$$

This completes the proof.

APPENDIX

The purpose of this appendix is to briefly indicate an alternative technique for determining sufficient conditions for the recoverability of s_1.

Instead of the assumptions stated in Section 3.4 suppose that for some real constant ψ_0:

$$\inf_{\omega \varepsilon \Omega} | D - B(\psi_0 - C)^{-1} A | > 0$$

$$\| (\psi_0 \mathbf{I} - \mathbf{C})^{-1} \| \, \eta(\psi_0) = \operatorname{ess\,sup}_{\omega} | (\psi_0 - C)^{-1} | \, \eta(\psi_0) = q < 1.$$

These inequalities imply that $\{\mathbf{PD} + \mathbf{PB}(\psi_0 \mathbf{I} - \mathbf{C})^{-1} \mathbf{A}\}$ possesses a bounded inverse on $\mathcal{B}(\Omega)$ and that for any $g \varepsilon \mathfrak{L}_{2R}$ the equation $\psi[w] = \mathbf{C}w + g$ possesses a unique solution $w \varepsilon \mathfrak{L}_{2R}$.

From

$$\psi[w] = \mathbf{C}w + \mathbf{A}s_1 , \qquad s_3 = \mathbf{PB}w + \mathbf{D}s_1 , \tag{29}$$

and $\psi[w] = \psi_0 w + \tilde{\psi}[w]$ we have

$$s_3 = \{\mathbf{PD} + \mathbf{PB}(\psi_0 \mathbf{I} - \mathbf{C})^{-1} \mathbf{A}\} s_1 - \mathbf{PB}(\psi_0 \mathbf{I} - \mathbf{C})^{-1} \tilde{\psi}[w]. \tag{30}$$

Equation (30) can be written as

$$s_1 = \mathbf{M}s_1 + \{\mathbf{PD} + \mathbf{PB}(\psi_0 \mathbf{I} - \mathbf{C})^{-1} \mathbf{A}\}^{-1} s_3$$

where

$$\mathbf{M}s_1 = \{\mathbf{PD} + \mathbf{PB}(\psi_0 \mathbf{I} - \mathbf{C})^{-1} \mathbf{A}\}^{-1} \mathbf{PB}(\psi_0 \mathbf{I} - \mathbf{C})^{-1} \tilde{\psi}[w].$$

Of course the dependence of the right-hand side on s_1 is through w.

Let $\bar{w}$ be the solution of $\psi[w] = \mathbf{C}w + \mathbf{A}s_1$ corresponding to $s_1 = \bar{s}_1$. Then by arguments similar to those leading to Theorem II,

$$\| w - \bar{w} \| \leqq \frac{1}{1 - q} \| (\psi_0 \mathbf{I} - \mathbf{C})^{-1} \, \mathbf{AP} \| \cdot \| s_1 - \bar{s}_1 \| .$$

Thus $\mathbf{M}$ is a contraction mapping of $\mathcal{B}(\Omega)$ into itself if

$$p = \| \{\mathbf{PD} + \mathbf{PB}(\psi_0\mathbf{I} - \mathbf{C})^{-1}\mathbf{A}\}^{-1}\mathbf{PB}(\psi_0\mathbf{I} - \mathbf{C})^{-1} \| \\ \eta(\psi_0)[1/(1 - q)] \| (\psi_0\mathbf{I} - \mathbf{C})^{-1}\mathbf{AP} \| < 1.$$

Hence if the received signal s_3 is known to be related to the transmitted signal $s_1 \varepsilon \mathcal{B}(\Omega)$ by (29), s_1 can be recovered if our assumptions are satisfied and if $p < 1$. Using arguments similar to those leading to Lemma I,

$$p = \operatorname*{ess\,sup}_{\omega \varepsilon \Omega} \left| \frac{B}{D(\psi_0 - C) + BA} \right| \eta(\psi_0) \frac{1}{1 - q} \operatorname*{ess\,sup}_{\omega \varepsilon \Omega} \left| \frac{A}{\psi_0 - C} \right| .$$

REFERENCES

1. Landau, H. J., and Miranker, W. L., The Recovery of Distorted Bandlimited Signals, J. Math. Anal. and Appl., **2,** February, 1961, pp. 97–104.
2. Zames, G. D., Conservation of Bandwidth in Nonlinear Operations, Quarterly Progress Report, MIT Res. Lab. Eng., October, 1959, No. 55.
3. Sandberg, I. W., On the Properties of Some Systems that Distort Signals—I, B.S.T.J., **42,** Sept. 1963, p. 2033.
4. Kolmogorov, A. N., and Fomin, S. V., *Elements of the Theory of Functions and Functional Analysis*, New York, Graylock Press, 1957.

Existence of Eigenvalues of a Class of Integral Equations Arising in Laser Theory

By D. J. NEWMAN and S. P. MORGAN

(Manuscript received March 19, 1963)

It is proved that the integral equation

$$\int_{-1}^{1} G(x)F(xy)H(y)f(y)\, dy = \lambda f(x)$$

has at least one nonzero eigenvalue if F is any integral function of finite order, G and H are any bounded functions on $[-1,1]$, and the trace of the kernel $G(x)F(xy)H(y)$ does not vanish. In particular, this theorem furnishes the first rigorous proof that the kernel exp $[ik(x - y)^2]$, which arises in the theory of the gas laser, has an eigenvalue for arbitrary complex k.

I. INTRODUCTION AND SUMMARY

In an idealized model of the gas laser or optical maser, as studied by Fox and Li[1,2] and others, electromagnetic radiation is reflected back and forth between two infinitely long metal strips which are mirror images of each other. A typical field quantity, such as the current density, at the surface of each reflector satisfies the integral equation

$$\int_{-1}^{1} \exp\{i[k(x - y)^2 - h(x) - h(y)]\}\, f(y)\, dy = \lambda f(x), \qquad (1)$$

where k is a dimensionless real parameter which depends on the width and spacing of the reflectors and the wavelength, and $h(x)$ is a real function specifying the departure of the reflecting surfaces from parallel planes.

The eigenfunctions of (1) represent the field distributions at the reflectors of the possible modes of oscillation of the laser, and the eigenvalue λ corresponding to a particular mode represents the complex factor by which the field strength is multiplied as a result of one reflection and transit between the reflectors. From the magnitude of λ one can deduce the amount of amplification which would have to be provided by an active medium between the reflectors in order just to sustain oscillations in the given mode, while the phase of λ determines admissible reflector spacings for oscillations at a particular frequency.

The mathematical interest of (1) centers around the fact that its kernel $K(x,y)$ is complex symmetric but not Hermitian;* that is,

$$K(x,y) = K(y,x) \qquad \text{but} \qquad K(x,y) \neq \overline{K(y,x)}. \tag{2}$$

The ordinary theory of Hermitian kernels does not even suffice to prove the existence of eigenvalues of complex symmetric kernels. Fox and Li[1] have made extensive calculations of the eigenvalues and eigenfunctions of (1) for $h(x) = 0$ by iterative numerical techniques up to about $k = 60$ (in applications k may be as large as a few hundred); but heretofore there has been no formal mathematical proof of the existence of solutions except† for $|k| \ll 1$, which is not a case of physical interest.

This paper contains a proof of the following

Theorem: Let $G(x)$ and $H(x)$ be any bounded functions on the interval $-1 \leqq x \leqq 1$, and let $F(z)$ be any integral function of finite order such that

$$\int_{-1}^{1} G(x)F(x^2)H(x)\, dx \neq 0. \tag{3}$$

Then the integral equation

$$\int_{-1}^{1} G(x)F(xy)H(y)f(y)\, dy = \lambda f(x) \tag{4}$$

has at least one nonzero eigenvalue.

As a corollary, it follows that the integral equation (1) has at least one eigenvalue for arbitrary complex k, provided only that

$$\int_{-1}^{1} e^{-2ih(x)}\, dx \neq 0. \tag{5}$$

Furthermore if $h(x)$ is an even function of x, then (1) has at least two eigenvalues for all but certain exceptional values of k, a particular exceptional value being $k = 0$.

The idea of the proof is quite simple. The assumption that $F(xy)$ in (4) is an integral function of finite order means that ultimately the coefficients of its Taylor series in powers of xy fall off with extreme rapidity.

* The kernel is normal in the special case $h(x) = kx^2$. The eigenfunctions of $\exp(-2ikxy)$ are prolate spheroidal wave functions, as pointed out in connection with lasers by Boyd and Gordon.[3]

† If $|k| \ll 1$ then $\exp[ik(x-y)^2]$ is nearly unity, and the existence of at least one eigenvalue follows from perturbation theory; see Sz.-Nagy.[4]

If we truncate the Taylor series after a finite number of terms, (4) is replaced by an integral equation with a kernel of finite rank. The eigenvalues of such a kernel are merely the latent roots of a finite matrix, and these are not all zero if their sum, which is the trace of the matrix, does not vanish. The limiting value of the trace is just the left side of (3), and does not vanish by hypothesis. By taking more and more terms of the series for $F(xy)$, we obtain a sequence of larger and larger matrices, whose elements ultimately vanish very rapidly with distance from the upper left corner. We show that it is possible to pick one eigenvalue from the set of eigenvalues of each succeeding matrix in such a way that the resulting sequence of numbers has a nonzero limit point. This limit point is an eigenvalue of the infinite matrix, and hence an eigenvalue of the original integral equation.

Details of the argument just sketched are given in a series of lemmas in the next section, followed by the proof of the main theorem. Since the existence proof makes heavy use of asymptotic inequalities, it does not generally provide a practical technique for obtaining numerical results. The important practical question of finding approximate expressions, valid for large k, for the eigenfunctions and eigenvalues of equations such as (1) is a separate problem, as is also the question whether any particular equation has a finite or infinite number of eigenvalues.

For a gas laser with finite (not strip) mirrors of arbitrary, dissimilar shape and size, the integral equation still has a complex symmetric kernel,[2] although the domain of integration is two-dimensional and the kernel is more complicated than that of (1). The existence of eigenvalues in the most general case still remains to be settled.

II. MATHEMATICAL DETAILS

We shall use the following notation referring to an $n \times n$ matrix:

$$A^{(n)} = (a_{ij}), \qquad i = 1, 2, \cdots, n; \qquad j = 1, 2, \cdots, n;$$

$$A^{(n)}(i) = \sum_{j=1}^{n} | a_{ij} |, \qquad i = 1, 2, \cdots, n; \tag{6}$$

$$S(A^{(n)}) = \sum_{i=1}^{n} A^{(n)}(i) = \sum_{i=1}^{n} \sum_{j=1}^{n} | a_{ij} |.$$

If the superscript is omitted, n is understood to be infinite.

Lemma 1:

$$| \mathit{det}\, A^{(n)} | \leqq \prod_{i=1}^{n} A^{(n)}(i). \tag{7}$$

Proof: Using Hadamard's inequality,

$$\begin{aligned} \mid \det A^{(n)} \mid &\leqq \prod_{i=1}^{n} \left[\sum_{j=1}^{n} \mid a_{ij} \mid^{2} \right]^{1/2} \\ &\leqq \prod_{i=1}^{n} \left[\left(\sum_{j=1}^{n} \mid a_{ij} \mid \right)^{2} \right]^{1/2} = \prod_{i=1}^{n} A^{(n)}(i). \end{aligned} \tag{8}$$

Lemma 2:

$$\begin{aligned} &\mid det(A^{(n)} + B^{(n)}) - det\, A^{(n)} \mid \\ &\qquad \leqq \prod_{i=1}^{n} [A^{(n)}(i) + B^{(n)}(i)] - \prod_{i=1}^{n} A^{(n)}(i). \end{aligned} \tag{9}$$

Proof: The lemma is obviously true for $n = 1$. To proceed by induction, assume it is true for all determinants of order $n - 1$, and expand the determinants in (9) by minors of the first row. Let C_{1j} be the algebraic complement of $a_{1j} + b_{1j}$ in $A^{(n)} + B^{(n)}$, and let A_{1j} be the algebraic complement of a_{1j} in $A^{(n)}$. Then

$$\begin{aligned} \det(A^{(n)} + B^{(n')}) &= \sum_{j=1}^{n} (a_{1j} + b_{1j})C_{1j} \\ &= \det A^{(n)} + \sum_{j=1}^{n} a_{1j}(C_{1j} - A_{1j}) + \sum_{j=1}^{n} b_{1j}C_{1j}\,. \end{aligned} \tag{10}$$

By Lemma 1,

$$\begin{aligned} \mid C_{1j} \mid &\leqq \prod_{i=2}^{n} \left[\sum_{k=1}^{n} \mid a_{ik} + b_{ik} \mid \right] \\ &\leqq \prod_{i=2}^{n} [A^{(n)}(i) + B^{(n)}(i)]. \end{aligned} \tag{11}$$

By the inductive hypothesis,

$$\mid C_{1j} - A_{1j} \mid \leqq \prod_{i=2}^{n} [A^{(n)}(i) + B^{(n)}(i)] - \prod_{i=2}^{n} A^{(n)}(i). \tag{12}$$

where we have used the fact that the right-hand side is increasing as a function of the $A^{(n)}(i)$ and $B^{(n)}(i)$. Hence (10) gives

$$
\begin{aligned}
| \det(A^{(n)} + B^{(n)}) - \det A^{(n)} | \\
&\leqq A^{(n)}(1) \left\{ \prod_{i=2}^{n} [A^{(n)}(i) + B^{(n)}(i)] - \prod_{i=2}^{n} A^{(n)}(i) \right\} \\
&\quad + B^{(n)}(1) \prod_{i=2}^{n} [A^{(n)}(i) + B^{(n)}(i)] \qquad (13) \\
&= \prod_{i=1}^{n} [A^{(n)}(i) + B^{(n)}(i)] - \prod_{i=1}^{n} A^{(n)}(i),
\end{aligned}
$$

and the induction is complete.

Now let $\mathfrak{B}$ be the Banach space* whose elements are all bounded sequences of complex numbers, e.g.,

$$x = (x_1, x_2, \cdots, x_i, \cdots) \qquad (14)$$

with norm

$$\| x \| = \sup_i | x_i |. \qquad (15)$$

Let A be a linear matrix operator on the space $\mathfrak{B}$, defined by

$$(Ax)_i = \sum_{j=1}^{\infty} a_{ij} x_j, \qquad i = 1, 2, \cdots. \qquad (16)$$

Ax will be an element of $\mathfrak{B}$ provided that $\sup_i A(i)$ is finite. The norm of A is defined by

$$\| A \| = \sup \{ \| Ax \|; \| x \| = 1 \}, \qquad (17)$$

and it is easy to show that

$$\| A \| = \sup_i A(i). \qquad (18)$$

Henceforth we shall restrict our attention to matrix operators for which

$$S(A) \equiv \sum_{i=1}^{\infty} A(i) < \infty. \qquad (19)$$

Such operators are completely continuous, because they can be approximated by the sequence $\{A^{(n)}\}$ of completely continuous operators which converges in norm to A. Here $A^{(n)}$ is a matrix whose elements co-

* The standard definitions and theorems which we shall require from functional analysis may be found in Kolmogorov and Fomin.[5]

incide with those of A for $1 \leqq i \leqq n$ and $1 \leqq j \leqq n$, and are zero otherwise.

A complex number λ is said to be in the *spectrum* of an operator A if the operator $A - \lambda I$ has no inverse. An *eigenvalue* of A is any value of λ for which there exists a nonzero x satisfying the homogeneous equation

$$Ax - \lambda x = 0. \tag{20}$$

If A is completely continuous and if λ ($\neq 0$) lies in the spectrum of A, then λ is an eigenvalue of A. In finite-dimensional space the eigenvalues are the latent roots of the matrix $A^{(n)}$; that is, they are the roots of the characteristic equation

$$\det (A^{(n)} - \lambda I^{(n)}) = 0. \tag{21}$$

Lemma 3: If $A^{(n)}$ has λ as an eigenvalue, then $A^{(n)} + B^{(n)}$ has λ', where

$$| \lambda - \lambda' | \leqq \left\{ \prod_{i=1}^{n} [A^{(n)}(i) + B^{(n)}(i) + | \lambda |] - \prod_{i=1}^{n} [A^{(n)}(i) + | \lambda |] \right\}^{1/n}. \tag{22}$$

Proof: Denote the eigenvalues of $A^{(n)} + B^{(n)}$ by $\lambda_1, \lambda_2, \cdots, \lambda_n$. Then

$$\begin{aligned} &| (\lambda - \lambda_1)(\lambda - \lambda_2) \cdots (\lambda - \lambda_n) | \\ &\quad = | \det (A^{(n)} + B^{(n)} - \lambda I^{(n)}) - \det (A^{(n)} - \lambda I^{(n)}) |, \end{aligned} \tag{23}$$

the second determinant being equal to zero because λ is an eigenvalue of $A^{(n)}$. Let

$$D^{(n)} = A^{(n)} - \lambda I^{(n)}, \tag{24}$$

so that

$$D^{(n)}(i) = \sum_{j=1}^{n} | a_{ij} - \lambda \delta_{ij} | \leqq A^{(n)}(i) + | \lambda |. \tag{25}$$

Then, using Lemma 2,

$$\begin{aligned} \prod_{k=1}^{n} | \lambda - \lambda_k | &\leqq \prod_{i=1}^{n} [D^{(n)}(i) + B^{(n)}(i)] - \prod_{i=1}^{n} D^{(n)}(i) \\ &\leqq \prod_{i=1}^{n} [A^{(n)}(i) + B^{(n)}(i) + | \lambda |] \\ &\qquad - \prod_{i=1}^{n} [A^{(n)}(i) + | \lambda |], \end{aligned} \tag{26}$$

since the right side of the first line is an increasing function of $D^{(n)}(i)$. It follows from (26) that for at least one of the factors $|\lambda - \lambda_k|$ the inequality (22) holds.

Lemma 4: Let A be an infinite matrix with $S(A) < \infty$. Suppose that from the eigenvalues of the sequence of finite matrices $\{A^{(n)}\}$ we can pick a sequence $\{\lambda^{(n)}\}$ such that $\lambda^{(n)}$ does not approach zero as $n \to \infty$. Then A has a nonzero eigenvalue.

Proof: The $\lambda^{(n)}$ are bounded, since in fact

$$|\lambda^{(n)}| \leqq \| A^{(n)} \| = \max_i A^{(n)}(i) \leqq S(A). \tag{27}$$

Also for sufficiently large n we can pick a subsequence which is bounded away from zero, and which therefore has at least one nonzero limit point. Suppose that the subsequence $\lambda^{(p)}$ converges to the limit point $\lambda \neq 0$, as p runs through some increasing sequence of integers. We assert that λ is an eigenvalue of A. If it were not so, then $(A - \lambda I)^{-1}$ would exist and therefore be bounded. Suppose $(A - \lambda I)^{-1}$ were bounded, and let $x^{(p)}$ be the characteristic vector of $A^{(p)}$ corresponding to $\lambda^{(p)}$. Then we would have

$$\begin{aligned} x^{(p)} &= (A - \lambda I)^{-1}(A - \lambda I)x^{(p)} \\ &= (A - \lambda I)^{-1}[A^{(p)}x^{(p)} - \lambda^{(p)}x^{(p)} \\ &\qquad + (A - A^{(p)})x^{(p)} - (\lambda - \lambda^{(p)})x^{(p)}] \\ &= (A - \lambda I)^{-1}[(A - A^{(p)})x^{(p)} - (\lambda - \lambda^{(p)})x^{(p)}], \end{aligned} \tag{28}$$

where in the last equation $A^{(p)}$ represents an infinite matrix which coincides with A in a square of side p in the upper left corner, and has zeros elsewhere. Taking norms, we have

$$\begin{aligned} \| x^{(p)} \| &\leqq \|(A - \lambda I)^{-1}\| \, \|(A - A^{(p)})x^{(p)} - (\lambda - \lambda^{(p)})x^{(p)}\| \\ &\leqq \|(A - \lambda I)^{-1}\| \, [\| A - A^{(p)} \| + |\lambda - \lambda^{(p)}|] \, \| x^{(p)} \|, \end{aligned} \tag{29}$$

or

$$\| (A - \lambda I)^{-1} \| \geqq \frac{1}{\| A - A^{(p)} \| + |\lambda - \lambda^{(p)}|}. \tag{30}$$

But since both $\| A - A^{(p)} \|$ and $|\lambda - \lambda^{(p)}|$ go to zero as $p \to \infty$, we derive a contradiction.

Theorem: Let A be an infinite matrix with $S(A) < \infty$ and with $Tr(A) \neq 0$. If

$$S(A) - S(A^{(n)}) < (c/n^{\epsilon})^n, \tag{31}$$

for some c, $\epsilon > 0$, then A has a nonzero eigenvalue.

Proof: Since $\mathrm{Tr}(A) \neq 0$ and $\mathrm{Tr}(A^{(n)}) \rightarrow \mathrm{Tr}(A)$, it follows that for $n \geqq n_1$ (say) and some $\delta > 0$, we have $|\mathrm{Tr}(A^{(n)})| \geqq \delta$. Since the trace is the sum of the eigenvalues, $A^{(n)}$ must have at least one eigenvalue $\lambda^{(n)}$ such that

$$| \lambda^{(n)} | \geqq \delta/n. \tag{32}$$

We shall in fact show that if n_1 is a sufficiently large fixed integer, and if

$$n_j = 2^{j-1} n_1, \qquad j = 1, 2, 3, \cdots \tag{33}$$

then for each j there exists an eigenvalue which is *uniformly* bounded away from zero, i.e.,

$$\lambda^{(n_j)} \geqq \delta/2n_1. \tag{34}$$

Then by Lemma 4 the theorem will be proved.

We substitute into Lemma 3 as follows:

$$\begin{aligned} n &= n_{j+1}, \\ |\lambda| &= |\lambda^{(n_j)}| = t, \\ A^{(n)} &= A^{(n_j)}, \\ B^{(n)} &= A^{(n_{j+1})} - A^{(n_j)}, \end{aligned} \tag{35}$$

where it is understood that $A^{(n_j)}$ now represents the original matrix $A^{(n_j)}$ augmented below and to the right with enough zeros to give it dimensions $n_{j+1} \times n_{j+1}$. Then (22) becomes

$$\begin{aligned} &|\lambda^{(n_j)} - \lambda^{(n_{j+1})}| \\ &\quad \leqq \left\{ \prod_{i=1}^{n_{j+1}} [A^{(n_{j+1})}(i) + t] - t^{n_{j+1}-n_j} \prod_{i=1}^{n_j} [A^{(n_j)}(i) + t] \right\}^{1/n_{j+1}} \\ &\quad \leqq \left\{ \prod_{i=1}^{n_{j+1}} [A(i) + t] - t^{n_{j+1}-n_j} \prod_{i=1}^{n_j} [A^{(n_j)}(i) + t] \right\}^{1/n_{j+1}}. \end{aligned} \tag{36}$$

Since

$$|\lambda^{(n_j)} - \lambda^{(n_{j+1})}| \geqq t - |\lambda^{(n_{j+1})}|, \tag{37}$$

we can rearrange (36) to get

$$|\lambda^{(n_{j+1})}| \geqq t - \left\{ \prod_{i=1}^{n_{j+1}} [A(i) + t] - t^{n_{j+1}-n_j} \prod_{i=1}^{n_j} [A^{(n_j)}(i) + t] \right\}^{1/n_{j+1}}. \tag{38}$$

Hence

$$\begin{aligned}\frac{|\lambda^{(n_{j+1})}|}{|\lambda^{(n_j)}|} &\geqq 1 - \left\{\prod_{i=1}^{n_{j+1}}\left[1 + \frac{A(i)}{t}\right] - \prod_{i=1}^{n_j}\left[1 + \frac{A^{(n_j)}(i)}{t}\right]\right\}^{1/n_{j+1}} \\ &\geqq 1 - \left\{\prod_{i=1}^{n_{j+1}}\left[1 + \frac{n_jA(i)}{\delta}\right] - \prod_{i=1}^{n_j}\left[1 + \frac{n_jA^{(n_j)}(i)}{\delta}\right]\right\}^{1/n_{j+1}}, \qquad (39)\end{aligned}$$

since we already know that $t \geqq \delta/n_j$.

Now consider

$$\prod_{i=1}^{n_j}\left[1 + \frac{n_jA(i)}{\delta}\right] \leqq \prod_{i=1}^{n_j} \exp\left[\frac{n_jA(i)}{\delta}\right] \leqq \exp\left[\frac{n_jS(A)}{\delta}\right]. \qquad (40)$$

Also

$$\begin{aligned}\prod_{i=n_j+1}^{n_{j+1}}\left[1 + \frac{n_jA(i)}{\delta}\right] &\leqq \exp\left[\frac{n_j}{\delta}\sum_{i=n_j+1}^{\infty} A(i)\right] \\ &\leqq \exp \frac{n_j}{\delta}\,[S(A) - S(A^{(n_j)})] \qquad (41) \\ &\leqq \exp\left[\frac{n_j}{\delta}\left(\frac{c}{n_j^{\epsilon}}\right)^{n_j}\right] \leqq 1 + \frac{2n_j}{\delta}\left(\frac{c}{n_j^{\epsilon}}\right)^{n_j},\end{aligned}$$

provided that n_1 and hence n_j are sufficiently large, where in the next to last step we have used (31) and in the last step we have used $e^x \leqq 1 + 2x$ for $0 \leqq x \leqq 1$, say. Finally,

$$\begin{aligned}&\prod_{i=1}^{n_j}\left[1 + \frac{n_jA^{(n_j)}(i)}{\delta}\right] \\ &\quad = \prod_{i=1}^{n_j}\left[1 + \frac{n_j\,A(i)}{\delta} - \frac{n_j}{\delta}\,\{A(i) - A^{(n_j)}(i)\}\right] \\ &\quad \geqq \prod_{i=1}^{n_j}\left[1 + \frac{n_jA(i)}{\delta}\right] \\ &\qquad\qquad - \frac{n_j}{\delta}\sum_{k=1}^{n_j}\left\{[A(k) - A^{(n_j)}(k)]\prod_{i=1}^{n_j}\left[1 + \frac{n_jA(i)}{\delta}\right]\right\} \qquad (42) \\ &\quad \geqq \prod_{i=1}^{n_j}\left[1 + \frac{n_jA(i)}{\delta}\right] \\ &\qquad\qquad - \frac{n_j^2}{\delta}\,[S(A) - S(A^{(n_j)})]\prod_{i=1}^{n_j}\left[1 + \frac{n_jA(i)}{\delta}\right] \\ &\quad \geqq \left\{1 - \frac{n_j^2}{\delta}\left(\frac{c}{n_j^{\epsilon}}\right)^{n_j}\right\}\prod_{i=1}^{n_j}\left[1 + \frac{n_jA(i)}{\delta}\right].\end{aligned}$$

Substituting (40), (41), and (42) into (39) yields

$$\begin{aligned}\frac{|\lambda^{(n_{j+1})}|}{|\lambda^{(n_j)}|} &\geqq 1 - \left\{\frac{n_j(2+n_j)}{\delta}\left(\frac{c}{n_j^{\epsilon}}\right)^{n_j} \exp\frac{n_jS(A)}{\delta}\right\}^{1/n_{j+1}} \\ &= 1 - \left[\frac{n_j(2+n_j)}{\delta}\right]^{1/2n_j} \frac{c_1}{n_j^{\epsilon/2}},\end{aligned} \tag{43}$$

where in the last step we have used the fact that $n_{j+1} = 2n_j$ and have set

$$c_1 = c^{1/2} \exp\,[S(A)/2\delta]. \tag{44}$$

If we assume in advance that

$$\delta \leqq 2, \qquad n_1 \geqq \max\,(2, 4/\epsilon), \tag{45}$$

then

$$\begin{aligned}\left[\frac{n_j(2+n_j)}{\delta}\right]^{1/2n_j} \frac{c_1}{n_j^{\epsilon/2}} &\leqq \left[\frac{2n_j^{\,2}}{\delta}\right]^{1/2n_j} \frac{c_1}{n_j^{\epsilon/2}} \\ &\leqq \frac{2c_1}{\delta n_j^{\epsilon/4}} = \frac{c_2}{2^{(j-1)\epsilon/4}} = c_2r^{j-1},\end{aligned} \tag{46}$$

where

$$c_2 = \frac{2c_1}{\delta n_1^{\epsilon/4}}, \qquad r = 2^{-\epsilon/4} < 1. \tag{47}$$

Hence (43) and (46) imply

$$\frac{|\lambda^{(n_{j+1})}|}{|\lambda^{(n_j)}|} \geqq 1 - c_2r^{j-1}, \tag{48}$$

and by induction

$$\frac{|\lambda^{(n_J)}|}{|\lambda^{(n_1)}|} \geqq \prod_{j=1}^{J-1} [1 - c_2r^{j-1}]. \tag{49}$$

But if $c_2 \leqq \tfrac{1}{2}$, say, then

$$\begin{aligned}\prod_{j=1}^{\infty} (1 - c_2r^{j-1}) &= \exp\left[\sum_{j=1}^{\infty} \log\,(1 - c_2r^{j-1})\right] \\ &\geqq \exp\left[-2\sum_{j=1}^{\infty} c_2r^{j-1}\right] = \exp\left[-\frac{2c_2}{1-r}\right] > \tfrac{1}{2},\end{aligned} \tag{50}$$

where the last step requires

$$c_2 < \tfrac{1}{2}(1-r)\log 2 = \tfrac{1}{2}(1 - 2^{-\epsilon/4})\log 2, \tag{51}$$

and by (47) this inequality can always be satisfied for large enough n_1. But (49) and (50) imply

$$\lambda^{(n_j)} \geqq \tfrac{1}{2}\lambda^{(n_1)} \geqq \delta/(2n_1) > 0 \tag{52}$$

for all j, and so the theorem follows from Lemma 4. Q.E.D.

An integral function of finite order ρ is a function $F(z)$ which has no singularities in any finite region of the z-plane, and whose maximum modulus $M(r)$ on the circle $|z| = r$ satisfies

$$\log M(r) < r^k \tag{53}$$

for all sufficiently large r when $k > \rho$, but not when $k < \rho$. Such a function may be expanded in a Taylor series,

$$F(z) = \sum_{n=0}^{\infty} a_n z^n, \tag{54}$$

which converges for all z, and whose coefficients satisfy[6]

$$|a_n| < 1/n^{n\epsilon} \tag{55}$$

for all sufficiently large n, where ϵ is any fixed number less than $1/\rho$. Alternatively, for any fixed $\epsilon < 1/\rho$, there exists a constant c such that for all $n > 0$

$$|a_n| \leqq \left[\frac{c}{(n+1)^{\epsilon}}\right]^{n+1}. \tag{56}$$

We are now ready to prove the result stated in Section I.

Theorem: Let $G(x)$ and $H(x)$ be any bounded functions on the interval $-1 \leqq x \leqq 1$, and let $F(z)$ be any integral function of finite order such that

$$\int_{-1}^{1} G(x)F(x^2)H(x)dx \neq 0. \tag{57}$$

Then the integral equation

$$\int_{-1}^{1} G(x)F(xy)H(y)f(y)dy = \lambda f(x) \tag{58}$$

has at least one nonzero eigenvalue.

Proof: Expand $F(xy)$ in a Taylor series, so that the integral equation becomes

$$\int_{-1}^{1} \sum_{n=1}^{\infty} [a_{n-1}^{1/2} G(x)x^{n-1}][a_{n-1}^{1/2} H(y)y^{n-1}]f(y)dy = \lambda f(x). \tag{59}$$

Let

$$f(x) = G(x) \sum_{n=1}^{\infty} f_n \, a_{n-1} x^{n-1}, \tag{60}$$

where $\{f_n\}$ is a bounded sequence of complex numbers; the a_n's tend to zero fast enough so that $f(z)/G(z)$ will be an integral function of finite order.

Since the powers of x are linearly independent, (59) is equivalent to the matrix equation

$$Af = \lambda f, \tag{61}$$

where

$$a_{ij} = a_{ji} = (a_{i-1}a_{j-1})^{1/2} \int_{-1}^{1} G(t)H(t)t^{i+j-2}dt, \tag{62}$$

$$i = 1, 2, \cdots; \qquad j = 1, 2, \cdots .$$

Since $G(x)$ and $H(x)$ are bounded in $-1 \leqq x \leqq 1$ and the Taylor coefficients of $F(z)$ satisfy (56), it is clear that

$$| \, a_{ij} \, | \leqq \frac{M}{i + j - 1} \left(\frac{c}{i^{\epsilon}}\right)^{i/2} \left(\frac{c}{j^{\epsilon}}\right)^{j/2} . \tag{63}$$

In preparation for an application of the preceding theorem, consider

$$\begin{aligned} S(A) - S(A^{(n)}) &\leqq 2 \sum_{i=n+1}^{\infty} \sum_{j=1}^{i} \frac{M}{i + j - 1} \left(\frac{c}{i^{\epsilon}}\right)^{i/2} \left(\frac{c}{j^{\epsilon}}\right)^{j/2} \\ &= 2M \sum_{i=n+1}^{\infty} \left[\left(\frac{c}{i^{\epsilon}}\right)^{i/2} \sum_{j=1}^{i} \frac{1}{i + j - 1} \left(\frac{c}{j^{\epsilon}}\right)^{j/2} \right]. \end{aligned} \tag{64}$$

Now $(c/j^{\epsilon})^{j/2}$ is bounded as $j \to \infty$, and

$$\sum_{j=1}^{i} \frac{1}{i + j - 1} \leqq \int_{i-1}^{2i-1} \frac{dx}{x} = \log \left[\frac{2i - 1}{i - 1}\right], \tag{65}$$

which is bounded for $i \geqq n + 1 \geqq 2$. Hence with a new bounding constant we have

$$S(A) - S(A^{(n)}) \leqq M_1 \sum_{i=n+1}^{\infty} \left(\frac{c^{1/2}}{i^{\epsilon/2}}\right)^{i} . \tag{66}$$

Choose $\log n \geqq (2 + \log c)/\epsilon$, so that $n^{\epsilon} \geqq ce^{2}$; then

$$\begin{aligned} \sum_{i=n+1}^{\infty} \left(\frac{c^{1/2}}{i^{\epsilon/2}}\right)^{i} &\leqq \int_{n}^{\infty} \left(\frac{c^{1/2}}{x^{\epsilon/2}}\right)^{x} dx \leqq \int_{n}^{\infty} \left(\frac{c^{1/2}}{n^{\epsilon/2}}\right)^{x} dx \\ &= -\frac{(c/n^{\epsilon})^{n/2}}{(\log c - \epsilon \log n)/2} \leqq \left(\frac{c^{1/2}}{n^{\epsilon/2}}\right)^{n}, \end{aligned} \tag{67}$$

and so from (66)

$$S(A) - S(A^{(n)}) \leqq \left(\frac{c_1}{n^{\epsilon_1}}\right)^n, \tag{68}$$

where c_1 is a new bounding constant and $\epsilon_1 = \epsilon/2$.

Finally we have

$$\begin{aligned} \operatorname{Tr}(A) &= \sum_{i=1}^{\infty} a_{ii} = \sum_{i=1}^{\infty} a_{i-1} \int_{-1}^{1} G(t)H(t)t^{2i-2}\,dt \\ &= \int_{-1}^{1} G(t)H(t)F(t^2)\,dt, \end{aligned} \tag{69}$$

and this does not vanish by hypothesis. Hence all the conditions of the previous theorem are satisfied, and the integral equation has a nonzero eigenvalue. Q.E.D.

Since exp $(-2ikz)$ is an integral function of finite order 1, it is an obvious corollary that the kernel exp $i[k(x-y)^2 - h(x) - h(y)]$ has a nonzero eigenvalue for arbitrary complex k, provided only that $h(x)$ is bounded and that

$$\int_{-1}^{1} e^{-2ih(x)}\,dx \neq 0. \tag{70}$$

Furthermore if $h(x)$ is an even function of x and if $f(x)$ is an even function which satisfies

$$\int_0^1 \exp\{i[k(x^2+y^2) - h(x) - h(y)]\} \cos(2kxy)f(y)dy = \tfrac{1}{2}\lambda f(x), \tag{71}$$

then $f(x)$ also satisfies (1). But the theorem just proved obviously holds for arbitrary finite limits of integration and applies to the kernel of (71), so (71) has at least one nonzero eigenvalue if

$$\int_0^1 \exp\{2i[kx^2 - h(x)]\} \cos(2kx^2)dx \neq 0. \tag{72}$$

Similarly if $h(x)$ is even and if $f(x)$ is an odd function which satisfies

$$\int_0^1 \exp\{i[k(x^2+y^2) - h(x) - h(y)]\} \sin(2kxy)f(y)dy = \tfrac{1}{2}i\lambda f(x), \tag{73}$$

then $f(x)$ also satisfies (1), and (73) has at least one nonzero eigenvalue if

$$\int_0^1 \exp\{2i[kx^2 - h(x)]\} \sin(2kx^2)dx \neq 0. \tag{74}$$

At least one of (72) and (74) will be satisfied whenever (70) holds. Except for certain particular values of k, one of which is evidently $k = 0$, both (72) and (74) will be satisfied, and (1) will have at least two distinct eigenfunctions corresponding to nonzero eigenvalues.

REFERENCES

1. Fox, A. G., and Li, T., Resonant Modes in a Maser Interferometer, B.S.T.J., **40,** March, 1961, pp. 453–488.
2. Fox, A. G., and Li, T., Modes in a Maser Interferometer with Curved and Tilted Mirrors, Proc. IEEE, **51,** January, 1963, pp. 80–89.
3. Boyd, G. D., and Gordon, J. P., Confocal Multimode Resonator for Millimeter through Optical Wavelength Masers, B.S.T.J., **40,** March, 1961, pp. 489–508.
4. Sz.-Nagy, B., Perturbations des Transformations Linéaires Fermées, Acta Sci. Math. Szeged, **14,** 1951, pp. 125–137.
5. Kolmogorov, A. N., and Fomin, S. V., *Elements of the Theory of Functions and Functional Analysis* (tr. Boron, L. F.), Graylock Press, Rochester, 1957, Chs. 3–4.
6. Copson, E. T., *Theory of Functions of a Complex Variable*, Oxford University Press, London, 1935, pp. 175–178.

Deposition of Tantalum Films with an Open-Ended Vacuum System

By J. W. BALDE, S. S. CHARSCHAN, and J. J. DINEEN

(Manuscript received July 19, 1963)

New devices using vacuum-deposited metal films require a high-speed, low-cost method of vacuum deposition. The capability of the open-ended multiple-chamber deposition equipment has been investigated to determine its suitability for depositing tantalum nitride thin films. This was accomplished by examining the measurable electrical properties of the deposited film and by determining the stability of resistors made from these films.

Tantalum films produced by the open-ended deposition system were found comparable to those produced by many bell-jar systems. It was possible to control the addition of nitrogen to the films, and tantalum nitride films of satisfactory stability were obtained. Because the open-ended deposition method can produce large quantities of suitable thin films, it is expected that this will be an important process in the manufacture of future products.

I. INTRODUCTION

Tantalum thin film circuit techniques developed at Bell Telephone Laboratories[1] can produce resistor and capacitor circuit elements and associated interconnections. Such tantalum film circuits have high stability and good reliability, superior to that of discrete components with their multiple interconnections.[2]

The Western Electric Company has developed a continuous open-ended vacuum system for deposition of these tantalum films. This system provides for the passage of substrates through a sequence of chambers which vary in pressure from atmospheric pressure to high vacuum and then back to atmospheric pressure. The design of this system and the details of its operation have been previously reported.[3]

This open-ended system has advantages for quantity deposition of thin films. All vacuum chambers remain at their operating pressures; no time is lost pumping down prior to deposition. Work chambers need not be exposed to room atmosphere and possible contamination. Degassing and preheating operations can be restricted to the substrates and associated carriers; repeated degassing of the system is unnecessary. Substrate motion is continuous through the system; no operator handling or manipulation is required.

The open-ended deposition process differs in a number of ways from earlier work with batch processes using bell-jar vacuum chambers. Chamber materials and hardware are very different from those developed for round bell-jar enclosures. Substrates move through the sputtering glow zone, continuously passing the cathode. This motion produces thermal gradients which result from the dynamic equilibrium conditions for a given substrate speed. Deposited films are the result of an integration of the effect of each part of the cathode, rather than the result of a static pattern of deposition. Film thickness can be controlled by the length of chamber and the speed of substrate motion as well as by deposition rate.

II. TEST PROCEDURE

To investigate the effects of these changed deposition conditions, the product of the open-ended machine was examined to ascertain whether the films have satisfactory properties, and also to determine that there was no adverse effect on the subsequent processing operations. The evaluation of the quality of film deposition in the open-ended system consisted of the following parts:

First, examination was made of the tantalum film deposited without any intentional nitrogen addition. The properties of tantalum film could be strongly altered by contaminant gases from atmospheric leaks or by outgassing of material in the sputtering chamber. Examination of this tantalum film quality should reveal any inadequate cleaning or adverse effect from the deposition method.

Second, the properties of the films were examined as a function of the amount of nitrogen added to the sputtering atmosphere. This establishes the ability to add sufficient nitrogen to produce useful resistor films, as demonstrated by stability, resistivity, and temperature coefficient measurements.

Third, the reproducibility and control of the tantalum nitride deposition process were examined by repeat depositions at the same operating point, and by the examination of many depositions which deviated only slightly from the operating point for most suitable film properties.

Fourth, an examination was made of uniformity of deposition over the width and length of the substrate.

III. MEASUREMENT PROCEDURE

Satisfactory film quality is judged initially by measuring three film properties: thickness (Å), specific resistivity (ρ), and the temperature coefficient of resistance (α). In order to insure that the variability of film properties is due to the machine processing system and not to errors in the *measurement* of the properties, the test details and procedures were evaluated.

A test pattern was developed to insure that all films would have their properties measured on the same effective area and at the same position on the substrate. The zigzag test pattern for a 1.5-inch by 3-inch substrate is shown in Fig. 1. It consists of 20 resistors with a nominal line width of 0.015 inch, each having a path length of 144 squares. The resistors are interconnected by a center stripe and have separate terminal tabs for each resistor. The test resistors are defined by using silk-screen techniques to apply a resist to a tantalum-coated substrate. The unwanted film is removed by etching.

3.1 *Film Thickness Measurement*

In preparing films for thickness measurements, hot sodium hydroxide is used to remove the unwanted tantalum film without appreciable etch

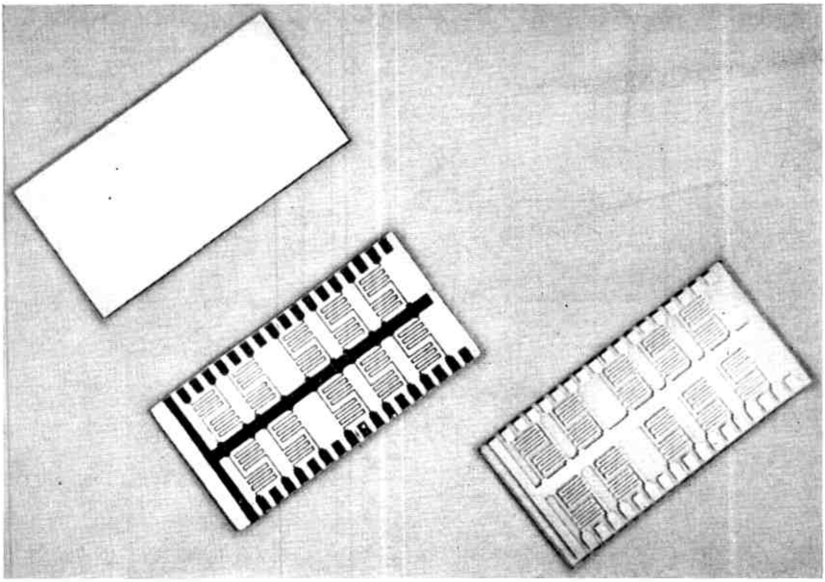


Fig. 1 — Resistor pattern for film property evaluation.

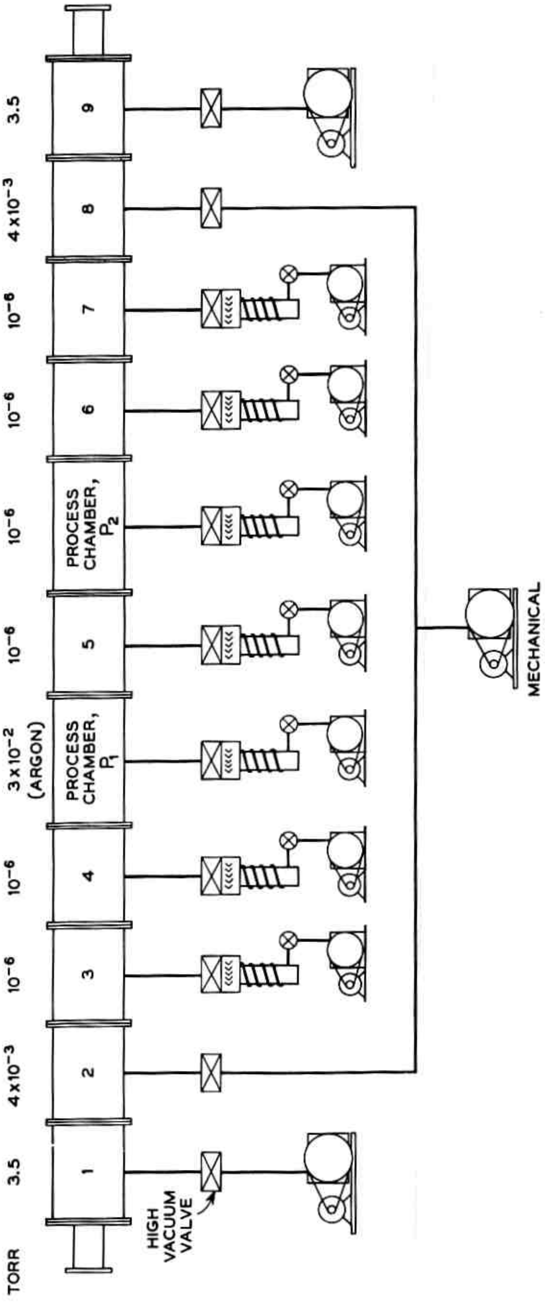


Fig. 2 — Open-ended vacuum system, dynamic operating conditions.

of the glass substrate surface beneath the deposited film. After the resist has been removed, the films are measured using a Talysurf instrument.[4] For thickness measurements of the 1200-Å films deposited in this open-ended vacuum system, the 1σ error of measurement is 56 Å.

3.2 *Specific Resistivity*

The specific resistivity is computed as follows

$$\rho = R_s \text{Å} \times 10^{-2} \text{ microohm-cm}$$

where R_s is sheet resistance in ohms per square and Å is thickness in angstroms.

The sheet resistance of an unetched film is determined by a four-point probe measurement in ohms per square. For convenience, these measurements are made using a simplified direct-reading meter of 1 per cent accuracy.

3.3 *Temperature Coefficient of Resistance Measurement*

After the test resistor pattern has been defined by etching, connections are made to the center stripe and the appropriate tab areas. The resistance is measured at 30°C and at 60°C. The temperature coefficient of resistance is then computed as follows:

$$\text{TCR}(\alpha) = \frac{R_{60} - R_{30}}{R_{30}\Delta T} \times 10^6 \text{ ppm/°C}$$

where R_{30} and R_{60} are in ohms and ΔT is in degrees centigrade. Error of measurement studies indicate a 1σ error of 3 ppm/°C.

IV. ANALYSIS OF UNDOPED TANTALUM FILM

In order to show that the machine process is reproducible at a useful quality level, a series of experiments were run. For this experimental work, one 1.5-inch by 3-inch coated lime glass slide was produced per minute. A carrier 5 inches in length was used to bring the substrate through the chambers. The chamber lengths were such that the carrier and substrate remained in the first four chambers for a total of 15 minutes of high temperature preheating at four decreasing pressure levels. The pressure levels used for this experiment are shown in Fig. 2. Table I gives the preheating power and the sputtering conditions used.

The results of these experiments, shown in Fig. 3, indicate that films deposited in this manner have a specific resistivity of 240 microohm-cm

TABLE I — EXPERIMENTAL OPERATING CONDITIONS

Preheat Stations	#1	#2	#3	#4
Preheat lamp input, watts	300	300	300	220
Sputtering potential, vdc	4500			
Sputtering current, ma	500			
Sputtering pressure, microns (gauge)	32			
Cathode-anode spacing, inches	2.0			
Experimental cathode area, in^2	158			
Deposition rate, Å/min	300			

and a temperature coefficient of resistivity of +56 ppm/°C at a nominal thickness of 1190 Å. The quality of these films is comparable to that obtained by batch processes using bell-jar systems.

4.1 *Process Controllability*

The process controllability for these films was estimated from control charts to have a standard deviation of 11 microohm-cm in specific resistivity and 27 ppm/°C in temperature coefficient of resistance. Film thickness was shown to be controllable, with a standard deviation of 50 Å about a mean of 1190 Å. Based on these results, the process was deemed to be controllable and reproducible for tantalum films.

V. NITROGEN DOPING

Tantalum films without intentional additives are used primarily to make capacitors. Work done by Gerstenberg and Mayer[5] has established that the *resistors* with the best stability were made when one to five per cent of nitrogen is added to the sputtering atmosphere, the amount depending on the pumping and geometry characteristics of the particular system. This nitrogen reacts with the tantalum, and the resulting film contains appreciable tantalum nitride. Having established that the open-ended vacuum deposition system could produce satisfactory tantalum films, it was next necessary to investigate the ability of the system to produce nitrided tantalum resistors with suitable component properties.

The properties of the films of tantalum nitride depend on the environment in the sputtering chamber. Geometry, voltage, current, pressure, gas composition, and gas thru-put all affect the film properties. Slight differences in chamber materials, glow region, gas flow paths, or thermal gradients can also have a major effect on the amount of nitrogen needed to produce film with satisfactory properties. It is customary, therefore, to investigate the relationships between film properties and nitrogen quantity in any new deposition system. This is done by experimentally

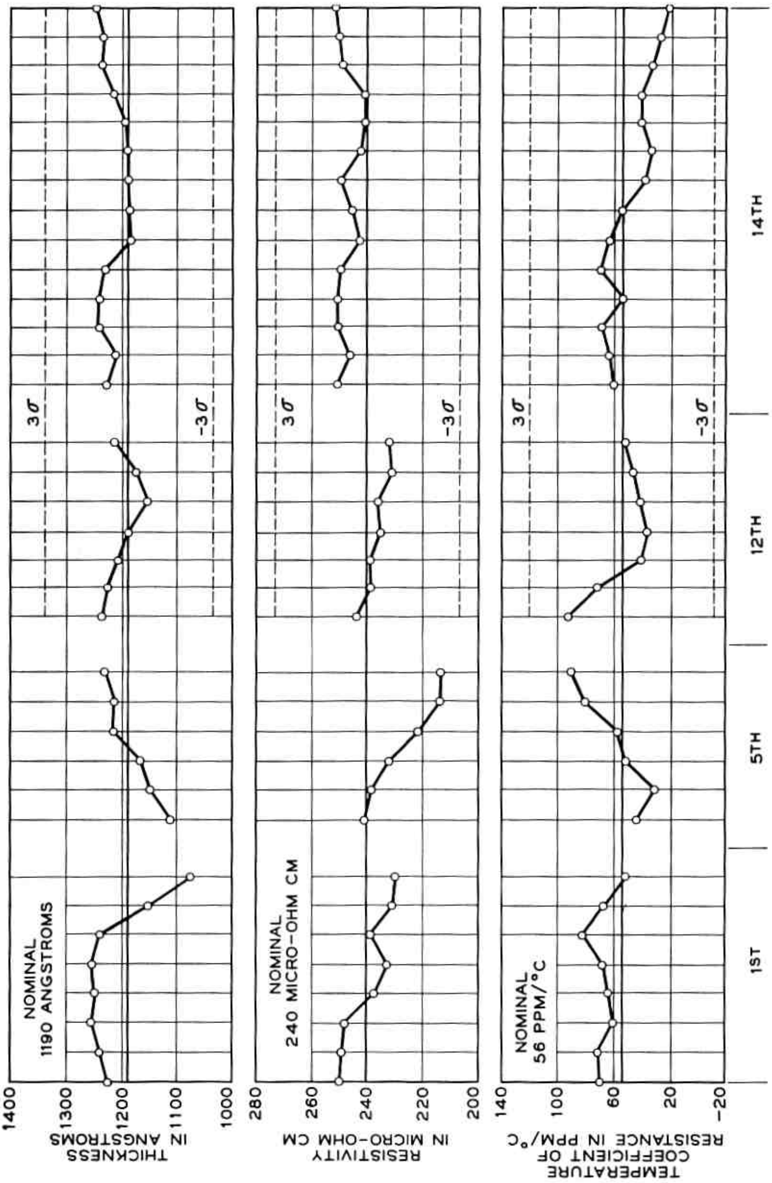


Fig. 3 — Characteristics of tantalum films deposited in the open-ended vacuum system.

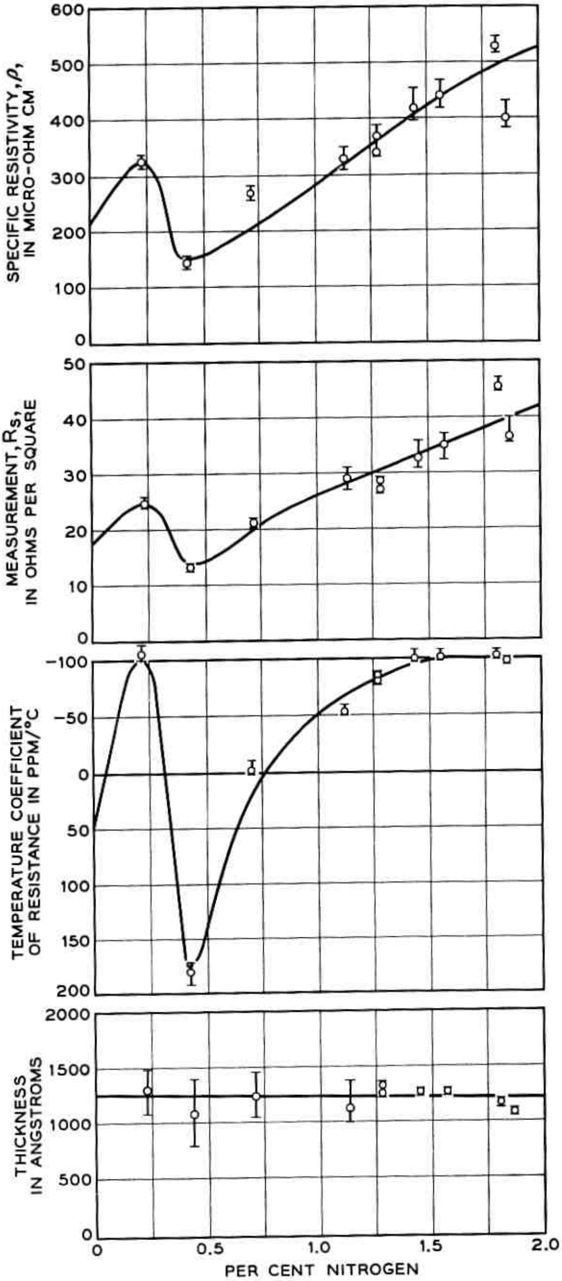


Fig. 4 — Nitrogen-doped product characteristics.

determining the characteristic curve for each of the important nitrogen-film property relationships. These characteristic curves must be determined for each vacuum system, and the proper operating point chosen for each. The influence of trace impurities of nitrogen in the open-ended vacuum system was therefore explored by a series of characterization experiments in the machine processing system. The experimental procedure did not materially differ from that used in the earlier undoped experiments. The operating conditions previously stated in Table I were again held in all cases. The only additive was the controlled flow of nitrogen gas, which was mixed with the argon prior to entering the sputtering chamber.

A single experiment, of the series used for this purpose, consisted of establishing an operating point by adjusting the flow of nitrogen gas until the sheet resistivity was some desired value, and holding it at that value to within ±1 ohm/square. Sample slides were sent through the machine at 10-minute intervals to determine that the sheet resistivity was in contıol, thus assuring that drifts were removed from the system. Then 20 consecutive slides were given a film deposition in the machine.

Each experimental lot was sampled as follows: four consecutive slides in the center of the lot were processed into resistors; four slides were used to determine the initial film characteristics; and four more were used to examine such physical properties as adhesion, visual defects, and the anodizability of these films. The remaining slides were held as spares for future exploratory studies.

5.1 *Nitrogen-Doped Film Characteristics*

The influence of nitrogen on the characteristics of these resistors after processing is shown by the curves in Fig. 4. The data presented here show that doped films from this machine processing system exhibit a characteristic form similar to that previously reported for tantalum nitride films produced in bell-jar systems.[6] Films with low resistivity and high positive temperature coefficient are formed in the vicinity of 0.30 to 0.40 per cent nitrogen.

5.2 *Accelerated Life Test Data*

The ultimate criteria for satisfactory films are the observed qualities of the circuit elements made from the films. Resistors made of tantalum and tantalum nitride should have a stability characteristic of less than 1 per cent drift in resistance in a 20-year lifetime. Accelerated aging tests, used by J. S. Fisher,[7] permit relative judgments to be made much earlier than 20 years — in fact, tests of standard pattern resistors at twice rated load can differentiate between performances of resistors in about 3 months.

The resistor pattern used for accelerated life testing consists of 24 resistors, each rated at 0.5 watt. This resistor pattern is shown in Fig. 5. Twelve resistors are arranged on each side of a common center strip on the 1.5-inch by 3-inch alkali-free glass substrate (Corning Code 7059). Each resistor is formed by a zig-zag pattern of lines 0.008 inch wide, containing 364 squares. The components are defined by using a conventional photo-resist (KMER)* and etched in a hydrofluoric-nitric acid mixture.

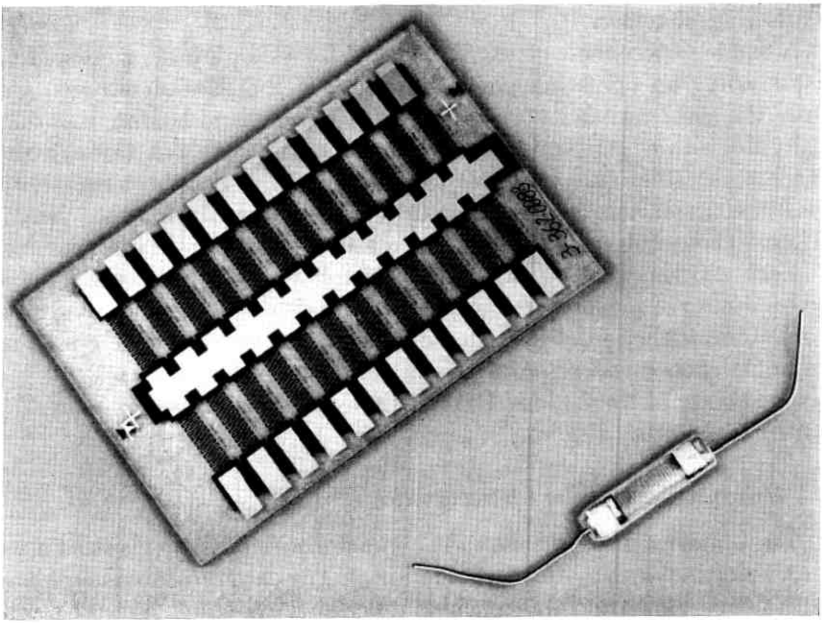


Fig. 5 — Product stability test pattern.

Nichrome and gold are evaporated in turn onto the terminal areas. The films are bath-anodized to 30 volts in citric acid.[8] Oven baking at 250°C in air for five hours is used to stabilize the films. Resistors are then separated into individual units and trim anodized to 15,000 ohms ±1 per cent wherever possible. For initial sheet resistance of greater than 40 ohms/square, it is necessary to trim anodize to a maximum of 20,000 ohms ±1 per cent.

The stability of resistors, for the range of nitrogen additive from 0.0 to 1.84 per cent, was studied by placing eight resistors under double-

* Kodak Metal Etch Resist, Eastman Kodak Company.

rated power life test, four from each of two slides in the center of the lot. This life test consists of a dc power load of one watt in ambient air at 30°C ± 5°C, and corresponds to 40 watts/in² of tantalum film.

The performance of these films under such conditions can be seen in Fig. 6. The stability characteristics change rapidly with slight varia-

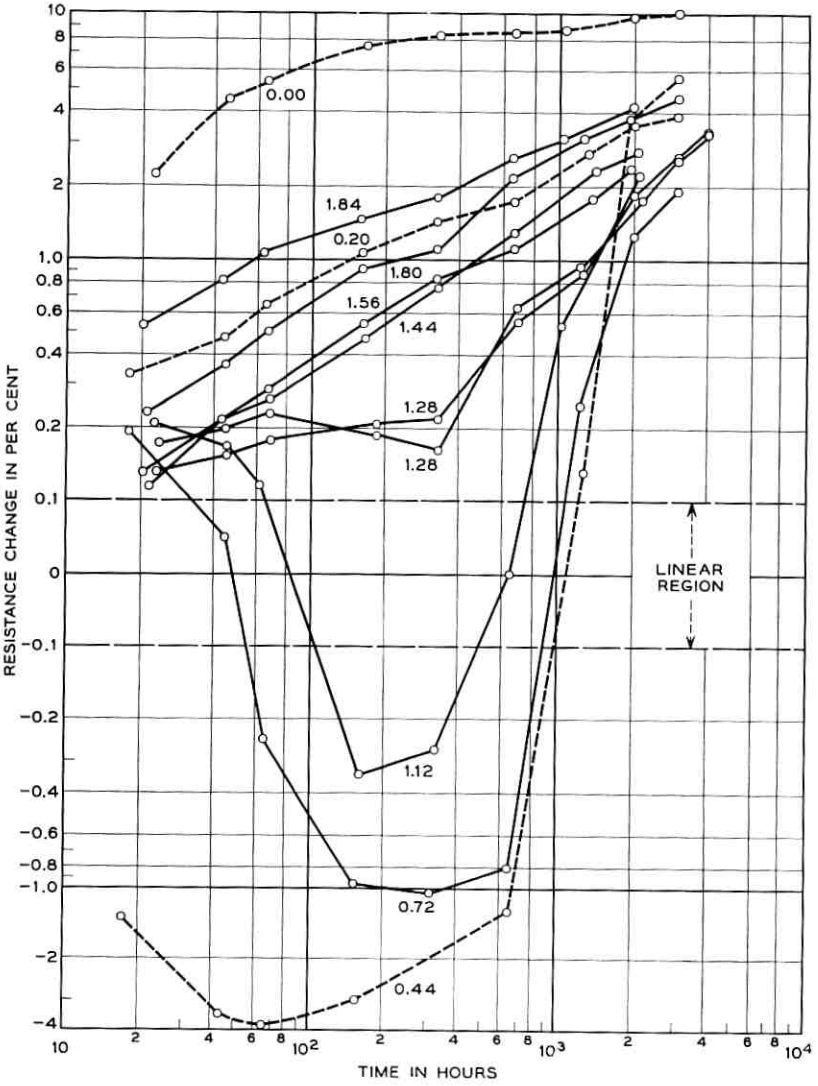


Fig. 6 — Accelerated life test of resistors with 0.0 to 1.84% nitrogen.

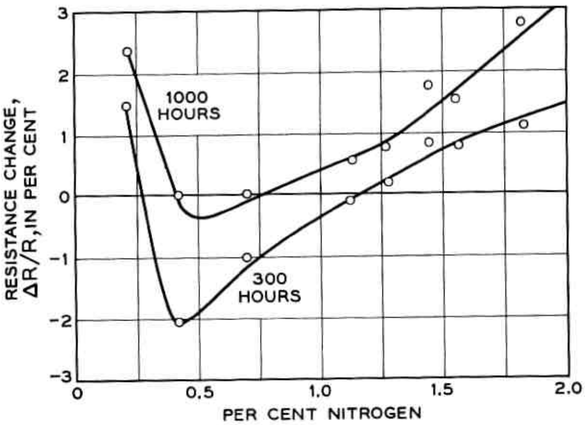


Fig. 7 — Life test time cross section.

tions in amounts of nitrogen doping. The data shown here for resistance change ($\Delta R/R$) were obtained on the same films whose nitrogen doped characteristics prior to life tests were shown in Fig. 4.

The 0.0 per cent nitrogen lot shows almost 9 per cent increase in 1000 hours. The 0.2 per cent nitrogen film appears to be more stable. The 0.44 per cent nitrogen film, at the bottom of the figure, exhibits a decrease in resistance in the first 1000 hours. However, as more nitrogen is added, the decrease in resistance is reduced until it has almost disappeared in the vicinity of 1.56 per cent nitrogen.

These data can be analyzed in a different manner by plotting time cross sections of the data against per cent nitrogen. Fig. 7 shows that this data-display technique produces a curve with the same characteristic form as the tantalum properties previously plotted. The dip in the curve occurs at the same per cent nitrogen for $\Delta R/R$ as it does for the other film properties. This minimum in each property has been previously observed in product produced in bell jars. It is believed that in the vicinity of the dip the product possessed greater metallic purity than at other nitrogen levels.

The films that were made with about 1 per cent nitrogen added to the sputtering atmosphere seem to provide the least total resistance change on this plot. Re-examination of Fig. 6 shows, however, that these films went through a large negative change in resistance before returning to original value. If films with consistent behavior are chosen instead, those with a nitrogen additive of about 1.48 per cent are to be preferred.

When changes in resistor films having 1.44 to 1.56 per cent nitrogen are examined on a log-log plot (as in Fig. 6), the drift behavior is found to be quite linear, with a trend line that can be defined by the equation:

$$\log_{10} \Delta R/R = -3.74 + 0.63 \log_{10} t.$$

This drift rate produces resistance changes at 1000 hours that are comparable to those reported from batch process bell-jar-deposited films.

Many research workers are expending considerable experimental work to establish equivalency of accelerated power aging rates to the aging rate of resistors when used at the more normal power dissipation of 20 watts/in^2. Such work indicates that the 1.48 per cent nitrogen resistors should have an average change of 0.4 per cent in 20 years under normal load. With allowance for the variability of film from run to run, this group of films should be processable into resistors with maximum aging change of less than 1 per cent. Of course, considerably more time must elapse and more correlations must be established before the exact equivalency of normal aging to such accelerated aging can be determined.

VI. NITROGEN DOPED FILM REPRODUCIBILITY

Since nitrogen doping adds a new and major variable to the operating conditions of the machine processing system, experimental runs were made to demonstrate the reproducibility of the doped film properties. Over a typical five-month period, for example, six runs were made at a particular nitrogen level of 1.28 per cent. The machine processing system was adjusted to the standard operating conditions previously mentioned. The average values of the three resistor characteristics α, ρ, and R_s for each run are shown in Table II.

6.1 *Reproducibility of Life Performance*

The stability of tantalum resistors was discussed previously in connection with the characterization curves of Fig. 6. To evaluate the ability

TABLE II — NITROGEN-DOPED FILM REPRODUCIBILITY

Sputtering Date	Temperature Coeff. of Resistance α ppm/°C	Specific Resistivity ρ $\mu\Omega$-cm	Sheet Resistance $R_s\Omega/\square$
10-2	−79	300	25.2
10-25	−81	375	26.6
11-1 A.M.	−82	334	26.5
11-1 P.M.	−87	374	27.9
1-23	−78	392	27.6
2-15	−73	318	28.1
Average	−80	349	27.0
Std. dev.	±5.5	±33	±1.1

(These standard deviations were estimated from the range of the data.)

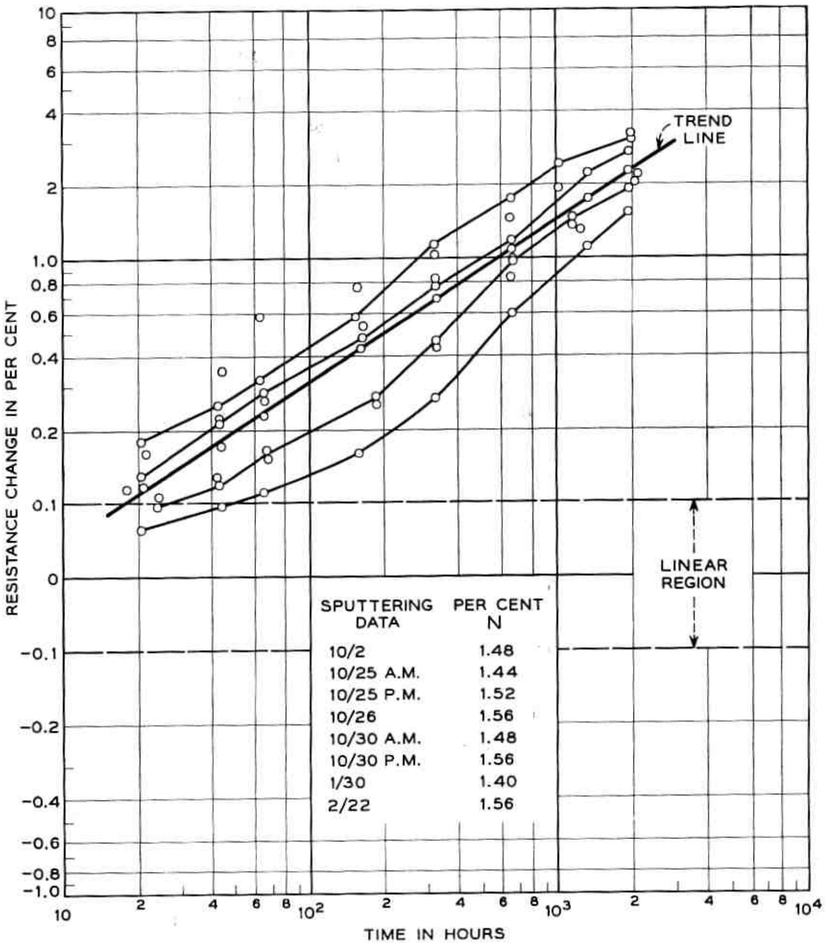


Fig. 8 — Accelerated life test of resistors with 1.40 to 1.56% nitrogen.

of this system to produce films of consistent stability, the aging characteristics of tantalum films with 1.48 ± 0.08 per cent nitrogen were examined. Resistors were processed from 8 separate runs of film having the previously mentioned nitrogen levels. The results of accelerated aging tests of these resistors are shown in Fig. 8. Sufficient power was applied to each resistor to produce a power dissipation of 40 watts per square inch of tantalum area. While there is some spread of resistance change due to the variation in nitrogen content, these resistors do consistently exhibit closely similar aging rates. The difference between films shows up as changes in resistance at the 20-hour measurement.

VII. FILM UNIFORMITY

Post-deposition processing of tantalum films requires that the resistor film be anodized to achieve stability and to adjust the resistance of the film to a required value.[8] Using etch techniques, multiple networks can be produced from a single substrate. Economical processing should be performed on the full substrate area, rather than on an individual resister or network. Economic production of large volumes of stable thin film circuits, then, requires not only that the deposition process produce a high output of film-coated substrates at a low cost, but also that the properties of the deposited films be uniform over the area of the substrate.

The resistance of the tantalum-nitride film produced in the open-ended system has a variation of 5 per cent over an effective length of 2.8 inches (see Fig. 9). This variation is comparable to that of bell-jar product, and makes possible production of resistor networks with a tolerance of ±3.0 per cent on the individual resistors. The resistance variation is not random, but has a definite pattern of higher resistance near the ends of the substrate. Since the substrate moves through the deposition zone at a constant speed, this suggests some effect of the substrate carrier on the film uniformity.

Typical tantalum-nitride film properties from a single open-ended system, under controlled production conditions, may vary 50 microohm-cm in resistivity, 100 Å in thickness, and 20 ppm/°C in temperature coefficient. This variability in film properties does not contribute significantly to the complexity of subsequent processes. However, if film deposi-

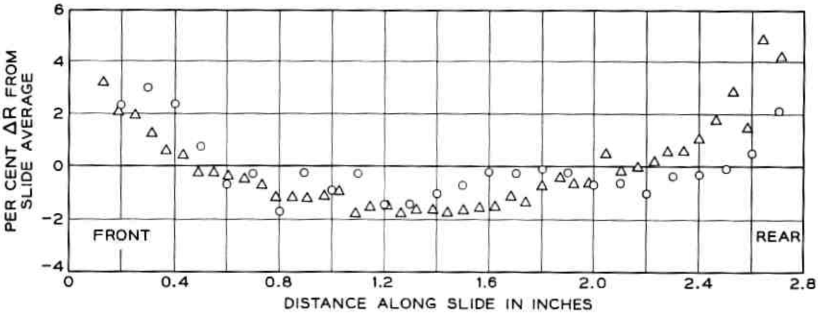


Fig. 9 — Resistance profile, two slides with 1.14% nitrogen.

tion is accomplished by using the larger number of bell-jar systems which would be required to meet the same production demand, the film properties would be influenced not only by the variability of a single chamber, but also by the chamber-to-chamber variability of the associated bell-jar systems. Compensation for this total variability will significantly influence the complexity and even the design of some of the subsequent process equipment and hence the over-all manufacturing cost of thin film resistor networks. The use of the open-end system to deposit tantalum should simplify quantity manufacture and reduce costs significantly.

VIII. CONCLUSION

At the present stage of the developmental work, it can be concluded that the open-ended in-line vacuum concept can be used to deposit large quantities of tantalum for thin film resistors. Each machine can coat two 5-inch by 5-inch substrates per minute. One such machine, on one-shift operation, can therefore produce approximately 4,000,000 square inches of metal film per year. Such films have exhibited the required stability, uniformity and reproducibility. Further work is in progress to optimize film characteristics. The work to date has established the feasibility of manufacturing production using this new deposition concept.

REFERENCES

1. McLean, D. A., Microcircuitry with Refractory Metals, I.R.E., Wescon Convention Record, 1959, Vol. III, Part 6, pp. 87–91.
2. Berry, R. W., Tantalum Thin Film Circuitry and Components, Bell Labs. Record, February, 1963, pp. 46–55.
3. Charschan, S. S., Glenn, R. W., and Westgaard, H., A Continuous Vacuum Processing Machine, W. E. Engineer, April, 1963, pp. 9–17.
4. Schwartz, N., and Brown, R., A Stylus Method for Evaluating the Thickness of Thin Films and Substrate Surface Roughness, Trans. Eighth Nat. Vac. Symp., AVC II, 1961, pp. 836–845.
5. Gerstenberg, D., and Mayer, E. H., Properties of Tantalum Sputtered Films, Proc. Elec. Comp. Conf., Washington, D. C., 1962, pp. 57–61.
6. Calbick, C. J., and Schwartz, N., Tantalum Crystal Chemistry in the Electron Microscope, Ninth Nat. Vac. Symp., AVC III, 1962, pp. 81–88.
7. Fisher, J. S., private communication.
8. Berry, R. W., and Schwartz, N., Thin Film Components Based on Tantalum, Proc. Nat. Conf. Mil. Elec., 1960, pp. 214–218.

Digital Troposcatter Transmission and Modulation Theory

By E. D. SUNDE

(Manuscript received May 20, 1963)

In tropospheric scatter transmission beyond the horizon, the amplitude, phase and frequency of a received sine wave exhibit random fluctuations owing to variable multipath transmission and noise. The probability of errors in digital transmission over such random multipath media has been dealt with in the literature on the premise of flat Rayleigh fading over the band occupied by the spectrum of transmitted pulses. This is a legitimate approximation at low transmission rates, such that the pulse spectrum is adequately narrow, but not at high digital transmission rates. The probability of errors is determined here also for high transmission rates, such that selective fading over the pulse spectrum band must be considered. Such selective fading gives rise to pulse distortion and resultant intersymbol interference that may cause errors even in the absence of noise.

Troposcatter transmission can be approximated by an idealized multipath model in which the amplitudes of signal wave components received over different paths vary at random and in which there is a linear variation in transmission delay with a maximum departure $\pm\Delta$ *from the mean delay. Various statistical transmission parameters are determined on this premise, among them the probability distribution of amplitude and phase fluctuations and of derivatives thereof with respect to time and with respect to frequency. The probability of errors in the absence of noise owing to such fluctuations is determined together with the probability of errors owing to noise, for digital transmission by binary PM and FM. Charts are presented, from which can be determined the combined probability of errors from various sources, as related to the transmission rate and certain basic parameters of troposcatter links.*

CONTENTS

INTRODUCTION

In tropospheric transmission beyond the horizon, narrow-beam transmitting and receiving antennas are used in a frequency range from about 400 to 10,000 megacycles. The received wave can be considered the sum of a large number of components of varying amplitudes, resulting from a multiplicity of reflections within the common volume at the intersection of the antenna beams. These various components arrive with different transmission delays owing to path-length differences, and each will exhibit a variation in amplitude owing to structural changes within the common volume, caused largely by winds. When a steady-state sine wave is transmitted, the received wave will consequently exhibit variations in its envelope and phase, commonly referred to as fading. When a signal wave is transmitted, its various frequency components will suffer unwanted amplitude and phase variations with resultant transmission impairments that depend on the particular carrier modulation method. These impairments are discussed herein for digital transmission by carrier phase and frequency modulation.

Various properties of the transmittance of troposcatter channels have been dealt with in several publications.[1,2,3,4] These properties include the expected average path loss and systematic seasonal variations from the average, together with the probability distributions of slow and rapid fading or fluctuations from the mean. Other important properties from the standpoint of systems design and performance are the distribution of duration of fades and the fading rapidity or rate.

The above various properties relate to transmittance variations with time at a particular frequency. Of basic importance is also the variation in transmittance with frequency at any instant, i.e., the amplitude and phase characteristics of trophospheric channels. These will be highly variable quantities, as illustrated in Fig. 1. At a fixed instant the characteristics may be as indicated in Fig. 1(a) and at a later instant as in Fig. 1(b). Such fluctuations will give rise to a distortion of the spectrum of received signals, with resultant transmission impairments of various kinds, depending on the modulation method. In addition, random noise at the receiver input must be considered as in conventional stable channels. Owing to the above random fluctuations, diversity transmission is ordinarily required to insure adequate performance.

At present, frequency modulation is used for transmission of multiplexed voice channels over troposcatter links. With this method, pronounced intermodulation noise is encountered[5,6] owing to the types of transmittance variations with frequency indicated in Fig. 1. With digital transmission, these variations will give rise to pulse distortion and resultant intersymbol interference that may severely limit the transmission rate.

In evaluation of error probabilities in digital transmission, it is necessary to consider variations in the average path loss over a convenient period, such as an hour, relative to the average over a much longer period, say a month. These slow fluctuations in loss are closely approximated by the log-normal law; i.e., the loss in db follows the normal law.[1] In addition, consideration must be given to rapid fluctuations in loss relative to the above hourly averages. These are closely approximated by the Rayleigh law, which also applies for the envelope of narrow-band random noise. They are ordinarily more important than slow fluctuations, particularly in digital transmission, in that they cannot be fully compensated for by automatic gain control. Nearly all theoretical analyses of error probabilities in digital transmission over fading channels are based on a Rayleigh distribution together with various other simplifying assumptions, as outlined below.

The simplest assumption is flat or nonselective Rayleigh fading over the channel band, in conjunction with a sufficiently slow fading rate such that changes over a few pulse intervals can be disregarded. These

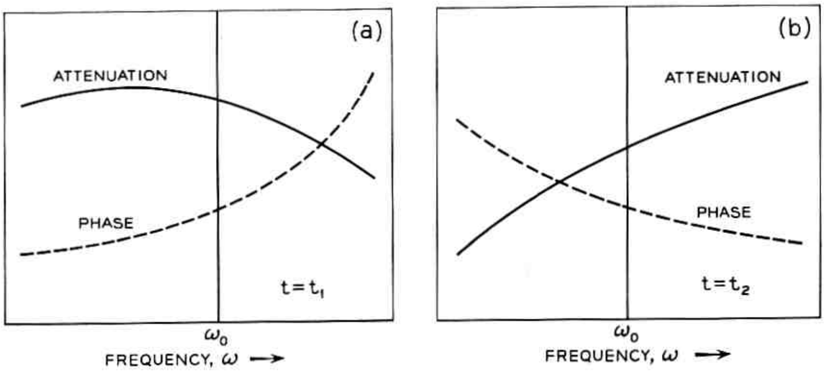


Fig. 1 — Illustrative variations in attenuation and phase characteristics with frequency at two instants t_1 and t_2.

are legitimate premises in transmission over line-of-sight radio links, where fading is much slower than on tropospheric links and is virtually nonselective over rather wide bands. With these simplifying assumptions Turin[7] has determined error probabilities in binary transmission over noisy channels with ideal synchronous (coherent) detection and envelope (noncoherent) detection. His analysis includes the effect of correlation between successive pulses and also postulates a nonfading signal component, such that the results in one limit also apply for nonfading channels.

On the same premise of slow, flat Rayleigh fading, Pierce[8] has determined the optimum theoretical diversity improvement for frequency shift keying with dual filter reception employing coherent and noncoherent detection of the filter outputs. Dual filter detection is ordinarily assumed in place of the usual method of frequency discriminator detection that does not lend itself as readily to theoretical analysis.

The error probability with two-phase and four-phase modulation with differential phase detection has been determined by Voelcker[9] on the premise of flat Rayleigh fading at such a rate that the change in phase over a pulse interval must be considered. Moreover, he considers the probability of both single and double digital errors, with both single and dual diversity transmission.

Voelcker's analysis is applicable to transmission at a sufficiently slow rate such that amplitude and phase distortion can be ignored over the relatively narrow band of the pulse spectra. However, it does not apply to high-speed digital transmission that requires sufficiently wide pulse spectra such that the amplitude and phase distortion indicated in Fig. 1 must be considered. For this case the duration of pulses will be so short that the phase changes considered by Voelcker can be disregarded. Instead, it now becomes necessary to take into account pulse distortion and resultant intersymbol interference caused by the erratic variations with frequency in the amplitude and phase characteristics illustrated in Fig. 1. An evaluation is made herein of error probabilities on the latter account, which has not been considered in previous publications.*

From the solutions for the above two limiting cases of low and high transmission rates, it is possible by simple graphical methods to estimate the error probability for the general case in which both time and frequency variations in the amplitude and phase characteristics must be considered. Charts are presented of error probabilities in digital transmission by binary PM and FM as related to various basic parameters of tropospheric scatter links and of the signals. Among these

* For reference to a recent related paper, see Section 8.9.

parameters are the average signal-to-noise ratio, the bandwidth of the pulse spectrum, the fading bandwidth of the troposcatter link, and the maximum departure from the mean transmission delay, which is related to the length of the link and the antenna beam angles.

The analysis shows that a principal source of pulse distortion and resultant transmission impairments is a component of quadratic phase distortion. On this premise, an evaluation has been made in a companion paper* of intermodulation distortion in analog transmission by FM and PM, that conforms well with the results of measurements.[5,6]

I. CHANNEL TRANSMISSION CHARACTERISTICS

1.1 *General*

Transmission performance with any modulation method depends on the statistical properties of the signals and of channel noise, together with various properties of the channel transmittance or transmission-frequency characteristic. When the latter varies with time, the usual methods of determining network response to specified input waves must be modified in various respects, that result in appreciable complications in the analytical methods[10] and in certain conceptual difficulties. However, when the time variations in transmittance are slow in relation to those in the input waves, it is legitimate to assume that the transmittances are constant over an appreciable number of pulse intervals. With relatively slow random fluctuations as encountered in troposcatter systems at representative transmission rates, it is thus permissible to determine the responses for various essentially time invariant transmittances that can be encountered. In evaluating transmission performance, the various transmittances that can be encountered must be weighted or averaged statistically in a manner that depends on the signal properties and the modulation method.

Among the statistical properties of troposcatter transmittances are the probability distribution of the envelope of received carrier waves together with the autocorrelation function of the envelope with respect to time and with respect to frequency. These are discussed here, while other statistical properties will be considered in later sections.

1.2 *Tropospheric Scatter Waves*

To determine an appropriate model for the random process in tropospheric scatter transmission, it is necessary to consider the physics

* See part 2 of this issue of the B.S.T.J., to appear.

of this phenomenon, as dealt with in various publications. Though these may differ in their assumptions regarding the exact mechanism of the reflections, they appear to agree that they occur as a result of heterogeneities within the common antenna volume indicated in Fig. 2. If the transmission medium were uniform, no reception would be possible. Owing to the numerous heterogeneities in the common volume, a very large number of reflections will occur, and the received wave can be considered the sum of a large number of components of different amplitudes and different transmission delays. Over any short interval, the envelope of a received sine wave will depend on the frequency, as will the phase. Because of variations in the heterogeneities caused largely by winds, the envelope and phase of a received carrier will vary with time.

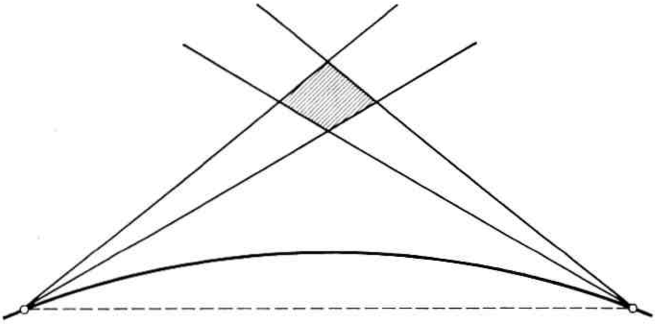


Fig. 2 — Illustrative antenna beams and common antenna volume.

The transmittance of troposcatter channels is dealt with here, based on an idealized model discussed further in the Appendix, and certain statistical parameters obtained from experimental data are discussed. Two limiting cases that permit simplified analysis are considered. In one case the transmission band is assumed sufficiently narrow, such that the attenuation characteristic can be considered constant and the phase characteristic linear over the narrow band. There will then be fluctuations with time in the attenuation accompanied by independent variations in the slope of the phase characteristic, a condition referred to as nonselective flat fading and ordinarily assumed in random multipath digital transmission theory. The other limiting case is that of digital transmission at a sufficiently high rate so that time variations in the transmittance can be disregarded over an appreciable number of pulse intervals. In this case it is necessary to consider erratic variations with frequency in both the attenuation and phase characteristics.

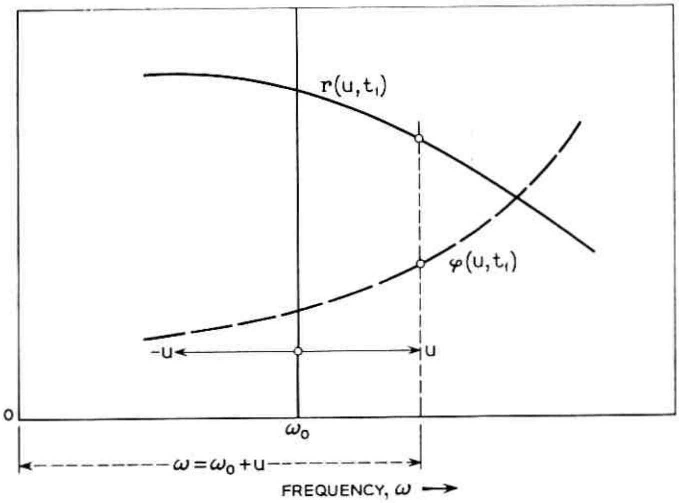


Fig. 3 — Illustrative dependence of envelope and phase of transmittance with frequency u from a reference frequency ω_0 at a specified time t_1 .

1.3 *Troposcatter Transmittance*

Let a sine wave of frequency ω be transmitted, and let $\omega = \omega_0 + u$, as indicated in Fig. 3, where ω_0 is a conveniently chosen reference frequency. In complex notation the received wave is then of the general form

$$e(u,t) = r(u,t) \exp[-i\varphi(u,t)] \exp(i\omega t) \tag{1}$$

where $r(u,t)$ and $\varphi(u,t)$ are random variables of the time t for a fixed ω or u, and of u for a fixed time t. The channel transmittance is then

$$T(u,t) = r(u,t) \exp[-i\varphi(u,t)]. \tag{2}$$

The following general relations apply

$$r(u,t) = [U^2(u,t) + V^2(u,t)]^{\frac{1}{2}} \tag{3}$$

$$\varphi(u,t) = \tan^{-1} [V(u,t)/U(u,t)]. \tag{4}$$

As shown in the Appendix, in the case of idealized tropospheric channels the functions U and V can be represented in the following form

$$U(u,t) = \sum_{j=-\infty}^{\infty} a_j(t) \frac{\sin (j\pi - u\Delta)}{j\pi - u\Delta} \tag{5}$$

$$V(u,t) = \sum_{j=-\infty}^{\infty} b_j(t) \frac{\sin (j\pi - u\Delta)}{j\pi - u\Delta} \tag{6}$$

where

Δ = maximum departure from mean transmission delay owing to path length differences.

In (5) and (6) the coefficients $a_j(t)$ and $b_j(t)$ vary at random with time t and for a given t vary at random with j. Owing to the latter variation with j, there will be a random variation in U and V with the frequency u taken in relation to the reference frequency ω_0.

Equations for an idealized troposcatter channel, as given in the Appendix, show that $a_j(t)$ is related to the sum $A(x,t) + A(-x,t)$ of two random processes and $b_j(t)$ to the difference $A(x,t) - A(-x,t)$. The two random processes $A(x,t)$ and $A(-x,t)$ will have equal rms amplitudes, in which case $a_j(t)$ and $b_j(t)$ will have zero correlation coefficient. They will then also be independent random variables, provided $A(x,t)$ and $A(-x,t)$ have a Gaussian probability distribution, which appears to be a legitimate approximation since each will be the sum of waves from a large number of reflections.

A further assumption underlying (5) and (6) is that there is an infinite number of transmission paths. An additional approximation that will be made in the following analysis is that there will be independent random fluctuations in the signal components received over the various paths. Actually there will be some correlation between the fluctuations, particularly for paths with small separation. In effect, there will be a limited number of essentially independently fading paths.

The above assumptions entail certain statistical properties of troposcatter channels, as outlined below for time and frequency variations.

1.4 *Transmission Loss Fluctuations*

On troposcatter links there is a certain average transmission loss over a year, which depends on the length of the link, on the properties of the terrain and on climatic conditions. Experimental data indicate that there will be systematic monthly and seasonal departures from this yearly average, owing principally to slow temperature changes. The average loss during a winter month may thus be up to 20 db greater than the average during a summer month. That is, the departure in transmission loss from the yearly mean may be ±10 db.

During each month there will be a more or less random fluctuation in the hourly average loss from the mean of the month. This fluctuation has been found to be almost independent of frequency and seems to be associated with the variations in average refraction of the atmosphere and resultant variation in the bending of beams. This fluctuation in the hourly average loss relative to the monthly average has been found to follow closely the log-normal law. That is to say, let the monthly median loss be

$$\alpha_m = -\ln \bar{r}_m^{\,2} \tag{7}$$

and the hourly average loss be

$$\alpha = -\ln \bar{r}^2 \tag{8}$$

where $\ln = \log_\epsilon$, $\bar{r}_m$ is the monthly rms amplitude of the envelope $r(u,t)$, and $\bar{r}$ the rms amplitude over an hour. (Other reference times could have been chosen, as will appear below.)

The probability that the average hourly loss exceeds a specified value $\alpha_1 = \ln \bar{r}_1^{\,2}$ is then given by

$$P(\alpha \geqq \alpha_1) = \frac{1}{2}\left[1 - \operatorname{erf} \frac{\alpha_1 - \alpha_m}{\sqrt{2}\sigma_\alpha}\right] \tag{9}$$

where erf is the error function and σ_α the standard deviation in transmission loss expressed in nepers, when α and α_m are expressed in nepers as above. For links 100 to 200 miles in length, a representative value of σ_α appears to be about 0.9 neper (8 db).

In addition to the above slow variations in the average hourly loss, there will be more rapid fluctuations in the envelope $r(u,t)$, owing to changes in the multipath transmission structure caused principally by winds. This type of fluctuation follows a Rayleigh distribution law. According to this law the probability that the instantaneous value r of the envelope exceeds a specified value r_1 is

$$P(r > r_1) = \exp(-r_1^{\,2}/\bar{r}^2) \tag{10}$$

where $\bar{r}$ is the hourly rms value referred to above.

It may be noted that while the log-normal law for slow variation has been determined solely by measurements, the Rayleigh law for rapid fluctuations follows by theory when the received wave is the sum of a large number of variable components.

The probability distribution (10) can be related to the monthly rms value of $r(u,t)$ with the aid of (9) by

$$P(r > r_1) = \int_0^\infty p(\bar{r}) \exp(-r_1^{\,2}/\bar{r}^2)\, d\bar{r} \tag{11}$$

where $p(\bar{r})$ is the probability density function corresponding to (9), which is

$$p(\bar{r}) = \frac{1}{\sqrt{2\pi}\sigma_\alpha \bar{r}} \exp\{-[\ln \bar{r}^2/r_m{}^2]^2/2\sigma_\alpha{}^2\}. \tag{12}$$

It will be recognized that (11) will yield the same result regardless of the period over which the rms value $\bar{r}$ is taken, since $\bar{r}$ simply plays the role of an intermediate parameter that disappears after integration.

The above probability functions relating to average loss or the distribution of the instantaneous values of $r(u,t)$, are independent of the frequency. In addition to the above distribution there are others which are important from the standpoint of transmission systems design and performance, as discussed in the following section.

1.5 *Time Autocorrelation Functions of Transmittance*

Expressions for the probabilities of rapid changes in the amplitude and phase of the transmittance with time will be considered in Section II. These involve the autocorrelation functions of the components U and V defined by (5) and (6), or the corresponding power spectra. Both have the same autocorrelation function and power spectrum, so that only $U(u,t)$ needs to be considered.

The time autocorrelation function of $U(u,t)$ depends on the variation in $a_j(t)$ with time. These are related to changes in the physical structure of the common volume and to resultant variations in the heterogeneities that are responsible for tropospheric transmission. The rate at which these occur depends on the velocity and directions of winds and on temperature changes. Under these conditions the autocorrelation function will vary with time, and it becomes necessary to consider a certain median autocorrelation function and corresponding power spectrum, as discussed in Section 1.6.

Let $\Psi(\tau)$ be the autocorrelation function of variations in $U(u,t)$ with t. The corresponding one-sided power spectrum is then

$$W(\gamma) = \frac{2}{\pi} \int_0^\infty \Psi(\tau) \cos \gamma\tau \, d\tau \tag{13}$$

where γ is used to designate the radian frequency of spectral components to avoid confusion with the frequency ω of the transmitted wave.

The autocorrelation function $\Psi(\tau)$ or the corresponding power spectrum $W(\gamma)$ of the components U and V cannot be determined as readily by measurements as the autocorrelation function $\Psi_r(\tau)$ of the envelope. The latter is related to $\Psi(\tau)$ by[11]

$$\Psi_r(\tau) = \Psi(0)\{2E[\kappa(\tau)] - [1 - \kappa^2(\tau)]K[\kappa(\tau)]\} \tag{14}$$

where

$$\kappa(\tau) = \Psi(\tau)/\Psi(0) \tag{15}$$

E = complete elliptic integral of second kind

K = complete elliptic integral of first kind.

For $\tau = 0$, $\Psi_r(0) = 2\Psi(0)$. Hence the autocorrelation coefficient of the envelope can be written

$$\kappa_r(\tau) = E[\kappa(\tau)] - \tfrac{1}{2}[1 - \kappa^2(\tau)]K[\kappa(\tau)]. \tag{16}$$

With the aid of (16), the autocorrelation coefficient $\kappa(\tau)$ of each quadrature component can be determined from measurements of $\kappa_r(\tau)$.

1.6 *Observed Time Autocorrelation*

Observations of the autocorrelation function of rapid fluctuations indicate that the autocorrelation function $\Psi(\tau)$ of the components U and V is nearly Gaussian and is given by

$$\Psi(\tau) = \Psi(0)\exp(-\sigma^2\tau^2/2). \tag{17}$$

The corresponding power spectrum obtained from (13) is

$$W(\gamma) = \Psi(0)(2/\pi\sigma^2)^{\frac{1}{2}}\exp(-\gamma^2/2\sigma^2) \tag{18}$$

where $\Psi(0)$ is the average power in each component as obtained with $\tau = 0$ in (17) .

The equivalent bandwidth of a flat power spectrum $W(\gamma) = W(0)$ is given by

$$\bar{\gamma} = \sqrt{(\pi/2)}\,\sigma \approx 1.25\sigma. \tag{19}$$

As noted in Section 1.5, there will be a certain median autocorrelation function and corresponding median values of the power spectrum, of σ and of γ. Measurements[2] indicate that these median values depend on the antenna beamwidths and that the fading rate is not quite proportional to frequency. Furthermore, there can be appreciable departure from the median values. From measurements of the median number of fades per minute, the median value of σ can be determined, with the aid of equation (26) in Ref. 2. These measurements indicate that for a particular antenna arrangement $\sigma \approx 0.1$ cps at 460 mc and about 1.3 cps at 4110 mc. The corresponding equivalent bandwidths of a flat power spectrum are thus $\bar{\gamma} \approx 0.125$ cps, or 0.8 radian/sec. at 460 mc, and $\bar{\sigma} \approx 1.6$ cps, or about 10 radians/sec. at 4110 mc. The measurements further indicate that there is a probability of about 0.01 that the fading rate exceeds the median value by a factor of about 7 at 460 mc and a factor of about 3.5 at 4110 mc.

1.7 *Frequency Correlation Function of Transmittance*

Returning to (5) and (6), let the time t be fixed, and consider variations in U and V with u. The coefficients a_j and b_j will then have certain values that vary with j, and there will be a certain variation in U and V with u. At a different time there will be another set of coefficients and a different variation with u. The form of (5) and (6) indicates that if u is regarded as a time variable and Δ as a frequency, $U(u)$ would be the variation in time owing to impulses of amplitudes a_j and b_j impinging at time intervals π on a flat low-pass filter of bandwidth Δ. That is to say, the autocorrelation function of components U and V for a difference $\nu = \omega_2 - \omega_1$ in frequency is

$$\Psi(\nu) = \Psi(0)(\sin \nu\Delta/\nu\Delta). \tag{20}$$

The corresponding power spectrum of the variation in U and V with frequency δ is

$$W(\delta) = \frac{2}{\pi}\int_0^\infty \Psi(\nu) \cos \nu\delta \, d\nu \tag{21}$$

$$\left.\begin{aligned} &= \Psi(0) \qquad \text{for} \qquad 0 < \delta < \Delta \\ &= 0 \qquad\quad \text{for} \qquad \Delta < \delta. \end{aligned}\right. \tag{22}$$

When $\Psi(\nu)$ is given, it is possible to determine the autocorrelation function $\Psi_r(\nu)$ for variations in $r(u,t)$ with u. Expression (14) applies with ν in place of τ, for the autocorrelation function of time variation with frequency.

For an autocorrelation function (20), the corresponding correlation coefficient is

$$\kappa(\nu) = (\sin \nu\Delta/\nu\Delta). \tag{23}$$

The corresponding autocorrelation coefficient of the envelope, as obtained from (16), is

$$\kappa_r(\nu) = E\left(\frac{\sin \nu\Delta}{\nu\Delta}\right) - \frac{1}{2}\left[1 - \frac{\sin^2 \nu\Delta}{(\nu\Delta)^2}\right] K\left(\frac{\sin \nu\Delta}{\nu\Delta}\right). \tag{24}$$

For various values of $\nu\Delta$ the correlation function of the envelope is given in Table I and is shown in Fig. 4.

TABLE I — AUTOCORRELATION FUNCTION OF ENVELOPE

$\nu\Delta = 0$	$\pi/2$	π	$3\pi/2$	∞
$\kappa_r(\nu) = 1$	0.9	$\pi/4$	0.78	$\pi/4$

The autocorrelation functions (23) and (24) apply for certain idealized conditions outlined in the Appendix and in Section 1.3. For one thing, the average power received over each elementary path is assumed the same. For another, a linear variation in the transmission delay with angular deviation from the mean paths is assumed, with maximum departures ±Δ from the mean delay. Furthermore, an infinity of transmission paths is assumed, with independent random fluctuations in the signal components received over the various paths, though there will be some correlation between the fluctuations in the signal components received over various paths.

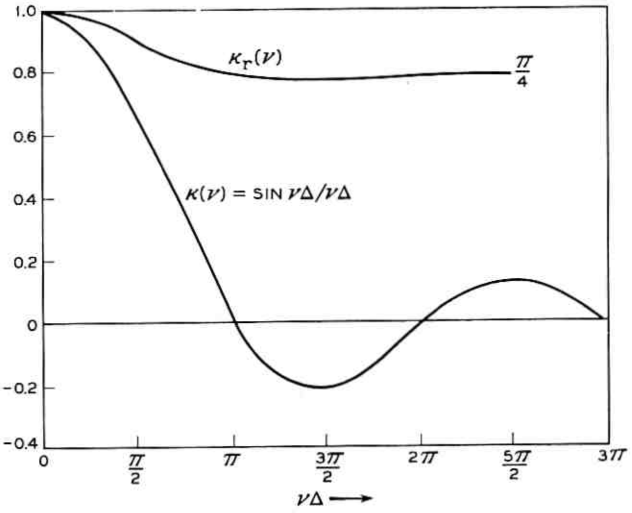


Fig. 4 — Frequency autocorrelation coefficient $\kappa_r(\nu)$ of envelope for autocorrelation coefficient $\kappa(\nu)$ of components U and V.

In spite of the various approximations, it appears possible to obtain a reasonably satisfactory conformance with the results of measurements of the autocorrelation functions of the envelope, as shown in Section 1.9.

1.8 *Differential Transmission Delay* Δ

Exact determination of the equivalent maximum departure from the mean transmission delay requires consideration of the beam patterns as affected by scattering. On the approximate basis of equivalent beam angles α, the following relation applies, with notation as indicated in Fig. 5

$$\Delta \approx \frac{L}{v}\frac{\alpha + \beta}{2}\left(\theta + \frac{\alpha + \beta}{2}\right) \tag{25}$$

where $\beta \leqq \alpha$, v is the velocity of propagation in free space, L is the length of the link, and

$$\theta = \frac{L}{2R} = \frac{L}{2R_0K} \tag{26}$$

where R_0 is the radius of the earth and the factor K is ordinarily taken as 4/3.

The equivalent beam angle α from midbeam to the 3-db loss point depends on the free-space antenna beam angle α_0 and on the effect of scatter, which is related in a complex manner to α_0 and the length L, or alternately θ. Narrow-beam antennas as now used in actual systems are loosely defined by $\alpha_0 \leqq 2\theta/3$. For these $\alpha \approx \alpha_0$ on shorter links, while on longer links $\alpha > \alpha_0$ owing to beam-broadening by scatter. Analytical determination of α for longer links appears difficult, and only

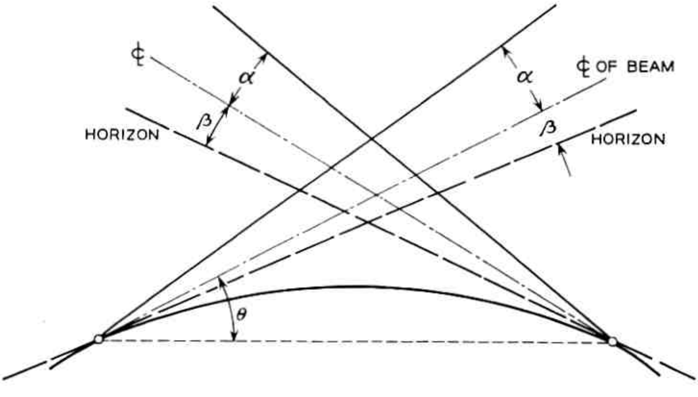


Fig. 5 — Definition of antenna beam angles α, take-off angle β and chord angle θ to midbeam. With different angles at the two ends, the mean angles are used in expressions for Δ. In applications to actual beams, α would be the angle to the 3-db loss point.

limited experimental data are available at present. For broad-beam antennas, $\alpha_0 > 2\theta/3$ and beam-broadening by scatter is in theory inappreciable.

By way of numerical example, let $L = 170$ miles and $K = 4/3$, in which case $\theta = 0.016$ radian. Since $\alpha_0 = 0.004$ radian $\ll 2\theta/3$, it is permissible to take $\alpha = \alpha_0$. With $\beta \approx \alpha_0$, (25) gives $\Delta = 0.08 \times 10^{-6}$ second.

1.9 *Observed Frequency Variations in Transmittance*

In Fig. 6 is indicated the shapes of the envelope vs frequency variations that can be obtained from (3) when the components U and V are given by (5) and (6). These fluctuations will vary with time but will have the characteristic shapes indicated in Fig. 6, which resemble shapes obtained in sweep-frequency measurements on a link of the length for which the above value of Δ applies.[2]

A better indication of the adequacy of the present idealized troposcatter model is obtained by comparing the autocorrelation coefficient of the envelope as given by (24) with the correlation coefficient derived from observations. In Fig. 7 is shown the theoretical coefficient for $\Delta = 0.08 \times 10^{-6}$ second together with coefficients obtained from three experimental runs considered representative.[2]

The bandwidth capability can be defined as the maximum baseband signal spectrum that can be received with some coherence between spectral components at the maximum and minimum frequencies. This

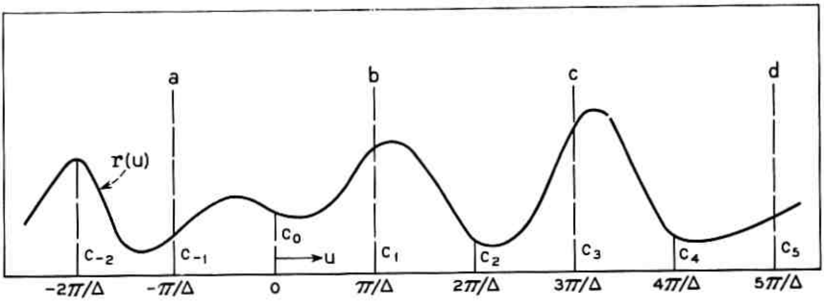


Fig. 6 — Illustrative rectified envelope vs frequency characteristic $r(u)$ obtained with expressions (5) and (6) in (3). The amplitudes c_j at the radian frequencies $u_j = j\pi/\Delta$ from the carrier are $c_j = (a_j^2 + b_j^2)^{\frac{1}{2}}$. The amplitude of the envelope at any intermediate frequency u depends on the amplitudes and phases of all c_j between $j = -\infty$ and $j = \infty$. In sweep-frequency measurements with a radian frequency sweep from $-\pi/\Delta$ to π/Δ from the carrier, the envelope variations might be like that in any of the intervals a-b, b-c, c-d, etc.

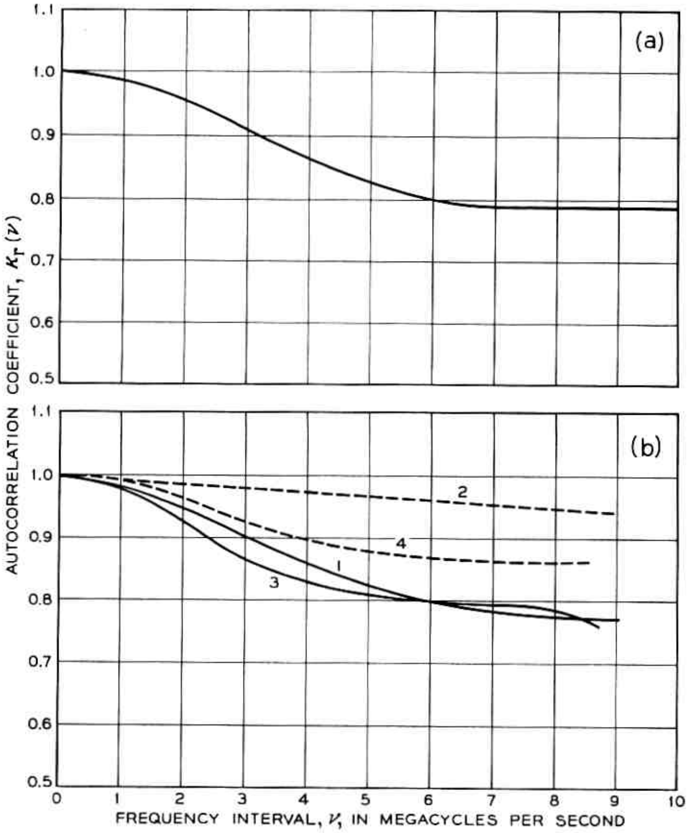


Fig. 7 — Theoretical vs observed envelope autocorrelation functions. Above: autocorrelation coefficient obtained from (24) with $\Delta = 0.08 \times 10^{-6}$ second. Below: autocorrelation coefficients given in Fig. 70 of Ref. 2 and derived from measurements of envelope variations with narrow-beam antennas on four days: 1. Sept. 13, 1957; 2. Sept. 30, 1957 (considered very unusual); 3. Oct. 15, 1957, and 4. Nov 8, 1957. The value of Δ derived from (25) for the experimental link is $\Delta = 0.08 \times 10^{-6}$ second.

bandwidth is equal to the separation between c_j and c_{j+1} in Fig. 6, which corresponds to the separation between null points in (23), for which $\kappa(\nu) = 0$ and $\kappa_r(\nu) = \pi/4$. It is given by $1/2\Delta$ cps and for $\Delta = 0.08 \times 10^{-6}$ second is 6.3 mc/second.

With a smaller spectral bandwidth, distortion will be reduced and transmission performance improved. A more realistic appraisal might be half the above maximum bandwidth, or 3.15 mc/second, for which $\kappa_r(\nu) = 0.9$. In Ref. 2 the criterion $k^2(\nu) = 0.6$ corresponding to $\kappa_r(\nu) =$

0.904 has been selected, and twice this spectrum bandwidth as required in double sideband transmission is quoted in Table VII of the reference.

The mathematical model represented by (3) to (6) is based on certain idealizations outlined in Section 1.7 and in the Appendix. It appears from the above that certain theoretical transmittance variations based on this model conform sufficiently well with observed variations for the model to be acceptable.* In order to determine expected performance with digital transmission, it is necessary to consider certain other statistical properties of tropospheric channels based on the above model, as discussed in sections that follow.

II. TRANSMITTANCE VARIATIONS WITH TIME

2.1 *General*

As discussed in Section 1.2, the transmission vs frequency characteristic of a tropospheric scatter channel is a highly variable quantity, as indicated in Fig. 1. One way of avoiding transmission impairments owing to variations in transmittance with frequency is to transmit by narrow-band modulation of a number of different carriers. The amplitude vs frequency characteristic can then be regarded as virtually constant over each narrow band, and the phase characteristic as linear, as indicated in Fig. 1. With this method, it is permissible to assume flat fading within each narrow band, but the various narrow channels will not fade independently. In addition to such flat fading there will be variations in the phase and frequency of each received carrier with time. Owing to the narrow bandwidth of each channel, the duration T of a signal or sampling interval may be relatively long, and it becomes necessary to consider the above amplitude, phase and frequency variations over this interval T. The probability distribution of these variations are basic to later considerations of various digital transmission methods and are discussed here. They can be obtained from expressions given by Rice for narrow-band random noise.[12]

2.2 *Amplitude and Phase Distributions*

Let the frequency ω and thus $u = \omega - \omega_0$ be fixed, and consider only time variations in r and φ. The probability density of φ is simply $p(\varphi) = 1/2\pi$, since each phase is equally probable. Since the components U and V are the sum of a very large number of independent random variables, in accordance with (5) and (6), each component U and V will have a

* This conclusion appears to be supported by the results of recent measurements of $\kappa(\nu)$ for a 100-mile path.[24]

normal law or Gaussian probability density. The probability density of the envelope in this case follows the Rayleigh law, and the probability that the envelope r exceeds a specified value r_1 is given by

$$P(r \geqq r_1) = \exp(-r_1^2/\bar{r}^2) \tag{27}$$

where $\bar{r}$ is the rms amplitude of the envelope or the transmittance taken over an appropriately long time.

The average received envelope power is in this case $\bar{r}^2 = \bar{S} = 2S$, where S is the average carrier power, i.e., the average power within the envelope. The probability that the received envelope power at any instant exceeds a specified value $\bar{S}_1 = 2S_1$ is

$$P(S > S_1) = \exp(-\bar{S}_1/\bar{S}) = \exp(-S_1/S). \tag{28}$$

The median value S_m of S is obtained from $P(S \geqq S_m) = \frac{1}{2}$, which gives $S_m = \bar{S} \ln 2$. Hence, in terms of the median value

$$P(S \geqq S_1) = \exp[-(S_1/S_m) \ln 2]. \tag{29}$$

The distribution represented by (28) or (29) is shown in Fig. 8.

The above distribution of rapid fades is to be distinguished from the distribution of slow variations in the envelope, or in attenuation, discussed in Section 1.4.

2.3 *Distribution of Envelope Slopes* (r')

One measure of the rapidity of the above amplitude variations is the fading bandwidth discussed in Section 1.6. From this fading bandwidth can be derived the probability distribution of the slope $r' = dr(t)/dt$ in the envelope.

The rapidity of changes in the envelope and phase depends on the time rate of change in the heterogeneities in the common volume — that is to say, the variations with respect to time of the coefficients $a_j(t)$ and $b_j(t)$ in (5) and (6). These changes are characterized by the autocorrelation function of $U(t)$ and $V(t)$, or by the corresponding power spectrum. When the power spectra of U and V are the same, and are specified, the probability distribution of $r' = dr(t)/dt$ and $\varphi' = d\varphi(t)/dt$ can be determined. These distributions are the same as for random noise of specified power spectrum. The probability that $|r'|$ exceeds a specified value $|r_1'|$ follows the normal law[12]

$$P(|r'| \geqq |r_1'|) = \operatorname{erfc}(k/2^{\frac{1}{2}}) \tag{30}$$

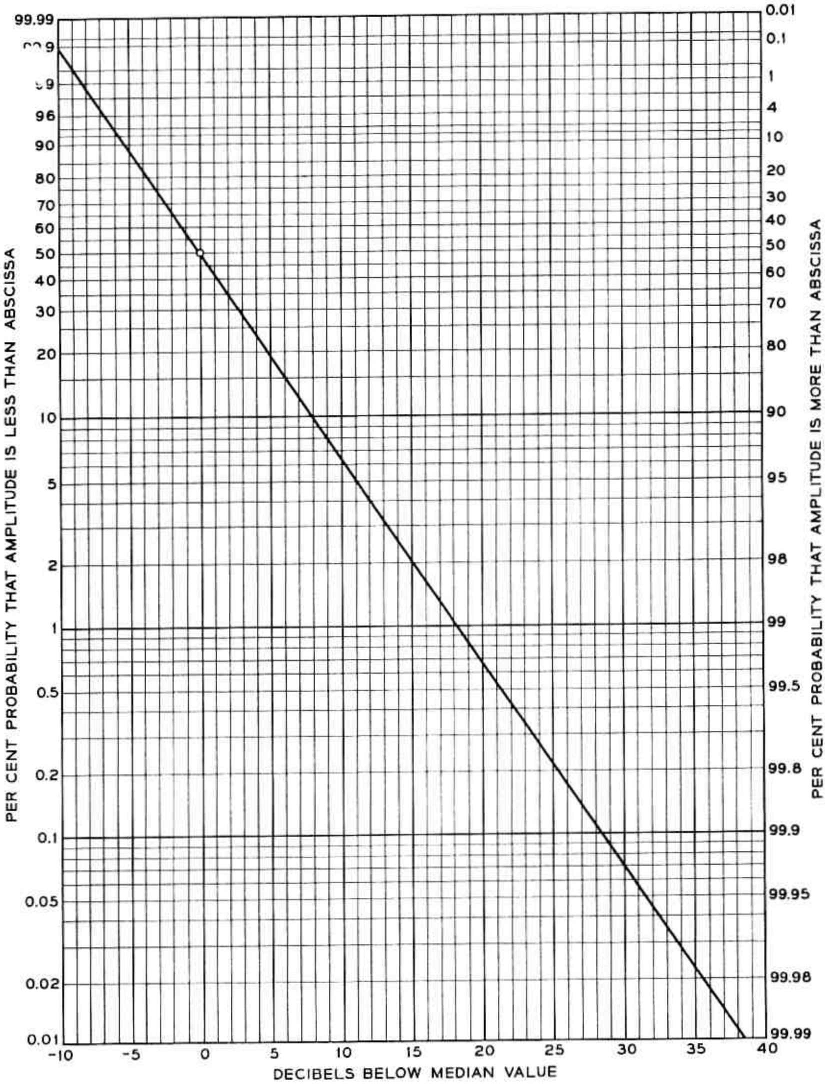


Fig. 8 — Rayleigh probability distribution of rapid fluctuations in envelope of a received carrier owing to multipath propagation.

in which

$$\begin{aligned} k &= r_1'/\bar{r}' \\ \bar{r}' &= \text{rms amplitude of } r' \\ &= [\tfrac{1}{2}(b_2 - b_1^2/b_0)]^{\frac{1}{2}} \end{aligned} \tag{31}$$

where

$$b_n = \int_0^\infty W(\gamma)\gamma^n \, d\gamma. \tag{32}$$

The above result (30) follows from equation (4.6) in Ref. 12 for $Q = 0$, by integration with respect to $R = r$ between 0 and ∞, and in turn with respect to $R' = r'$ between r_1' and ∞.

Expression (30) can alternatively be written

$$P[\,|\,r'\,|\, \geqq k\bar{r}'] = \text{erfc } (k/2^{\frac{1}{2}}). \tag{33}$$

In the particular case of flat power spectrum $W(\gamma) = W$ of bandwidth $\hat{\gamma}$, (32) gives

$$b_0 = W\hat{\gamma}; \qquad b_1 = W\hat{\gamma}^2/2; \qquad b_2 = W\hat{\gamma}^3/3$$

and (31) becomes

$$\bar{r}' = \bar{r}\hat{\gamma}/6^{\frac{1}{2}} \approx 0.405\bar{r}\hat{\gamma}. \tag{34}$$

The fading bandwidth in the above case is $\hat{\gamma}$ radians/second.

With a Gaussian spectrum (17) expression (32) gives

$$b_0 = \Psi(0); \qquad b_1 = \sigma(2/\pi)^{\frac{1}{2}}\Psi(0); \qquad b_2 = \sigma^2\Psi(0)$$

and (31) becomes

$$\begin{aligned} \bar{r}' &= \bar{r}\sigma\left(\frac{1}{2} - \frac{1}{\pi}\right)^{\frac{1}{2}} \\ &\approx 0.42\bar{r}\sigma \approx 0.34\bar{r}\bar{\gamma} \end{aligned} \tag{35}$$

where $\bar{\gamma}$ is the equivalent bandwidth given by (19).

2.4 *Distribution of Phase Derivative* (φ')

In considering a small phase change $\Delta\varphi$, and over a small interval $\Delta\tau$, it is legitimate to use the probability distribution of the phase derivative $\varphi' = \Delta\varphi/\Delta\tau$, which is given by [Section 5 of Ref. 12]

$$P(\,|\,\varphi'\,|\, \geqq |\,\varphi_1'\,|) = 1 - \frac{k}{\sqrt{1 + k^2}} \tag{36}$$

in which

$$k = (b_0/b_2)^{\frac{1}{2}}\varphi_1' = (b_0/b_2)^{\frac{1}{2}}(\Delta\varphi_1/\Delta\tau) \tag{37}$$

where b_0 and b_2 are given by (32).

Expression (36) can alternatively be written

$$\begin{aligned} P(|\varphi'| \geqq k(b_2/b_0)^{\frac{1}{2}}) &= 1 - \frac{k}{\sqrt{1+k^2}} \\ &\approx \frac{1}{2k^2} \qquad \text{for} \qquad k \gg 1. \end{aligned} \tag{38}$$

For a flat power spectrum $W(\gamma) = W$ of bandwidth $\hat{\gamma}$

$$(b_2/b_0)^{\frac{1}{2}} = \hat{\gamma}/3^{\frac{1}{2}} \approx 0.58\hat{\gamma}. \tag{39}$$

For a Gaussian spectrum (17)

$$(b_2/b_0)^{\frac{1}{2}} = \sigma \approx 0.8\bar{\gamma} \tag{40}$$

where $\bar{\gamma}$ is the equivalent bandwidth given by (19).

2.5 *Distribution of Frequency Derivative* (φ'')

The probability of exceeding a small variation $\Delta\omega$ in frequency over a brief interval $\Delta\tau$ can be determined from the probability distribution of $\varphi'' = \Delta\omega/\Delta\tau$.

The probability that φ'' exceeds a specified value φ_1'' is given by

$$\begin{aligned} P(|\varphi''| \geqq |\varphi_1''|) &= P(|\varphi''| \geqq kb_0/b_2) \\ &= 1 - \frac{2k}{\pi}\int_0^\infty \frac{dx}{[g(x) + k^2]g(x)} \\ &\quad - \frac{2}{\pi}\int_0^\infty \frac{\tan^{-1}(k/g^{\frac{1}{2}}(x))}{(1+x^2)^{\frac{1}{2}}}\,dx \end{aligned} \tag{41}$$

where

$$k = b_0\varphi_1''/b_2 \tag{42}$$

$$g(x) = (a - 1 + 4x^2)(1 + x^2) \tag{43}$$

$$a = b_0b_4/b_2^2. \tag{44}$$

Expression (41) is obtained from relation (6.10) of Ref. 12 for $p(r,\varphi,\varphi',\varphi'')$ for $Q = 0$, by integration with respect to r, φ and φ', between 0 and ∞, 0 and 2π and $-\infty$ and $+\infty$, respectively, and in turn by integration with respect to φ'' between φ_1'' and ∞. Considerable simplification is required to obtain (41).

For very large values of k the following approximation applies

$$P(|\varphi''| \geqq kb_2/b_0) \approx \frac{2}{\pi k}\left[1 + \ln\left(\frac{k}{2} + 1\right)\right] \qquad (45)$$

where $\ln = \log_\epsilon$.

For a flat spectrum $W(\gamma) = W$ of bandwidth $\hat{\gamma}$

$$a = 9/5 \qquad \text{and} \qquad b_2/b_0 = \hat{\gamma}^2/3. \qquad (46)$$

For a Gaussian power spectrum (18)

$$a = 3 \qquad \text{and} \qquad b_2/b_0 = \sigma^2. \qquad (47)$$

The quantity $(b_2/b_0)^{\frac{1}{2}}$ is the rms frequency of the power spectrum and b_2/b_0 is the "variance."

The probability distribution (41) as obtained by numerical integration is shown in Tables II and III for flat and Gaussian power spectra. For large values of k, approximation (45) is shown in parentheses. These probability distributions are shown in Fig. 9.

III. TRANSMITTANCE VARIATIONS WITH FREQUENCY

3.1 *General*

In the previous section a sufficiently narrow signal band spectrum was assumed such that amplitude and phase distortion over the narrow band could be neglected. In this case it was necessary to consider time fluctuations in the transmittance over a pulse duration T that would be relatively long owing to the narrow spectrum bandwidth.

The other extreme of wideband transmission will now be considered, in which the duration of a pulse would be short enough for fluctuations in transmittance over a pulse interval to be disregarded. In this case it becomes necessary to consider variations in the transmittance with frequency over the much greater signal spectrum band. The variations in the amplitude and phase characteristics with frequency will fluctuate with time, so that it becomes necessary to determine the resultant

TABLE II — PROBABILITY DISTRIBUTION $P(|\varphi''| > k\hat{\gamma}^2/3)$ FOR FLAT POWER SPECTRUM

$k = 0$	1	2	3	4	5	10	20	50	100
1	.538	.381	.321	.269	.238	.158	.100	.051	.031 (.03)

Table III — Probability Distribution $P(|\varphi''| > k\sigma^2)$ for Gaussian Power Spectrum

$k = 0$	1	2	3	4	5	10	20	50	100
1	.595	.447	.369	.317	.280	.182	.113	.057	.033 (.03)

transmission impairments on the basis of certain probability distributions.

In a first approximation the departure from a constant amplitude vs frequency characteristic will be a characteristic with a linear slope, as indicated in Fig. 10, that will vary with time. Similarly the departure from a constant transmission delay over the channel band can be approxi-

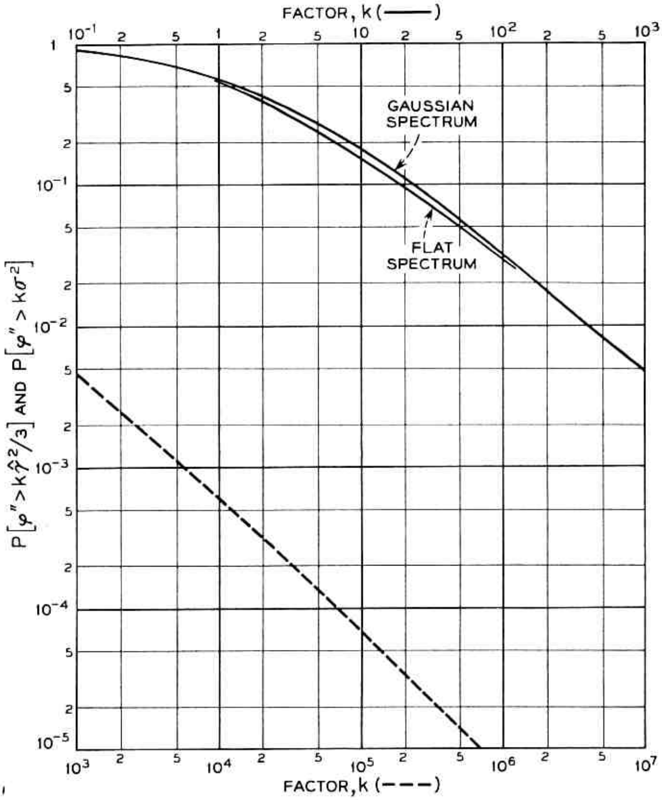


Fig. 9 — Probability that $\ddot{\varphi}$ exceeds "variance" of fading power spectrum by factor k for flat power spectrum with bandwidth $\hat{\gamma}$ and "variance" $\hat{\gamma}^2/3$ and for Gaussian power spectrum with "variance" σ^2.

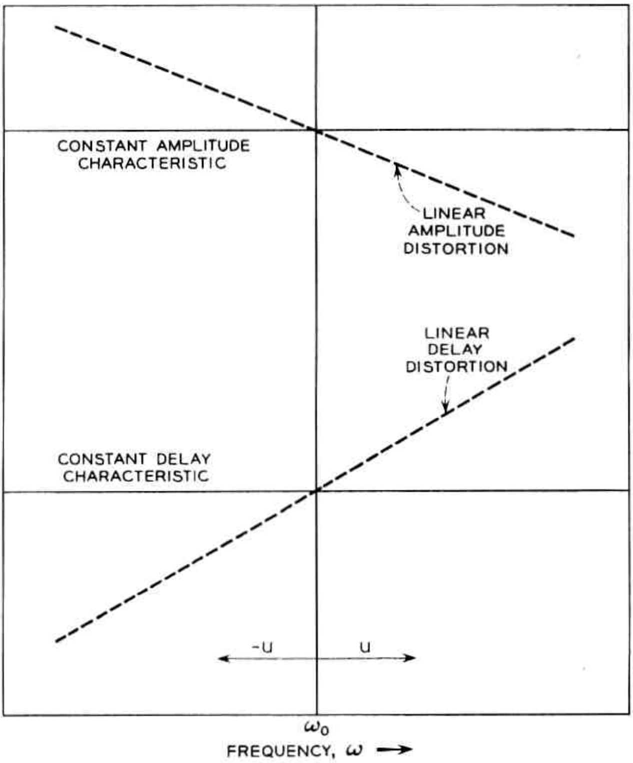

CONSTANT AMPLITUDE CHARACTERISTIC
LINEAR AMPLITUDE DISTORTION
LINEAR DELAY DISTORTION
CONSTANT DELAY CHARACTERISTIC
-U
U
ω_0
FREQUENCY, ω →

Fig. 10 — First approximations to random departures from constant amplitude and delay characteristics are represented by linear variations with frequency.

mated by a linear variation in transmission delay. The probability distributions of the slopes of these linear variations in the amplitude and delay characteristics are the same as for corresponding variations with time, with appropriate modification of the basic parameters, as discussed in the following.

3.2 *Amplitude and Phase Distributions*

Let the time t be fixed, and consider only variations in r and φ with the frequency ω of a number of transmitted sine waves.

Each sine wave could be regarded as a spectral component of a carrier pulse of very short duration with an essentially flat and continuous spectrum about the carrier frequency. In this case u rather than t is changed in expressions (5) and (6) for the two components $U(u,t)$ and $V(u,t)$. There will in this case be a particular variation with u for each time t. When observations are made for a sufficiently large number of specified times, the resultant probability distribution of the amplitude and phase will be the same as discussed in Section 2.2 for variation in time for a given frequency u.

3.3 *Slope in Amplitude Characteristic* ($\dot{r}$)

At a particular time, the envelopes $r(u,t)$ of the received sine waves will vary with frequency u. The slope of the envelope will be designated $dr(u,t)/du = \dot{r}$. It will have a probability distribution as given by (30) for the time rate of change in $r(u,t)$. This probability distribution is

$$P(\,|\dot{r}| > |\dot{r}_1|\,) = P(\,|\dot{r}| \geqq k\dot{\bar{r}}) = \operatorname{erfc}\,(k/2^{\frac{1}{2}}) \tag{48}$$

where erfc is the error function complement and

$$\begin{aligned} k &= \dot{r}_1/\dot{\bar{r}} \\ \dot{\bar{r}} &= \text{rms value of } \dot{r} \\ &= [\tfrac{1}{2}(b_2 - b_1^2/b_0)]^{\frac{1}{2}} \end{aligned} \tag{49}$$

except that now

$$b_n = \int_0^\infty W(\delta)\delta^n \, d\delta \tag{50}$$

where $W(\delta)$ is the power spectrum given by (21). When $W(\delta)$ is given by (22), (50) gives $b_0 = \Psi(0)/\Delta$; $b_1 = \Psi(0)\Delta^2/2$; $b_3 = \Psi(0)\Delta^3/3$ and (49) yields

$$\dot{\bar{r}} = \bar{r}\Delta/6^{\frac{1}{2}} \tag{51}$$

where $\bar{r} = \Psi(0)^{\frac{1}{2}}$ is the rms amplitude of the envelope.

3.4 *Envelope Delay Distribution*

The envelope delay at a particular time t and frequency u is given by $\dot{\varphi} = d\varphi(u,t)/du$. The probability distribution of this delay $\dot{\varphi}$ is given by (36) or (38). Thus

$$\begin{aligned} P(\,|\dot{\varphi}| > |\dot{\varphi}_1|\,) &= P[\,|\dot{\varphi}| \geqq k(b_2/b_0)^{\frac{1}{2}}] \\ &= 1 - \frac{k}{\sqrt{1+k^2}} \end{aligned} \tag{52}$$

where as before

$$k = (b_0/b_2)^{\frac{1}{2}}\dot{\varphi}_1 \tag{53}$$

where b_0 and b_2 are given by (50).

For a flat power spectrum (22)

$$(b_2/b_0)^{\frac{1}{2}} = \Delta/3^{\frac{1}{2}} \approx 0.58\Delta. \tag{54}$$

3.5 *Distribution of Linear Delay Distortion*

The slope $\ddot{\varphi} = d\dot{\varphi}/du$ at a particular time represents linear delay distortion. The probability that $\ddot{\varphi}$ exceeds a specified value $\ddot{\varphi}$, is given by (41), or

$$\begin{aligned} P(|\ddot{\varphi}| > |\ddot{\varphi}_1|) &= F(|\ddot{\varphi}| \geqq kb_2/b_0) \\ &= 1 - \frac{2k}{\pi}\int_0^\infty \frac{dx}{(g(x) + k^2)g(x)} \\ &\qquad - \frac{2}{\pi}\int_0^\infty \frac{\tan^{-1}(k/g^{\frac{1}{2}}(x))}{(1 + x^2)^{\frac{3}{2}}}\, dx. \end{aligned} \tag{55}$$

For very large values of k (45) applies, or

$$P(|\ddot{\varphi}| \geqq kb_2/b_0) \approx \frac{2}{\pi k}\left[1 + \ln\left(\frac{k}{2} + 1\right)\right] \tag{56}$$

where now

$$k = b_0\ddot{\varphi}_1/b_2 \tag{57}$$

$$g(x) = (a - 1 + 4x^2)(1 + x^2) \tag{58}$$

$$a = b_0b_4/b_2^2 \tag{59}$$

and b_n is given by (50).

For a flat power spectrum (22)

$$b_2/b_0 = \Delta^2/3. \tag{60}$$

The probability distribution (55) as a function of k is given previously in Table II for a flat power spectrum and is shown in Fig. 9.

IV. ERRORS FROM TRANSMITTANCE VARIATIONS WITH FREQUENCY

4.1 *General*

As discussed later, the error probability in digital transmission over noisy channels with selective Rayleigh fading can be approximated by combining the probability of errors from three basic sources. One of these is errors from random noise determined in the presence of flat Rayleigh fading. The second source is errors from time variations in the transmittance, which is important at low transmission rates. The third source is errors from transmittance variations with frequency, which becomes important at high transmission rates and puts an upper bound on the transmission rate for a specified error probability. In this section an approximate evaluation is made of errors on the latter account.

As a first approximation, the statistical properties of transmittance variations with frequency, ordinarily referred to as selective fading, can be represented by the probability distribution (48) of $\dot{r}$ and (55) of $(\ddot{\varphi})$. The first of these represents a linear slope on the amplitude vs frequency characteristics, and the second represents a linear variation in transmission delay. Errors will occur even in the absence of noise, when $\dot{r}$ or $\ddot{\varphi}$ exceeds certain maximum values. These maxima will depend on the spectrum of pulses in the absence of distortion, on the pattern of transmitted pulses and on the carrier modulation method. After these maximum values are determined, it is possible to determine the probability of encountering them with the aid of the probability distributions of $\dot{r}$ and $\ddot{\varphi}$ given in Section III.

4.2 *Carrier Pulse Transmission Characteristics*

It will be assumed that a carrier pulse of rectangular or other suitable envelope is applied at the transmitting end of a bandpass channel. The received pulse with carrier frequency ω_0 can then be written in the general form[13]

$$P_0(t) = \cos(\omega_0 t - \psi_0)R_0(t) + \sin(\omega_0 t - \psi_0)Q_0(t) \tag{61}$$

$$= \cos[\omega_0 t - \psi_0 - \varphi_0(t)]\bar{P}_0(t), \tag{62}$$

where

$$\bar{P}_0(t) = [R_0^2(t) + Q_0^2(t)]^{\frac{1}{2}}, \tag{63}$$

$$\varphi_0(t) = \tan^{-1}[Q_0(t)/R_0(t)], \tag{64}$$

$$R_0(t) = \bar{P}_0(t)\cos\varphi_0(t), \tag{65}$$

$$Q_0(t) = \bar{P}_0(t)\sin\varphi_0(t). \tag{66}$$

In the above relations R_0 and Q_0 are the in-phase and quadrature components of the received carrier pulse and $\bar{P}_0(t)$ the resultant envelope. The time t is taken with respect to a conveniently chosen origin, for example the midpoint of a pulse interval or the instant at which $R_0(t)$ or $\bar{P}_0(t)$ reaches a maximum value.

Let $S_0(u)$ be the spectrum of received pulses at the output of the receiving filter, i.e., at the detector input, and $\psi_0(u)$ the phase function of the spectrum, as illustrated in Fig. 11. The functions $R_0(t)$ and $Q_0(t)$ are then given by[13]

$$R_0 = R_0^- + R_0^+, \qquad Q_0 = Q_0^- - Q_0^+,$$

$$R_0^- = \frac{1}{\pi}\int_0^{\omega_0} S_0(-u) \cos [ut + \Psi_0 (-u)]\, du, \tag{67}$$

$$R_0^+ = \frac{1}{\pi}\int_0^{\infty} S_0(u) \cos [ut - \Psi_0(u)]\, du, \tag{68}$$

$$Q_0^- = \frac{1}{\pi}\int_0^{\omega_0} S_0(-u) \sin [ut + \Psi_0(-u)]\, du, \tag{69}$$

$$Q_0^+ = \frac{1}{\pi}\int_0^{\infty} S_0(u) \sin [ut - \Psi_0(u)]\, du. \tag{70}$$

The upper limit ω_0 can ordinarily be replaced by ∞, since $S_0(-\omega_0) = 0$.

Let $S(u)$ be the spectrum in the absence of amplitude distortion, and $A(u)$ the amplitude characteristic of the channel. The received spectrum is then, for a time invariant channel

$$S_0(u) = S(u)A(u). \tag{71}$$

4.3 *Ideal Pulse Spectra and Pulse Shapes*

In carrier pulse transmission over an ideal channel the sideband spectrum of carrier pulses at the detector input will be symmetrical

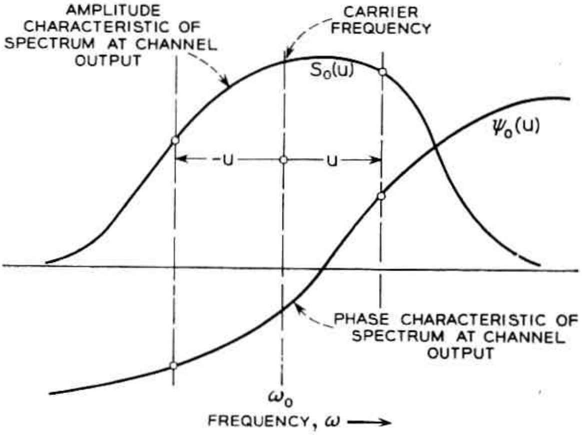

AMPLITUDE CHARACTERISTIC OF SPECTRUM AT CHANNEL OUTPUT
CARRIER FREQUENCY
$S_0(u)$
$\psi_0(u)$
-u
u
PHASE CHARACTERISTIC OF SPECTRUM AT CHANNEL OUTPUT
ω_0
FREQUENCY, ω →

Fig. 11 — Amplitude and phase functions of pulse spectrum at channel output, i.e., detector input.

about the carrier frequency. As discussed elsewhere,[14] it is possible to realize optimum performance in binary transmission by AM, PM and FM with an infinite variety of pulse spectra at the detector input, with the general properties illustrated in Fig. 12. With all of these spectra, pulses can be transmitted without intersymbol interference at intervals

$$T = \pi/\Omega = 1/2B \tag{72}$$

where B is the mean bandwidth in cps to each side of the carrier frequency, as indicated in Fig. 12.

A desirable pulse spectrum in various respects is a raised cosine spectrum as illustrated in Fig. 13, given by

$$S(u) = S(-u) = \frac{\pi}{\Omega} \cos^2 \frac{\pi}{4} \frac{u}{\Omega}. \tag{73}$$

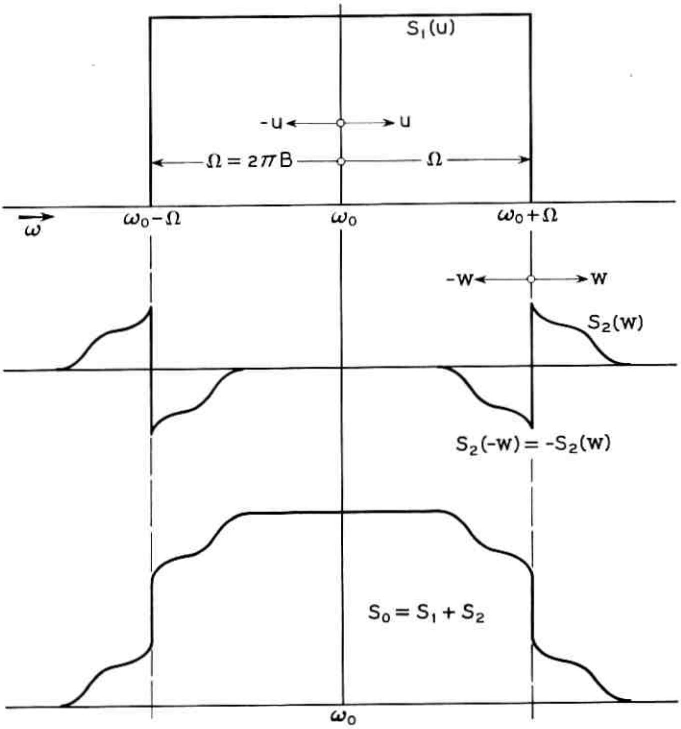


Fig. 12 — General properties of ideal spectra of carrier pulses at channel output (detector input) that permit pulse transmission without intersymbol interference at intervals $T = \pi/\Omega = 1/2B$.

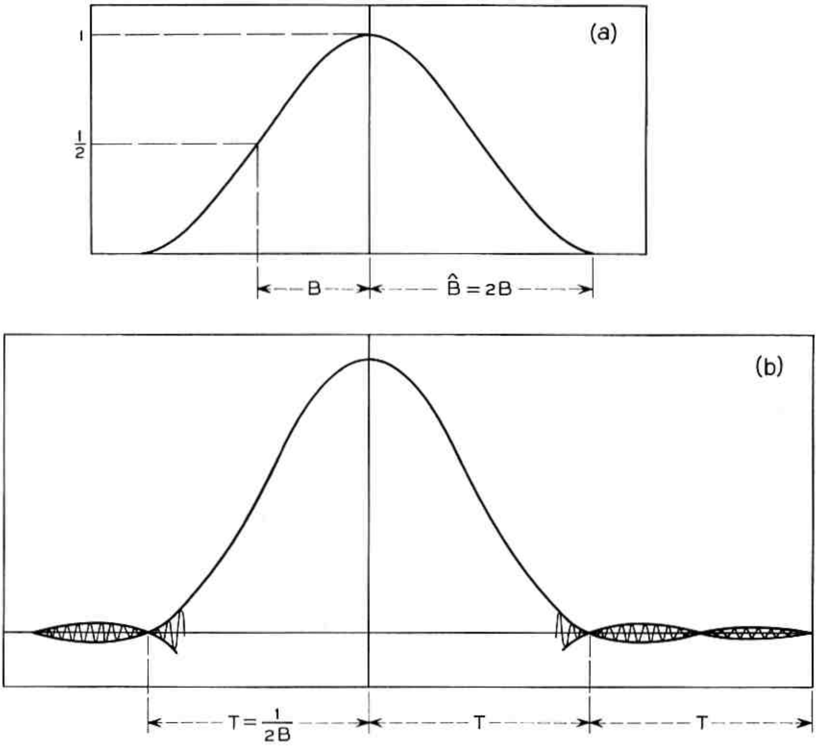


Fig. 13 — (a) Raised cosine bandpass pulse spectrum and (b) carrier pulse transmission characteristic, i.e., envelope of a carrier pulse.

The corresponding carrier pulse at the detector input as shown in Fig. 13 is given by

$$P_0(t) = \bar{P}_0(t) \cos (\omega_0 t - \varphi_0) \tag{74}$$

where

$$\bar{P}_0(t) = R_0(t) = \frac{\sin \Omega t}{\Omega t} \frac{\cos \Omega t}{1 - (\Omega t/\pi)^2}. \tag{75}$$

4.4 *Linear Variation in Amplitude Characteristic*

Let $\psi_0(u) = 0$ and

$$A(u) = 1 + cu \tag{76}$$

where c is a constant. In this case (71) becomes

$$S_0(u) = S(u)(1 + cu). \tag{77}$$

When the received spectrum in the absence of distortion has even symmetry about the carrier frequency ω_0, such that $S(-u) = S(u)$, (77) in (67) to (70) gives

$$R_0(t) = \frac{2}{\pi} \int_0^\infty S(u) \cos \omega t \, du \tag{78}$$

$$Q_0(t) = -\frac{c^2}{\pi} \int_0^\infty uS(u) \sin ut \, du \tag{79}$$

$$= c \frac{d}{dt} R_0(t) = cR_0'(t). \tag{80}$$

In the case of a raised cosine spectrum, $R_0(t)$ is given by (75) and (80) yields

$$Q_0(t) = c2\Omega \frac{\cos 2\Omega t}{2\Omega t[1 - (2\Omega t/\pi)^2]} - c2\Omega \frac{\sin 2\Omega t}{(2\Omega t)^2[1 - (2\Omega t/\pi)^2]^2} \tag{81}$$

$$= 0 \qquad \text{for} \qquad t = 0. \tag{82}$$

At the first sampling points before and after $t = 0$, $t = \pm \mathrm{T} = \pm(\pi/\Omega)$ and(81) yields

$$Q_0(\pm T) = \pm c\Omega/3\pi. \tag{83}$$

At the next sampling points $t = \pm 2T = \pm 2\pi/\Omega$

$$Q_0(\pm 2T) = \pm c\Omega/30\pi. \tag{84}$$

From (83) and (84) it appears that only the first sampling points $t = \pm T$ need to be considered in determining the effect of linear amplitude distortion.

4.5 *Probability of Errors from Linear Amplitude Distortion*

The rms amplitude of the component $Q_0(\pm T)$ is given by

$$\bar{Q}_0(\pm T) = \bar{c}\Omega/3\pi = \bar{c}\hat{B}/3 \tag{85}$$

where $\hat{B} = 2\Omega/2\pi$ and $\bar{c}$ is the rms amplitude of $\dot{r}$ as given by (51) or

$$\bar{c} = \bar{\dot{r}} = \bar{r}\Delta/6^{\frac{1}{2}}. \tag{86}$$

Thus (85) becomes

$$\bar{Q}_0(\pm T) = \bar{r}(\hat{B}\Delta/3\cdot 6^{\frac{1}{2}}). \tag{87}$$

The rms amplitude of $R_0(0)$ is $\bar{r}$. Hence

$$\bar{\eta} = \frac{\bar{Q}_0(T)}{\bar{R}_0(0)} = \frac{\hat{B}\Delta}{3\cdot 6^{\frac{1}{2}}}. \tag{88}$$

This is the ratio of rms intersymbol interference at the first sampling points to the rms value of the peak pulse amplitude.

The probability of exceeding the above ratio by a factor k is, in accordance with (48)

$$P(\eta \geqq k\bar{\eta}) = \operatorname{erfc}\,(k/2^{\frac{1}{2}}). \tag{89}$$

The probability of error will depend on the carrier modulation method. In general, however, the approximate allowable peak value of η in the absence of noise is

$$\hat{\eta} \approx \tfrac{1}{2}. \tag{90}$$

The probability of exceeding this value, corresponding to $k = 3\cdot 6^{\frac{1}{2}}/2\hat{B}\Delta$ is

$$P_e = \operatorname{erfc}\,(3\cdot 3^{\frac{1}{2}}/2\hat{B}\Delta) \approx \operatorname{erfc}\,(2.6/\hat{B}\Delta). \tag{91}$$

This probability is much smaller than that resulting from a linear variation in delay over the transmission band. For example, if $\hat{B} = 10^6$ cps and $\Delta = 10^{-7}$ sec, $1/\hat{B}\Delta = 10^{-1}$ and $P_e = \operatorname{erfc}\,(26)$, which is negligible.

4.6 *Linear Variation in Envelope Delay*

It will be assumed that the phase distortion component is given by

$$\Psi_0(u) = cu^2, \tag{92}$$

which corresponds to a linear delay distortion given by

$$\Psi_0'(u) = 2cu. \tag{93}$$

In this case expressions (67) to (70) give for a raised cosine spectrum

$$R_0(-t) = R_0(t) = \frac{4}{\pi}\int_0^{\pi/2} \cos^2 x \cos ax \cos bx^2\,dx \tag{94}$$

$$Q_0(-t) = Q_0(t) = \frac{4}{\pi}\int_0^{\pi/2} \cos^2 x \cos ax \sin bx^2\,dx, \tag{95}$$

where

$$a = 4(t/T), \qquad b = (4/\pi)\,(d/T); \qquad T = (1/\hat{B})$$

in which the delay d is defined as in Fig. 14.

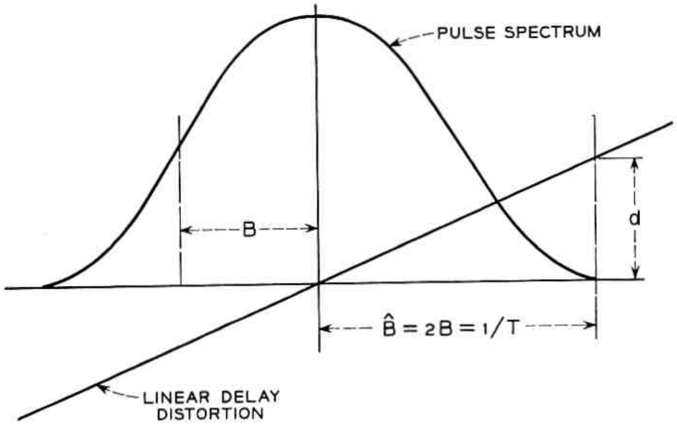


Fig. 14 — Raised cosine pulse spectrum with linear delay distortion.

The above integrals have been evaluated by numerical integration and are tabulated elsewhere.[13] The functions $R_0(t)$ and $Q_0(t)$ are shown in Fig. 15, as a function of $t/T = t\hat{B}$ for various values of $d/T = d\hat{B}$. The phase has been adjusted to 0 at $t = 0$, hence the notation R_{00} and Q_{00}.

4.7 *Maximum Tolerable Linear Delay Distortion*

Intersymbol interference at sampling points owing to linear delay distortion is significantly greater than that resulting from a linear slope in the amplitude characteristic. Moreover, pulse patterns that cause maximum intersymbol interference with linear delay distortion will not give rise to intersymbol interference from a linear slope in the amplitude characteristic, and conversely. For this reason it suffices to consider the more important component, i.e., linear delay distortion.

The reduction in tolerable noise power owing to linear delay distortion has been determined elsewhere[13] for binary AM with envelope detection, binary PM with synchronous detection, and binary FM with frequency discriminator detection. For these methods the reduction in noise margin is shown in Fig. 16 as a function of the parameter $\lambda = d/T = d \cdot \hat{B}$. In the same figure is shown the reduction in noise margin for two-phase and four-phase modulation, with differential phase detection as determined by methods similar to those for the other modulation methods in the above reference. These methods essentially consist in determining the maximum intersymbol interference that can be encountered, considering the pulse shapes shown in Fig. 15 and all possible pulse patterns over the number of pulse intervals that contribute

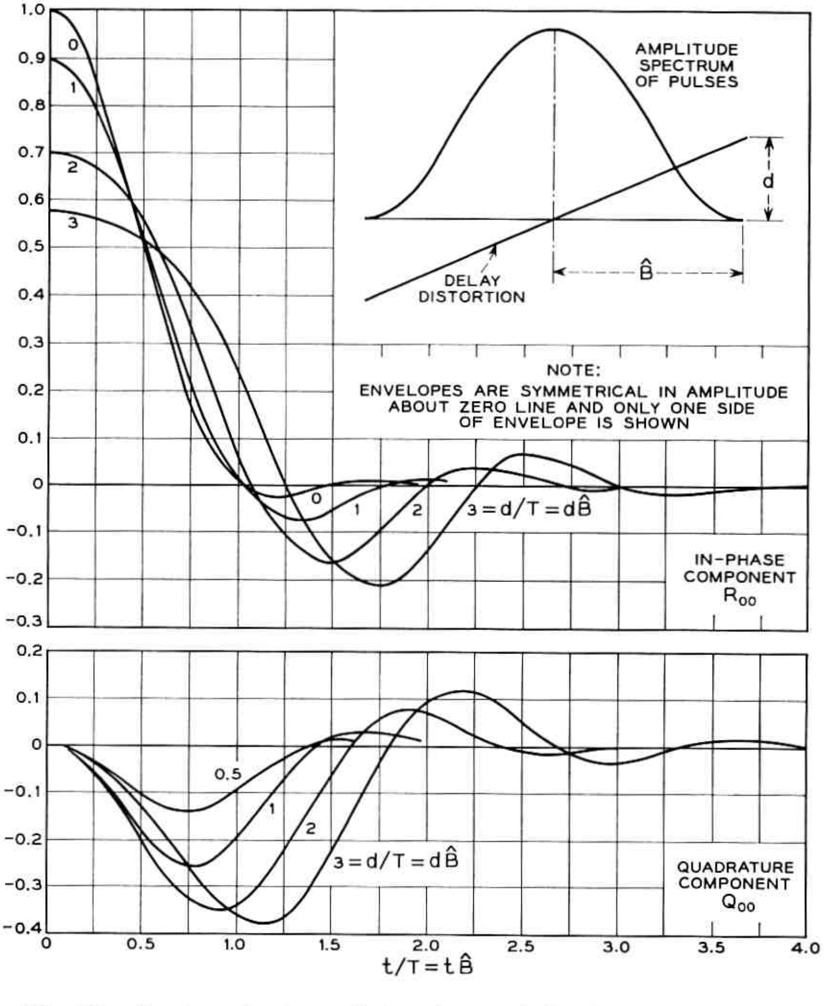


Fig. 15 — Carrier pulse transmission characteristics for raised cosine pulse spectrum and linear delay distortion. For negative values of $t/T = t \cdot \hat{B}$ the characteristics are the same as shown for positive values.

significantly to intersymbol interference. Exact analytic determination of the maximum impairments does not appear feasible, and it becomes necessary to resort to trials for selection of the worst condition. It should be noted that with binary PM with differential phase detection the optimum threshold level differs from zero owing to a bias component in the demodulator output.[13] The curve in Fig. 16 and the analysis that follows assume automatic adjustment to the optimum threshold level,

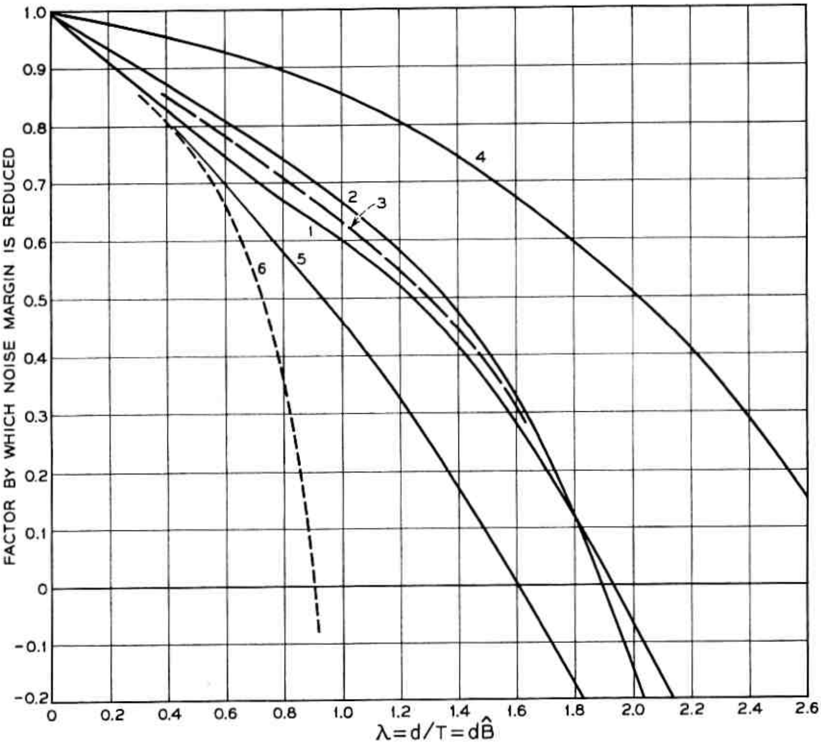


Fig. 16 — Maximum reduction in noise margin owing to linear delay distortion: 1, binary AM with envelope detection; 2, binary FM with frequency discriminator detection; 3, binary PM with differential phase detection; 4, binary PM with synchronous detection; 5, four-phase modulation with synchronous detection; 6, four-phase modulation with differential phase detection.

and a significantly greater error probability would be encountered with zero threshold level.

It will be noted that the noise margin is reduced to zero for certain values λ_0 of λ. These values apply for certain combinations of baseband pulses in about four pulse positions. The probability of this and other pulse patterns must be considered in evaluating error probability as discussed below.

4.8 *Probability of Errors from Linear Delay Distortion*

As λ is increased slightly above the value λ_0 mentioned above, intersymbol interference increases rapidly. Thus errors will occur for a value λ_e of λ only slightly greater than λ_0, for certain combinations of two pulses, occurring at times $-T$ and $+T$ relative to the sampling instant $t = 0$. There are four possible combinations of these two pulses. For one of these (say 1, 1), an error will occur if $\lambda \geqq \lambda_e$. For another (say $-1, -1$), an error will occur if $\lambda \leqq -\lambda_e$. For the other combinations $(-1, 1)$ and $(1, -1)$, intersymbol interference will cancel so that the probability of error is zero. The probability of error is thus

$$\begin{aligned} P_e &= \tfrac{1}{2}(\tfrac{1}{4} + \tfrac{1}{4})P(\,|\lambda| \geqq |\lambda_e|\,) \\ &= \tfrac{1}{4}P(\,|\lambda| \geqq |\lambda_e|\,) \end{aligned} \tag{96}$$

where $P(\,|\lambda| \geqq |\lambda_e|\,)$ is the probability that the absolute value of λ is greater than λ_e.

For a given value $\lambda_e = d_e\hat{B}$ the corresponding slope $\ddot{\varphi}$ of the linear delay distortion is

$$\begin{aligned} \ddot{\varphi}_e &= d_e/2\pi\hat{B} \\ &= \lambda_e/2\pi\hat{B}^2. \end{aligned} \tag{97}$$

The following relation applies

$$P(\,|\lambda| \geqq |\lambda_e|\,) = P(\,|\ddot{\varphi}| \geqq |\ddot{\varphi}_e|\,). \tag{98}$$

The probability distribution represented by the right-hand side of (98) is given by (55) with $\ddot{\varphi}_1 = \ddot{\varphi}_e$. For small probabilities (56) applies, so that in view of (96) and (98) the error probability is

$$\begin{aligned} P_e &= \tfrac{1}{4}P(\,|\ddot{\varphi}| \geqq |\ddot{\varphi}_e|\,) \\ &= \frac{1}{2\pi k_e}\left[1 + \ln\left(\frac{k_e}{2} + 1\right)\right] \end{aligned} \tag{99}$$

where

$$\begin{aligned} k_e &= 3\ddot{\varphi}_e/\Delta^2 \\ &= 3\lambda_e/2\pi\Delta^2\hat{B}^2. \end{aligned} \tag{100}$$

With (100) in (99)

$$P_e = \frac{\Delta^2\hat{B}^2}{3\lambda_e}\left[1 + \ln\left(1 + \frac{3\lambda_e}{4\pi\Delta^2\hat{B}^2}\right)\right]. \tag{101}$$

From Fig. 16 it will be noted that for binary AM and FM, and for binary PM with differential phase detection, $\lambda_0 \approx 1.8$. For these cases it appears a legitimate approximation to take $\lambda_e = 2$. On this premise the error probabilities given in Table IV are obtained for various values of the parameter $\Delta\hat{B}$.

TABLE IV — PROBABILITY OF ERRORS IN A DIGIT OWING TO LINEAR DELAY DISTORTION IN ABSENCE OF NOISE FOR BINARY AM, FM AND PM (WITH DIFFERENTIAL PHASE DETECTION)

$\Delta\hat{B} = 10^{-4}$	10^{-3}	10^{-2}	10^{-1}
3.1×10^{-8}	2.4×10^{-6}	1.6×10^{-4}	8×10^{-3}

The above error probabilities are shown in Fig. 17 as a function of $\Delta\hat{B}$. If, for example, $\Delta = 10^{-7}$ second and $\hat{B} = 10^5$ cps, then $\Delta\hat{B} = 10^{-2}$ and $P_e = 1.6 \times 10^{-4}$. Pulses could in this case be transmitted at a rate of 100,000 per second with a minimum error probability $P_e = 1.6 \times 10^{-4}$. In the presence of noise the error probability will be greater, as discussed in a later section.

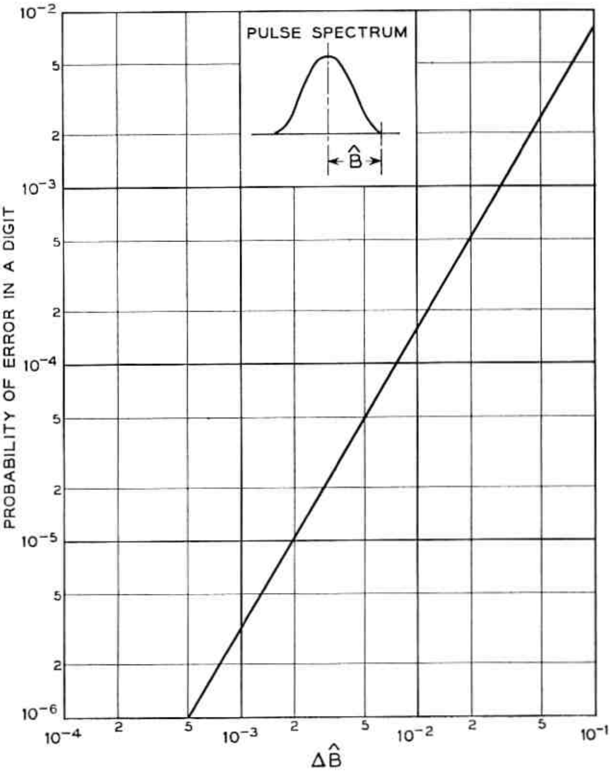


Fig. 17 — Error probability in binary AM, FM and PM owing to linear delay distortion for maximum departure Δ (seconds) from mean transmission delay.

The error probability with four-phase modulation and differential phase detection can be determined in a similar way. In this case $\lambda_0 \approx 0.9$ and $\lambda_e \approx 1$ in (101).

V. ERRORS FROM TRANSMITTANCE VARIATIONS WITH TIME

5.1 *General*

As mentioned in Section 4.1, transmittance variations with time is a second basic source of error in digital transmission. In transmission at low rates the bandwidth $\hat{B}$ of the pulse spectra will be narrow, so that fading can be regarded as constant over the spectrum band. Errors from selective fading, as considered in Section IV, can then be disregarded. On the other hand, the duration of a signal interval T may then be sufficiently long so that consideration must be given to random fluctuations in the amplitude, phase and frequency of the carrier between one signal interval and the next. Errors may occur owing to such fluctuations even in the absence of noise. The probability of errors in this account is evaluated here.

5.2 *Amplitude Variations*

The amplitude of a received wave will fluctuate with a Rayleigh distribution (10). Because of the great range of fluctuation, it is essential to provide automatic gain control at the receiver to prevent overloading and resultant adverse effects. Such gain control is activated by circuitry that integrates the received wave over a number of signal intervals T. With FM and PM only a few pulse intervals are required, for the reason that the received carrier wave is essentially independent of the pulse patterns. It is thus possible to provide effective gain control against rapid variations in the received carrier wave that occurs over a few signal intervals. Moreover, with FM and PM the distinction between marks and spaces is made by positive and negative deviations from zero threshold level in the detection process. This permits the use of limiters at the input to the detectors, to prevent the adverse effect of rapid fluctuations in the amplitude of the received carrier wave owing to fading. These advantages in applications to fading channels are not shared by AM, for reasons outlined below.

In binary AM or on-off carrier transmission, the received wave may be absent over a large number of consecutive signal intervals T. Hence automatic gain control must be activated by circuitry that integrates the received pulse train over a very large number of signal intervals T;

otherwise gain would be increased during long spaces, regardless of the fading condition. For this reason automatic gain control is inherently slow, in relation to the duration of a signal interval. It may thus be ineffective as applied to transmission at slow rates. With transmission at high rates, however, such that variations in the received wave owing to fading are inappreciable even over a large number of signal intervals, it may be possible to implement effective gain control.

At low transmission rates, such that fading is virtually constant over the band of the pulse spectrum, intersymbol interference can be made inappreciable. In this case it is possible to employ limiting prior to detection, and this method may then be more effective than automatic gain control, or could be used in conjunction with it. The limiter would slice the received wave at an appropriately selected level L. In the choice of the optimum slicing level it is necessary to consider the probability of errors during a mark owing to fading such that the received wave is less than L. In accordance with (10) this probability is

$$\begin{aligned} P(r \leqq L) &= 1 - \exp\left(-L^2/\bar{r}^2\right) \\ &\approx L^2/\bar{r}^2. \end{aligned} \tag{102}$$

A second consideration in the choice of L is the probability of errors owing to noise during a space, which is increased as L is reduced. The optimum threshold level considering both effects is determined in Section 6.9.

Owing to even small intersymbol interference, the use of a limiter as postulated above may be precluded in actual systems. For example, let L be 10 per cent of the rms signal amplitude $\bar{r}$, and let intersymbol interference be 5 per cent of L when the received signal is just equal to L. When the received signal is increased by a factor 20, intersymbol interference would be increased correspondingly and would be equal to L. Hence errors would occur even in the absence of noise. This is the inherent reason why limiting is generally ineffective as applied to binary AM. However, even if intersymbol interference could be disregarded, the error probability in the presence of noise will be greater than with binary PM or FM, as shown in Section 6.9.

5.3 *Carrier Frequency Variations*

In transmission over troposcatter links, random fluctuations will occur in the carrier frequency, which may be important from the standpoint of receiver implementation with any modulation method. Such fluctuations can be limited at the input to the IF filter with the aid of signal-tracking oscillators for demodulation of the received radio frequency wave. The frequency of such oscillators may be controlled by feedback from the mixer output or from the detector output. The following expressions apply for the probability distribution of carrier frequency fluctuations without such frequency control at the receiver.

The probability distribution of frequency variations is given by (38). For a Gaussian fading power spectrum, the probability that the frequency variation $\varphi' = \Delta\omega$ exceeds $k\sigma$ is thus

$$P(\mid \Delta\omega \mid \geqq k\sigma) \approx (1/2k^2). \tag{103}$$

The equivalent fading bandwidth is in accordance with (19) $\bar{\gamma} \approx 1.25\sigma$. The probability that $\Delta\omega$ exceeds $k\bar{\gamma}$ is thus

$$P(\mid \Delta\omega \mid \geqq k\bar{\gamma}) \approx (1/3k^2). \tag{104}$$

Since σ and $\bar{\gamma}$ are nearly proportional to the carrier frequency, it follows that the frequency fluctuations encountered with a specified probability will be nearly proportional to the carrier frequency. By way of example let $\bar{\gamma} \approx 2$ radians/second or about 0.3 cps. The probability that the frequency fluctuation exceeds 30 cps is in this case obtained from (104) with $k = 100$ and is 3×10^{-5}. It appears that for bandwidths of the pulse spectra in excess of about 5000 cps, frequency fluctuations will not be important. However, for narrow band spectra the random frequency excursions may become excessive and give rise to errors, particularly with frequency modulation, as discussed below.

5.4 *Frequency Variations over a Signal Interval*

It will be assumed that the carrier frequency excursion is limited with the aid of a signal-tracking oscillator, or that a demodulation process is used in binary FM in which the change from mark to space is based on comparison of the frequencies in adjacent signal intervals of duration T. If the separation between mark and space frequencies is $2\Omega_{01}$, an error will occur if the frequency is changed by $+\Omega_{01}$ for a space and by $-\Omega_{01}$ for a mark.

From (41) it is possible to determine the probability of errors owing to frequency changes $\pm\Omega_{01}$ over a signal interval of duration T. The maximum permissible value of φ'' is determined from

$$\varphi_{\max}''T = \pm\Omega_{01} \tag{105}$$

where the positive sign applies for a space and the negative sign for a mark.

With an ideal pulse spectrum the pulse interval is given by $T = \pi/\Omega$, so that (105) can be written

$$\varphi''_{\max} = \pm\Omega_{01}\Omega/\pi. \tag{106}$$

5.5 *Error Probability in Binary FM*

The error probability is in this case

$$P_e = \tfrac{1}{2}P(\,|\,\varphi''\,| \geqq |\,\varphi_{\max}''\,|\,) \tag{107}$$

where the factor $\frac{1}{2}$ occurs when the probability functions is defined in terms of the absolute values as in (41).

The parameter k defined by (42) in this case becomes

$$\begin{aligned} k_{\max} &= \varphi_{\max}''/\sigma^2 \\ &= \Omega_{01}\Omega/\pi\sigma^2. \end{aligned} \tag{108}$$

With frequency discriminator detection, $\Omega_{01} = \Omega$. For a raised cosine spectrum, $\hat{B} = 2B = \Omega/\pi$ and

$$k_{\max} = \pi\hat{B}^2/\sigma^2. \tag{109}$$

Employing (45), the probability (107) of an error becomes

$$P_e = \left(\frac{\sigma}{\pi\hat{B}}\right)^2 \left[1 + \ln\left(1 + \frac{\pi\hat{B}^2}{2\sigma^2}\right)\right]. \tag{110}$$

In the above relation, σ is in radians/second while $\hat{B}$ is in cps. The equivalent fading bandwidth is, in accordance with (19), $\bar{\gamma} \approx 1.25\sigma$. The ratio of the maximum bandwidth $\hat{B}$ in cps to $\bar{\gamma}$ in cps is thus

$$\mu = \frac{\hat{B}}{\bar{\gamma}/2\pi} = \frac{2\pi\hat{B}}{1.25\sigma} \approx \frac{5\hat{B}}{\sigma}. \tag{111}$$

The probability of error (110) is given in Table V for various ratios μ. These error probabilities are shown in Fig. 18.

TABLE V — ERROR PROBABILITIES WITH BINARY FM FROM FLAT RAYLEIGH FADING IN ABSENCE OF NOISE

$\mu = 10$	100	1000	10000
$\hat{B}/\sigma = 2$	20	200	2000
6×10^{-3}	9.3×10^{-5}	1.4×10^{-6}	1.8×10^{-8}

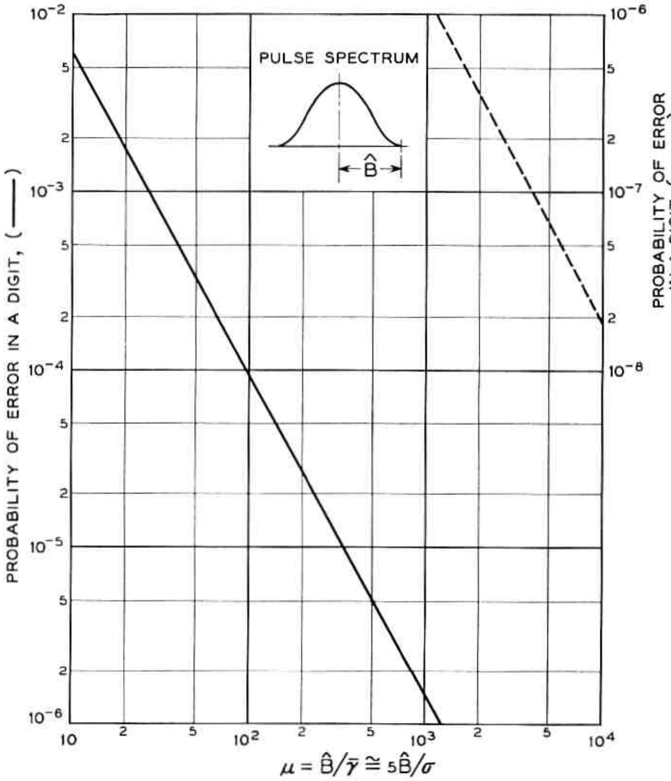


Fig. 18 — Error probability in binary FM in absence of noise, owing to frequency variations over a pulse interval T resulting from flat Rayleigh fading.

5.6 *Phase Variations over a Signal Interval*

The probability density of the carrier phase is $1/2\pi$, such that any phase may be encountered unless the carrier phase wander is limited by phase tracking oscillators in the demodulation process. In a digital phase modulation system where appreciable phase wander may be expected, the preferable demodulation method is differential phase detection. With this method the phase error will be limited to that encountered over a signal interval T.

From (36) it is possible to determine the probability of an error for a given maximum tolerable phase change θ over an interval T. For $k \gg 1$ the following relation applies

$$P(|\varphi'| \geqq |\varphi_1'|) = \frac{1}{2k^2} \tag{112}$$

$$= \frac{b_2}{2b_0} \frac{T^2}{\theta^2}. \tag{113}$$

With a Gaussian fading power spectrum (40) applies and

$$P[|\varphi'| \geqq (\varphi_1')] = (\sigma^2 T^2/2\theta^2). \tag{114}$$

5.7 *Error Probabilities in PM*

With two-phase modulation $\theta = \pm(\pi/2)$, while with four-phase modulation $\theta = \pm(\pi/4)$. Hence the probability of error with these methods as obtained from (114) is, for two-phase modulation

$$P_e \approx (2/\pi^2)\sigma^2 T^2 \approx 0.2\sigma^2 T^2 \tag{115}$$

and for four-phase modulation

$$P_e \approx (8/\pi^2)\sigma^2 T^2 \approx 0.82\sigma^2 T^2. \tag{116}$$

These expressions apply provided the signal duration is sufficiently short so that the change in phase is small and can be considered linear over the interval. More accurate expressions that do not involve this assumption have been derived by Voelcker[9] for the error probability. Thus, with two-phase modulation the error probability is actually

$$P_e = \tfrac{1}{2}[1 - \kappa(T)] \tag{117}$$

and with four-phase modulation

$$P_e = \frac{1}{2} - \frac{2}{\pi}\kappa(T)[2 - \kappa^2(T)]^{-\frac{1}{2}} \tan^{-1} \frac{\kappa(T)}{[2 - \kappa^2(T)]^{\frac{1}{2}}} \tag{118}$$

where $\kappa(T) = \kappa(\tau)$ for $\tau = T$, i.e., the autocorrelation function for each quadrature component as defined by (15).

For a Gaussian fading spectrum, $\kappa(T)$ as obtained from (17) is

$$\kappa(T) = \exp(-\sigma^2 T^2/2). \tag{119}$$

For $\sigma T \ll 1$:

$$\kappa(T) \approx 1 - \sigma^2 T^2/2. \tag{120}$$

With the latter approximation in (117) and (118), the error probability with two-phase modulation becomes

$$P_e \approx \tfrac{1}{4}\sigma^2 T^2 = 0.25\sigma^2 T^2 \tag{121}$$

and with four-phase modulation

$$P_e = \left(\frac{1}{2} + \frac{1}{\pi}\right)\sigma^2T^2 \approx 0.82\sigma^2T^2 \tag{122}$$

which are to be compared with (115) and (116), respectively. The somewhat greater inaccuracy with two-phase than with four-phase modulation comes about since the phase change $\pm(\pi/2)$ cannot be considered small as required for (114) to apply.

In the above relations T is the interval between phase changes, which is related to the bandwidth of the baseband pulse spectrum. With idealized spectra of the type shown in Fig. 12, the interval is

$$T = 1/2B \text{ (two-phase)} \tag{123}$$

$$= 1/4B \text{ (four-phase)} \tag{124}$$

where B is the equivalent mean bandwidth.

In the particular case of pulses with a raised cosine spectrum, the maximum bandwidth is

$$\hat{B} = 2B \tag{125}$$

so that

$$\begin{aligned} T &= 1/\hat{B} \text{ (two-phase)} \\ &= 1/2\hat{B} \text{ (four-phase).} \end{aligned} \tag{126}$$

In terms of the above bandwidth the error probabilities (115) and (116) are thus the same for both two-phase and four-phase modulation and are given by

$$P_e \approx 0.05(\sigma/B)^2 \tag{127}$$

$$\approx 0.2(\sigma/\hat{B})^2. \tag{128}$$

The above relations apply for any number of phases. For this reason the capacity of a noiseless channel could be increased indefinitely by increasing the number of phases. There will, however, be certain limitations in this respect owing to intersymbol interference, as in stable channels.

The above error probability is shown in Table VI for various values of $\hat{B}/\sigma$ and $\mu = 5\hat{B}/\sigma$, where μ is the ratio defined by (111). It will be noted that these error probabilities are somewhat smaller than with binary FM as given in Table V.

The above probabilities of an error in a single digit are shown in Fig. 19, as a function of μ.

TABLE VI — ERROR PROBABILITIES WITH DIFFERENTIAL PM FROM FLAT RAYLEIGH FADING IN ABSENCE OF NOISE

$\mu = 10$	100	1000	10000
$\hat{B}/\sigma = 2$	20	200	2000
2×10^{-3}	2×10^{-5}	2×10^{-7}	2×10^{-9}

As noted in Section 1.6, there will be a certain median value of $\bar{\gamma}$ and thus a certain median value of μ and corresponding median error probability. During certain intervals, the error probabilities will be significantly smaller or significantly greater than the median error probabilities.

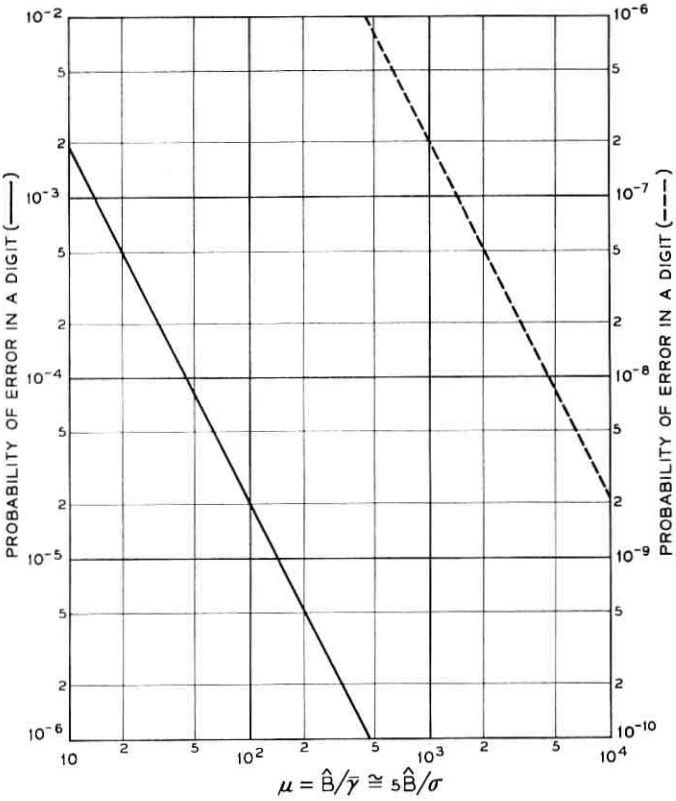


Fig. 19 — Error probability in binary PM with differential phase detection in absence of noise, owing to phase variations over pulse interval T resulting from flat Rayleigh fading.

VI. ERRORS FROM NOISE WITH FLAT RAYLEIGH FADING

6.1 *General*

As mentioned in Section 4.1, a third basic source of errors in troposcatter transmission is random noise. The probability of errors from noise depends on the modulation and detection methods and on their implementation. For optimum performance it is in the first place necessary to have appropriate pulse spectra such that intersymbol interference is avoided in transmission over ideal channels. Moreover, the error probability depends on the division of spectrum shaping between transmitting and receiving filters. The minimum error probabilities with various modulation and detection methods as quoted here are based on optimum design in the above and various other respects, such as accurate sampling of pulse trains. The probability of errors from noise in actual systems will be greater owing to various imperfections in implementation.

6.2 *Signal-to-Noise Ratios*

In carrier pulse transmission over an ideal channel, the sideband spectrum of the carrier pulses at the detector input will be symmetrical about the carrier frequency. As discussed elsewhere,[14] it is possible to realize optimum performance in binary transmission by AM, PM and FM with an infinite variety of pulse spectra at the detector input with the general properties discussed in Section 4.3.

The error probability in digital transmission over noisy channels is ordinarily specified in terms of the average signal-to-noise ratio at the input to the receiving filter that ordinarily precedes the detector. This signal-to-noise ratio is ordinarily taken as

$$\rho = S/N$$

S = average carrier power at detector input

N = average noise power in a flat band $B = 1/2T$ at input-to-receiving filter.

When S represents the average signal power in a fading channel, the designation $\bar{\rho} = S/N$ will be used in place of ρ.

The above reference band B is the minimum possible bandwidth in baseband pulse transmission without intersymbol interference. The minimum possible bandwidth in double sideband transmission as used in binary AM, PM and FM is $2B$.

The error probability as related to ρ will depend on the division of spectrum shaping between transmitting filters and the receiving filter at the detector input. With optimum division, the error probability is the same as for transmission over a flat band B to each side of the carrier frequency.[14] Such a flat channel band is ordinarily assumed or implied in theoretical analyses, though not feasible in actual systems.

6.3 *Error Probabilities with Flat Rayleigh Fading*

Let r be the signal amplitude and $P_e^0(r)$ the error probability of errors owing to random noise in transmission over a stable channel with signal amplitude r. In the presence of fading, let the probability density of various signal amplitudes be $p(r)$. The error probability in transmission over fading channels is then

$$P_e = \int_0^\infty P_e^0(r)p(r)\,dr. \tag{129}$$

With Rayleigh fading the probability density $p(r)$ is the derivative of (27) with respect to r_1. With r in place of r_1 the probability density is

$$p(r) = (2r/\bar{r}^2)\exp(-r^2/\bar{r}^2) \tag{130}$$

$$= (r/S)\exp(r^2/2S) \tag{131}$$

where $S = \bar{r}^2/2$ is the average signal power.

6.4 *Binary PM with Synchronous Detection*

In binary PM, marks and spaces are transmitted by phase reversals. With ideal coherent or synchronous detection the error probability in transmission over a stable channel is

$$P_e^0 = \tfrac{1}{2}\operatorname{erfc}(\rho/2)^{\frac{1}{2}}. \tag{132}$$

The error probability with Rayleigh fading as obtained from (129) is, in this case[7,9]

$$P_e = \frac{1}{2}\left[1 - \left(\frac{\bar{\rho}}{\bar{\rho}+1}\right)^{\frac{1}{2}}\right] \approx \frac{1}{4\bar{\rho}} \tag{133}$$

where $\bar{\rho} = S/N$ = ratio of average received signal power with Rayleigh fading to average noise power as previously defined.

6.5 *Binary PM with Differential Phase Dection*

With binary PM and differential phase detection the error probability in transmission over a stable channel is[15]

$$P_e^0 = \tfrac{1}{2}e^{-\rho}. \tag{134}$$

The error probability with Rayleigh fading is, in this case[9]

$$P_e = 1/2(\bar{\rho} + 1). \tag{135}$$

6.6 *Binary FM with Dual Filter Detection*

With this method two receiving filters are used, centered on the space and mark frequencies ω_1 and ω_2, as indicated in Fig. 20, with sufficient separation to avoid mutual interference between the space and mark channels. Complementary binary amplitude modulation is used at the two carrier frequencies, and the two baseband filter outputs are combined with reversal in the polarity of one.

The error probability in transmission over stable channels with coherent detection is[16]

$$P_e^0 = \tfrac{1}{2} \operatorname{erfc} (\rho^{\frac{1}{2}}/2) \tag{136}$$

and with noncoherent detection is[16]

$$P_e^0 = \tfrac{1}{2} \exp (-\rho/2). \tag{137}$$

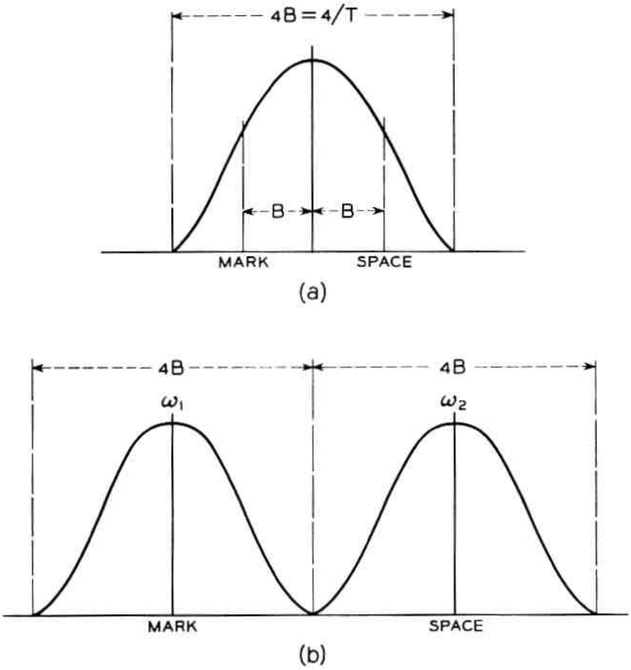


Fig. 20 — Comparison of channel bandwidth requirements in binary FM with (a) frequency discriminator detection and (b) dual filter detection.

Comparison of (136) with (132) shows that the error probability P_e with Rayleigh fading is obtained by replacing in (133) $\bar{\rho}$ with $\bar{\rho}/2$. This yields for coherent detection

$$P_e = \frac{1}{2}\left[1 - \left(\frac{\rho}{\bar{\rho}/2}\right)^{\frac{1}{2}}\right] \approx \frac{1}{2\bar{\rho}}. \tag{138}$$

Comparison of (137) with (134) shows that P_e is obtained by replacing in (135) $\bar{\rho}$ with $\bar{\rho}/2$, in which case, for noncoherent detection

$$P_e = 1/(\bar{\rho} + 2). \tag{139}$$

6.7 *Binary FM with Frequency Discriminator Detection*

With this method a single receiving filter is used, with space and mark frequencies as indicated in Fig. 20. Pulse transmission without intersymbol interference over a channel of the same bandwidth as required for double-sideband AM is in this case possible for certain ideal amplitude and phase characteristics of the channels, as shown elsewhere.[14]

The error probabilities in the absence of fading depends on the characteristics of the bandpass channel filters and the post-detection low-pass filter, and are difficult to determine exactly. Approximate evaluations[14] indicate that for a given error probability, about 4 db greater signal-to-noise ratio would be required than for binary PM with coherent detection, when no post-detection low-pass filter is used. Recent exact evaluations by Bennett and Salz,[17] indicate 3 to 4 db increase in the required signal-to-noise ratio over a variety of filter shapes. With an optimum post-detection low-pass filter, a small improvement may be realized, such that about 3 db increase over binary PM with coherent detection would be expected. On this basis it appears that the error probability will be virtually the same as for binary FM with dual filter coherent detection, such that the principal advantage over the latter method is a two-fold reduction in bandwidth.

6.8 *Binary AM with Ideal Gain Control*

It will be assumed that the receiver can be implemented with ideal automatic gain control, such that the output in the presence of a mark would have a fixed level l and in the presence of a space would be zero. This condition can be approached at sufficiently high transmission rates, such that the received wave prior to gain control changes insignificantly over a large number of pulse intervals of duration T. Under this condition the fading bandwidth is negligible relative to the bandwidth of the baseband pulse spectrum.

On the above premise and with ideal coherent (or synchronous) detection, the optimum threshold level for decision between marks and spaces would be $l/2$. The tolerable peak noise amplitude before an error occurs would be $l/2$, as compared with l for binary PM, resulting in 6 db reduction in noise margin. On the other hand, the average transmitter power is 3 db less than with binary PM. Hence this method would have a 3 db disadvantage compared to binary PM with synchronous detection.

Accordingly, (132) would be replaced by

$$P_e^0 = \tfrac{1}{2} \operatorname{erfc} (\rho/4)^{\frac{1}{2}} \tag{140}$$

and (133) would be replaced by

$$P_e = \frac{1}{2}\left[1 - \left(\frac{\bar{\rho}}{\bar{\rho} + 2}\right)^{\frac{1}{2}}\right]. \tag{141}$$

The above relations are the same as (136) and (138) for binary FM with dual filter coherent detection, and (141) is virtually the same as (135) for binary PM with differential phase detection. Hence binary AM offers no advantage in signal-to-noise ratio even at sufficiently high transmission rates such that ideal gain control could be implemented.

6.9 *Binary AM with Optimum Fixed Threshold Detection*

At low transmission rates, such that the received wave can change appreciably over a few pulse intervals owing to fading, gain control cannot be effectively implemented, as discussed in Section 5.2. Without effective gain control, there will be a certain optimum threshold for distinction between marks and spaces. This optimum level and the corresponding signal-to-noise ratio is determined here on the premise that no gain control is used. This threshold level could be implemented by either a predetection or a postdetection limiter. Assume a probability $\frac{1}{2}$ of a mark being present; in the absence of noise, the probability of errors in marks is, in view of (102)

$$P_e(r \leqq L) = \tfrac{1}{2}[1 - \exp(-L^2/2S)] \tag{142}$$

where L is the threshold level. In the presence of noise the error probability will be only slightly greater than (142).

A second consideration in the choice of L is the probability of errors during a space. This error probability is obtained from (137) with $\rho = L^2/N$ and is

$$P_e(n \geqq L) = \tfrac{1}{2} \exp(-L^2/2N) \tag{143}$$

where n is the instantaneous noise amplitude and N the average noise power.

The combined error probability is

$$P_e = \tfrac{1}{2}[1 - \exp(-\mu/2) + \exp(-\bar{\rho}\mu/2)] \tag{144}$$

where

$$\mu = L^2/S; \qquad \bar{\rho} = S/N. \tag{145}$$

The optimum L or μ is obtained from the condition $dP_e/d\mu = 0$. This yields the following relation for the optimum value μ_0

$$\exp(-\mu_0/2) = \bar{\rho} \exp(-\bar{\rho}\mu_0/2) \tag{146}$$

or

$$\mu_0 = \frac{2 \ln \bar{\rho}}{\bar{\rho} - 1} = \frac{4.606 \log_{10} \bar{\rho}}{\bar{\rho} - 1}. \tag{147}$$

In practicable systems $\bar{\rho} \gg 1$, in the order of 100 or more, and $\mu_0 \ll 1$. With (147) in (144), the following approximation is obtained for the minimum error probability

$$P_{e,\min} \approx \frac{1}{2}\left[\frac{\ln \bar{\rho}}{\bar{\rho} - 1} + \exp(-\ln \bar{\rho})\right]. \tag{148}$$

The above error probability is significantly greater than with binary PM or FM. The error probability (148) is thus greater than for binary FM with dual filter coherent detection by a factor of at least $\ln \bar{\rho}$. For $\bar{\rho} = 1000$ (30 db) this factor is about $\ln \bar{\rho} \approx 7$. Hence about $10 \log_{10} 7 \approx 8.5$ db greater average signal power would be required than with binary FM. This assumes that excessive intersymbol interference is avoided, which may not be feasible for reasons mentioned in Section 5.2. Since it is evident that binary AM is at a considerable disadvantage in signal-to-noise ratio as compared to binary PM and FM, it will not be considered further herein.

6.10 *Combined Rayleigh and Slow Log-Normal Fading*

In the previous determination of error probabilities, rapid Rayleigh fading was assumed, with a fixed mean signal-to-noise ratio $\bar{\rho}$ over the interval under consideration. It will now be assumed that in this interval there is a slow log-normal variation in path loss and thus in signal-to-noise ratio at the receiver, in conjunction with rapid Rayleigh fading.

Let P_e be the error probability with Rayleigh fading as previously related to the mean signal-to-noise ratio $\bar{\rho} = \bar{s}^2/\bar{n}^2$, where $\bar{s}$ is the rms signal amplitude and $\bar{n}$ the rms noise amplitude. If $p(\bar{s})$ is the probability density of the rms amplitudes with slow fading, the probability of error in an interval during which the rms amplitude exceeds $\bar{s}_1$ is

$$P_{e,1} = \int_{\bar{s}_1}^{\infty} P_e(\bar{s})p(\bar{s})\, d\bar{s}. \tag{149}$$

For $\bar{\rho} \gg 1$, the expression for $P_e(\bar{s})$ is of the general form

$$P_e(\bar{s}) \approx c/\bar{\rho} = \frac{c}{\bar{s}^2/\bar{n}^2}. \tag{150}$$

For binary PM with differential phase detection and for binary PM with coherent dual filter detection, $c = \frac{1}{2}$.

The probability density $p(\bar{s})$ is given by (12), or in the present notation

$$p(\bar{s}) = \frac{1}{\sqrt{2\pi}} \frac{1}{\sigma\bar{s}} \exp\left[-(\ln \bar{s}/\bar{s}_0)^2/2\sigma^2\right] \tag{151}$$

where $\bar{s}_0$ is the median rms amplitude and σ is the standard deviation of the fluctuation in $\bar{s}$.

With (150) and (151) in (144)

$$P_{e,1} = c\frac{1}{\sqrt{2\pi}}\frac{1}{\sigma}\int_{\bar{s}_1}^{\infty}\frac{1}{\bar{s}^2/\bar{n}^2}\frac{1}{\bar{s}}\exp\left[-(\ln \bar{s}/\bar{s}_0)^2/2\sigma^2\right] d\bar{s} \tag{152}$$

$$= \frac{c}{2}\frac{1}{\sqrt{2\pi}}\int_{\rho_1}^{\infty}\frac{1}{\rho^2}\exp\left[-(\tfrac{1}{2}\ln \rho/\rho_0)^2/2\sigma^2\right] d\rho \tag{153}$$

where $\rho_0 = \bar{s}_0^{\,2}/\bar{n}^2$ on $\rho_1 = \bar{s}_1^{\,2}/\bar{n}^2$.

Solution of (153) yields the relation

$$P_{e,1} = P_e \cdot \eta(\sigma, \kappa) \tag{154}$$

where

$$\kappa = \rho_1/\rho_0 \tag{155}$$

and

$$\eta(\sigma, \kappa) = \tfrac{1}{2}\exp(2\sigma^2)\operatorname{erfc}\left\{\frac{1}{\sqrt{8}\sigma}\left[4\sigma^2 + \ln \kappa\right]\right\}. \tag{156}$$

For $\rho_1 = 0$, $\ln \kappa = -\infty$ and $\operatorname{erfc}(-\infty) = 2$. Hence for this case

$$\eta = \exp(2\sigma^2). \tag{157}$$

This is the factor by which the error probability taken over a long interval is greater than without a log-normal variation in signal-to-noise ratio and only rapid Rayleigh fading.

Instead of modifying the error probability as above, an alternative method is to use an equivalent mean signal-to-noise ratio $\bar{\rho}_e$ that is smaller than $\bar{\rho}$ by the factor $\exp(-2\sigma^2)$. Thus

$$\bar{\rho}_e = \bar{\rho} \exp(-2\sigma^2). \tag{158}$$

When $\bar{\rho}_e$, $\bar{\rho}$ and σ are all expressed in db, expression (158) can alternatively be written

$$\bar{\rho}_{e,db} = \bar{\rho}_{db} - \sigma_{db}^2/8.69. \tag{159}$$

For example, with a representative value $\sigma_{db} = 8$ db, the last term in (159) is 7.4 db. Thus the charts in the later Figs. 21 and 22 apply when $\bar{\rho}$ is taken 7.4 db less than the median signal-to-noise ratios with log-normal fading.

VII. COMBINED ERROR PROBABILITY

7.1 *General*

In Sections IV to VI, three basic sources of errors in digital transmission over troposcatter links were discussed, and expressions were given for the probability of error from each of these sources in the absence of the others. In a first approximation, the error probability considering all three sources can be evaluated by taking the sum of the three error probabilities. Approximate expressions are given here for the resultant error probabilities, together with charts that facilitate determination of error probability as a function of the binary pulse transmission rate, when the basic system parameters are known. These are the average signal-to-noise ratio $\bar{\rho}$, the mean fading bandwidth $\bar{\gamma}$, and the maximum departure Δ from the mean transmission delay. The error probability for a given transmission rate can be reduced by various means that may or may not entail an increase in total transmitter power or bandwidth or both. For a given total transmitter power and bandwidth, the most effective means to this end is diversity transmission over independently fading paths, as discussed briefly herein.

7.2 *Combined Error Probability*

As a first approximation, the error probability is given by

$$P_e \approx P_e^{(1)} + P_e^{(2)} + P_e^{(3)} \tag{160}$$

where

$P_e^{(1)}$ = probability of errors in the absence of noise owing to intersymbol interference caused by frequency selective Rayleigh fading (Section IV)

$P_e^{(2)}$ = probability of errors in the absence of noise owing to random variations in carrier phase or frequency (Section V)

$P_e^{(3)}$ = probability or error owing to random noise with nonselective Rayleigh fading (Section VI).

As will be evident from the preceding discussion, and from charts that follow, $P_e^{(1)}$ can be disregarded when $P_e^{(2)}$ must be considered, and conversely, for error probabilities $P_e^{(3)}$ in the range of practical interest. Hence in actual applications (160) will take one of the following forms

$$P_e \approx P_e^{(1)} + P_e^{(3)} \tag{161}$$

$$P_e \approx P_e^{(2)} + P_e^{(3)}. \tag{162}$$

In addition, there are intermediate cases in which $P_e \approx P_e^{(3)}$.

In an exact determination of the error probability (161) it is necessary to consider the net effect of random intersymbol interference on the probability of errors owing to random noise, and similarly an exact determination of the error probability (162) the probability distribution of random phase deviations is involved. Intersymbol interference at a particular sampling instant may reduce or increase the tolerance to noise, and the net effect considering all pulse patterns may be such that (161) is a legitimate approximation. Similarly, random fluctuations in the slope of the phase characteristic may decrease or increase the tolerance to noise at a particular sampling instant, and the net effect considering all sampling instants may be such that (162) is a valid approximation. This is evidenced by the following exact relation derived by Voelcker[9] in place of (162) for binary PM with differential phase detection

$$P_e = [\bar{\rho}/(\bar{\rho} + 1)]P_e^{(2)} + P_e^{(3)}. \tag{163}$$

Since $\bar{\rho}$ would ordinarily exceed 100 (20 db), it follows that in this case (162) is a very good approximation to (163).

The exact error probability (161) depends on the probability distribution of phase distortion in conjunction with the probability distribution of intersymbol interference, which involves consideration of all pulse patterns. The combined probability distribution, and in turn the exact error probability, would be very difficult to determine, and hence the inaccuracy involved in (161) cannot readily be assessed. However, if the probability distribution of intersymbol interference were the same as that of the reduction in tolerance to noise owing to random fluctuations in the slope of the phase characteristic, the inaccuracy in (161) would be no greater than that indicated by (162) versus (163). In most engineering applications, substantially greater inaccuracy would be permissible in the estimation of error probability, such that (161) and hence (160) can be considered permissible approximations in the present context.

The above expression (160) is applied below to binary PM and FM.

7.3 *Binary PM with Differential Phase Detection*

For binary PM with differential phase detection $P_e^{(1)}$ is given by (101) with $\lambda_e = 2$ or

$$P_e^{(1)} = \frac{\Delta^2 \hat{B}^2}{6}\left[1 + \ln\left(1 + \frac{3}{2\pi\Delta^2\hat{B}^2}\right)\right]. \tag{164}$$

This error probability is given in Table IV as a function of $\Delta\hat{B}$.

The error probability $P_e^{(2)}$ is obtained from (117), or approximation (121)

$$P_e^{(2)} = \tfrac{1}{2}[1 - \kappa(T)] \tag{165}$$

$$\approx 0.25(\sigma T)^2 \approx 0.06(\sigma/\hat{B})^2 \tag{166}$$

$$\approx 0.039(\bar{\gamma}/\hat{B})^2. \tag{167}$$

The error probability $P_e^{(3)}$ is given by (135) or

$$P_e^{(3)} = 1/2(\bar{\rho} + 1). \tag{168}$$

7.4 *Error Probability Charts for Binary PM*

In Fig. 21 are shown the error probabilities $P_e^{(1)}$, $P_e^{(2)}$ and $P_e^{(3)}$ as a function of the transmission rate, for a raised cosine spectrum. The error probability $P_e^{(1)}$ depends on the maximum deviation Δ from the mean transmission delay, and curves are shown for a number of values of Δ. The probability $P_e^{(2)}$ depends on the mean fading bandwidth $\bar{\gamma}$, and curves applying for several values of $\bar{\gamma}$ are shown. Finally, the error probability $P_e^{(3)}$ depends on $\bar{\rho}$, and is shown for a number of different values of $\bar{\rho}$.

By way of illustration, the combined error probability obtained from (170) is shown by the dashed line in Fig. 20 for the particular case in which $\Delta = 10^{-7}$ second, $\bar{\gamma} = 2$ cps and $\bar{\rho} = 10^4$ (40 db).

The error probability as a function of transmission rate shown by this dashed line could apply to a variety of tropospheric scatter links,

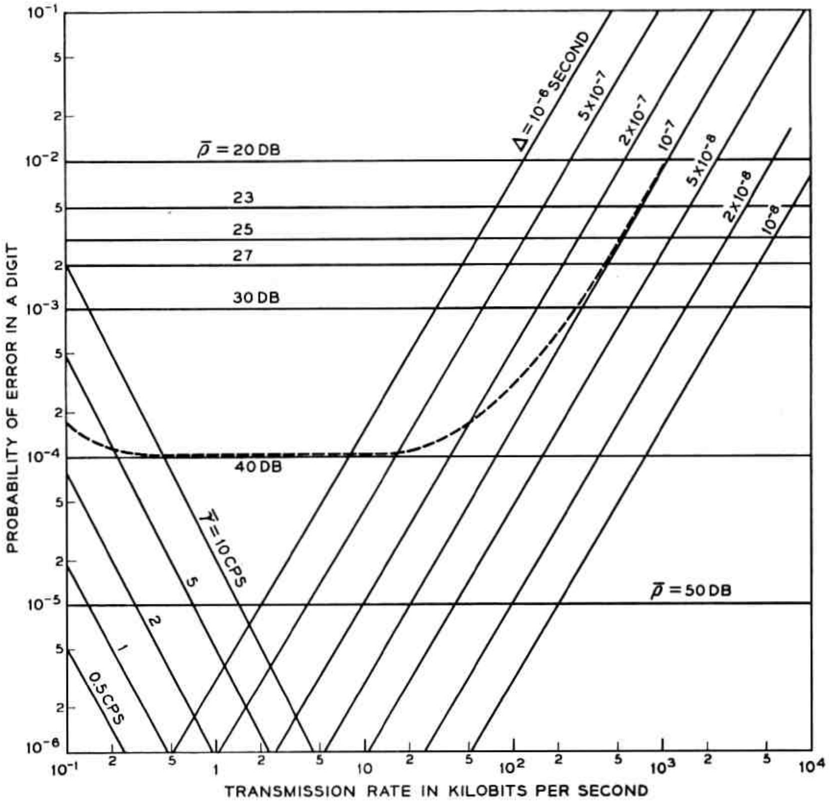


Fig. 21 — Probabilities of errors in binary PM with differential phase detection: 1, curves for various departures from mean delay show error probabilities in absence of noise owing to pulse distortion from selective fading; 2, curves for various mean fading bandwidths $\bar{\gamma}$ show error probabilities in absence of noise owing to random phase variations caused by flat fading; 3, curves for various mean signal-to-noise ratios $\bar{\rho}$ show error probabilities owing to noise for flat Rayleigh fading; 4, dashed curve shows approximate combined error probability for $\bar{\rho} = 40$ db, $\Delta = 10^{-7}$ second, and $\bar{\gamma} = 2$ cps.

since Δ depends on the length of the link and on the antenna beam angles. Moreover, $\bar{\rho}$ depends on the transmitter power, the length of the link, and the antenna beam angles. Hence, given values of Δ and $\bar{\rho}$ can be realized for a great variety of conditions.

7.5 *Binary FM with Frequency Discriminator Detection*

With frequency discriminator detection, the minimum required bandwidth for a given pulse transmission rate is the same as for binary PM, and half as great as that required with dual filter detection.

The error probability $P_e^{(1)}$ is in a first approximation the same as (161) for binary PM with differential phase detection. For the error probability $P_e^{(2)}$, approximation (110) applies, or

$$P_e^{(2)} = \left(\frac{\sigma}{\pi\hat{B}}\right)^2 \left[1 + \ln\left(1 + \frac{\pi\hat{B}^2}{2\sigma^2}\right)\right]. \tag{169}$$

This error probability is given in Table V as a function of $\hat{B}/\sigma$.

The probability of error owing to noise is, in a first approximation, the same as given by (139) for dual filter detection with coherent detection, or

$$P_e^{(3)} \approx 1/2\bar{\rho}. \tag{170}$$

7.6 *Error Probability Charts for Binary FM*

In Fig. 22 are shown the error probability $P_e^{(1)}$, $P_e^{(2)}$ and $P_e^{(3)}$ for binary FM as a function of the transmission rate. The curves apply for a raised cosine pulse spectrum, and the same basic parameters σ, $\bar{\gamma}$ and $\bar{\rho}$ as shown in Fig. 21 for binary PM. The error probability for the particular set of parameters previously assumed in Section 7.4 is shown by the dashed curve.

Comparison of the curves in Figs. 21 and 22 shows that the error probabilities are the same with both methods except at very low transmission rates. This applies only as a first approximation and with ideal implementation of both methods.

7.7 *Diversity Transmission Methods*

In diversity transmission, either space, frequency or time diversity can be used. The performance would be the same with these methods, and is an optimum when there is no correlation between the diversity paths. This entails adequate separation of receiving antennas in space diversity, adequate frequency separation in frequency diversity, or adequate time intervals between repetition of signals in time diversity.

With any one of the above three methods, different combining or decision procedures can be used at the receiver, as discussed in considerable detail by Brennan.[17] The optimum method from the standpoint of minimum required signal power for a specified error probability is known as "maximal ratio combining," in which the gain of the receiver in each path is made proportional to the input signal-to-noise ratio. This method is difficult to implement, and a simpler but somewhat less efficient method is "equal gain combining," in which the various receivers have equal gain and the demodulator baseband output are combined linearly.

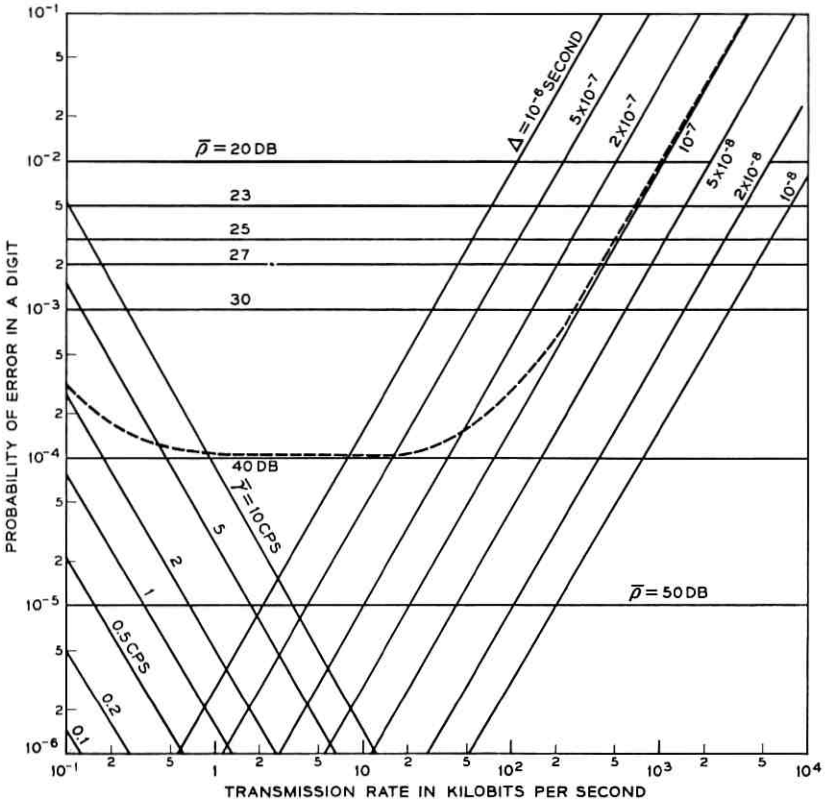


Fig. 22 — Probabilities of errors in binary FM with frequency discriminator detection: 1, curves for various departures Δ from mean delay show error probabilities in absence of noise owing to pulse distortion from selective fading; 2, curves for various mean fading bandwidths $\bar{\gamma}$ show error probabilities in absence of noise owing to random frequency variations caused by flat fading; 3, curves for various mean signal-to-noise ratios $\bar{\rho}$ at detector input show error probabilities owing to noise for flat Rayleigh fading; 4, dashed curve shows approximate combined error probability for $\bar{\rho} = 40$ db, $\Delta = 10^{-7}$ second and $\bar{\gamma} = 2$ cps.

This entails a demodulator in each diversity channel and common gain control of the various channels. The need for a demodulator in each diversity channel and common gain control is avoided with "selection diversity," in which the receiver having the largest signal is selected. Though this method is somewhat less efficient than equal gain combining, it has greater flexibility in that it can be used in conjunction with both linear and nonlinear modulation and detection methods, with path selection on the basis of predetection as well as post detection signals.

The principal diversity techniques would thus be space, frequency or time diversity, in conjunction with "equal gain combining" or "selection diversity." The error reduction afforded by the two latter methods is discussed below.

7.8 *Error Probabilities with Equal Gain Diversity*

The error reduction afforded by equal gain diversity transmission has been determined by Pierce[8] for binary FM with coherent and noncoherent dual filter detection, on the premise of sufficiently slow flat Rayleigh fading, such that errors from noise alone need to be considered. For binary PM with differential phase detection, the error probability with equal gain diversity transmission has been determined by Voelcker,[9] considering both errors from noise $[P_e^{(3)}]$ and errors from time variations in the transmittance $[P_e^{(2)}]$. Voelcker has also determined the error probability with dual diversity transmission for four-phase modulation with differential phase detection, considering errors from transmittance variations with time alone. For all of the above cases, the following approximation applies for the probability of single digit errors with dual diversity transmission over independently fading paths

$$P_{e,2} \approx 3P_{e,1}^{2} \tag{171}$$

where $P_{e,1}$ is the error probability for transmission over a single path (no diversity). For four-phase modulation, Voelcker's more exact expression, when reduced to small error probabilities, gives a factor $4\pi(3 + \pi)/(2 + \pi)^2 \approx 3.13$ in place of 3 in (171).

The mechanism responsible for error reduction by diversity transmission in the above cases also applies to transmission over channels with selective fading when the errors are caused principally by intersymbol interference. With independently fading transmission paths there will be no correlation between intersymbol interference in the various channels, even though the signals are the same. Hence relation (171) would also be expected to apply for the combined error probability P_e given by (160).

For small error probabilities, the following approximate expression is given by Pierce[8] for the error probability owing to noise with flat Rayleigh fading for binary FM and multidiversity transmission

$$P_{e,m} \approx \frac{(2m-1)!}{m!\,(m-1)!} P_{e,1}^{m} \tag{172}$$

$$P_{e,2} \approx 3P_{e,1}^{2} \tag{173}$$

$$P_{e,3} \approx 10P_{e,1}^{3} \tag{174}$$

$$P_{e,4} \approx 35P_{e,1}^{4}. \tag{175}$$

The optimum number of diversity paths will depend on a variety of considerations, among them the available bandwidth and transmitter power, system complexity, and the source of errors. When the errors are caused by noise it is possible to realize a certain minimum total average signal power for a specified error probability $P_{e,m}$, by appropriate choice of m. As shown by Pierce[18] and Harris,[19] the minimum total average signal power is attained for any specified error probability when m is so chosen that in each diversity channel $\bar{\rho} \approx 3$, or about 5 db, for binary FM with dual filter noncoherent detection. The number of diversity paths required to realize the minimum total average signal power is rather large, and the signal power reduction that can be realized with more than four paths is fairly small. For example, Pierce[18] shows that for an error probability $P_{e,m} = 10^{-4}$, the minimum average signal power is realized with $m = 16$, for which the total signal-to-noise ratio is 16.7 db, corresponding to a signal-to-noise ratio per channel of 4.7 db ($\bar{\rho} = 2.95$). With $m = 1$ the average signal-to-noise ratio is 40 db and with $m = 4$ is 19.4 db. Hence only a small additional reduction in signal power is realized when the number of diversity paths is increased from $m = 4$ to $m = 16$.

7.9 *Error Probabilities with Selection Diversity*

Equal gain diversity as considered above entails a linear addition of the baseband outputs of the various demodulators, and would be less effective in conjunction with nonlinear demodulation methods, such as binary FM with frequency discriminator detection. With the latter method, switch or selection diversity reception would probably be preferable, in which only the receiver having the largest signal is selected. With this method the following relations apply for m-diversity transmission when the errors are caused by noise and when receiver selection is based on the largest carrier signal at the detector input[8]

$$P_{e,m} \approx 2^{m-1} m! P_{e,1}{}^{m} \tag{176}$$

$$P_{e,2} \approx 4P_{e,1}{}^{2} \tag{177}$$

$$P_{e,3} \approx 24P_{e,1}{}^{3} \tag{178}$$

$$P_{e,4} \approx 192P_{e,1}{}^{4}. \tag{179}$$

For equal error probability, the average signal power with selection diversity must be greater than with optimum diversity by a factor equal to the mth root of the ratio of the factors in (176) and (172). The power must thus be increased by 0.62, 1.27 and 1.85 db for $m = 2$, 3 and 4, respectively.

7.10 *Multiband Digital Transmission*

The curves in Figs. 21 and 22 suggest that for a given total transmitter power and channel bandwidth, the error probability can be reduced by transmitting at a slower rate over each of a number of narrower channels in parallel. An approximate optimum bandwidth for each channel would be such that $P_e^{(1)} + P_e^{(2)}$ is minimized. This can be accomplished with separate transmitters and receivers for each channel, such that mutual interference between channels is avoided. Hence the adverse effects of selective fading can be overcome with the aid of more complicated terminal equipment, without the need for increased signal power or channel bandwidth.

An alternative method that is simpler in implementation is to transmit the combined digital wave from the parallel channels by frequency or phase modulation of a common carrier, as ordinarily used for transmission of voice channels in frequency division multiplex. This method entails some mutual interference between channels, as well as greater channel bandwidth and carrier power than with direct digital carrier modulation, as discussed below.

With the above method, the spectrum of the modulated carrier wave will have greater bandwidth than with direct digital carrier modulation. To avoid excessive transmission distortion of the combined wave, the bandwidth between transmitter and receiver must be at least twice that with digital carrier modulation. Hence, at least 3 db greater average carrier power is required in order that the noise threshold level of the common channel be comparable with that of direct digital carrier modulation.

With such multiband transmission, intersymbol interference owing to selective fading is avoided, in exchange for mutual interference between the various channels owing to intermodulation distortion caused by selective fading. Such intermodulation distortion is dealt with elsewhere (this issue, part 2) for a modulating wave with the properties of random noise, which is approximated with a large number of binary channels in frequency division multiplex. The results indicate that under this condition intermodulation distortion will cause less transmission impairment than does intersymbol interference in direct digital transmission. Hence multiband transmission by common carrier modulation permits a reduction in error probability in exchange for at least a twofold increase in bandwidth and carrier power. However, this reduction in error probability may be less than can be realized with direct digital carrier modulation in conjunction with a twofold increase in bandwidth and signal power with dual diversity.

Error probabilities in binary multiband transmission by frequency modulation of a common carrier are dealt with by Barrow[21] on the premise of slow flat fading over the combined band, so that only errors owing to noise need be considered and intermodulation distortion can be disregarded.

VIII. SUMMARY

The objective of this analysis has been to develop a transmission and modulation theory for troposcatter systems, applicable to digital transmission by AM, FM and PM at any speed and based on a realistic idealization of troposcatter transmittance properties. The basic model, together with the analytical procedure and certain basic assumptions, are reviewed here.

8.1 *Troposcatter Transmittance*

Based on certain physical considerations, an idealized multipath transmittance model is developed in which the received component waves vary at random in amplitude and phase and have transmission delays owing to path length differences which vary linearly with angular deviation from the mean path with maximum deviations $\pm\Delta$ from the mean delay. With this type of model, a Rayleigh probability distribution is obtained for the envelope of a received carrier wave in conformance with observations.

To facilitate determination of transmission performance, two basic statistical parameters are required aside from the signal-to-noise ratio at the receiver. One of these is the autocorrelation function of envelope variations with time at a given frequency. The other is the autocorrelation function with respect to frequency at a fixed time.

The first of these, the time autocorrelation function, depends on the rapidity of changes in the atmospheric structure within the common antenna volume. It has been determined by a number of observations with some theoretical support, as given in certain publications.

The second basic parameter, the autocorrelation function with respect to frequency, has been determined by observation on a particular link. These observations conform well with the autocorrelation function determined analytically herein on the premise that the maximum delay deviation $\pm\Delta$ noted above is given by the path length differences based on the beam angles between the 3-db loss points.*

With the aid of this idealized model, endowed with the above basic parameters, as determined by observation or theory, it is possible in

* This conclusion appears to be supported by the results of recent measurements on a 100-mile path.[24]

principle to determine analytically the associated idealized transmission performance with any modulation method. Though an exact solution is possible in principle, it appears intractable and is not essential for engineering purposes. An approximate solution for transmission at any digital rate is derived herein. To this end certain basic statistical parameters are determined from the above two autocorrelation functions.

8.2 *Variations in Transmittance with Time*

In Section II, distributions are given for the time rate of change in the envelope and for the first and second derivatives of the phase function. These probability distributions permit approximate evaluation of changes in the envelope, phase and frequency over a signal or pulse interval for narrow-band signal spectra.

8.3 *Variations in Transmittance with Frequency*

The corresponding probability distributions with respect to variations in transmittance with frequency are given in Section III and permit approximate determination of random attenuation and phase distortion over the band of the signal spectra owing to the selectivity of fading. From these random variations it is possible to determine the corresponding pulse distortion together with resultant intersymbol interference in carrier pulse trains and error probability in the absence of noise.

8.4 *Errors from Selective Fading*

As a next step in the determination of error probability, an approximate evaluation is made in Section IV of the probability of errors from intersymbol interference with selective Rayleigh fading in the absence of noise. In a first approximation it turns out that attenuation distortion can be neglected in comparison with phase distortion. Furthermore, the latter can be approximated by a component of quadratic phase distortion, or corresponding linear delay distortion. Intersymbol interference owing to quadratic phase distortion is determined for various carrier modulation methods, and an approximate relation is derived for the resultant error probability in the absence of noise.

8.5 *Errors from Nonselective Rayleigh Fading*

With transmission at sufficiently slow rates, errors can occur in the absence of noise, owing to changes in amplitude, phase or frequency over a pulse interval, caused by nonselective Rayleigh fading. The probability of errors on this account is determined in Section V on the approximate basis that changes over a pulse interval are proportional to the time derivatives of the amplitude, phase or frequency, depending on the modulation method. Comparison with available exact solutions for phase modulation shows that the inaccuracy resulting from this approximation is inappreciable.

8.6 *Errors from Random Noise*

In Section VI expressions are given for the probability of errors from random noise with flat Rayleigh fading, as derived in various publications for different digital carrier modulation methods. In addition, an expression is derived for error probability with rapid Rayleigh fading in conjunction with slow log-normal fading, as encountered on troposcatter links.

8.7 *Combined Error Probability*

In the final Section VII the combined error probability is determined on the approximate basis that it is the sum of the error probabilities for the three basic sources assumed above. Charts are presented from which can be determined the approximate combined error probabilities for binary phase and frequency modulation over a single path, and approximate expressions are given for the error probability with diversity transmission over independently fading paths.

8.8 *Basic Approximations*

The idealized model of troposcatter transmission assumed herein is of course an approximation, as are the idealizations regarding the performance of the carrier modulation methods. Even with exact mathematical analysis based on this model, the predicted performance would not conform entirely with that observed on actual systems.

In determining error probability from the idealized model, two basic approximations were used to obtain numerical results. One is that the maximum departures $\pm\Delta$ from the mean transmission delay can be determined from the beam angles taken between 3-db loss points. On short links with narrow-beam antennas, these are virtually equal to the free-space antenna beam angles, but for long links are greater owing to beam broadening by scatter. The second approximation is that errors from distortion owing to selective fading are caused principally by a quadratic component of phase distortion. This is the first component that gives rise to distortion in a power series expansion of a nonlinear phase characteristic as a function of the frequency from the carrier.

The same two basic approximations have been used in a companion paper (this issue, part 2) in a determination of intermodulation noise in analog transmission by FM of signals with the properties of random noise. Theoretical predictions based on free-space beam angles are in this case in reasonable agreement with measurements on two troposcatter links 185 and 194 miles in length, with narrow-beam antennas. Measurements on links 340 and 440 miles long give intermodulation noise that would correspond to beam angles and maximum delay differences $\pm\Delta$ that are greater than for free space by factors of about 1.35 and 2.15, respectively.

The above measurements also show that as the bandwidth increases, actual intermodulation noise will be progressively smaller than predicted on the premise of quadratic phase distortion. Translated to digital transmission, the error probabilities $P_e^{(1)}$ owing to selective fading as determined here on the premise of quadratic phase distortion would represent an upper bound, that should conform well with actual error probabilities when the latter do not exceed about 10^{-2} in Figs. 21 and 22.

8.9 *Comparison with Recent Related Publications*

Since the completion of the galley proof of this paper an article by Bello and Nelin[22] has appeared, dealing with errors in binary transmission owing to frequency selective fading by a different analytical procedure than used here. Numerical results are presented for error probabilities in dual and quadruple diversity transmission by binary FM with dual filter incoherent detection and binary PM with differential phase coherent detection. These results are based on an assumed Gaussian correlation function, or power spectrum, of the selectivity of fading with frequency. A comparison is made below of the above numerical results with those obtained on similar premises from relations presented here.

For a Gaussian power spectrum of correlation bandwidth B_c as used in the above paper, the corresponding value of σ^2 in (18) is $\sigma^2 = 2(\pi B_c)^{-2}$. Expression (55) applies with $b_2/b_0 = \sigma^2$ in place of $\Delta^2/3$. With this substitution and with $T = \hat{B}^{-1}$, expression (101) and Fig. 17 apply, with $\Delta \cdot \hat{B} = 0.79(B_cT)^{-1}$, where $(B_cT)^{-1}$ is the parameter appearing in Figs. 5 and 9 of the above paper for the irreducible error probabilities.

Binary FM with dual filter detection as assumed in the above paper can be considered equivalent to ideal complementary binary AM over each of two channels. When the frequency selectivity of fading is sufficient to cause errors in one or the other of these channels, the above method is essentially equivalent to dual diversity transmission by AM over two independently fading channels. On this basis, binary FM with dual diversity and dual filter noncoherent detection is approximately equivalent to binary AM with quadruple diversity. The error probabilities determined on the latter premise with $\Delta \cdot \hat{B} = 0.79(B_cT)^{-1}$ in (101), or in Fig. 17, in conjunction with (172) for $m = 4$, conform reasonably well with those given in Fig. 5 for dual diversity with $\psi = 0$ and $n = 1$. Complete agreement is not possible for the reason that the results in Fig. 5 assume a rectangular shape of undistorted pulses, whereas the present analysis is based on a more realistic pulse shape with a raised cosine spectrum, as indicated in Fig. 13.

In the case of binary PM with differential phase detection, the relations presented here with $\Delta \cdot \hat{B} = 0.79(B_cT)^{-1}$ yield error probabilities that are significantly smaller than those given in Fig. 9 of the above paper. This is to be expected, since the present relations are based on detection with an optimum threshold level, whereas those in the above paper assume zero threshold, which is not the optimum owing to the presence of a substantial bias component in the demodulator output, when pulse distortion is pronounced.[13] Moreover, the shapes of the undistorted pulses are different, as noted above.

It is evident from the above considerations that apparently unrelated and possibly misleading results can be obtained unless comparisons are made of binary modulation methods of equal bandwidths with optimum implementation of each, as was done in Fig. 17.

The above article called attention to another paper[23] by the same writers that refines Voelcker's original analysis[9] of errors in transmission over narrow-band channels owing to transmittance variations with time. Their results show that for a Gaussian power spectrum of the fading rate as assumed herein, Voelcker's analysis is exact, though this is not true for all forms of power spectra.

IX. ACKNOWLEDGMENTS

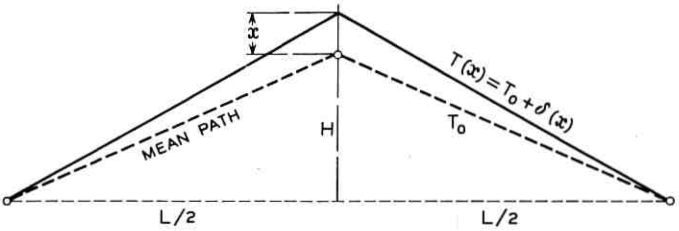
Results have been quoted herein from a number of papers dealing with troposcatter transmission properties and with error probabilities owing to random noise in conjunction with Rayleigh fading. The principal new results pertain to error probabilities at sufficiently high digital rates for selective fading to be important. In determining these error probabilities, advantage was taken of results published by S. O. Rice on the probability densities of the first and second time derivatives of

the phase of random narrow-band noise. The writer is also indebted to him for helpful suggestions resulting in certain mathematical simplifications. He also had the advantage of a discussion with I. Jacobs and D. S. Bugnolo, who pointed out certain basic limitations of the present idealized statistical model of troposcatter transmission, and he is also indebted to several other associates for helpful critical comments.

APPENDIX

Transmittance of Troposcatter Channels

Owing to the differences in path length from transmitter to receiver via the various heterogeneities in the common volume, the various components of the received wave arrive with different delays. For analytical purposes it is convenient to assume a certain mean reference path with delay T_0 and to express the transmission delay via other paths relative to the delay T_0. Actually there will be a large number of paths with the same delay T_0 as the mean path and a large number of paths for each other delay. In the present analysis the approximate model indicated below is assumed, with a single vertical scatter plane midway between transmitter and receiver.



The amplitude of the wave component arriving over a path at the distance x above the mean path is taken as $A(x,t)$ and the delay over this path as

$$T(x) = T_0 + \delta(x).$$

The wave component arriving via this path is then

$$e_x(\omega,t) = A(x,t) \cos \omega[t - T_0 - \delta(x)]. \tag{180}$$

Let L be the distance between transmitter and receiver and H the height of the mean path. In this case

$$\delta(x) = s(x)/v \tag{181}$$

where v is the velocity of propagation and $s(x)$ the path length difference given by

$$s(x) = \left[\frac{L^2}{4} + (H + x)^2\right]^{\frac{1}{2}} - \left(\frac{L^2}{4} + H^2\right)^{\frac{1}{2}}. \tag{182}$$

In actual systems $H \ll L$. Furthermore, the maximum value $\hat{x}$ of x is ordinarily much smaller than H. On these premises the following approximation applies

$$\delta(x) = (2H/Lv)x = x/c \tag{183}$$

where $c = vL/2H$.

It will further be assumed that there is an infinite number of paths, in which case the received wave becomes

$$e(\omega,t) = \int_{-\hat{x}}^{\hat{x}} A(x,t) \cos \omega(t - T_0 - x/c)\, dx \tag{184}$$

$$\begin{aligned} &= \cos \omega(t - T_0) \int_0^{\hat{x}} [A(x,t) + A(-x,t)] \cos (\omega x/c)\, dx \\ &\quad + \sin \omega(t - T_0) \int_0^{\hat{x}} [A(x,t) - A(-x,t)] \sin (\omega x/c)\, dx. \end{aligned} \tag{185}$$

It will now be assumed that

$$\int_0^{\hat{x}} [A(x,t) + A(-x,t)]\, dx = 0. \tag{186}$$

This appears to be an appropriate physical requirement, for the reason that reflections occur as a result of variations in the electrical properties of an elementary volume, relative to that of the common volume. No reflections occur with a uniform common volume. In a heterogeneous common volume, each positive reflection must be accompanied by an equal negative reflection, which is reflected in condition (186). Moreover, under this condition there is no reflection along the mean path of the transmitted beam. That is, with $x = 0$ in (185), $e(t) = 0$ provided (186) applies.

Condition (186) can be insured if the following Fourier series representations are used for $x \leqq \hat{x}$

$$A(x,t) + A(-x,t) = \sum_{m=1}^{\infty} a(m,t) \cos m\pi x/\hat{x} \tag{187}$$

and

$$A(x,t) - A(-x,t) = \sum_{m=1}^{\infty} b(m,t) \sin m\pi x/\hat{x}. \tag{188}$$

With $m = 1, 2, 3$, etc., as above, the area under each harmonic component vanishes, such that condition (186) is satisfied.

With (187) and (188) in (185), the following relation is obtained

$$e(\omega,t) = \cos \omega(t - T)U(\omega,t) + \sin \omega(t - T)V(\omega,t) \qquad (189)$$

where

$$U(\omega,t) = \sum_{m=1}^{\infty} a(m,t) \int_0^{\hat{x}} \cos m\pi x/\hat{x} \cos \omega x/c \, dx \qquad (190)$$

$$V(\omega,t) = \sum_{m=1}^{\infty} b(m,t) \int_0^{\hat{x}} \sin m\pi x/\hat{x} \sin \omega x/c \, dx \qquad (191)$$

Evaluation of the integrals yields the following expressions

$$U(\omega,t) = \sum_{m=1}^{\infty} \tfrac{1}{2} A(m,t) \left[\frac{\sin (m\pi - \omega\Delta)}{m\pi - \omega\Delta} + \frac{\sin (m\pi + \omega\Delta)}{m\pi + \omega\Delta}\right] \qquad (192)$$

$$V(\omega,t) = \sum_{m=1}^{\infty} \tfrac{1}{2} B(m,t) \left[\frac{\sin (m\pi - \omega\Delta)}{m\pi - \omega\Delta} - \frac{\sin (m\pi + \omega\Delta)}{m\pi + \omega\Delta}\right] \qquad (193)$$

where

$$\begin{aligned} A(m,t) &= \hat{x}a(m,t) \\ B(m,t) &= \hat{x}b(m,t) \\ \Delta &= \hat{x}/c. \end{aligned} \qquad (194)$$

It will be noted that Δ is the maximum departure from the mean delay T_0.

In evaluation of (192) and (193) it is convenient to introduce a new reference frequency ω_0 in place of 0, and to choose this reference frequency such that

$$\omega_0\Delta = n\pi. \qquad (195)$$

Thus

$$\omega\Delta = n\pi + u\Delta \qquad (196)$$

where $-\pi < u\Delta < \pi$, and u is the deviation in frequency from ω_0.

The functions (192) and (193) are then replaced by

$$\begin{aligned} U(u,t) = \sum_{m=1}^{\infty} \tfrac{1}{2} A(m,t) &\left\{\frac{\sin [(m - n)\pi - u\Delta]}{(m - n)\pi - u\Delta}\right. \\ &\left. + \frac{\sin [(m + n)\pi + u\Delta]}{\sin (m + n)\pi + u\Delta}\right\} \end{aligned} \qquad (197)$$

$$V(u,t) = \sum_{m=1}^{\infty} \tfrac{1}{2}B(m,t) \left\{ \frac{\sin [(m - n)\pi - u\Delta]}{(m - n)\pi - u\Delta} - \frac{\sin [(m + n)\pi + u\Delta]}{\sin (m + n)\pi + u\Delta} \right\}. \tag{198}$$

In troposcatter transmission it turns out that m is of the order of 100 to 1000. For this reason the second terms in the above series, in $(m + n)\pi$, can be neglected. With this simplification and with $m - n = j$, expressions (5) and (6) are obtained.

Expression (189) can then be written in the form

$$e(\omega,t) = r(u,t) \cos [\omega(t - T) - \varphi(u,t)] \tag{199}$$

where r and φ are given by (3) and (4).

The channel transmittance is accordingly given by (2).

REFERENCES

1. Bullington, K., Radio Propagation Fundamentals, B.S.T.J., **36**, May, 1957, p. 593.
2. Crawford, A. B., Hogg, D. C., and Kummer, W. H., Studies in Tropospheric Propagation Beyond the Horizon, B.S.T.J., **38**, September, 1959, p. 1067.
3. Ortwein, N. R., Hopkins, R. U. F., and Pohl, J. E., Properties of Tropospheric Scatter Fields, Proc. IRE, **49**, April, 1961, p. 788.
4. Rice, S. O., Distribution of the Duration of Fades in Radio Transmission, B.S.T.J., **37**, May, 1958, p. 581.
5. Clutts, C. E., Kennedy, R. N., and Trecker, J. M., Results of Bandwidth Tests on the 185-Mile Florida-Cuba Scatter Radio System, IRE Trans. on Comm. Systems, **9**, December, 1961, p. 434.
6. Beach, C. D., and Trecker, J. M., A Method of Predicting Interchannel Modulation Due to Multipath Propagation in FM and PM Tropospheric Radio Systems, B.S.T.J., **42**, January, 1963, p. 1.
7. Turin, G. L., Error Probabilities for Binary Symmetric Ideal Reception Through Nonselective Slow Fading and Noise, Proc. IRE, **46**, September, 1958, p. 1603.
8. Pierce, J. N., Theoretical Diversity Improvement in Frequency Shift Keying, Proc. IRE, **46**, May, 1958, p. 903.
9. Voelcker, H. B., Phase-Shift Keying in Fading Channels, JIEE, **107**, January, 1960, p. 31.
10. Zadeh, L. A., Frequency Analysis of Variable Networks, Proc. IRE, **38**, March, 1950, p. 291.
11. Price, R., A Note on the Envelope and Phase-Modulated Components of Narrowband Gaussian Noise, IRE Trans. on Information Theory, **1**, September, 1955, p. 9.
12. Rice, S. O., Properties of Sine Wave Plus Random Noise, B.S.T.J., **27**, January, 1948, p. 109.
13. Sunde, E. D., Pulse Transmission by AM, FM and PM in Presence of Phase Distortion, B.S.T.J., **40**, March, 1961, p. 353.
14. Sunde, E. D., Ideal Binary Pulse Transmission by AM and FM, B.S.T.J., **38**, November, 1959, p. 1357.
15. Lawton, J. G., Comparison of Binary Data Transmission Systems, Proc. of the Second National Conference on Military Electronics, 1958.
16. Reiger, S., Error Probabilities in Binary Data Transmission Systems in Presence of Random Noise, Convention Record of IRE, Part 8, 1953, p. 72.

17. Bennett, W. R., and Salz, J., Binary Data Transmission by FM over a Real Channel, B.S.T.J., **42,** September, 1963, p. 2387.
18. Brennan, D. G., Linear Diversity Combining Techniques, Proc. IRE, **47,** June, 1959, p. 1075.
19. Pierce, J. N., Theoretical Limitations on Frequency and Time Diversity for Fading Binary Transmissions, IRE Trans. on Comm. Systems, **9,** June, 1961, p. 186.
20. Harris, D. P., Techniques for Incoherent Scatter Communications, IRE Trans. on Comm. Systems, **10,** June, 1962, p. 154.
21. Barrow, B. B., Error Probabilities for Telegraph Signals Transmitted on a Fading FM Carrier, Proc. IRE, **48,** September, 1960, p. 1613.
22. Bello, P. A., and Nelin, D. B., The Effect of Frequency-Selective Fading on the Binary Error Probabilities of Incoherent and Differentially Coherent Matched Filter Receivers, IEEE Trans. on Comm. Syst. **11,** June, 1963 (issued in October), p. 170.
23. Bello, P. A., and Nelin, D. B., The Influence of Fading Spectrum on the Binary Error Probabilities of Incoherent and Differentially Coherent Matched Filter Receivers, IRE Trans. on Comm. Syst., **10,** June 1962, p. 160.
24. Patrick, W. S., and Wiggins, M. J., Experimental Studies of the Correlation Bandwidth of the Tropospheric Scatter Medium at Five Gigacycles, IEEE Trans. Aerosp. and Nav. Elect., June, 1963.

Au-n-Type GaAs Schottky Barrier and Its Varactor Application

By D. KAHNG

(Manuscript received July 19, 1963)

Evidence is presented to show that Au-n-type GaAs rectifying contacts are majority carrier rectifiers of the Schottky type. These diodes may be characterized by a Richardson constant of 20–60 amp/cm^{2}deg^2 and barrier heights of 1.03, 0.97 and 0.91 volts, corresponding to the $\langle 111 \rangle$, $\langle \overline{111} \rangle$ and $\langle 110 \rangle$ orientations of GaAs substrate.

GaAs Schottky barrier varactor diodes constructed on epitaxial films may be designed to yield a high cutoff frequency. Performance calculations in a practical case yield a "dynamic quality factor" of 50 at 6 gc under favorable conditions. A "dynamic quality factor" of about 20 at 6 gc should be obtainable with present fabrication technology.

I. INTRODUCTION

It has been demonstrated that under suitable conditions a metal-to-semiconductor rectifying contact may exhibit characteristics predictable from the simple theories advanced by Schottky[1] and Bethe.[2] An example of this type of system is the Au-n-type Si Schottky barrier which was reported earlier.[3] In the present paper evidence is presented to show that Au-n-type GaAs is also such a case.

The main features of a metal-to-semiconductor contact are that it may be designed as a majority carrier rectifier, i.e., noninjecting rectifying junction, and that the junction is accurately describable in terms of an ideal step junction. The first feature implies that the frequency response of the diode is limited only by *RC* charging time or transit time rather than by minority carrier lifetime. High cutoff frequency can be achieved through the use of an epitaxial structure. Such diodes may find application in high-speed switching, microwave detection and mixing, harmonic generation, or parametric amplification using the diode as a varactor. The first of these applications, fast switching, has been discussed elsewhere.[4]

The second feature, the ideal step junction, makes the Schottky barrier highly promising as a varactor. The step junction configuration when combined with epitaxy yields advantageous varactor performance in that its capacitive sensitivity with voltage is much higher than that of a graded junction; yet no loss in Q and breakdown voltage results from the high capacitive sensitivity. The case of a retrograded junction[5] is less favorable.

The choice of GaAs as the semiconductor part of the Schottky barrier varactor is based on two facts. First, its electron mobility is the highest among the common semiconductors available, thus allowing realization of minimum RC product while maintaining the capacitance of the unit small to facilitate diode broadband coupling to a microwave circuit. Secondly, doping close to degeneracy permits its operation at a low temperature without deterioration in performance due to carrier freeze-out.

In the following, the physical properties of the Au-n-type GaAs Schottky barrier are examined and a simple theory of a varactor design on the basis of the barrier properties is presented. The theory is used to calculate the expected performance of the varactor subject to practical considerations such as the thickness of the epitaxial layer, parasitic resistances arising from the wafer and the contact, and available pump power.

II. PHYSICAL PROPERTIES OF Au-n-TYPE GaAs SCHOTTKY BARRIER

Vacuum deposition of gold 1000 Å thick confined to a circular area of 2×10^{-3} cm² on suitably etched n-type GaAs surfaces results in diodes whose typical forward characteristics are as shown in Fig. 1. Notice that the characteristics follow the equation

$$I_f = I_s \exp\left[(q/kT)V\right] \tag{1}$$

very closely, indicating nearly ideal Schottky barrier behavior. Here I_f is the forward current, I_s the saturation current, q the electronic charge, k the Boltzmann constant, T the absolute temperature, and V the forward voltage.

Note also that I_s depends on the substrate orientation. I_s is smallest for a $\langle 111 \rangle$-directed* substrate and increases for the $\langle \overline{111} \rangle$ and $\langle 110 \rangle$ directions in that order. This suggests that the barrier height is sensitive to GaAs orientation.

* The $\langle 111 \rangle$ direction is defined to be perpendicular to the surface which gives a smoother appearance after an etch.

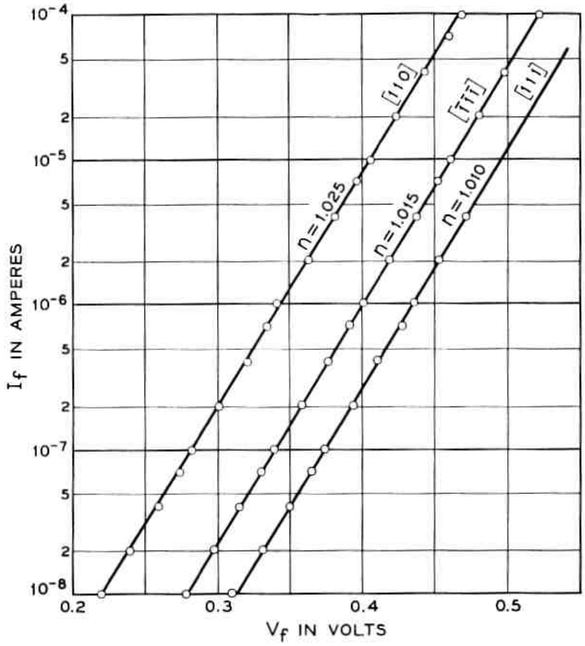


Fig. 1 — Semilog plot of typical forward characteristics for three substrate orientations; n is the slope parameter, namely,

$$\frac{d(\ln I_f)}{dV_f} = \frac{1}{n}\frac{q}{kT}.$$

For a uniformly doped substrate, the barrier capacity depends on the reverse voltage in accordance with the well-known equation

$$\frac{C}{A} = \left(\frac{\epsilon q N}{2V_T}\right)^{\frac{1}{2}} \tag{2}$$

where C is the capacity, A the junction area, ϵ the permittivity, N the donor concentration, and V_T the total voltage across the junction including the built-in voltage, V_D. This is demonstrated when $1/C^2$ vs V_R (applied voltage, reverse direction positive) plots are made as shown in Fig. 2. Such plots should be linear if (2) is closely followed, and they yield information on the diffusion voltage (built-in voltage) of the barrier as well as on the ionized donor density. Table I shows data for the three orientations mentioned earlier. Two separate evaporation runs were made for each orientation. Each set of N and V_D corresponds to a single diode. For the narrow range of donor concentrations measured, the

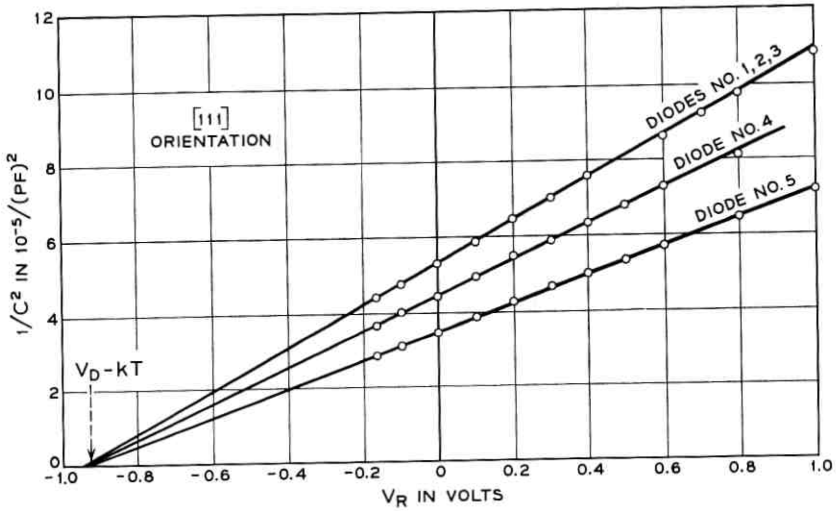


Fig. 2 — $1/C^2$ vs applied voltage for diodes constructed on ⟨111⟩-oriented GaAs surface.

equilibrium Fermi level of the substrate is about 2 kT below the conduction band edge. The energy difference of these two levels is denoted by E_{FC}. The barrier height, φ, is determined from

$$\varphi = V_D + E_{FC} \tag{3}$$

where $V_D = V_{\text{int}} + kT/q$ (V_{int} is the measured voltage intercept from Fig. 2. For details of this procedure see Ref. 3). Since I_s in (1) can be written as

$$I_s = A_R T^2 \exp - (q\varphi/kT) \tag{4}$$

one may proceed to calculate A_R, the Richardson constant, to check the validity of the model which led to (1) and (2). I_s can be determined from the forward characteristics by plotting $[\ln I_f - (qV/kT)]$ vs I_f. The resulting calculated A_R's are shown in the last column of Table I. The expected A_R is of the order of 100 amp/cm^{2}deg^2. Since the calculation of A_R is very sensitive to φ values, the results may be deemed to be in satisfactory agreement with this expectation.

It is of interest here to calculate the minority carrier contribution to the forward conduction. The hole injection efficiency, γ, can be written as[6]

$$\gamma \approx \frac{j_p}{j_s} = \frac{qp_n}{j_s} (D_p/\tau_p)^{\frac{1}{2}} \tag{5}$$

TABLE I

Orientation	N 10^{16} cm^{-3}	V_D (volts)	Φ (volts)	Ave Φ (volts)	A_R (amp/cm^2 deg^2)
111	5.8	0.95	1.03	1.03	45
	5.8	0.95	1.03		
	5.8	0.95	1.03		
	7.1	0.95	1.03		
	9.02	0.94	1.02		
$\overline{111}$	7.2	0.93	1.02	0.97	20
	7.2	0.87	0.95		
	7.2	0.88	0.96		
	8.4	0.90	0.98		
	8.4	0.88	0.96		
110	5.0	0.84	0.92	0.91	20
	5.0	0.84	0.92		
	5.0	0.83	0.91		
	5.3	0.83	0.91		
	6.2	0.80	0.88		
	6.4	0.82	0.90		
	7.6	0.89	0.97		

where j_p is the hole current density, j_s the electron saturation current density, p_n the equilibrium minority carrier density of the substrate, D_p the diffusion constant of holes and τ_p the hole lifetime. The upper limit of γ estimated, using $D_p = 20$ cm^{2}sec^{-1}, $\tau_p = 10^{-12}$ sec, and $j_s = 2 \times 10^{-11}$ amp/cm^2 for n-type GaAs of 10^{16} carrier concentration, is 5×10^{-4}. Indeed, the assumption of $\tau_p = 10^{-12}$ sec implies that the holes do not diffuse any appreciable distance. If one makes an assumption of longer hole lifetime, γ then would be even lower than the value above. The γ calculated above applies, strictly speaking, only at the origin of the V-I curve. For high forward current range, the calculation ought to be modified to include hole drift as well as diffusion.[7]

The Au-n-type GaAs Schottky barrier then can be characterized by the set of physical parameters φ and A_R as given in Table I for the various substrate orientations. It can also be treated as a noninjecting rectifier, at least for small forward currents.

III. EPITAXIAL SURFACE BARRIER VARACTOR PERFORMANCE

Assume that the surface barrier diode is constructed on an epitaxial film of thickness d grown on a substrate material of a resistivity ρ_s. For the sake of simplicity assume that for the maximum applied reverse voltage V_m, the space charge just extends through the entire thickness d of the epitaxial n region so that

$$d = [(2\epsilon/qN)V_m]^{\frac{1}{2}} = [(2\epsilon/qN)(V_0 - V_1)]^{\frac{1}{2}}. \tag{6}$$

Here V_0 is the dc bias voltage including the built-in voltage V_D, and V_1 is the pump amplitude.

The series resistance, R_s at a voltage $V < V_m$ is given by

$$R_s = \frac{\rho_e(d - s)}{A} + R_{ss} = \frac{\rho_e}{A}(2\epsilon/qN)^{\frac{1}{2}}(V_m^{\frac{1}{2}} - V^{\frac{1}{2}}) + R_{ss} \tag{7}$$

where ρ_e is the resistivity of the epitaxial film, A is the junction area, R_{ss} is the contribution from the substrate and contacts, and s is the space charge width corresponding to V given by

$$s = [(2\epsilon/qN)V]^{\frac{1}{2}}. \tag{8}$$

The assumption used in arriving at (6) does not lead to loss of generality, since the series resistance due to unswept-out epitaxial region may be incorporated into R_{ss} in (7). The performance may now be calculated in terms of the "dynamic quality factor," $\tilde{Q}$, of the diode as defined by Kurokawa and Uenohara.[8] This formulation is based on the assumption that the undesired sidebands are open-circuited. Experimental results are in closer agreement with the open-circuit assumption than with the closed-circuit assumption.[9]

The figure of merit $\tilde{Q}$ as defined in Ref. 8 may be modified to include the variation of the resistance, (7), to give

$$\tilde{Q} = \frac{1}{2\omega}\frac{D_1}{R_0} \tag{9}$$

where D_1 is the Fourier coefficient of the first harmonic of the elastance, $1/C$, ω is the operating frequency, and R_0 is the zero-order term of the Fourier expansion of R_s, [cf. (7)]. Equation (9) may be rewritten in combination with (2) and (7) as

$$\tilde{Q} = \frac{1}{2\omega}\,\frac{\dfrac{1}{A}(2\epsilon/qN)^{\frac{1}{2}}\mathfrak{F}_1(V^{\frac{1}{2}})}{\dfrac{\rho_e}{A}(2\epsilon/qN)^{\frac{1}{2}}\mathfrak{F}_0(V_m^{\frac{1}{2}} - V^{\frac{1}{2}}) + R_{ss}} \tag{10}$$

where the symbols $\mathfrak{F}_0$ and $\mathfrak{F}_1$ are used to indicate the zero- and first-order terms of the Fourier expansion of the expression inside the brackets following the symbols. Since

$$V = V_0 + V_1 \cos \omega_p t \tag{11}$$

and

$$V_m = V_0 + V_1 \tag{12}$$

where ω_p is the angular frequency of the pump, (10) can be expressed as

$$\frac{1}{\tilde{Q}} = 2\omega\epsilon\rho_e \frac{1 - [1/(1+\alpha)]^{\frac{1}{2}}\mathfrak{F}_0(\sqrt{1+\alpha\cos\omega_p t})}{[1/(1+\alpha)]^{\frac{1}{2}}\mathfrak{F}_1(\sqrt{1+\cos\omega_p t})} + \frac{2\omega A R_{ss}(\epsilon q N/2V_m)^{\frac{1}{2}}}{[1/(1+\alpha)]^{\frac{1}{2}}\mathfrak{F}_1(\sqrt{1+\alpha\cos\omega_p t})} \tag{13}$$

where

$$\alpha = V_1/V_0. \tag{14}$$

The first term of (13) is the $\tilde{Q}$ associated with the average loss in the epitaxial film region, and the second is the $\tilde{Q}$ associated with the external loss. We have

$$\frac{1}{\tilde{Q}} = \frac{1}{\tilde{Q}_i} + \frac{1}{\tilde{Q}_e}. \tag{15}$$

Fig. 3 shows the pertinent values for $\mathfrak{F}_0$ and $\mathfrak{F}_1$ of $\sqrt{1+\alpha\cos\omega_p t}$ as functions of α. Since these quantities show weak variations with α, one may take the values at $\alpha = 1$. (By definition α is never greater than unity.) Then

$$\tilde{Q}_i \cong \frac{0.58}{\omega}\frac{1}{\epsilon\rho_e} \tag{16}$$

$$\tilde{Q}_e = \frac{0.21}{\omega A R_{ss}}(2V_m/\epsilon q N)^{\frac{1}{2}} = \frac{0.21}{\omega}\frac{1}{R_{ss}C_m} = 0.21\frac{f_m}{f} \tag{17}$$

where C_m is the minimum capacity corresponding to V_m, f_m is the cutoff frequency corresponding to C_m, and f is the operating frequency.

More accurate calculation of $\tilde{Q}_i$ and $\tilde{Q}_e$ is possible whenever the pumping condition is specified. Namely, when V_0, the sum of the built-in voltage and the dc bias, and the pump amplitude are specified, the value of α is fixed. Now, corresponding to this α, more accurate numerical factors in (16) and (17) can be obtained from Fig. 3.

It is interesting to note that $\tilde{Q}$ is a function of α but not of V_0 or V_1 separately, provided the change in R_{ss} due to changes in V_0 or V_1 is taken into account. Nonuniform epitaxial film doping would not allow the use of Fig. 3 for the numerical values in (16) and (17). However, the essential form of these equations is retained and the appropriate values of the numerical factors are calculable once the doping profile is specified.

The optimum $\tilde{Q}_i$ is determined by smallest ρ_e one can practically use

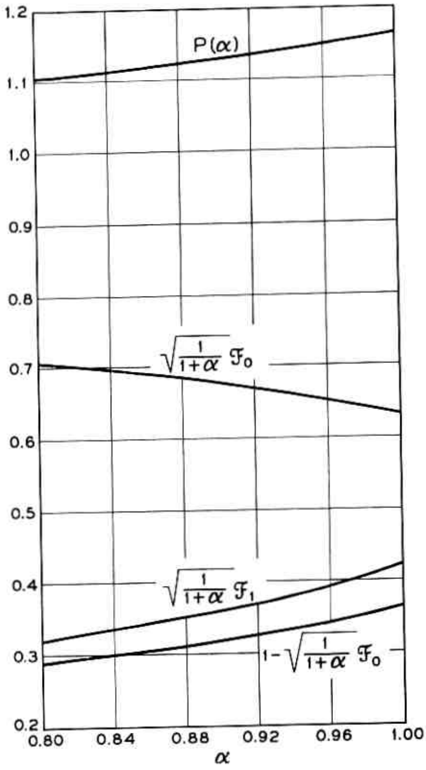


Fig. 3 — Pertinent Fourier coefficients.

$$P(\alpha) = \frac{\left(\frac{1}{1+\alpha}\right)^{\frac{1}{2}} \mathfrak{F}_1}{1 - \left(\frac{1}{1+\alpha}\right)^{\frac{1}{2}} \mathfrak{F}_0} .$$

subject to the maximum static capacity for circuit matching requirement. We now define the static capacity of the unit as

$$\bar{C} = \frac{1}{\mathfrak{F}_0(1/C)} \approx 2.8C_m \propto \frac{1}{V_m^{\frac{1}{2}}} . \tag{18}$$

Equation (18) indicates that V_m should be made as large as possible for this purpose. The extent to which V_m can be made large depends on two quantities, the breakdown voltage corresponding to a given doping level, N, and the pump amplitude. Let us examine the case where the maximum conductivity usable is limited by the breakdown voltage and the epitaxial film thickness. The relationship between the breakdown field, E_b, assumed here to be a constant for simplicity, and the maximum space charge thickness, (or the epitaxial layer thickness), d, is

$$E_b \geqq (q/\epsilon)Nd. \tag{19}$$

If d_m is the smallest thickness of epitaxial film practically attainable, then

$$1/\rho_e = \mu q N \leqq (\mu\epsilon E_b/d_m) \tag{20}$$

where μ is the electron mobility. For $E_b \cdot \epsilon \approx 5 \times 10^{-7}$ volt-fd/cm^2 and $d_m = 10^{-4}$ cm, (20) yields an optimum doping level of 3×10^{16} cm^{-3}, which corresponds to $\rho_e \approx 0.04$ ohm-cm, assuming $\mu = 5000$ cm^2/volt-sec. These figures will lead to $\tilde{Q}_i \approx 390$ at 6 gc.

Now let us calculate $\tilde{Q}_e$, using the doping level obtained above for $A = 2 \times 10^{-5}$ cm^2 (0.002-inch diameter circle). Also assume that $R_{ss} \approx 0.5$ ohm. Then (17) yields $\tilde{Q}_e \approx 57$, and (15) gives a $\tilde{Q}$ of 50.

The above calculation of dynamic quality factor was made assuming no limitations on the pump amplitudes and ideal breakdown voltage of about 25 volts. If one now assumes that only one-half of the epitaxial layer is penetrable, due to high leakage current, then $\tilde{Q}_e$ becomes 24 and $\tilde{Q} = 22$. If one is able to reduce the epitaxial thickness to 5×10^{-5} cm, the improvement is not very significant, in that $\tilde{Q}_e$ becomes 29 and $\tilde{Q} = 27$. In addition, if $R_{ss} = 0.8$ ohm this would affect $\tilde{Q}$ drastically, yielding $\tilde{Q}$ of only 17. These figures for $\tilde{Q}$ would undoubtedly deteriorate in actual cases because the package capacity is not taken into account, although the additional external circuit loss (for instance, the cavity loss) may be incorporated in R_{ss}.

Clearly, the ultimate value of $\tilde{Q}$ attainable is more heavily dependent on $\tilde{Q}_e$ than on $\tilde{Q}_i$. $\tilde{Q}_e$ is determined by R_{ss} and C_m. In a low-noise amplifier V_m may be advantageously made small, say about 10 volts or less. V_m should also be such that no appreciable reverse current flows. This means that the epitaxial layer thickness should be slightly larger than that dictated by (20), although $\tilde{Q}_e$ is somewhat sacrificed. The relaxation on V_m leads to a higher optimum epitaxial layer doping than that previously calculated. This is compatible with the necessity of having the layer thickness in excess of that dictated by V_m. Equation (20) gives optimum doping of 8×10^{16} cm^{-3} or 0.02 ohm-cm for $V_m = 10$ volts and a corresponding layer thickness of 0.4μ. If the total layer thickness is 1μ (compatible with present technology), then there is a contribution to R_{ss} from the 0.6 μ thick unswept-out layer. This could be partially compensated for by reducing the capacitance through use of a smaller junction area. The smallest junction area usable is, in turn, limited by the package capacity. Choice of an 0.001-inch diameter circular area leads to an unswept-out layer resistance of 0.2 ohm and C_m, corresponding to V_m, of 0.13 pf. The total R_{ss} then is approximately 0.8 ohm, which leads to $\tilde{Q}_e$ of 52 at 6 gc. $\tilde{Q}_i$ is increased to 780 by virtue of lowered epitaxial resistivity, yielding an over-all $\tilde{Q}$ of 50 at 6 gc. These figures are optimistic, since the influence of package capacitance is again neglected.

IV. CONCLUSIONS

The Au-n-type GaAs Schottky barrier can be characterized by the physical parameters, barrier height φ, and Richardson's constant A_R. The values of these parameters were found to be A_R = 20–60 amp/cm^2 deg^2 and φ of 1.03, 0.97 and 0.91 volts, corresponding to $\langle 111 \rangle$, $\langle \overline{111} \rangle$ and $\langle 110 \rangle$ orientation. It was shown that the barrier is essentially noninjecting for small forward currents.

The combination of the surface barrier rectifying junction with a GaAs epitaxial structure may lead to a dynamic quality factor, $\tilde{Q}$, of 20 at 6 gc with the presently available technology. In fact, one may look forward to achieving $\tilde{Q}$ of as much as 50 at 6 gc, either for low-voltage varactors ($V_m \leqq 10$ volts) or high-voltage units ($V_m \approx 25$ volts). The latter may be useful for high-power applications such as harmonic generation, as opposed to low-noise operation, for which the former is more suitable.

V. ACKNOWLEDGMENT

The author wishes to thank M. Uenohara, with whom the applicability of the device structure was first discussed, and R. M. Ryder and J. C. Irvin for stimulating discussions. He is also grateful to A. Loya and E. W. Chase for their assistance in fabrication and measurements.

REFERENCES

1. Schottky, W., Physik, **118,** 1942, pp. 539–592.
2. Bethe, H. A., Theory of the Boundary Layer of Crystal Rectifiers, MIT Radiation Lab Report, 43-12, November 23, 1942.
3. Kahng, D., Conduction Properties of the Au-n-type Si Schottky Barrier, Solid-State Electronics, **6,** 1963, p. 281.
4. Kahng, D., and D'Asaro, L. A., B.S.T.J., this issue, p. 225.
5. Chang, J. J., Forster, J. H., and Ryder, R. M., Semiconductor Junction Varactors with High Voltage Sensitivity, IEEE Trans. Electron Devices, **ED-10,** 1963, p. 281.
6. Henisch, H. K., *Rectifying Semiconductor Contacts*, Oxford University Press, 1957, p. 229.
7. Scharfetter, D. L., Anomalously High Minority Carrier Injection in Schottky Diodes, presented to 1963 IEEE Solid-State Device Research Conference at Michigan State University, June 12–14, 1963.
8. Kurokawa, K., and Uenohara, M., Minimum Noise Figure of the Variable-Capacitance Amplifier, BSTJ, **40,** 1961, p. 695.
9. Uenohara, M., private communication.

Gold-Epitaxial Silicon High-Frequency Diodes

By D. KAHNG and L. A. D'ASARO

(Manuscript received July 19, 1963)

A diode based on the properties of an evaporated gold contact on n-type epitaxial silicon has speed comparable to point contact diodes. The space charge region at zero bias can be designed to penetrate up to the impurity tail at the interface, thus reducing series resistance. An encapsulated diode was made with a 1-mil diameter gold contact on an epitaxial layer 1.5 microns thick having a surface doping of 1 × 10^{15} donors per cm^3. The zero-bias RC product of this diode is less than 1 × 10^{-12} second. Under forward bias the electron transit time through the epitaxial layer is less than 2 × 10^{-11} second. The breakdown voltage of experimental diodes is greater than 10 volts. Stress aging experiments in an inert atmosphere show no deterioration of electrical properties at temperatures up to the gold-silicon eutectic (370°C). This diode was used as a harmonic generator at 11 gc with an efficiency comparable to that of a gallium arsenide point contact diode.

I. INTRODUCTION

The metal-semiconductor rectifying contact in a variety of configurations called "point contact" has long been used for microwave rectification and amplification. This investigation shows that metal-semiconductor diodes can be designed and fabricated by large-area techniques with speeds adequate for application as fractional nanosecond switches or microwave mixers. In particular, a gold n-type silicon contact will be considered here. An estimate of the response time can be obtained from a calculation of the transit time of electrons through the space charge region and the *RC* time. The series resistance and capacitance of the diode are made small by using an epitaxial structure. Since the hole injection in these diodes at low currents is negligibly small, the response time can be independent of hole lifetime. In what follows, design of these diodes will be discussed, and the predictions of the preliminary design will be compared with experiment.

II. DIODE STRUCTURE AND FABRICATION

The structure of the diode is shown in Fig. 1. An epitaxial layer of n-type silicon is grown on an n^+ substrate. A layer of gold is evaporated in a small dot over the epitaxial layer. The metal-semiconductor contact formed in this way has an internal potential which results in a space charge region in the silicon near the gold. The doping and thickness of the silicon is chosen so that at zero bias the space charge region of thickness w occupies most of the epitaxial layer. The remaining portion, s, is a region of high doping due to diffusion of impurities from the substrate.[1,2]

Experimental diodes were fabricated as follows. Silicon wafers of resistivity 4×10^{-3} ohm-cm with faces perpendicular to the $\langle 111 \rangle$ direction were deposited with epitaxial layers of silicon by the hydrogen reduction of silicon tetrachloride.[1,3] The film thickness in a typical diode is 1.5 microns. The surface doping of the n-type layers is 2×10^{14} to 1×10^{15} donors per cm^3. The undeposited side of the wafers was provided with gold-antimony evaporated and alloyed ohmic contacts. These wafers were then subjected to cleaning consisting of oxidation and oxide removal steps. The wafers were cleaned immediately prior to gold evaporation. Gold evaporation was carried out in a vacuum of less than 2×10^{-6} mm Hg. Gold was evaporated through a molybdenum mask, confining the gold to a circular area 1 mil in diameter. After evaporation some of the diodes were etched, using the gold dots as masks. The etching removes the epitaxial region outside of the gold dots, thus preventing formation of large-area channels near the gold dots.

III. RESPONSE TIME

The low-current response time is determined by the transit time of electrons through the space charge region and the RC charging time. The transit time is given approximately by $\tau_t = w/v_s$, where w is the space charge width and v_s is the average scattering limited velocity in the space charge region. The RC charging time can be estimated from the resist-

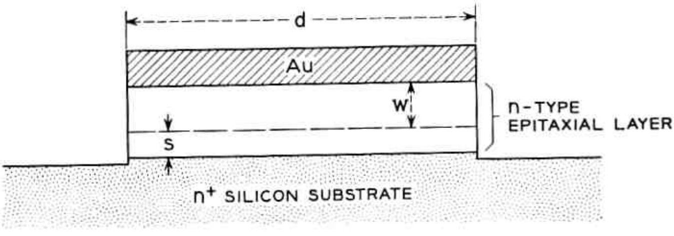


Fig. 1 — Structure of a gold-silicon epitaxial barrier diode.

ance of the unswept-out region of the epitaxial layer plus the spreading resistance in the substrate and the capacitance of the contact

$$RC = C_a \int_{\substack{\text{region}\\ s}} \rho_e \, dx + \frac{C_a \rho_s d}{2} \tag{1}$$

where C_a is the capacitance per unit area of the diode, ρ_e is the resistivity of the epitaxial layer in region s, ρ_s is the resistivity of the substrate and d is the diameter of the contact.

Calculation of the response time can be made for a case where the donor distribution in the epitaxial layer is known. In layers a few microns thick, the effect of diffusion from the substrate and the effect of the process of epitaxial growth on the distribution of impurities[1] need to be considered. The doping profile (concentration N versus distance x) may be approximately characterized by the form[1,2]

$$N = \frac{N_s}{2} \operatorname{erfc} \frac{x}{2\sqrt{Dt}} + N_0^* \, e^{-\phi x} + A(1 - e^{-\phi x}) \tag{2}$$

where the first term is due to diffusion from the substrate of doping N_s with an effective diffusion coefficient D for a time t (an approximation), the second term is the substrate contribution to the film doping through the exchange of dopant between the solid and gas phase with parameters N_0^* and ϕ, and the last term is the gas phase contribution to the film doping with an asymptotic value A for thick films. An example of an impurity distribution obtained in the fabrication of experimental gold-silicon epitaxial diodes is given in Fig. 2. The diffusion and exchange contributions to the doping are much larger than the gas phase contribution in the thicknesses used here. Within the lower doped region, one may approximate by a uniform doping for estimates of performance, since the film thickness is smaller than $1/\phi$.

The width of the space charge region at equilibrium in a uniformly doped material is given by

$$w = \left(\frac{2\epsilon V_D}{qN}\right)^{\frac{1}{2}} \tag{3}$$

where ϵ is the dielectric constant, V_D is the diffusion potential (shown in Fig. 3), q is the electron charge, and N is the donor concentration. In a typical case for these diodes the donor concentration in the region in which the exchange contribution dominates may be 1×10^{15}. The barrier potential for the gold-silicon contact (V_0 in Fig. 3) is known from measurements of the forward and reverse characteristics and the

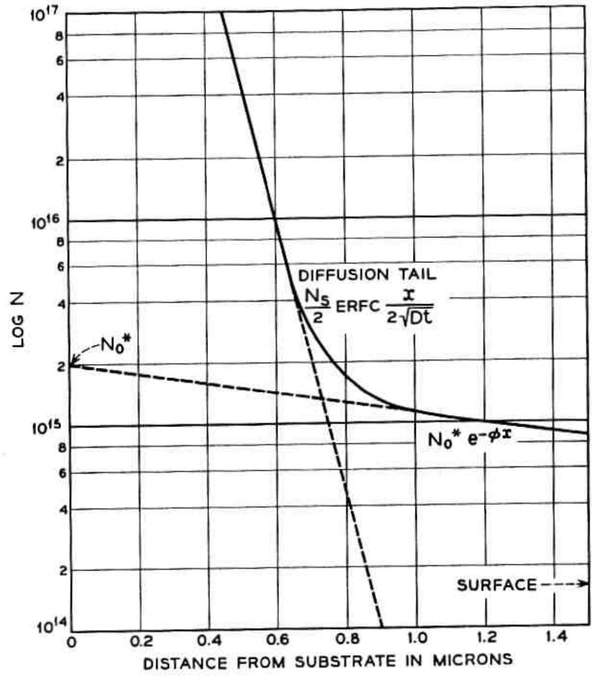


Fig. 2 — Impurity profile components for an epitaxial silicon film.

capacitance-voltage relation,[4] and is 0.79 ± 0.02 ev for silicon dopings from 0.1 to 10 ohm-cm. At $N_d = 1 \times 10^{15}$, the Fermi level is 0.25 volt below the conduction band, leading to $V_D = 0.54$ volt, and $w = 0.67$ micron. Since the edge of the space charge region falls in the diffusion tail, the series resistance of the diode is due to the doping in this tail. Integration over the doping distribution in Fig. 2 yields a zero-bias series resistance of 4.0 ohms.

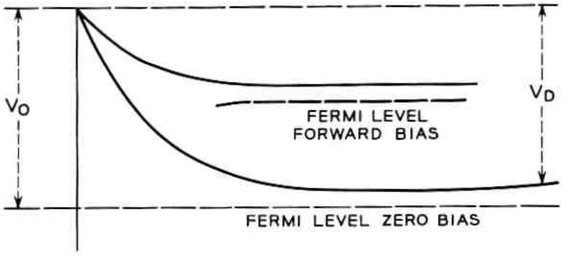


Fig. 3 — Shape of the potential barrier under zero and forward bias.

The zero-bias capacitance can be found from

$$C = (\epsilon/w)A \tag{4}$$

where A is the diode area. For a 1-mil diameter diode, the expected zero-bias capacitance is about 0.05 pf. The capacitance of the encapsulation raises the total to about 0.3 pf, making the zero-bias RC product equal to 1.2×10^{-12} second for the diodes with a series resistance of 4 ohms.

The transit time of majority carriers through the space charge region at zero bias leads to an upper limit on the response time. For the case given above under zero bias, the transit time obtained from an assumed scattering limited velocity of 5×10^{6} cm/sec is 2×10^{-11} second. Under forward bias the width of the space charge region decreases, and hence the response time may be shorter than this estimate.

IV. HOLE INJECTION CONSIDERATIONS

The hole injection ratio is defined as

$$\gamma = j_p/(j_p + j_n) \tag{5}$$

where j_p is the hole current and j_n is the electron current crossing the junction. Diffusion theory[5] allows this expression to be written as

$$\gamma = qD_p p_n/L_p j_{ns} \tag{6}$$

where D_p is the diffusion constant for holes, p_n is the equilibrium concentration of holes in n-type material, L_p is the diffusion distance for holes, and j_{ns} is the saturation value of the electron current, which can be obtained in terms of "diode" theory[6] as

$$j_{ns} = AT^2 e^{-\beta V_0}. \tag{7}$$

For $N_d = 1 \times 10^{15}$ and the experimental values of A (=40) and V_0 (=0.79 ev) from Ref. 4 one obtains $\gamma \approx 1 \times 10^{-7}$. Under low-current conditions the hole injection will not have a significant effect on the response time.

With increasing forward bias, the series resistance increases as the space charge region moves towards the gold-silicon junction. In the case of an extreme forward bias, the assumptions used earlier are not valid, and the hole current increases.[7] The series resistance may then be conductivity modulated and falls with continuously increasing current.

V. BREAKDOWN VOLTAGE

The avalanche breakdown voltage can be roughly estimated from the published ionization rate of electrons.[8] One may obtain the breakdown voltage in terms of empirically derived constants a and b as

$$V_B = bw/\ln aw \tag{8}$$

which gives $V_B = 36$ volts with $w = 0.9$ micron. Experimental diodes show breakdown voltages which occasionally approach this value. Newer data based on microplasma free junctions would predict higher values.[9]

VI. ELECTRICAL MEASUREMENTS

Experimental diodes in encapsulations typically show the following properties: breakdown voltage at 10 μamps, 25 volts; series resistance at 100 ma, 3 ohms; zero-bias capacitance, 0.35 pf. These diodes have a forward V-I characteristic given in Fig. 4. The forward characteristic can be described by the empirical relation

$$I = I_s \exp \frac{q}{nkT} (V\text{-}IR) \tag{9}$$

in which n is an empirical quantity and R is a series resistance. The "diode" theory[6] predicts the forward characteristics of the form of (9) with $n = 1$. The departure of n from unity may be attributable to currents generated at traps within the space charge region.[4] Experiments on diodes of larger diameter suggest that these traps are located around the periphery of the diode, at the gold-silicon interface. In general, n is a continuously varying quantity with the current. The series resistance may decrease in the high current density region due to increased minority carrier injection.[7] Characteristics of other diodes normalized to 1-mil diameter mesas are given for comparison in Fig. 4.

VII. RESPONSE TIME MEASUREMENTS

The response time of the experimental diodes was examined by a pulse recovery measurement. No storage time as large as the resolving time of the equipment, which is 1 nanosecond, was found.

A further measurement of an experimental diode was made by A. F. Dietrich using a method previously described for generating carrier pulses at a frequency of 11 gc.[10] In this method the RF pulses are generated directly from the harmonics of the envelope frequency that is found at the beginning or the end of the pulse transient of the diode. The power output at 11 gc was comparable to that previously obtained with a silicon snap-back diode (FD-100) or a GaAs point contact diode. These

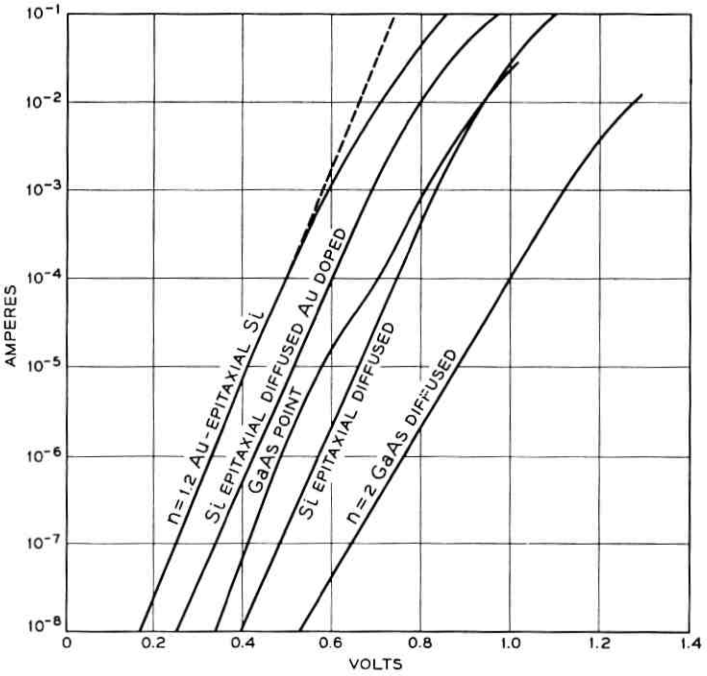


Fig. 4 — Forward bias voltage-current characteristics of a gold-epitaxial silicon diode, in comparison with other diodes. Diode diameters are 1 mil, except for the GaAs point contact. The dotted line has a slope of $n = 1.2$.

results indicate that the response time of the diode under forward bias of 60 ma is roughly 0.1 nanosecond.

VIII. STRESS AGING EXPERIMENT

A group of eight diodes was subjected to stress aging in an effort to establish the expected reliability of the gold-silicon contact. These diodes were all mounted on the same header in order to provide an equal stress condition. Heating them in an inert atmosphere for one-hour periods at increasing temperatures up to the gold-silicon eutectic temperature (370°C) produced no significant degradation in their forward or reverse characteristics. Another group of eight diodes was heated at 360°C for 64 hours. These diodes also showed no significant degradation in their *V-I* characteristics. In another experiment, diodes heated in air showed rapid degradation above 200°C. These experiments indicate that the gold-silicon contact can probably be made adequately stable for device use.

IX. CONCLUSIONS

The design described above has been found to yield experimental devices which are sufficiently fast and stable to be useful as computer diodes or as microwave mixer diodes. Another design in which the space charge region penetrates part way through the epitaxial layer may also be of interest as a varactor. One may expect that the large-area techniques used in the design and fabrication of these diodes will lead to more reproducible and stable devices than point contact diodes with similar frequency response.

REFERENCES

1. Thomas, C. O., Kahng, D., and Manz, R. C., Impurity Distribution in Epitaxial Silicon Films, J. Electrochem. Soc., **109**, No. 11, November, 1962, p. 1055.
2. Kahng, D., Thomas, C. O., and Manz, R. C., Anomalous Impurity Diffusion in Epitaxial Silicon Near the Substrate, J. Electrochem. Soc., **109**, No. 11, November, 1962, p. 1106.
3. Theurer, H. C., Epitaxial Silicon Films by Hydrogen Reduction of $SiCl_4$, J. Electrochem. Soc., **108**, No. 7, July, 1961, p. 649.
4. Kahng, D., Conduction Properties of the Au-n-type Si Schottky Barrier, Solid-State Electronics, **6**, 1963, p. 281.
5. Henisch, H. K., *Rectifying Semiconductor Contacts*, Oxford University Press, 1957, p. 229.
6. Bethe, H. A., Theory of the Boundary Layer of Crystal Rectifiers, MIT Radiation Lab Report 43/12, November 23, 1942.
7. Scharfetter, D. L., Anomalously High Minority Carrier Injection in Schottky Diodes, IRE-AIEE Solid State Device Research Conference, Michigan State University, June, 1963.
8. Maserjian, J., Determination of Avalanche Breakdown in p-n Junctions, J. Appl. Phys., **30**, No. 10, October, 1959, p. 1613.
9. Lee, C. A., Logan, R. A., Batdorf, R. L., Kleimack, J. J., and Wiegmann, W., to be published.
10. Dietrich, A. F., 8- and 11-GC Nanosecond Carrier Pulses Produced by Harmonic Generation, Proc. I.R.E., **49**, May, 1961, p. 972.

On the Discrete Spectral Densities of Markov Pulse Trains

By R. D. BARNARD

(Manuscript received August 12, 1963)

General formulae and existence criteria are derived for the discrete power spectral densities of first-order Markov pulse trains, viz., infinite pulse trains in which each pulse corresponds to one member of a finite set of specified waveforms and depends statistically on the previous pulse alone. These results are obtained through a distribution theoretic decomposition of the spectral formulation given for such pulse trains by Huggins and Zadeh.

I. INTRODUCTION

An important problem related to first-order Markov pulse trains is that of calculating the discrete and continuous power spectral densities of such processes. The spectral formulation first given by Huggins[1] and later extended by Zadeh[2] is perhaps the most appropriate and straightforward solution of this problem, the results being conveniently expressed in terms of the customary flow diagrams and recurrent event relations associated with Markov systems. As regards discrete spectra, however, their formulation lacks complete generality in two respects: (*i*) the limit notions of distribution theory, although essential for discrete components, are not incorporated; (*ii*) discrete components do not appear explicitly. In this paper we reformulate the Huggins-Zadeh result on a distribution theoretic basis, and derive both explicit relations and existence criteria for the discrete spectral densities. It is intended also that the analysis illustrate the distribution theoretic techniques required in cases involving more general spectral formulations.

II. BACKGROUND

The infinite pulse trains under discussion are treated as first-order Markov processes in that each pulse is assumed to correspond in waveshape to one member of a finite set (alphabet) of real time functions

$g_i(t)$, and to depend statistically on the previous pulse alone. More precisely, we consider random processes of the form

$$x(t) = \sum_{n=-\infty}^{\infty} d_n(t - t_n), \qquad t \,\varepsilon\, (-\infty, \infty) \tag{1}$$

$$t_n < t_{n+1} \tag{2}$$

where

$$d_n(t) \,\varepsilon\, \{g_i(t) \mid g_i \,\varepsilon\, L_1(-\infty, \infty); i = 1, 2, \cdots, M\} \tag{3}$$

$$P\{d_n = g_i \mid d_{n-1} = g_j ; d_{n-2} = g_k ; \cdots\} = P\{d_n = g_i \mid d_{n-1} = g_j\} \tag{4a}$$

$$P\{(t_{n+1} - t_n) \leqq \tau \mid d_n = g_i ; d_{n+1} = g_j ; \tau \geqq 0\} \equiv c_{ij}(\tau) \tag{4b}$$

with t_n signifying the nth occurrence time, and c_{ij} the cumulative transition distributions.* For fixed i and j, c_{ij} gives independently of n (i.e., the pulse position) the conditional probability of a direct transition from pulse g_i to pulse g_j within τ seconds after the occurrence of the former. As in related studies, the statistical and combinatorial structure of (1) is represented by the usual flow diagram of Fig. 1 in which nodes, or "states," symbolize pulses g_i, and directed links indicate possible transitions.†

The flow diagram in conjunction with signal flow graph techniques yields directly the more complex probability functions of general interest.‡ Most important to the development here are the cumulative distributions for first occurrences or recurrences, viz.

$$\begin{aligned} P\{(t_{n+m} - t_n) \leqq \tau \text{ for some } m \geqq 1 \mid d_{n+m} = g_j ; d_n = g_i ; \\ d_{n+\bar{m}} \neq g_j(\bar{m} = 1, \cdots, m-1); \tau \geqq 0\} \equiv q_{ij}(\tau). \end{aligned} \tag{5}$$

As indicated, q_{ij} denotes the conditional probability of a first occurrence (recurrence if $i = j$) of state j within τ seconds after an occurrence of state i. Although less basic than c_{ij}, functions q_{ij} are entirely sufficient for the calculation of spectral densities; consequently, in this paper the set $\{q_{ij}\}$ is regarded as initially specifying the Markov process in

* As applied here, the terms "cumulative distribution" and "distribution" pertain to probability theory and distribution theory, respectively.

† Zadeh[2] identifies the occurrence of state i with the generation of a unit impulse at node i, the impulse in turn functioning as the input to a linear filter with impulse response g_i; the corresponding responses due to all the nodes of the system are added directly to give the original pulse train.

‡ The expositions by Huggins[1] and Aaron[3] illustrate in detail the various flow diagram methods by which transition and recurrent event probabilities of higher order are calculated.

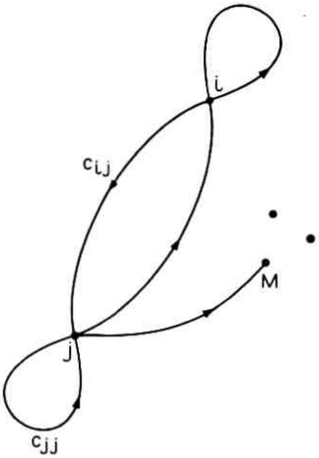


Fig. 1 — Flow diagram.

accordance with the following constraints:

(*i*) To comply with the usual probability conventions, we assume q_{ij} to be monotonically increasing, sectionally continuous, and such that

$$\begin{aligned} 0 \leqq q_{ij}(\tau) \leqq 1, &\qquad \tau \,\varepsilon\, [0, \infty) \\ q_{ij}(\tau) = 0, &\qquad \tau \,\varepsilon\, (-\infty, 0). \end{aligned} \tag{6}$$

Under these conditions both q_{ij} and the probability densities $f_{ij}(\tau) \equiv c_{ij}'(\tau)$ exist as distributions, or generalized functions.* (Earlier investigations have used f_{ij} exclusive of q_{ij}.)[1,3]

(*ii*) For pulses to occur with certainty and at distinct times ($t_n < t_{n+1}$), it is required that

$$q_{ij}(\tau) \to 1 \quad (\tau \to \infty) \tag{7}$$

$$q_{ij}(0) = q_{ij}(0^+) = 0. \tag{8}$$

Condition (7) merely asserts that every state is accessible from every other state, i.e., that the system is irreducible.

Assuming the specification of pulse trains $x(t)$ by either q_{ij} or f_{ij} and denoting the spectral density of $x(t)$ by $S_{xx}(f)$, we prove below that

* Briefly, an ordinary function $f(t)$ is an element of the space of distributions, or generalized functions, provided $[1 + t^2]^{-N}f(t) \,\varepsilon\, L_1(-\infty,\infty)$ for some $N \geqq 0$; moreover, for such functions as $f(t)$ there exist distribution derivatives of all orders and generalized Fourier transforms.[4,5,6]

$$S_{xx}(f) = \lim_{\alpha \to 0^+}{}^{(\mathrm{D})} \left\{ \sum_i \sum_j G_i(\bar{s}) G_j(s) \left[p_i \left(\frac{F_{ij}(s)}{1 - F_{jj}(s)} + \delta_{ij} \right) + p_j \left(\frac{F_{ji}(\bar{s})}{1 - F_{ii}(\bar{s})} \right) \right] \right\} \qquad (9)^*$$

where

$$G_i(s) = \int_0^\infty g_i(\tau) e^{-s\tau} d\tau = \mathfrak{L} \cdot g_i$$

$$F_{ij}(s) = \int_0^\infty e^{-s\tau}\, dq_{ij}(\tau) \equiv \int_0^\infty e^{-s\tau} f_{ij}(\tau)\, d\tau = \mathfrak{L} \cdot f_{ij}$$

$$s = \alpha + 2\pi i f, \qquad \bar{s} = \alpha - 2\pi i f, \qquad i = \sqrt{-1}, \qquad f = \text{frequency}$$

$$p_i = \left[\int_0^\infty \tau dq_{ii}(\tau) \right]^{-1} = \lim_{\substack{s \to 0 \\ \alpha > 0}} \left[\frac{s}{1 - F_{ii}(s)} \right] = -\frac{1}{F_{ii}'(0)}$$

$$\delta_{ij} = \begin{cases} 1 & (i = j) \\ 0 & (i \neq j) \end{cases}$$

and $\lim^{(\mathrm{D})} \{\cdot\}$ signifies a distribution limit (cf. Ref. 4, p. 107, and Ref. 5, p. 183). The presence of $\lim^{(\mathrm{D})}$ and the conjugated variable $\bar{s}$ in relation (9) is especially significant, both features constituting the essential modification of the spectral density expression given by Zadeh (cf. Ref. 2, Eq. 9, and Ref. 1, Eq. 10b). These two formulations prove equivalent, however, relative to continuous spectra. Specifically, if f is such that $F_{ii}(2\pi i f) \neq 1$, then the distribution limit reduces to an ordinary limit, and S_{xx} represents the same point value of the continuous spectral density as results from Zadeh's expression. On the other hand, analyzing discrete spectra† requires a proper interpretation of functions

$$\frac{1}{1 - F_{ii}(s)}$$

in the vicinity of points $s = 2\pi i f$ for which $F_{ii}(2\pi i f) = 1$; hence, the notion of distribution limits is in general necessary. Another item to be noted in (9) is the functional form of g_i . Although it is assumed that $g_i \varepsilon L_1$, one can relax this restriction in certain cases by first considering an infinite sequence of functions $g_i^{(m)} \varepsilon L_1$ such that $g_i^{(m)} \to g_i \not\varepsilon L_1$ $(m \to \infty)$, and then performing a second limit operation on the corre-

* The quantity $[F_{ij}(1 - F_{jj})^{-1} + \delta_{ij}] \equiv U_{ij}(s)$ in (9) corresponds to the Laplace transform of what Huggins terms the "expectation density" [cf. Ref. 1, Eq. (10b), p. 80].

† The term "discrete" relates to both the discrete power spectrum and the line spectral density composed of Dirac delta functions.

sponding density functions $S_{xx}^{(m)}$. An example illustrating this approach appears in Appendix A.

The following development deals primarily with the distribution theoretic formulation of (9) and its decomposition into discrete and continuous components. A detailed proof of this formulation and an analysis of the two types of components are given in Sections III and IV, respectively. Discrete spectral density expressions for the basic classes of first-order Markov pulse trains are derived in Sections 4.3, 4.4, 4.5, and 4.6 (cf. Theorems II–VI).

III. THE HUGGINS-ZADEH SPECTRAL DENSITY FORMULATION

In deriving S_{xx}, we find it convenient first to decompose $x(t)$ into M separate pulse trains which consist individually of identical pulses; i.e., we set

$$x(t) = \sum_{n=-\infty}^{\infty} d_n(t - t_n) = \sum_{i=1}^{M} x_i(t) \tag{10}$$

where

$$\begin{aligned} x_i(t) &= \sum_{m=-\infty}^{\infty} g_i(t - t_m^{(i)}) \\ t_m^{(i)} &\ \varepsilon\ \{t_n \mid d_n = g_i\} \\ t_m^{(i)} &< t_{m+1}^{(i)} \\ t_m^{(i)} &< 0 \qquad (m < 0) \\ t_m^{(i)} &\geqq 0 \qquad (m \geqq 0). \end{aligned}$$

Therefore, by standard spectral theory[7] S_{xx} can be written as

$$S_{xx}(f) = \sum_i \sum_j S_{x_i x_j}(f) \tag{11}$$

where

$$\begin{aligned} S_{x_i x_j}(f) &= \lim_{T\to\infty}{}^{(\mathrm{D})} \frac{1}{2T} E\{[\overline{\mathfrak{F}\cdot x_{iT}}][\mathfrak{F}\cdot x_{jT}]\} \\ x_{iT}(t) &= \sum_{m=M_i}^{N_i} g_i(t - t_m^{(i)}) \\ N_i &= \sup\ \{m \mid t_m^{(i)}\ \varepsilon\ [-T, T]\} \\ M_i &= \inf\ \{m \mid t_m^{(i)}\ \varepsilon\ [-T, T]\} \\ \mathfrak{F}\cdot &\equiv \int_{-\infty}^{\infty} dt\, e^{-2\pi i f t} \qquad (i = \sqrt{-1}). \end{aligned}$$

It is noted here that $S_{x_i x_j}$, the cross-spectral density of x_i and x_j, holds for both stationary and nonstationary processes.

Combined with the relation

$$\mathfrak{F} \cdot x_{iT} = G_i(2\pi i f) \sum_{M_i}^{N_i} \exp\left(-2\pi i f t_m^{(i)}\right) \tag{12}$$

(11) reduces to

$$S_{xx}(f) = \sum_i \sum_j G_i(-2\pi i f) G_j(2\pi i f) S_{ij}(f) \tag{13}$$

where

$$S_{ij}(f) = \lim_T{}^{(\mathrm{D})} \frac{1}{2T} E\left\{\sum_{M_i}^{N_i} \sum_{M_j}^{N_j} \exp\left[-2\pi i f(t_n^{(j)} - t_m^{(i)})\right]\right\}. \tag{14}$$

To transform the summation indices in (14), we let

$$t_n^{(j)} - t_m^{(i)} = \tau_{m,k}^{(ij)} > 0 \tag{15}$$

where integer $k \geqq 1$ indicates the number of occurrences of state j in the interval $(t_m^{(i)}, t_n^{(j)}]$; further, to eliminate the variation of summation indices across the ensemble, we define a weighting factor $\eta_{m,k}^{(ij)}$ such that

$$\eta_{m,k}^{(ij)} = \begin{cases} 1; & t_m^{(i)} \text{ and } t_n^{(j)} \,\varepsilon\, [-T,T], \quad t_m^{(i)} < t_n^{(j)} \\ 0; & t_m^{(i)} \text{ or } t_n^{(j)} \not\varepsilon\, [-T,T], \quad t_m^{(i)} < t_n^{(j)}. \end{cases} \tag{16}$$

These definitions along with condition (8) relating to distinct occurrence times yield

$$\begin{aligned} \sum_{M_i}^{N_i} \sum_{M_j}^{N_j} \exp\left[-2\pi i f(t_n^{(j)} - t_m^{(i)})\right] &= \sum_{k=1}^{\infty} \sum_{m=-\infty}^{\infty} \eta_{m,k}^{(ij)} \exp\left(-2\pi i f \tau_{m,k}^{(ij)}\right) \\ &+ \sum_{k=1}^{\infty} \sum_{m=-\infty}^{\infty} \eta_{m,k}^{(ji)} \exp\left(2\pi i f \tau_{m,k}^{(ji)}\right) + \delta_{ij} N_{iT} \end{aligned} \tag{17}$$

with N_{iT} equal to the number of occurrences of state i in the interval $[-T,T]$.

As random variables for the time difference between occurrences, $\tau_{m,k}^{(ij)}$ are characterized statistically by the cumulative distributions q_{ij}. In particular, (15) and (5) imply that

$$P\{\tau_{m,1}^{(ij)} \leqq \tau\} = q_{ij}(\tau). \tag{18}$$

Moreover, since the quantity

$$q_{ij}(\tau - \tau')[q_{jj}(\tau' + \Delta\tau) - q_{jj}(\tau')]$$

gives the approximate probability of two specific occurrences of state j within τ seconds after that of state i, it follows that the total probability of all such mutually exclusive events is expressed as

$$P\left\{\tau_{m,2}^{(ij)} \leqq \tau\right\} = \int_0^\tau q_{ij}(\tau - \tau')dq_{jj}(\tau') \equiv q_{ij}^{(2)}(\tau). \tag{19}$$

Generally

$$P\left\{\tau_{m,k}^{(ij)} \leqq \tau\right\} = \int_0^\tau q_{ij}^{(k-1)}(\tau - \tau')\, dq_{jj}(\tau') \equiv q_{ij}^{(k)}(\tau) \quad (k \geqq 2)$$
$$q_{ij}^{(1)}(\tau) \equiv q_{ij}(\tau). \tag{20}$$

At this point we introduce a basic device with which to simplify the summations in (17) as well as justify the interchange of various limit operations employed below. If functions q_{ij} are specified so as to vanish not only for $\tau \leqq 0$ [cf. (6)] but also in an arbitrarily small neighborhood $(-\epsilon, \epsilon)$, then there can be only a finite number of states in any finite time interval (i.e., $P\{-T \leqq t_m^{(i)} \leqq T\} = 0$ for all $|m|$ sufficiently large), and the summations in (17) remain finite. Despite this initial restriction on q_{ij}, the spectral density proves continuous in ϵ; consequently, the resultant spectral formulation is viewed as having a final, nonexplicit limit corresponding to $\epsilon \to 0$. Such a limiting procedure is entirely sufficient for physical pulse trains.

For evaluating the expectation in (14), we first define

$$P_m^{(i)}(t) = P\{t_m^{(i)} \leqq t\} \tag{21}$$

$$\mu(x) = \begin{cases} 1 & (x \geqq 0) \\ 0 & (x < 0) \end{cases} \tag{22}$$

$$\delta(x) = \frac{d\mu(x)}{dx}. \tag{23}$$

Hence, for any state i

$$\begin{aligned}
&\lim_T{}^{(\mathrm{D})} \frac{1}{2T} \sum_m \int_{-T}^{T-\tau} dP_m(t) \\
&= \lim_T{}^{(\mathrm{D})} \frac{1}{2T} \sum_m \int_{-\infty}^{\infty} [\mu(T - \tau - t) - \mu(-T - t)]\, dP_m(t) \\
&= \lim_T{}^{(\mathrm{D})} \frac{1}{2T} E\left\{\sum_m \int_{-T}^{T-\tau} \delta(t' - t_m)\, dt'\right\} = \lim_T \frac{1}{2T} E\{N_{iT}\} \\
&= [E\{t_m^{(i)} - t_{m-1}^{(i)}\}]^{-1} = \left[\int_0^\infty \tau dq_{ii}(\tau)\right]^{-1}.
\end{aligned} \tag{24}$$

On the other hand, since

$$\frac{1 - e^{-s\tau}}{s} \to \tau \qquad (s \to 0)$$

$$\left| \frac{1 - e^{-s\tau}}{s} \right| \leqq \tau \qquad (\mathrm{Re}\; s = \alpha \geqq 0)$$

the dominated convergence theorem[8] yields

$$\begin{aligned} \lim_{\substack{s\to 0\\ \alpha>0}} \frac{s}{1 - F_{ii}(s)} \equiv p_i &= \left[\lim_{\substack{s\to 0\\ \alpha>0}} \int_0^\infty \left(\frac{1 - e^{-s\tau}}{s} \right) dq_{ii}(\tau) \right]^{-1} \\ &= \left[\int_0^\infty \tau dq_{ii}(\tau) \right]^{-1} = \lim_T \frac{1}{2T} E\{N_{iT}\}. \end{aligned} \tag{25}$$

Thus, again by the convergence theorem, there results

$$\begin{aligned} &p_i \int_0^\infty e^{-2\pi i f\tau}\, dq_{ij}^{(k)}(\tau) \\ &\quad = \lim_T{}^{(\mathrm{D})} \frac{1}{2T} \int_0^{2T} \left[\sum_m \int_{-T}^{T-\tau} dP_m(t) \right] e^{-2\pi i f\tau}\, dq_{ij}^{(k)}(\tau) \\ &\quad = \lim_T{}^{(\mathrm{D})} \frac{1}{2T} \sum_m \int_0^{2T} \left[\int_{-T}^{T-\tau} dP_m(t) \right] e^{-2\pi i f\tau}\, dq_{ij}^{(k)}(\tau). \end{aligned} \tag{26}$$

Fundamental to the analysis of (26) is the following distribution theoretic identity, a detailed proof of which appears in Appendix B:

$$\lim_{N\to\infty}{}^{(\mathrm{D})} \sum_{k=1}^{N} \int_0^\infty e^{-2\pi i f\tau}\, dq_{ij}^{(k)}(\tau) = \lim_{\alpha\to 0+}{}^{(\mathrm{D})} \frac{F_{ij}(s)}{1 - F_{jj}(s)}. \tag{27}$$

From (26) and (27) it is found that

$$\begin{aligned} &\lim_N{}^{(\mathrm{D})} \sum_{k=1}^{N} p_i \int_0^\infty e^{-2\pi i f\tau}\, dq_{ii}^{(k)}(\tau) \\ &= \lim_N{}^{(\mathrm{D})} \sum_{k=1}^{N} \lim_T{}^{(\mathrm{D})} \frac{1}{2T} \sum_{m=-\infty}^{\infty} \int_0^{2T} \left[\int_{-T}^{T-\tau} dP_m(t) \right] e^{-2\pi i f\tau}\, dq_{ij}^{(k)}(\tau) \\ &= \lim_T{}^{(\mathrm{D})} \frac{1}{2T} \sum_{k=1}^{\infty} \sum_{m=-\infty}^{\infty} \int_0^{2T} \left[\int_{-T}^{T-\tau} dP_m(t) \right] e^{-2\pi i f\tau}\, dq_{ij}^{(k)}(\tau) \\ &= \lim_T{}^{(\mathrm{D})} \frac{1}{2T} E\left\{ \sum_k \sum_m \eta_{m,k}{}^{(ij)} \exp\left(-2\pi i f \tau_{m,k}{}^{(ij)}\right) \right\} \\ &= p_i \lim_{\alpha\to 0+}{}^{(\mathrm{D})} \frac{F_{ij}(s)}{1 - F_{jj}(s)}. \end{aligned} \tag{28}$$

Hence, (13), (14), (17), (25), and (28) combine to give

$$S_{xx}(f) = \lim_{\alpha \to 0^+}{}^{(\mathrm{D})} \left\{ \sum_i \sum_j G_i(\bar{s}) G_j(s) \right. \\ \left. \cdot \left[p_i \left(\frac{F_{ij}(s)}{1 - F_{jj}(s)} + \delta_{ij} \right) + p_j \left(\frac{F_{ji}(\bar{s})}{1 - F_{ii}(\bar{s})} \right) \right] \right\}. \tag{29}$$

IV. DISCRETE AND CONTINUOUS SPECTRA

The evaluation of the distribution limit in relation (9), as shown below, centers mainly on analyzing the asymptotic behavior of functions

$$\frac{F_{ij}(s)}{1 - F_{jj}(s)} \qquad (\mathrm{Re}\ s = \alpha \geqq 0) \tag{30}$$

as the variable s approaches singular points along the frequency axis, viz., points $s = 2\pi i f$ for which $F_{jj}(2\pi i f) = 1$; the results of this analysis together with certain general properties of F_{ij} serve to resolve S_{xx} into discrete and continuous components.

Considering singularities of (30) first, one notes that

$$F_{jj}(0) = \int_0^\infty dq_{jj}(\tau) = \lim_{\tau \to \infty} q_{jj}(\tau) - q_{jj}(0) = 1 \tag{31}$$

$$| F_{jj}(s) | \leqq \int_0^\infty e^{-\alpha\tau} dq_{jj}(\tau) = \alpha \int_0^\infty e^{-\alpha\tau} q_{jj}(\tau) d\tau \\ < \alpha \int_0^\infty e^{-\alpha\tau} d\tau = 1 \qquad (\mathrm{Re}\ s > 0) \tag{32}$$

$$F_{jj}(-2\pi i f) = \bar{F}_{jj}(2\pi i f). \tag{33}$$

Consequently, for all processes point $s = 0$ is singular, points in the open half plane Re $s > 0$ are nonsingular, and the existing singularities on the frequency axis occur in conjugate pairs. In establishing notation, we define

$$\left.\begin{aligned} & s_{j,n} \ \varepsilon \ \{ s \mid F_{jj}(s) = 1; \qquad \mathrm{Re}\ s = 0 \} \\ & s_{j,n} = 2\pi i f_{j,n} = \bar{s}_{j,-n} \\ & f_{j,n} < f_{j,n+1} \\ & f_{j,0} = 0 \end{aligned}\right\} \tag{34}$$

$$\left.\begin{aligned} & p_{j,n} = \left[\int_0^\infty \tau \exp\,(-s_{j,n}\,\tau)\; dq_{jj}(\tau) \right]^{-1} = -\frac{1}{F'(s_{j,n})} \\ & p_{j,n} = \bar{p}_{j,-n} \\ & p_{j,0} = p_j \end{aligned}\right\} \tag{35}$$

Then, as in (25)

$$\frac{1}{1 - F_{jj}(s)} = \left[\int_0^\infty \exp\,(-s_{j,n}\tau)\,dq_{jj}(\tau) - \int_0^\infty e^{-s\tau}\,dq_{jj}(\tau)\right]^{-1}$$

$$= \frac{1}{s - s_{j,n}}\left\{\int_0^\infty \left[\frac{1 - \exp\,[-(s - s_{j,n})\tau]}{s - s_{j,n}}\right] \cdot \exp\,(-s_{j,n}\tau)\,dq_{jj}(\tau)\right\}^{-1} \tag{36}$$

$$\sim \frac{1}{s - s_{j,n}}\,p_{j,n} \qquad (s \to s_{j,n}\,,\ \text{Re } s > 0)$$

On the basis of this asymptotic result it is found convenient to rearrange (30) as

$$\frac{F_{ij}(s)}{1 - F_{jj}(s)} = Q_{ij}(s) + R_{ij}(s) \tag{37}$$

where

$$Q_{ij}(s) = \frac{F_{ij}(s)}{2}\sum_n p_{j,n}\left[\frac{1}{\bar{s} + s_{j,n}} + \frac{1}{s - s_{j,n}}\right] \tag{38}$$

$$R_{ij}(s) = S_{ij}(s) - \sum_n p_{j,n}T_n^{(ij)}(s) \tag{39}$$

$$S_{ij}(s) = \frac{F_{ij}(s)}{1 - F_{jj}(s)} - F_{ij}(s)\sum_n \frac{p_{j,n}}{s - s_{j,n}} \tag{40}$$

$$T_n^{(ij)}(s) = \frac{F_{ij}(s)}{2}\left[\frac{1}{\bar{s} + s_{j,n}} - \frac{1}{s - s_{j,n}}\right]. \tag{41}$$

The summations in (37) are considered for the moment to be finite and to involve only those singularities present in a frequency interval $(-f_A\,, f_A)$.

4.1 *Functions Q_{ij} and R_{ij}*

It is shown next that for $f \varepsilon (-f_A\,, f_A)$ functions Q_{ij} and R_{ij} can be identified as contributing respectively to the discrete and continuous spectra:

(*i*) That functions Q_{ij} give rise to only discrete components follows immediately from the relation

$$
\begin{aligned}
&\operatorname*{Lim}_{\alpha\to 0^+}{}^{(\mathrm{D})}\, G_i(\bar{s})G_j(s)Q_{ij}(s) \\
&\quad = \tfrac{1}{2}[G_i(-2\pi if)G_j(2\pi if)F_{ij}(2\pi if)] \\
&\qquad\qquad\qquad \cdot\sum_n p_{j,n} \operatorname*{lim}_{\alpha}{}^{(\mathrm{D})} \frac{2\alpha}{[\alpha^2 + 4\pi^2(f - f_{j,n})^2]} \\
&\quad = \tfrac{1}{2}[\bar{G}_iG_jF_{ij}]_f \sum_n p_{j,n} \operatorname*{lim}_{\alpha}{}^{(\mathrm{D})} \cdot\mathfrak{F}\cdot\exp\,[-(\alpha\mid t\mid) + 2\pi if_{j,n}t] \qquad (42) \\
&\quad = \tfrac{1}{2}[\bar{G}_iG_jF_{ij}]_f \sum_n p_{j,n}\mathfrak{F}\cdot\operatorname*{lim}_{\alpha}{}^{(\mathrm{D})} \cdot\exp\,[-(\alpha\mid t\mid) + 2\pi if_{j,n}t] \\
&\quad = \tfrac{1}{2}[\bar{G}_iG_jF_{ij}]_f \sum_n p_{j,n}\mathfrak{F}\cdot\exp\,(2\pi if_{j,n}t) \\
&\quad = \tfrac{1}{2}[\bar{G}_iG_jF_{ij}]_f \sum_n p_{j,n}\delta(f - f_{j,n}).
\end{aligned}
$$

(*ii*) As regards functions R_{ij}, we first determine the behavior of functions S_{ij} in the neighborhood of points $s_{j,n}$. Substituting definition (35) into (40) yields

$$
\begin{aligned}
s_{ij}(s) &\sim F_{ij}\left[\frac{1}{1 - F_{jj}} - \frac{p_{j,n}}{s - s_{j,n}}\right] \\
&= \frac{p_{j,n}F_{ij}}{1 - F_{jj}}\left\{\int_0^\infty \left[\tau - \frac{1 - \exp\,[-(s - s_{j,n})\tau]}{s - s_{j,n}}\right]\right. \\
&\qquad\qquad \left.\cdot\exp\,(-s_{j,n}\tau)\,dq_{jj}(\tau)\right\} \qquad (43) \\
&\to \frac{p_{j,n}{}^2F_{ij}(s_{j,n})}{2}\int_0^\infty \tau^2\exp\,(-s_{j,n}\tau)\,dq_{jj}(\tau) \\
&\qquad\qquad\qquad (s \to s_{j,n},\ \mathrm{Re}\ s > 0)
\end{aligned}
$$

which implies that functions S_{ij} are both bounded and integrable in $(-f_A, f_A)$, and that points $s_{j,n}$ correspond to simple poles with residues $p_{j,n}F_{ij}(s_{j,n})$. Since functions S_{ij} are integrable, they can contribute to only the continuous portion of the power spectrum. Regarding functions $T_n{}^{(ij)}$ next, we note that

$$\begin{aligned}
&\lim_{\alpha\to 0^+}{}^{(\mathrm{D})}\, G_i(\bar{s})G_j(s)T_n{}^{(ij)}(s) \\
&\qquad = \tfrac{1}{2}[\bar{G}_iG_jF_{ij}]_f \lim_{\alpha}{}^{(\mathrm{D})} \left[\frac{4\pi i(f - f_{j,n})}{\alpha^2 + 4\pi^2(f - f_{j,n})^2}\right] \\
&\qquad = \tfrac{1}{2}[\bar{G}_iG_jF_{ij}]_f \lim_{\alpha}{}^{(\mathrm{D})}\cdot\mathfrak{F}\cdot[(\mu(-t) - \mu(t)) \\
&\qquad\qquad \cdot \exp\,(-\alpha\,|\,t\,| + 2\pi i f_{j,n}t)] \\
&\qquad = \tfrac{1}{2}[\bar{G}_iG_jF_{ij}]_f\mathfrak{F}\cdot[(\mu(-t) - \mu(t))\exp\,(2\pi i f_{j,n}t)] \\
&\qquad = \tfrac{1}{2}[\bar{G}_iG_jF_{ij}]_f\left[-\frac{1}{2\pi i(f - f_{j,n})}\right] \to \infty \qquad (f \to f_{j,n}).
\end{aligned} \tag{44}$$

Hence, in a deleted neighborhood of $s_{j,n}$, functions $T_n{}^{(ij)}$ appear to predominate all other terms of S_{xx}. For showing that functions $T_n{}^{(ij)}$ in fact sum so as to remain bounded, we set all pulses equal to zero except one, viz., g_i. If under this condition S_{xx} becomes unbounded as $f \to f_{j,n}$, then (44) and (9) give

$$S_{xx}(f) \sim p_j\{|\,G_i\,|^2[p_{j,n}F_{jj} - \bar{p}_{j,n}\bar{F}_{jj}]\}_f\left[\frac{-1}{2\pi i(f - f_{j,n})}\right], \qquad (f \to f_{j,n}) \tag{45}$$

However, since the factor in braces is continuous at $f_{j,n}$, the sign reversal of the unbounded factor indicates that S_{xx} assumes, contrary to definition, arbitrarily large negative values; therefore,

$$p_{j,n}F_{jj}(2\pi i f_{j,n}) - \bar{p}_{j,n}F_{jj}(-2\pi i f_{j,n}) = 0$$

which by (34) becomes

$$p_{j,n} = \bar{p}_{j,n} = p_{j,-n} \tag{46}$$

[The trivial case $p_{j,n} = 0$ need not be considered inasmuch as the associated terms in (37)–(41) vanish identically under this condition]. Condition (46) is sufficient as well as necessary for the ratio

$$\begin{aligned}
\frac{F_{jj}(2\pi i f) - F_{jj}(-2\pi i f)}{2\pi i(f - f_{j,n})} &= [F_{jj}{}'(s_{j,n}) - F_{jj}{}'(\bar{s}_{j,n})] + 0(f - f_{j,n}) \\
&= \left[\frac{1}{p_{j,n}} - \frac{1}{\bar{p}_{j,n}}\right] + 0(f - f_{j,n}) \\
&= 0(f - f_{j,n}) \quad (f \to f_{j,n})
\end{aligned} \tag{47}$$

to be bounded in a neighborhood of point $f_{j,n}$. Similarly, allowing two pulses to be nonzero and arbitrary yields

$$
\begin{aligned}
S_{xx}(f) \sim p_i p_{j,n}[\bar{G}_i G_j F_{ij} - \bar{G}_j G_i F_{ij}]_f \left[\frac{-1}{2\pi i(f - f_{j,n})}\right] \\
+ p_j\,, p_{i,m}[\bar{G}_j G_i F_{ji} - \bar{G}_i G_j \bar{F}_{ji}]_f \left[\frac{-1}{2\pi i(f - f_{i,m})}\right] \qquad (f \to f_{j,n})
\end{aligned}
\tag{48}
$$

where the second term is present provided $f_{j,n} = f_{i,m}$. It is evident that with the second term absent and both g_i and g_j arbitrary the first term cannot be made to vanish identically at $f_{j,n}$; thus

$$s_{j,n} = s_{i,m} = s_{i,n} \equiv s_n \tag{49}$$

and

$$
\begin{aligned}
\{\bar{G}_i G_j[p_i p_{j,n} F_{ij} - p_j p_{i,n} \bar{F}_{ji}] \\
+ \bar{G}_j G_i[p_j p_{i,n} F_{ji} - p_i p_{j,n} \bar{F}_{ij}]\}_{f_n} = 0.
\end{aligned}
\tag{50}
$$

Again because of arbitrary g_i and g_j there results

$$p_i p_{j,n} F_{ij}(2\pi i f_n) = p_j p_{i,n} F_{ji}(-2\pi i f_n). \tag{51}$$

As in (47), this is a necessary and sufficient condition that (48) be bounded in a neighborhood of point $f = f_{j,n} = f_n$; thus, for $f \varepsilon (-f_A, f_A)$ functions $T_n^{(ij)}$, S_{ij}, and sums R_{ij} contribute to only the continuous spectrum. It is important to note that although the use of R_{ij} is necessary for an appropriate decomposition of S_{xx}, the complete continuous spectrum can be obtained directly from relation (9) with $f \neq f_n$ [cf. (9) et seq.]. Nevertheless, from a computational standpoint functions R_{ij} might be more suitable.

4.2 *General Formulation for Discrete Spectra*

At this point we consider in detail both formulae and existence criteria for the discrete spectral density. With respect to the complete spectral density, the substitution of definition (37) into (9) gives at once the decomposition

$$
\begin{aligned}
S_{xx}(f) = \lim_{\alpha \to 0^+}{}^{(\mathrm{D})} \left\{\sum_i \sum_j G_i(\bar{s}) G_j(s)[p_i Q_{ij}(s) + p_j Q_{ji}(\bar{s})]\right\} \\
+ \lim_{\alpha \to 0^+}{}^{(\mathrm{D})} \left\{\sum_i p_i \mid G_i(s) \mid^2 + \sum_i \sum_j [p_i R_{ij}(s) + p_j R_{ji}(\bar{s})]\right\}
\end{aligned}
\tag{52}
$$

where according to the properties of functions Q_{ij} and R_{ij} [cf., (42), (51) et seq.] the first term in braces consists of discrete components only, and the second is bounded for $f \varepsilon (-f_A, f_A)$. Consequently, on letting

$S_{xx}^{(d)}(f)$ denote the discrete spectral density in the interval $(-f_A, f_A)$, we obtain

$$S_{xx}^{(d)}(f) = \lim_{\alpha \to 0^+}{}^{(\mathrm{D})} \left\{ \sum_i \sum_j G_i(\bar{s}) G_j(s) [p_i Q_{ij}(s) + p_j Q_{ji}(\bar{s}) \right\} \quad (53)$$

which by (42), (46), (49), and (51) becomes

$$\begin{aligned} S_{xx}^{(d)}(f) &= \tfrac{1}{2} \sum_i \sum_j \bar{G}_i G_j \Big[p_i F_{ij} \sum_n p_{j,n} \delta(f - f_n) \\ &\qquad + p_j \bar{F}_{ji} \sum_n p_{i,n} \delta(f + f_n) \Big] \\ &= \tfrac{1}{2} \sum_i \sum_j \bar{G}_i G_j \Big[p_i F_{ij} \sum_n p_{j,n} \delta(f - f_n) \\ &\qquad + p_j \bar{F}_{ji} \sum_n p_{i,-n} \delta(f + f_{-n}) \Big] \\ &= \sum_i \sum_j \bar{G}_i G_j F_{ij} \sum_n p_i p_{j,n} \delta(f - f_n) \\ &= \sum_n \Big[\sum_i \sum_j p_i p_{j,n} G_i(-2\pi i f) \\ &\qquad \cdot G_j(2\pi i f) F_{ij}(2\pi i f) \Big] \delta(f - f_n). \end{aligned} \quad (54)$$

Since the interval $(-f_A, f_A)$ is arbitrary, the sum over n in (54) can be extended as a distribution limit to include all the singular points along the frequency axis; hence, this expression represents the general formula for the discrete spectral density. In the sections immediately following, formula (54) is applied to the two fundamental classes of first-order Markov pulse trains: entirely random and stochastically uniform pulse trains.

4.3 *Discrete Spectra of Entirely Random Pulse Trains*

We define the processes under discussion to be entirely random if for at least one state i

$$\begin{gathered} q_{ii}(\tau) = \hat{q}_{ii}(\tau) + \sum_k \alpha_k^{(ii)} \mu(\tau - \tau_k^{(ii)}) \\ f_{ii}(\tau) = q_{ii}'(\tau) = \hat{q}_{ii}'(\tau) + \sum_k \alpha_k^{(ii)} \delta(\tau - \tau_k^{(ii)}) \\ 0 \leqq \alpha_k^{(ii)} \leqq 1 \\ \hat{q}_{ii}(\infty) + \sum_k \alpha_k^{(ii)} = 1 \end{gathered} \quad (55)$$

where $\hat{q}_{ii}$ is either continuous and strictly increasing in some interval (τ_A, τ_B), i.e.

$$\hat{q}_{ii}'(\tau) > 0 \qquad \tau \in (\tau_A, \tau_B) \tag{56}$$

or $\hat{q}_{ii}$ vanishes identically and the set of parameters $\tau_k^{(ii)}$ consists of two or more incommensurate elements. Processes of this class are characterized more completely by the following theorem:

Theorem I: A pulse train is entirely random if and only if for any state i

$$\begin{aligned} F_{ii}(2\pi if) &\neq 1 \qquad (f \neq 0) \\ F_{ii}(0) &= 1. \end{aligned} \tag{57}$$

For such processes all first recurrence distributions q_{ii} have the same form.

Proof: The second condition of (57) is merely a restatement of the general result given by (31). To establish the sufficiency of the first condition, we consider the only possible form for q_{ii} not representable by (55), viz.

$$\begin{aligned} q_{ii}(\tau) &= \sum_{k-1}^{\infty} \alpha_k^{(ii)} \mu(\tau - kT_i) \\ f_{ii}(\tau) &= \sum_k \alpha_k^{(ii)} \delta(\tau - kT_i). \end{aligned} \tag{58}$$

This yields

$$F_{ii}(2\pi if) = \sum_k \alpha_k^{(ii)} e(-2\pi ifkT_i) \tag{59}$$

whence

$$F_{ii}\left(2\pi i \frac{n}{T_i}\right) = 1 \qquad (n = 0, \pm 1, \cdots). \tag{60}$$

Therefore, any q_{ii} satisfying (57) must be representable by (55), and the process entirely random. To establish necessity, we consider (55) to be satisfied for at least one state i. Under condition (56)

$$\begin{aligned} \left| \int_{\tau_A}^{\tau_B} e^{-2\pi if\tau} \, d\hat{q}_{ii}(\tau) \right| &= \left| \int_{\tau_A}^{\tau_B} e^{-2\pi if\tau} \hat{q}_{ii}' \, d\tau \right| \\ &< \int_{\tau_A}^{\tau_B} \hat{q}_{ii}' \, d\tau = \int_{\tau_A}^{\tau_B} d\hat{q}_{ii}(\tau) \qquad (f \neq 0) \end{aligned}$$

whence

$$| F_{ii}(2\pi if) | < \int_0^{\infty} d\hat{q}_{ii}(\tau) + \sum_k \alpha_k^{(ii)} = \int_0^{\infty} dq_{ii}(\tau) = 1 \qquad (f \neq 0).$$

On the other hand, with $\hat{q}_{ii} \equiv 0$ and $\tau_k^{(ii)}$ incommensurate

$$| F_{ii}(2\pi if) | = | \sum_k \alpha_k^{(ii)} \exp (-2\pi if\tau_k^{(ii)}) | < 1 \qquad (f \neq 0).$$

Thus, (57) is necessary for state i. Finally, since $F_{ii}(2\pi if_{i,n}) = 1$ and $f_{i,n} = f_n$ for all i [cf., (31), (34), and (49)], the realization of (57) for any q_{ii} necessarily implies the same realization and consequently the same form for all q_{ii}.

Theorem I, although essential to the treatment of discrete spectra, is not the only test for identifying entirely random processes; a somewhat more direct test is afforded by the cumulative distributions c_{ij}. In particular, functions q_{ij} have form (55) provided at least one of the functions c_{ij} does also. This fact follows from a basic property of irreducible processes, viz., the property that each density $f_{ij} \equiv q_{ij}'(\tau)$ equals a specific combination of positive sums and convolutions of all the densities $c_{ij}'(\tau)$.[1,3]

As regards singular points s_n and discrete spectra, it is clear from Theorem I and (34) that the point $s = s_0 = 0$ constitutes the only singularity of entirely random processes; therefore, the formulation given by (54) becomes

$$\begin{aligned} S_{xx}^{(d)}(f) &= \Big[\sum_i \sum_j p_i p_{j,0} G_i(0) G_j(0) F_{ij}(0)\Big]\delta(f) \qquad (61)\\ &= \Big[\sum_i \sum_j p_i p_j G_i(0) G_j(0)\Big]\delta(f). \end{aligned}$$

This expression leads immediately to the following result:
Theorem II: The discrete spectral density of entirely random pulse trains is given by

$$S_{xx}^{(d)}(f) = \left\{\int_{-\infty}^{\infty}\left[\sum_i p_i g_i(t)\right] dt\right\}^2 \delta(f) \qquad (62)$$

which vanishes if and only if

$$\int_{-\infty}^{\infty}\left[\sum_i p_i g_i(t)\right] dt = 0. \qquad (63)$$

Comparing (62) with (54), we note that Theorem II applies to the $\delta(f)$, or dc, component of all the processes treated in this paper.

4.4 *Discrete Spectra of Stochastically Uniform Pulse Trains*

Processes not classified as entirely random are defined here to be stochastically uniform. It is evident that the only first recurrence distributions representing the uniform process, i.e., satisfying neither definition (55) nor the criteria of Theorem I, must be of the form

$$q_{ii}(\tau) = \sum_{k=1}^{\infty} \alpha_k^{(ii)} \mu(\tau - kT_i)$$
$$0 \leqq \alpha_k^{(ii)} \leqq 1 \tag{64}$$
$$\sum_k \alpha_k^{(ii)} = 1$$

where parameters T_i are assumed to have the largest values possible. Under this specification

$$F_{ii}(2\pi if) = \sum_{k=1}^{\infty} \alpha_k^{(ii)} \exp(-2\pi ifkT_i) \tag{65}$$

Hence, on letting i_0 denote the state for which

$$T_i \leqq T_{i_0} \qquad (i = 1, \cdots, M) \tag{66}$$

we find that all the singular values f_n satisfying

$$F_{i_0 i_0}(2\pi if_n) = 1 \tag{67}$$

are given by

$$f_n = \frac{n}{T_{i_0}} \qquad (n = 0, \pm 1, \cdots). \tag{68}$$

Furthermore, since

$$F_{ii}(2\pi if_n) = 1 \tag{69}$$

for all states [cf. (34) and (49)], then

$$T_{i0} = T_i \equiv T \qquad (i = 1, \cdots, M) \tag{70}$$

which in turn implies that all F_{ii} are periodic over an interval of length T^{-1}, and all functions q_{ii} have the basic form

$$q_{ii}(\tau) = \sum_{k=1}^{\infty} \alpha_k^{(ii)} \mu(\tau - kT). \tag{71}$$

Considering also relations (65), (68), and (35) it is seen that

$$p_{i,n} = \left[\sum_k \tau \alpha_k^{(ii)}\right]^{-1} = p_{i,0} = p_i \,. \tag{72}$$

Finally, results (68), (70), and (72) combine with (54) to give the following theorem:

Theorem III: The discrete spectral density of stochastically uniform pulse trains is given by

$$S_{xx}^{(d)}(f)$$
$$= \left[\sum_i \sum_j p_i p_j G_i(-2\pi i f) G_j(2\pi i f) F_{ij}(2\pi i f) \right] \sum_{n=-\infty}^{\infty} \delta(f - n/T) \qquad (73)$$
$$T = n/f_n$$
$$F_{ii}(2\pi i f_n) = 1$$

which vanishes if and only if

$$\left[\sum_i \sum_j p_i p_j \bar{G}_i G_j F_{ij} \right]_{n/T} = 0 \qquad (n = 0, \pm 1, \cdots) \qquad (74)$$

or if

$$\left[\sum_i \sum_j p_i p_j \bar{G}_i G_j F_{ij} \right]_f = 0 \qquad (-\infty < f < \infty). \qquad (75)$$

At this point we consider a special but very important subclass of uniform pulse trains, namely, that of uniformly positioned pulses.

4.5. *Discrete Spectra of Uniformly Positioned Pulse Trains*

Pulse trains are defined to be uniformly positioned over a reference interval of length T_0 if the time intervals between successive pulses can assume only the discrete values $kT_0 (k = 1, 2, \cdots)$, i.e., if function q_{ij} take the form

$$q_{ij}(\tau) = \sum_{k=1}^{\infty} \alpha_k^{(ij)} \mu(\tau - kT_0) \qquad (i, j = 1, \cdots, M)$$
$$0 \leqq \alpha_k^{(ij)} \leqq 1 \qquad (76)$$
$$\sum_k \alpha_k^{(ij)} = 1$$

where T_0 constitutes the maximum value for which this representation is valid. With q_{ij} so specified there results

$$F_{ij}(2\pi i f) = \sum_k \alpha_k^{(ij)} \exp(-2\pi i f k T_0) \qquad (77)$$

Consequently, for a particular state i the condition

$$\alpha_{Kk'}^{(ii)} \geqq 0 \qquad (k' = 1, 2, \cdots)$$

$$\alpha_k^{(ii)} = 0 \qquad (k \neq Kk') \tag{78}$$

holds for some maximum $K \geqq 1$, the corresponding function F_{ii} is periodic over an interval of length $(KT_0)^{-1}$, and the singular values f_n satisfying (69) are given by

$$f_n = \frac{n}{KT_0} = \frac{n}{T}. \tag{79}$$

In addition, as values f_n are independent of i, condition (78) must for all states hold for the same value of K, the specific value in any particular case being determined either from one set of coefficients $\alpha_k^{(ii)}$, from (79), or from the recurrence pattern associated with one node of the flow graph. For all $K \geqq 1$, relations (77) and (79) yield the general conditions

$$\left.\begin{aligned} F_{ij}\left(2\pi i \frac{n}{T_0}\right) &= 1 \\ F_{ii}\left(2\pi_i \frac{n}{KT_0}\right) &= F_{ii}(2\pi i f_n) = 1 \\ F_{ij}\left(2\pi i \frac{n+K}{KT_0}\right) &= F_{ij}\left(2\pi i \frac{n}{KT_0}\right) \end{aligned}\right\} \begin{matrix} (K \geqq 1; \quad i,j = 1, \cdots, M; \\ n = 0, \pm 1, \cdots). \end{matrix} \tag{80}$$

Combining these conditions with (79) and Theorem III, we obtain

$$\begin{aligned} S_{xx}^{(d)}(f) &= \left[\sum_i \sum_j p_i p_j \bar{G}_i G_j F_{ij}\right]_f \sum_{n=-\infty}^{\infty} \delta\left(f - \frac{n}{KT_0}\right) \\ &= \left[\sum_i \sum_j p_i p_j \bar{G}_i G_j F_{ij}\right]_f \sum_{k=0}^{K-1} \sum_{n=-\infty}^{\infty} \delta\left(f - \frac{n}{T_0} - \frac{k}{KT_0}\right) \\ &= \left[\sum_i \sum_j p_i p_j \bar{G}_i G_j\right]_f \sum_{n=-\infty}^{\infty} \delta\left(f - \frac{n}{T_0}\right) \\ &\quad + \sum_{k=1}^{K-1} \left\{\left[\sum_i \sum_j p_i p_j G_i(-2\pi i f) G_j(2\pi i f)\right.\right. \\ &\quad \left.\left. \cdot F_{ij}\left(2\pi i \frac{k}{KT_0}\right)\right] \sum_{n=-\infty}^{\infty} \delta\left(f - \frac{n}{T_0} - \frac{k}{KT_0}\right)\right\}. \end{aligned} \tag{81}$$

The following theorem is based on this last expression:

Theorem IV: The discrete spectral density of pulse trains uniformly positioned over a reference interval of length T_0 is given by

$$S_{xx}^{(d)}(f) = \left| \sum_i p_i G_i(2\pi i f) \right|^2 \sum_{n=-\infty}^{\infty} \delta\left(f - \frac{n}{T_0}\right)$$
$$+ \sum_{k=1}^{K-1} \left\{ \left[\sum_i \sum_j p_i p_j G_i(-2\pi i f) G_j(2\pi i f) F_{ij}\left(2\pi i \frac{k}{KT_0}\right) \right] \cdot \sum_{n=-\infty}^{\infty} \delta\left(f - \frac{n}{T_0} - \frac{k}{KT_0}\right) \right\}$$

$$\left.\begin{aligned} K &= \frac{n}{T_0 f_n} \\ F_{ii}(2\pi i f_n) &= 1 \end{aligned}\right\} (K \geqq 1;\ i, j = 1, \cdots, M;\ n = 0, \pm 1, \cdots) \tag{82}$$

which vanishes if

$$\sum_i p_i g_i(t) = 0 \tag{83}$$

$$\sum_i \sum_j p_i p_j F_{ij}\left(2\pi i \frac{k}{KT_0}\right) \int_{-\infty}^{\infty} g_i(\tau) g_j(\tau + t)\, d\tau = 0 \qquad (k = 1, \cdots, K - 1). \tag{84}$$

A special case of Theorem IV is noted as follows:

Theorem V: The discrete spectral density of uniformly positioned pulse trains corresponding to $K = 1$ is given by

$$\begin{aligned} S_{xx}^{(d)}(f) &= \left| \sum_i p_i G_i(2\pi i f) \right|^2 \sum_{n=-\infty}^{\infty} \delta\left(f - \frac{n}{T_0}\right) \\ f_n &= \frac{n}{T_0} \end{aligned} \tag{85}$$

which vanishes if

$$\sum_i p_i g_i(t) = 0. \tag{86}$$

Titsworth and Welch[9] have proved Theorem V for special pulse trains in which pulses are nonoverlapping and transitions occur every T_0 seconds. This theorem is also implicit in the classic work of Bennett on synchronous pulse trains [cf. Ref. 10, Eq. (35), p. 1509].

4.6. *Aaron's Discrete Spectral Formulation for Special Classes of Pulse Trains*

The analysis in Sections 4.3 and 4.5 yields the following theorem, a result first obtained by M. R. Aaron:[3]

Theorem VI: The discrete spectral density of entirely random pulse trains *and uniformly positioned pulse trains for which* $K = 1$ [*cf.* (78) *et seq.*] *is given by*

$$S_{xx}^{(d)}(f) = \sum_n \left\{ \operatorname*{Res}_{s_n} \left[\sum_i G_i(s) U_{ji}(s) \right] \right\}^2 \delta(f - f_n) \qquad (87)^*$$

where

$$U_{ji} = F_{ji}[1 - F_{ii}]^{-1} + \delta_{ji} \qquad (88)$$

and $\operatorname*{Res}_{s_n} [\cdot]$ *denotes the residue of the quantity in brackets at* $s = s_n = 2\pi i f_n$.

Proof: From relations (36), (72) and Theorem I we find that

$$\operatorname*{Res}_{s_n} \left[\frac{G_i(s) F_{ji}(s)}{1 - F_{ii}(s)} \right] = p_i G_i(2\pi i f_n) F_{ji}(2\pi i f_n) \qquad (89)$$

for either the entirely random or $K = 1$ case. On the other hand

$$F_{ji}(2\pi i f_n) = 1 \qquad (i, j = 1, \cdots, M) \qquad (90)$$

in both cases [cf., (79) and (80)]; thus,

$$\operatorname*{Res}_{s_n} \left[\sum_i G_i U_{ji} \right] = \sum_i p_i G_i(2\pi i f_n). \qquad (91)$$

Inserting this expression into either (61) or (85) gives formula (87).

V. SUMMARY

Theorems I through VI, which constitute the principal results of the preceding sections, give explicitly the discrete spectra of first-order Markov pulse trains. As presented, these theorems provide fundamental existence criteria for not only the analysis but also the synthesis of such processes. It is important to emphasize again that the distribution theoretic techniques employed in extracting discrete components from the Huggins-Zadeh formulation are applicable also to more general spectral formulations.

VI. ACKNOWLEDGMENTS

The author wishes to express his appreciation to M. R. Aaron, B. K. Kinariwala, and M. Karnaugh for several fruitful discussions and comments during the course of this work.

* Huggins has shown that the sum $\sum_i G_i U_{ji}$ represents the Laplace transform of the average signal following the occurrence of state j [cf., Ref. 1, Eq. (23a), p. 82].

APPENDIX A

Entirely Random Square Waves

For illustrating the techniques that often apply to cases in which $g_i \epsilon L_1$, we consider a random square wave process of the form

$$x(t) = a \sum_n (-1)^{n-1} [\mu(t - t_{n-1}) - \mu(t - t_n)] \tag{92}$$

$$x'(t) \equiv y(t) = 2a \sum_n (-1)^n \delta(t - t_n) \tag{93}$$

where y represents a two-state pulse train with pulses related by

$$\begin{gathered} g_1 = -g_2 = 2a\delta(t) \epsilon L_1 \\ a = \text{constant} > 0 \end{gathered} \tag{94}$$

and an entirely random statistical structure (cf. Section 4.3) specified by c_{12}, c_{21}, and

$$c_{11} = c_{22} = 0. \tag{95}$$

(Note that states 1 and 2 can be identified with the $+a$ and $-a$ portions of the square wave x.) Thus, in accordance with definitions (4b) and (5)

$$\begin{gathered} q_{12} = c_{12}, \qquad q_{21} = c_{21} \\ q_{11} = \int_0^\infty c_{12}(\tau - \tau') dc_{21}(\tau') = q_{22} \end{gathered} \tag{96}$$

whence

$$\begin{gathered} F_{11} = F_{22} = F_{12}F_{21} \\ p_1 = \int_0^\infty \tau \, dq_{11}(\tau) = -\frac{1}{F_{11}'(0)} = p_2 \equiv p. \end{gathered} \tag{97}$$

We next construct a set of "smooth" approximations to x; i.e., we smooth out the corners and discontinuities of each of the pulse trains x into a sequence $\{x_m(t)\}$ of continuous waveforms such that

$$\begin{gathered} S_{xx}(f) = \lim_{m \to \infty}{}^{(D)} S_{x_m x_m}(f) \qquad (m = 1, 2, \cdots) \\ x_m'(t) \equiv y_m(t) = \sum_n (-1)^n g^{(m)}(t - t_n) \end{gathered} \tag{98}$$

where

$$\begin{aligned} g^{(m)} &\;\varepsilon\; L_1 \\ \lim_m{}^{(\mathrm{D})}\, g^{(m)} &= 2a\delta(t) \\ g^{(m)} &= g_1^{(m)} = -g_2^{(m)}. \end{aligned} \tag{99}$$

Since pulse trains y_m and y have the same transition properties and therefore the same statistical specification c_{ij}, the former process is classified as entirely random; it then follows from the condition

$$\sum_i p_i g_i^{(m)} = p(g_1^{(m)} + g_2^{(m)}) = 0$$

and from Theorem II [cf. (62)] relating to entirely random pulse trains that $S_{y_m y_m}$ has no discrete components. Consequently, relations (9), (97), (98), and (99) yield

$$\begin{aligned} 4\pi^2 f^2 S_{xx}(f) &= \lim_m{}^{(\mathrm{D})} [4\pi^2 f^2 S_{x_m x_m}(f)] = \lim_m{}^{(\mathrm{D})} S_{y_m y_m}(f) \\ &= \lim_m{}^{(\mathrm{D})} \left\{ 2p \mid G^{(m)}(2\pi i f) \mid^2 \right. \\ &\qquad \left. \cdot \mathrm{Re}\left[\frac{(1 - F_{12})(1 - F_{21})}{1 - F_{12}F_{21}} \right]_f \right\} \\ &= 8pa^2\, \mathrm{Re}\left[\frac{(1 - F_{12})(1 - F_{21})}{1 - F_{12}F_{21}} \right]_f . \end{aligned} \tag{100}$$

The most general function S_{xx} satisfying this last expression is given by

$$S_{xx}(f) = \frac{2pa^2}{\pi^2 f^2}\, \mathrm{Re}\left[\frac{(1 - F_{12})(1 - F_{21})}{1 - F_{12}F_{21}} \right]_f + K_1\delta(t) = K_2\delta'(f) \tag{101}$$

where the first term on the right represents a continuous component, and constants K_1 and K_2 are to be determined. As spectral densities must be even functions, $K_2 = 0$. Regarding the discrete term, constant K_1 is the square of the dc, or average, component of x; hence, with

$$\begin{aligned} \mathrm{ave}\,[x(t)] &= \frac{a \int_0^\infty \tau\, dc_{12}(\tau) - a \int_0^\infty \tau\, dc_{21}(\tau)}{\int_0^\infty \tau\, dq_{11}(\tau)} \\ &= ap \left\{ \int_0^\infty \tau\, d[q_{12}(\tau) - q_{21}(\tau)] \right\} \\ &= ap\, [F_{21}'(0) - F_{12}'(0)] \end{aligned} \tag{102}$$

(101) becomes

$$S_{xx}(f) = \frac{2pa^2}{\pi^2 f^2} \operatorname{Re}\left[\frac{(1 - F_{12})(1 - F_{21})}{1 - F_{12}F_{21}}\right]_f + a^2p^2[F_{21}'(0) - F_{12}'(0)]^2\delta(f). \tag{103}$$

It is important to note here that the discrete component in (103) arises from the pulse structure of x and not from the singularities of $[1 - F_{ii}]^{-1}$. A more extensive treatment of this particular pulse train has been given by Aaron.[11]

APPENDIX B

A Distribution Identity

Essential to the formulation of the spectral density is the relationship between functions F_{ij} and the limit of

$$\sum_{k=1}^{N} q_{ij}^{(k)}(\tau) \equiv y_N(\tau) \tag{104}$$

as $N \to \infty$ [cf. (11) and (18)]. It is convenient to consider initially the integral

$$\int_0^{\tau} y_N(\tau)\, d\tau \equiv z_N(\tau). \tag{105}$$

Inasmuch as functions $q_{ij}^{(k)}$ and, consequently, y_N are sectionally continuous, then

$$z_N'(\tau) = y_N(\tau) \tag{106}$$

almost everywhere in the classical sense or identically in the distribution sense. Also, with $q_{ij}^{(k)} \geqq 0$ [cf. (20)] function $y_N \geqq 0$, and

$$\begin{aligned} 0 &\leqq z_N(\tau) \leqq z_N(\tau + \Delta\tau) \qquad (\Delta\tau > 0) \\ 0 &\leqq z_N(\tau) \leqq z_{N+1}(\tau). \end{aligned} \tag{108}$$

Considering the limit conditions on sequence $\{z_N\}$, we note first from definition (20) and the properties of Stieltjes convolution[12] that

$$\begin{aligned} \int_0^{\infty} e^{-s\tau}\, dz_N(\tau) &= \sum_{k=1}^{N} \int_0^{\infty} e^{-s\tau}\, d\left[\int_0^{\tau} q_{ij}^{(k)}(\tau)\, d\tau\right] \\ &= \sum_{k=1}^{N} \frac{1}{s} F_{ij}(s) F_{jj}^{k-1}(s) \\ &= \sum_{k=1}^{N} \frac{F_{ij}}{s}\left[\frac{1 - F_{jj}^{N}}{1 - F_{jj}}\right] \qquad (\operatorname{Re} s = \alpha > 0). \end{aligned} \tag{109}$$

Therefore, the inverse Stieltjes transform[12] yields

$$\begin{aligned} z_N(\tau) = \frac{1}{2\pi i} \int_{\alpha-i\infty}^{\alpha+i\infty} \frac{1}{s^2} \left[\frac{F_{ij}}{1 - F_{jj}} \right] e^{s\tau}\, ds \\ - \frac{1}{2\pi i} \int_{\alpha-i\infty}^{\alpha+i\infty} \frac{1}{s^2} \left[\frac{F_{ij} F_{jj}^{\,N}}{1 - F_{jj}} \right] e^{s\tau}\, ds. \end{aligned} \tag{110}$$

Finally, since (6), (8) and (9) imply

$$\begin{aligned} | F_{ij}(s) | \leqq \int_0^\infty e^{-\alpha\tau}\, dq_{ij}(\tau) = \alpha \int_0^\infty e^{-\alpha\tau} q_{ij}(\tau)\, d\tau \\ < \alpha \int_0^\infty e^{-\alpha\tau}\, d\tau = 1 \qquad (\alpha > 0; \quad i, j = 1, \cdots, M) \end{aligned} \tag{111}$$

then

$$\begin{aligned} \left| \frac{1}{2\pi i} \int_{\alpha-i\infty}^{\alpha+i\infty} \frac{1}{s^2} \left[\frac{F_{ij}}{1 - F_{jj}} \right] e^{s\tau}\, ds \right| \leqq \sup_f \left| \frac{F_{ij}(s)}{1 - F_{jj}(s)} \right| \\ \cdot \int_{-\infty}^{\infty} \frac{df}{\alpha^2 + 4\pi^2 f^2} < \infty \qquad (\alpha > 0) \end{aligned} \tag{112}$$

$$\begin{aligned} \left| \frac{1}{2\pi i} \int_{\alpha-i\infty}^{\alpha+i\infty} \frac{1}{s^2} \left[\frac{F_{ij} F_{jj}^{\,N}}{1 - F_{jj}} \right] e^{s\tau}\, ds \right| \leqq \sup_f \left| \frac{F_{ij}(s) F_{jj}^{\,N}(s)}{1 - F_{jj}(s)} \right| \int_{-\infty}^{\infty} \frac{df}{\alpha^2 + 4\pi^2 f^2} \\ \xrightarrow[N \to \infty]{} 0 \qquad (\alpha > 0) \end{aligned} \tag{113}$$

and, hence, the limit

$$\lim_{N\to\infty} z_N(\tau) = \frac{1}{2\pi i} \int_{\alpha-i\infty}^{\alpha+i\infty} \frac{1}{s^2} \left[\frac{F_{ij}}{1 - F_{jj}} \right] e^{s\tau}\, ds \equiv z(\tau) \qquad (\alpha > 0) \tag{114}$$

exists. Relative to the asymptotic properties of function z we obtain from (25), (114), and (107) the conditions

$$\int_0^\infty e^{-s\tau}\, dz(\tau) = \frac{1}{s} \frac{F_{ij}(s)}{1 - F_{jj}(s)} \sim \frac{p_j}{s^2} \qquad (s \to 0, \alpha > 0) \tag{115}$$

$$z(\tau) \leqq z(\tau + \Delta\tau) \qquad (\Delta\tau > 0) \tag{116}$$

which by Karamata's Tauberian Theorem[12] give

$$z(\tau) \sim \frac{p_j}{2} \tau^2 \qquad (\tau \to \infty). \tag{117}$$

This asymptotic result together with (112) and (114) implies that

$$[1 + \tau^2]^{-2} z(\tau)\ \varepsilon\ L_1(-\infty, \infty). \tag{118}$$

Thus, function z is a proper distribution, or generalized function (cf. footnote, Section II and Ref. 6, pp. 21–23). In addition, since

$$0 \leqq z_N(\tau) \leqq z_{N+1}(\tau) \leqq z(\tau) \tag{119}$$

then

$$\lim_{N\to\infty}{}^{(\mathrm{D})}\, z_N(\tau) = z(\tau). \tag{120}$$

The functional properties of z as given by (112) and (117) imply also that

$$\lim_{\alpha\to 0^+}{}^{(\mathrm{D})}\, e^{-\alpha\tau} z(\tau) = z(\tau) \qquad (\alpha > 0). \tag{121}$$

In combining (104), (105), (106), and (120), there results

$$\begin{aligned} \mathfrak{F}\cdot z''(\tau) &= \lim_N{}^{(\mathrm{D})}\cdot\mathfrak{F}\cdot z_N''(\tau) = \lim_N{}^{(\mathrm{D})}\cdot\mathfrak{F}\cdot y_N{}' \\ &= \lim_N{}^{(\mathrm{D})} \sum_{k=1}^{N} \int_0^\infty e^{-2\pi i f\tau}\, dq_{ij}{}^{(k)}(\tau). \end{aligned} \tag{122}$$

On the other hand, (114) and (121) give

$$\begin{aligned} \mathfrak{F}\cdot z''(\tau) &= \mathfrak{F}\cdot\frac{d^2}{d\tau^2}\cdot\lim_\alpha{}^{(\mathrm{D})}\, [e^{-\alpha\tau}z(\tau)] \\ &= \mathfrak{F}\cdot\lim_\alpha{}^{(\mathrm{D})} \left\{\left(\frac{d^2}{d\tau^2} + 2\alpha\frac{d}{d\tau} + \alpha^2\right)[e^{-\alpha\tau}z(\tau)]\right\} \\ &= \lim_\alpha{}^{(\mathrm{D})}\, \{[(2\pi i f^2 + 2\alpha(2\pi i f) + \alpha^2]\mathfrak{F}\cdot[e^{-\alpha\tau}z(\tau)]\} \\ &= \lim_\alpha{}^{(\mathrm{D})}\, \{s^2\mathfrak{F}\cdot[e^{-\alpha\tau}z(\tau)]\} \\ &= \lim_\alpha{}^{(\mathrm{D})}\, \frac{F_{ij}(s)}{1 - F_{jj}(s)}. \end{aligned} \tag{123}$$

We finally obtain from (122) and (123) the following identity

$$\begin{aligned} \lim_{N\to\infty}{}^{(\mathrm{D})} \sum_{k=1}^{N} \int_0^\infty e^{-2\pi i f\tau}\, dq_{ij}{}^{(k)}(\tau) &= \lim_N{}^{(\mathrm{D})}\, F_{ij}(2\pi i f)\left[\frac{1 - F_{jj}{}^N(2\pi i f)}{1 - F_{jj}(2\pi i f}\right] \\ &= \lim_{\alpha\to 0+}{}^{(\mathrm{D})}\, \frac{F_{ij}(s)}{1 - F_{jj}(s)}. \end{aligned} \tag{124}$$

APPENDIX C

Definitions of symbols

$x(t)$	— cf. equation (1)
$x_i(t)$	— (10)
$d_n(t)$	— (1)
t_n	— (1)
$t_m^{(i)}$	— (10)
$g_i(t)$	— (3)
$G_i(s)$	— (9)
$s,\bar{s}$	— (9)
$s_{j,n} = s_n$	— (34), (49)
α	— (9)
f	— (9)
$f_{j,n} = f_n$	— (34), (49)
$c_{ij}(\tau)$	— (4b)
$q_{ij}(\tau)$	— (5)
$q_{ij}^{(k)}(\tau)$	— (20)
$F_{ij}(s)$	— (9)
$S_{xx}(f)$	— (9), (11)
$S_{xx}^{(d)}(f)$	— (53)
p_i	— (9)
$p_{i,n}$	— (35)
$\mathfrak{F}$	— (11)
$\mathfrak{L}$	— (9)
$\mu(x)$	— (22)
$\delta(x) = \mu'(x)$	— (23)
δ_{ij}	— (9)
$Q_{ij}(s)$	— (38)
$R_{ij}(s)$	— (39)
$S_{ij}(s)$	— (40)
$T_n^{(ij)}(s)$	— (41)
T	— (73)
T_0	— (76)
K	— (78), (82)
$U_{ij}(s)$	— (88).

REFERENCES

1. Huggins, W. H., Signal Flow Graphs and Random Signals, Proc. I.R.E., **45**, January, 1957, pp. 74–86.
2. Zadeh, L. A., Signal Flow Graphs and Random Signals, Letter to the Editor, Proc. I.R.E., **45**, October, 1957, pp. 1413–1414.
3. Aaron, M. R., Notes on the Computation of Power Spectra from Signal Flow Graphs, to be published.
4. Kolmagorov, A. N., and Fomin, S. V., *Functional Analysis*, Vol. 1, Graylock Press, Rochester, New York, 1957, pp. 105–109.
5. Temple, G., Generalized Functions, Proc. Roy. Soc. (London), A228, 1955, pp. 175–190.
6. Lighthill, M. J., *Fourier Analysis and Generalized Functions*, Cambridge University Press, London, 1959.
7. Middleton, D., *Statistical Communication Theory*, McGraw-Hill Book Company, Inc., New York, 1960, pp. 141–145.
8. Burkill, J. C., *The Lebesque Integral*, Cambridge University Press, London, 1961, p. 74.
9. Titsworth, R. C., and Welch, L. R., Power Spectra of Signals Modulated by Random and Pseudorandom Sequences, Jet Propulsion Laboratories Technical Report No. 32–140, October 10, 1961.
10. Bennett, W. R., Statistics of Regenerative Digital Transmission, B.S.T.J., **37**, November, 1958, pp. 1501–1542.
11. Aaron, M. R., unpublished work.
12. Widder, D. V., *The Laplace Transform*, Princeton University Press, Princeton, 1946.

Imperfections in Active Transmission Lines

By H. E. ROWE

(Manuscript received July 30, 1963)

The effect of discrete imperfections on the behavior of active transmission lines (i.e., lines with distributed gain) is considered. Two cases are studied:

1. Lines with identical, equally spaced reflectors. The transmission and reflection gains versus frequency are studied as functions of the magnitude of the reflectors. Limits on the magnitude of the reflectors to guarantee stability are investigated.

2. Lines with r ındom reflectors, having random position and/or magnitude. The statistics of the transmission are studied; in particular, the average value and the variance and covariance of the transmission are determined for small reflections. If the reflections become large enough, instability may occur, and these calculations may become invalid. Stability of active distributed systems is studied in a companion paper.[1]

I. INTRODUCTION

In the present paper we consider the theory of active transmission lines (i.e., lines with gain) with discrete imperfections. Both equally spaced, identical imperfections and random imperfections will be considered. This study was suggested by R. Kompfner as a rough mathematical model for the effects of imperfections in certain types of optical maser amplifiers, in which the optical signal is reflected back and forth through the active medium on essentially nonoverlapping paths by an array of mirrors. A. G. Fox has suggested that this mathematical model will also provide a description of a one-dimensional active medium (e.g., maser) with (one-dimensional) random inhomogeneities.

Consider an active transmission line that provides exponential gain to both forward and backward waves, and further provides distortionless amplification. The voltage (and current) then vary as

$$\begin{aligned} &e^{-\Gamma z} \text{ — forward wave,} \\ &e^{+\Gamma z} \text{ — backward wave,} \end{aligned} \tag{1}$$

$$\Gamma = -\alpha + j\beta. \tag{2}$$

Since the line has gain,

$$\alpha > 0. \tag{3}$$

Since we assume distortionless transmission, the propagation constant β is related to the angular frequency ω by

$$\beta = \omega/v \tag{4}$$

where the velocity of propagation v is a constant independent of the frequency ω. Further, the gain constant α is independent of ω. We may thus interpret β either as the propagation constant or as the normalized frequency.

Consider a line with N discrete reflectors, as illustrated in Fig. 1. The wave traveling to the right at a distance z is denoted by $W_0(z)$, the wave traveling to the left by $W_1(z)$, as indicated in this figure. We take $W_0(L_k+)$ and $W_1(L_k+)$ as the right- and left-traveling waves just to the right of the kth reflector c_k, $W_0(L_k-)$ and $W_1(L_k-)$ as the right- and left-traveling waves just to the left of the kth reflector.

Each reflector is characterized by a scattering matrix relating incident and reflected waves. Thus for the typical reflector illustrated in Fig. 2 we have

$$\begin{bmatrix} W_1(L_k-) \\ W_0(L_k+) \end{bmatrix} = S_k \begin{bmatrix} W_0(L_k-) \\ W_1(L_k+) \end{bmatrix} \tag{5}$$

$$S = \begin{bmatrix} s_{11} & s_{12} \\ s_{12} & s_{22} \end{bmatrix}. \tag{6}$$

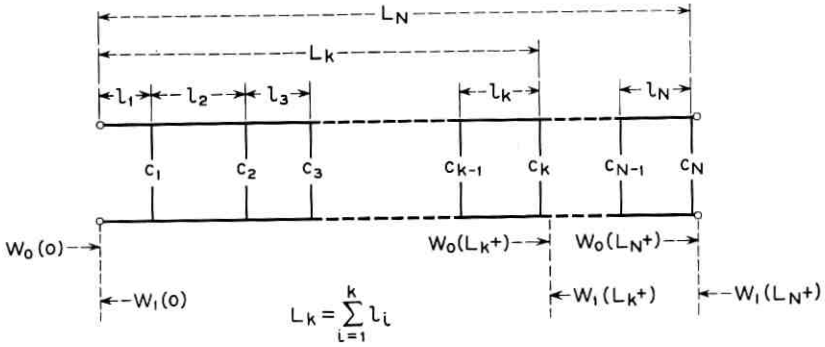


Fig. 1 — Line with N discrete reflectors.

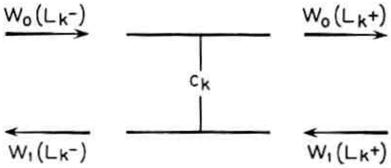


Fig. 2 — Typical reflector.

The incident and reflected wave amplitudes are assumed normalized so that the power in any wave is simply the square of its absolute magnitude. For example, if the reflected wave is absent at the left of the obstacle in Fig. 2 the power in the incident wave is $| W_0(L_k -)|^2$; similarly, if the incident wave is absent the power in the reflected wave is $| W_1(L_k -)|^2$. We make the following assumptions:

1. The powers in the forward and backward waves are additive; for example, the total power P flowing in the $+z$ direction at the left of Fig. 2 is given by

$$P = | W_0(L_k -)|^2 - | W_1(L_k -)|^2. \tag{7}$$

2. The reflectors are lossless, and consequently have unitary scattering matrices.[2] For a reflector of a given magnitude there is a single arbitrary phase parameter in the scattering matrix; this phase has been chosen in such a way as to yield a scattering matrix for the obstacle of the following form:

$$S = \begin{bmatrix} jc & \sqrt{1 - c^2} \\ \sqrt{1 - c^2} & jc \end{bmatrix}, \qquad 0 \leqq | c | \leqq 1. \tag{8}$$

c is a measure of the magnitude of the reflection; for $c = 0$ the reflection is zero and the guide is perfect. c is assumed to be independent of frequency, although this assumption is not compatible with physical realizability. We note that the matrix of (8) is correct only for ω (or β) > 0. For ω (or β) < 0 the signs of the diagonal terms of the matrix must be changed, so that the various responses will be real, even though unrealizable; alternately, we may change the sign of c for negative ω (or β).

Next consider the cascade connection of reflectors and ideal guide sections shown in Fig. 1. We require the wave matrix A corresponding to the scattering matrix of (8) for an obstacle. Referring to Fig. 2,

$$\begin{bmatrix} W_0(L_k-) \\ W_1(L_k-) \end{bmatrix} = A_k \begin{bmatrix} W_0(L_k+) \\ W_1(L_k+) \end{bmatrix}, \tag{9}$$

$$A_k = \frac{1}{\sqrt{1-c_k^2}} \begin{bmatrix} 1 & -jc_k \\ +jc_k & 1 \end{bmatrix}. \tag{10}$$

The wave matrix for the kth line section of length l_k between reflectors c_{k-1} and c_k is given by

$$\begin{bmatrix} W_0(L_{k-1}+) \\ W_1(L_{k-1}+) \end{bmatrix} = \begin{bmatrix} e^{\Gamma l_k} & 0 \\ 0 & e^{-\Gamma l_k} \end{bmatrix} \begin{bmatrix} W_0(L_k-) \\ W_1(L_k-) \end{bmatrix}. \tag{11}$$

Thus the matrix X_k for the cascade connection of the kth line section of length l_k and the kth reflector is given by

$$\begin{gathered} \begin{bmatrix} W_0(L_{k-1}+) \\ W_1(L_{k-1}+) \end{bmatrix} = X_k \begin{bmatrix} W_0(L_k+) \\ W_1(L_k+) \end{bmatrix}, \\ X_k = \frac{1}{\sqrt{1-c_k^2}} \begin{bmatrix} e^{+\Gamma l_k} & -jc_k e^{+\Gamma l_k} \\ +jc_k e^{-\Gamma l_k} & e^{-\Gamma l_k} \end{bmatrix}. \end{gathered} \tag{12}$$

The over-all wave matrix $\bar{X}$ for the line consisting of N sections in Fig. 1 is

$$\begin{bmatrix} W_0(0) \\ W_1(0) \end{bmatrix} = \bar{X} \begin{bmatrix} W_0(L_N+) \\ W_1(L_N+) \end{bmatrix}, \qquad \bar{X} = X_1 X_2 \cdots X_N = \prod_{k=1}^{N} X_k . \tag{13}$$

Setting

$$\bar{X} = \begin{bmatrix} x_{11} & x_{12} \\ x_{21} & x_{22} \end{bmatrix} \tag{14}$$

and referring to Fig. 1, the (complex) transmission and reflection losses $\mathbf{L}_T$ and $\mathbf{L}_R$ or corresponding (complex) gains $\mathbf{G}_T$ and $\mathbf{G}_R$ are given as follows:

$$\mathbf{L}_T = \frac{1}{\mathbf{G}_T} = \frac{W_0(0)}{W_0(L_N+)} = x_{11} \tag{15}$$

$$\mathbf{L}_R = \frac{1}{\mathbf{G}_R} = \frac{W_0(0)}{W_1(0)} = \frac{x_{11}}{x_{21}}. \tag{16}$$

$W_0(0)$, $W_1(0)$ and $W_0(L_N+)$, the incident, reflected, and transmitted waves for the entire structure, are illustrated in Fig. 1.

It has been necessary to state the above analysis in terms of wave matrices that give the input as a function of the output (instead of vice versa) because the boundary conditions are known at the output. The output is assumed to be matched, so that in Fig. 1

$$W_1(L_N+) = 0. \tag{17}$$

In contrast, the reflection coefficient at the input is not known in advance, and so it is not convenient to express the output $\begin{bmatrix} W_0(L_N+) \\ W_1(L_N+) \end{bmatrix}$ as a matrix product times the input $\begin{bmatrix} W_0(0) \\ W_1(0) \end{bmatrix}$.

We consider below two cases of interest:

(a) Identical, equally spaced reflectors,

(b) Independent reflectors with random magnitude and/or position.

II. IDENTICAL, EQUALLY SPACED REFLECTORS

We now assume that all reflectors have identical magnitude and equal spacing. Setting

$$c_k = c, \qquad l_k = l$$

in (12), from (13) and (14) the over-all wave matrix becomes

$$\bar{X} = \begin{bmatrix} x_{11} & x_{12} \\ x_{21} & x_{22} \end{bmatrix} = \frac{1}{(1 - c^2)^{N/2}} \begin{bmatrix} e^{+\Gamma l} & -jc\, e^{+\Gamma l} \\ +jc\, e^{-\Gamma l} & e^{-\Gamma l} \end{bmatrix}^N . \tag{18}$$

By the usual methods we find:

$$x_{11} = \frac{1}{(1 - c^2)^{N/2}(K_+ - K_-)} (K_+\alpha_+^{\,N} - K_-\alpha_-^{\,N}), \tag{19}$$

$$x_{21} = \frac{1}{(1 - c^2)^{N/2}(K_+ - K_-)} (\alpha_+^{\,N} - \alpha_-^{\,N}), \tag{20}$$

$$\alpha_\pm = \cosh \Gamma l \pm \sqrt{\sinh^2 \Gamma l + c^2}, \tag{21}$$

$$K_\pm = \frac{jc\, e^{+\Gamma l}}{e^{+\Gamma l} - \alpha_\pm} = \frac{\alpha_\pm - e^{-\Gamma l}}{jc\, e^{-\Gamma l}}. \tag{22}$$

With the help of (15) and (16) the transmission and reflection gains or losses may be determined.

Consider the various x_{ij} of (18), and in particular x_{11} and x_{21} of (19) and (20), to be functions of $j\beta l$, where we recall from (4) that β is proportional to the angular frequency ω. We recall from the discussion following (8) that these results are valid only for positive frequencies,

$\beta > 0$. The x_{ij} have certain general properties of interest. First, we have:

$$x_{ij}[j(\beta l + \pi)] = (-1)^N x_{ij}[j\beta l], \qquad \beta \geqq 0, \tag{23}$$

$$x_{ij}[j(\pi - \beta l)] = (-1)^{N+i+j} x_{ij}^*[j\beta l], \qquad 0 \leqq \beta l \leqq \pi. \tag{24}$$

Further,

$$x_{ij}[-j\beta l] = x_{ij}^*[j\beta l]. \tag{25}$$

Equation (23) shows that x_{ij} is periodic in the normalized frequency β, of period $2\pi/l$. Equation (25) guarantees that the over-all response to a real input is real. Taken together, (23) and (24) show that the magnitudes of the losses $|\mathbf{L}_T|$ and $|\mathbf{L}_R|$ of (15) and (16) are periodic in β of period π/l, and are symmetric about the points $\beta l = 0, \pi/2, \pi, 3\pi/2, \cdots$. Consequently in studying the magnitudes of these losses at real frequencies we need consider only the range $0 \leqq \beta l \leqq \pi/2$.

Next, from (19)–(22) it might appear that the various functions x_{ij} have branch points in the complex frequency plane because of the radicals in these equations. This is not true, however; a little study of these equations shows that the radicals really disappear for all (integral) N. Alternately, by considering the matrix multiplication of (18) it becomes clear that all the x_{ij} are single-valued functions of Γ, and that no branch points can appear.

We may thus determine the exact expression for the transmission or reflection gain via either (19)–(22) or direct matrix multiplication in (18). However, we shall most often be interested in cases where the reflection parameter c is small in some suitable sense; application of perturbation theory to (19)–(22) greatly simplifies these relations and permits a useful interpretation of these results.

Consider the radical in (21). If

$$|c| \ll |\sinh \Gamma l| \tag{26}$$

then we may expand the radical in a power series and retain only the first correction term. Since

$$|\sinh \Gamma l|^2 = \sinh^2 \alpha l + \sin^2 \beta l \geqq \sinh^2 \alpha l, \tag{27}$$

(26) will be satisfied for all β if

$$|c| \ll \sinh \alpha l. \tag{28}$$

Therefore

$$\sqrt{\sinh^2 \Gamma l + c^2} \approx \sinh \Gamma l + \frac{c^2}{2 \sinh \Gamma l}. \tag{29}$$

Then (21) and (22) become:

$$\alpha_{\pm} \approx e^{\pm\Gamma l} \pm \frac{c^2}{2 \sinh \Gamma l}, \tag{30}$$

$$K_{+} \approx -j\, 2e^{\Gamma l} \frac{\sinh \Gamma l}{c}, \tag{31a}$$

$$K_{-} \approx j\, \tfrac{1}{2}e^{\Gamma l} \frac{c}{\sinh \Gamma l}. \tag{31b}$$

Substituting (30) and (31) into (19) and (20) and neglecting various small quantities, we obtain the following approximate results:

$$x_{11} = \frac{1}{(1 - c^2)^{N/2}} e^{N\Gamma l} [1 + F], \tag{32a}$$

$$F = \left(\frac{c}{2 \sinh \Gamma l}\right)^2 (e^{-2N\Gamma l} - 1), \tag{32b}$$

$$x_{21} = \frac{jc}{(1 - c^2)^{N/2}} e^{-\Gamma l} \frac{\sinh N\Gamma l}{\sinh \Gamma l}. \tag{33}$$

We make one further assumption, often used below, that the total gain in the absence of reflectors ($c = 0$) is large; i.e., referring to (2) and (3),

$$e^{N\alpha l} \gg 1. \tag{34}$$

Then (32b) becomes

$$F = \left(\frac{c}{2 \sinh \Gamma l}\right)^2 e^{-2N\Gamma l}, \qquad e^{N\alpha l} \gg 1. \tag{35}$$

So far we have ignored the question of stability; it is clear that such an active device can oscillate under some conditions. If the device does oscillate, our present results for loss (or gain) lack physical significance, for reasons discussed below. Instability can occur only if the gain functions of (15) and (16) have poles in the right-half complex frequency plane; if all poles of $\mathbf{G}_T$ and $\mathbf{G}_R$ are in the left-half plane the device will be stable. Since from (15–16) the poles of the **G**'s are the zeros of x_{11}, we investigate the zeros of x_{11} as given by the approximate expressions of (32a) and (35).

For $c = 0$, i.e., with reflections absent, the device will be stable, and consequently the zeros of x_{11} lie in the left-half plane. It seems obvious on physical grounds that the device remains stable for small enough values of $|c|$, and will oscillate only when $|c|$ exceeds some critical value. Assuming this to be true, we determine the conditions for stability by finding the minimum value of $|c|$ for which a zero of x_{11} appears on the real frequency axis, i.e., for some value of β.

From (32a) the zeros of x_{11} occur when

$$F = -1. \tag{36}$$

Equivalently,

$$|F| = 1; \tag{37a}$$

$$\angle F = \pm\pi, \pm 3\pi, \cdots . \tag{37b}$$

Noting that

$$\sinh^2 \Gamma l = \sinh^2 (-\alpha + j\beta) l = (\sinh^2 \alpha l + \sin^2 \beta l)\, e^{-j2\varphi}, \tag{38a}$$

$$\varphi = \tan^{-1} \frac{\tan \beta l}{\tanh \alpha l}, \tag{38b}$$

where the principal value of $\tan^{-1}$ is implied, we have from (35)–(37) the following approximate relation for a zero of x_{11} lying on the real frequency axis.

$$F = \frac{c^2}{4(\sinh^2 \alpha l + \sin^2 \beta l)}\, e^{2N\alpha l}\, e^{-j(2N\beta l - 2\varphi)} = -1. \tag{39}$$

Thus

$$N\beta l = \varphi + (\pi/2) + m\pi; \qquad m = 0, \pm 1, \pm 2, \cdots \tag{40a}$$

$$\frac{c^2}{4(\sinh^2 \alpha l + \sin^2 \beta l)}\, e^{2N\alpha l} = 1. \tag{40b}$$

φ is given by (38b). We now fix αl and find the smallest value of $|c|$ for which (40) has a solution. Equation (40a), together with (38b), can be readily seen to have $2(N-1)$ roots $(\beta l)_j$ for $0 < \beta l < 2\pi$. For each of these roots there is a corresponding solution $c = \pm |c_j|$ for (40b). It is obvious that the smallest of these $|c_j|$ corresponds to the smallest $(\beta l)_j$, which is that root lying closest to $\beta l = 0$ and which we denote $(\beta l)_1$.

For convenience we summarize the approximate results derived above in the present section.

$$x_{11} = \frac{1}{(1 - c^2)^{N/2}} e^{N\Gamma l} [1 + F] \tag{41a}$$

$$F = \left(\frac{c}{2 \sinh \Gamma l}\right)^2 e^{-2N\Gamma l}$$

$$= \frac{c^2}{4(\sinh^2 \alpha l + \sin^2 \beta l)} e^{2N\alpha l} e^{-j(2N\beta l - 2\varphi)},$$

$$\varphi = \tan^{-1} \frac{\tan \beta l}{\tanh \alpha l}. \tag{41b}$$

Conditions:

$$| c | \ll \sqrt{\sinh^2 \alpha l + \sin^2 \beta l} \tag{41c}$$

$$e^{N\alpha l} \gg 1. \tag{41d}$$

The results of (41a) and (41b) will be valid for all β if the condition of (41c) is replaced by the more restrictive condition of (42):

$$| c | \ll \sinh \alpha l. \tag{42}$$

The maximum value of the reflection coefficient magnitude $| c |$ that yields a stable amplifier is given as follows, subject to the conditions of (41d) and (42)

$$N(\beta l)_1 = \tan^{-1} \frac{\tan (\beta l)_1}{\tanh \alpha l} + \frac{\pi}{2} \qquad \text{(principal value of } \tan^{-1}) \tag{43a}$$

$$| c |_{\max} = 2e^{-N\alpha l} \sqrt{\sinh^2 \alpha l + \sin^2 (\beta l)_1}\,. \tag{43b}$$

In deriving (43) we required that the results of (41a) and (41b) be valid for all β. Consequently the more restrictive condition of (42) must hold; however, it is not obvious in advance that (42) will end up being satisfied in all cases. However, it is easy to show that this is indeed so, so that the approximate limits on $| c |$ imposed by the requirement of stability are indeed given by (43), so long as (41d) is satisfied (i.e., the high-gain case). From (43a) we have

$$(\beta l)_1 < \pi/N. \tag{44}$$

From (41d) and (44)

$$(\beta l)_1 \ll \alpha l \tag{45}$$

and consequently

$$\sin^2 (\beta l)_1 \ll \sinh^2 \alpha l. \tag{46}$$

Equation (43b) thus guarantees that the more restrictive bound of (42) will always be satisfied in the high-gain case.

The general behavior of the gain-vs-frequency (or βl) curve is readily seen from (41a) and (41b). In the second line of (41a) the first factor and φ vary slowly with βl, while the factor $e^{-j2N\beta l}$ varies rapidly. The angle of F increases steadily as βl increases from 0 to 2π; the magnitude of F is largest at $\beta l = 0, \pi, 2\pi, \cdots$, and decreases rapidly away from these points. Therefore the gain $\mathbf{G}_T$ of (15) plotted vs βl (or frequency) will have an oscillatory behavior, with the magnitude of oscillation greatest near $\beta l = 0, \pi, 2\pi, \cdots$, and quite small elsewhere. The larger N, the more rapid will be the rate of oscillation.

It is instructive to consider a few numerical examples. We consider the following two cases:

$$\begin{aligned}
20 \log_{10} e^{N\alpha l} &\equiv 20 \log_{10} e^{\alpha L_N}\\
&= 30 \text{ db, total gain in } (i) \text{ and } (ii) \text{ below}\\
(i) \qquad 20 \log_{10} e^{\alpha l} &= 1 \text{ db, gain per section}\\
N &= 30\text{, number of sections}\\
(180/\pi)\cdot(\beta l)_1 &= 4.05^\circ\text{, phase shift per section at oscillation}\\
\lvert c \rvert_{\max} &= 0.00860\text{, maximum value of reflection coefficient}\\
& \text{for stability}\\
(ii) \qquad 20 \log_{10} e^{\alpha l} &= 0.1 \text{ db, gain per section}\\
N &= 300\text{, number of sections}\\
(180/\pi)\cdot(\beta l)_1 &= 0.405^\circ\text{, phase shift per section at oscillation}\\
\lvert c \rvert_{\max} &= 0.000860\text{, maximum value of reflection co-}\\
& \text{efficient for stability.}
\end{aligned}$$

The total gain in both cases is large, and hence $\lvert c \rvert_{\max}$ has been computed by (43). The transmission gain $\mathbf{G}_T$ plotted versus the normalized frequency βl for these two cases is shown in Figs. 3 and 4 respectively for several values of c. These results are computed by direct matrix multiplication [see (18)] rather than via (19)–(22) or via the approximate results of (41). Figs. 3(a) and 4(a) show the gain vs normalized frequency for three values of $\lvert c \rvert$ less than $\lvert c \rvert_{\max}$ as well as for $c = \lvert c \rvert_{\max}$ [computed via the approximate results of (43)], which corresponds to the limiting case of stability. It is readily seen how the device approaches instability as c approaches $\lvert c \rvert_{\max}$. Figs. 3(b) and 4(b) show computed curves of the "gain" versus frequency for a value of c greater than $\lvert c \rvert_{\max}$. Under these conditions the device is unstable, so that these curves have little direct physical significance; however, these curves do not look too different from the stable ones of Figs. 3(a) and 4(a). This should provide explicit warning against taking any such computed curve seriously without first investigating stability.

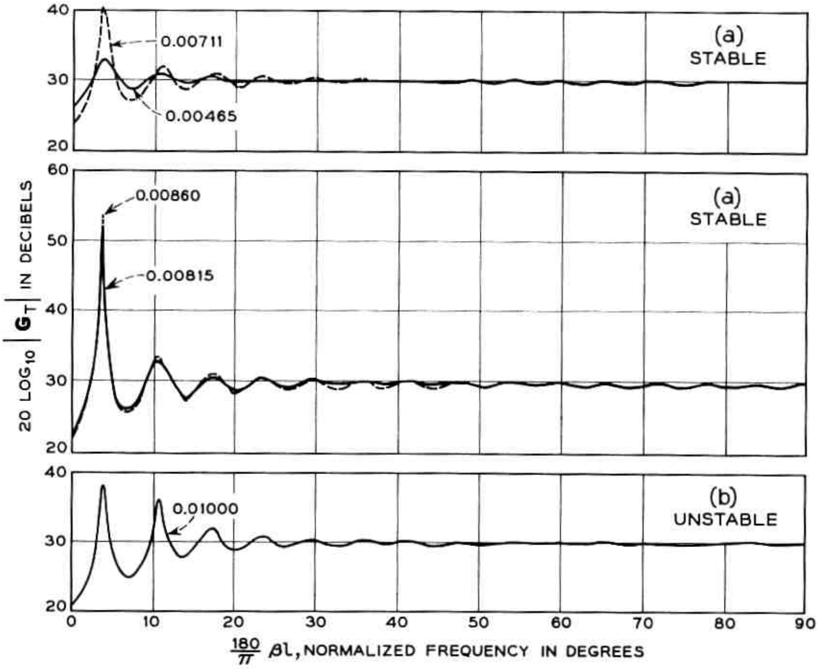


Fig. 3 — Transmission gain vs normalized frequency for one-dimensional active medium with identical, equally-spaced reflectors. $N = 30$, number of sections; $20 \log_{10} e^{\alpha l} = 1$ db, gain per section; total gain = 30 db; c = magnitude of reflectors, parameter indicated on curves.

A detailed picture of the behavior of these devices could be worked out in terms of the poles of the gain function in the complex plane. For small $|c|$ the poles lie in the left-half plane. As $|c|$ is increased the poles move toward the j-axis, causing greater oscillation in the gain-frequency curve. As $|c| \rightarrow |c|_{\max}$ the closest pole touches the j-axis, causing the gain to approach infinity at one frequency. Finally, as $|c|$ becomes greater than $|c|_{\max}$ this pole moves to the right-half plane and the "gain"-frequency curve becomes finite. As $|c|$ increases further the first peak decreases, but the next pole approaches the j-axis, so that the second peak increases, approaches infinity, and eventually decreases. The different peaks in the gain-frequency curve behave in a similar manner as the various poles cross the j-axis in succession.

Figs. 5 and 6 show similar curves for the reflection gain $\mathbf{G}_R$. $\mathbf{G}_R$ approaches infinity for the same values of $|c|$ and βl as does $\mathbf{G}_T$; this must be so, since for the limiting case of stability, power must emerge from both ends of the device in the absence of any incident wave. As in

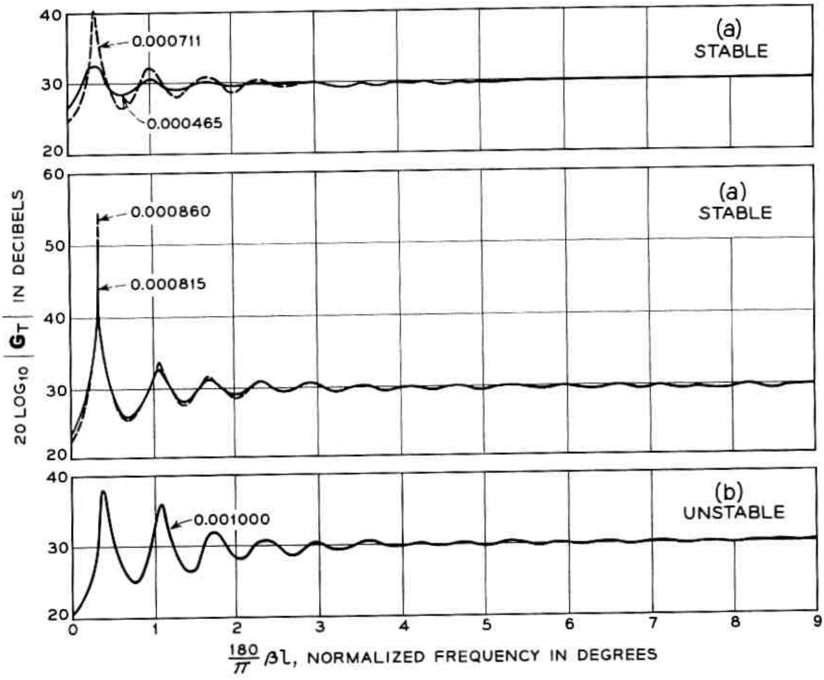


Fig. 4 — Transmission gain vs normalized frequency for one-dimensional active medium with identical, equally-spaced reflectors. $N = 300$, number of sections; $20 \log_{10} e^{\alpha l} = 0.1$ db, gain per section; total gain = 30 db; c = magnitude of reflectors, parameter indicated on curves.

Figs. 3(b) and 4(b), the curves of Figs. 5(b) and 6(b) correspond to instability and hence lack direct physical significance.

If the total gain in the absence of reflectors is not large, then the above results of (43) are not valid, and the approximate results of (41) are not valid over the entire range of permissible values of c. It is interesting to examine the exact computer solutions for one such case.

$$
\begin{aligned}
(iii) \qquad 20 \log_{10} e^{\alpha l} &= 0.1 \text{ db, gain per section} \\
N &= 50, \text{ number of sections} \\
20 \log_{10} e^{N\alpha l} &\equiv 20 \log_{10} e^{\alpha L_N} \\
&= 5 \text{ db, total gain} \\
(180/\pi)\cdot(\beta l)_1 &= 5^\circ, \text{ phase shift per section at oscillation} \\
|\, c \,|_{\max} &= 0.065, \text{ maximum value of reflection coefficient for stability.}
\end{aligned}
$$

Gain-frequency curves for several values of c are shown in Figs. 7 and 8. The values of $(\beta l)_1$ and $|\, c \,|_{\max}$ given above have been determined

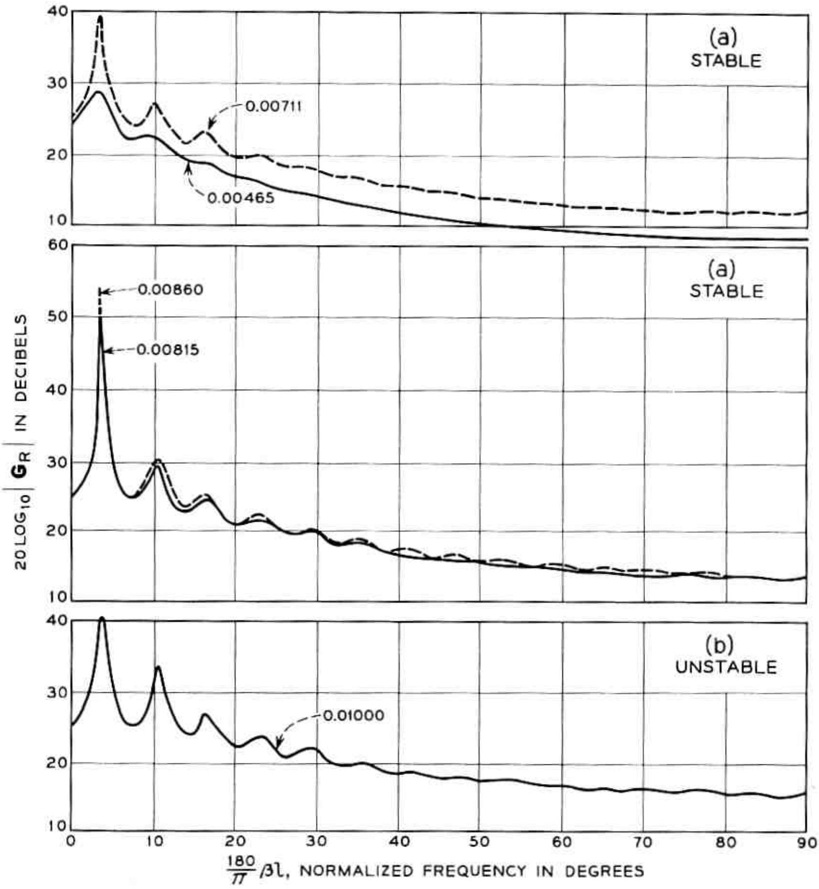


Fig. 5 — Reflection gain vs normalized frequency for one-dimensional active medium with identical, equally-spaced reflectors. $N = 30$, number of sections; $20 \log_{10} e^{\alpha l} = 1$ db, gain per section; total gain = 30 db; c = magnitude of reflectors, parameter indicated on curves.

from these curves. As above, Figs. 7(a) and 8(a) show the transmission and reflection gains for the stable case, $| c | \leqq | c |_{\max}$, while Figs. 7(b) and 8(b) show the "gains" for an unstable case. The general comments given above for examples (*i*) and (*ii*) apply also to this case. The approximation of (43), which was valid in examples (*i*) and (*ii*) above, would have predicted $(\beta l)_1 = 3.37°$, $| c |_{\max} = 0.0135$ for the oscillation conditions; this approximation is quite inaccurate in the present low-gain case, particularly for $| c |_{\max}$.

Straightforward calculation based on (18) or (19)–(22) in the peri-

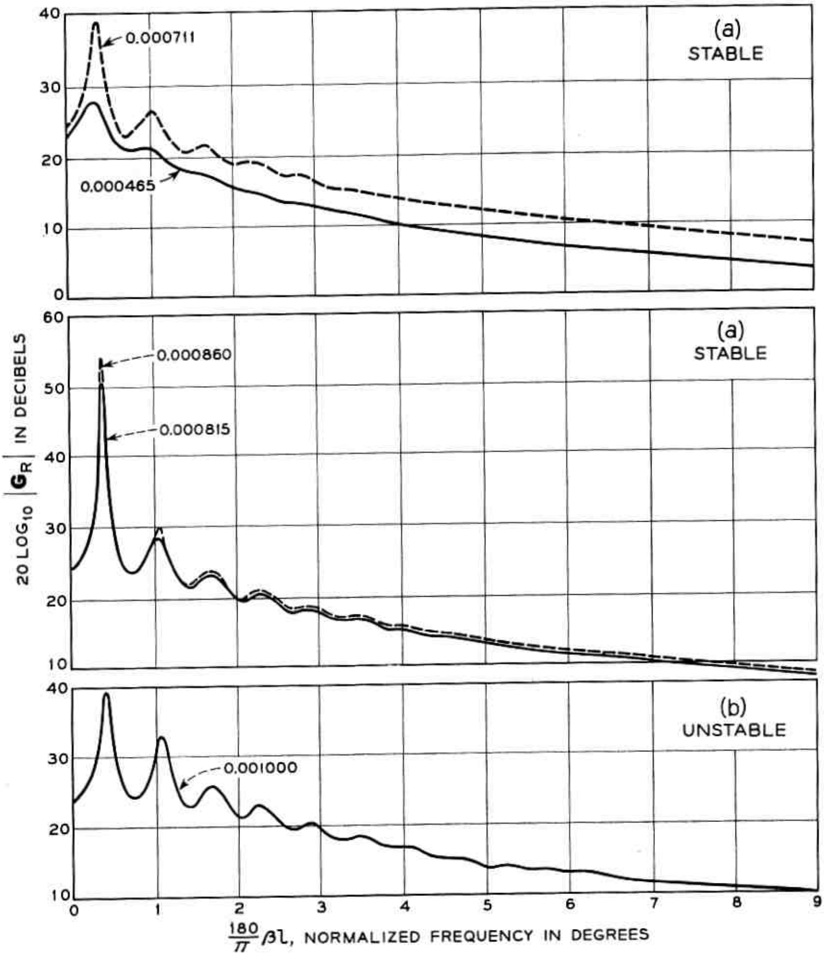


Fig. 6 — Reflection gain vs normalized frequency for one-dimensional active medium with identical, equally-spaced reflectors. $N = 300$, number of sections; $20 \log_{10} e^{\alpha l} = 0.1$ db, gain per section; total gain = 30 db; c = magnitude of reflectors, parameter indicated on curve.

odic case, or (12) and (13) in the general case, will of course always lead to some definite result for x_{11} as a function of frequency, whether or not the device is stable. However, only if we are assured that the device is stable will x_{11} have the desired physical significance of the steady-state loss function $\mathbf{L}_T$. If the device is unstable it will of course oscillate, and ultimately the linear behavior assumed here must break down. However,

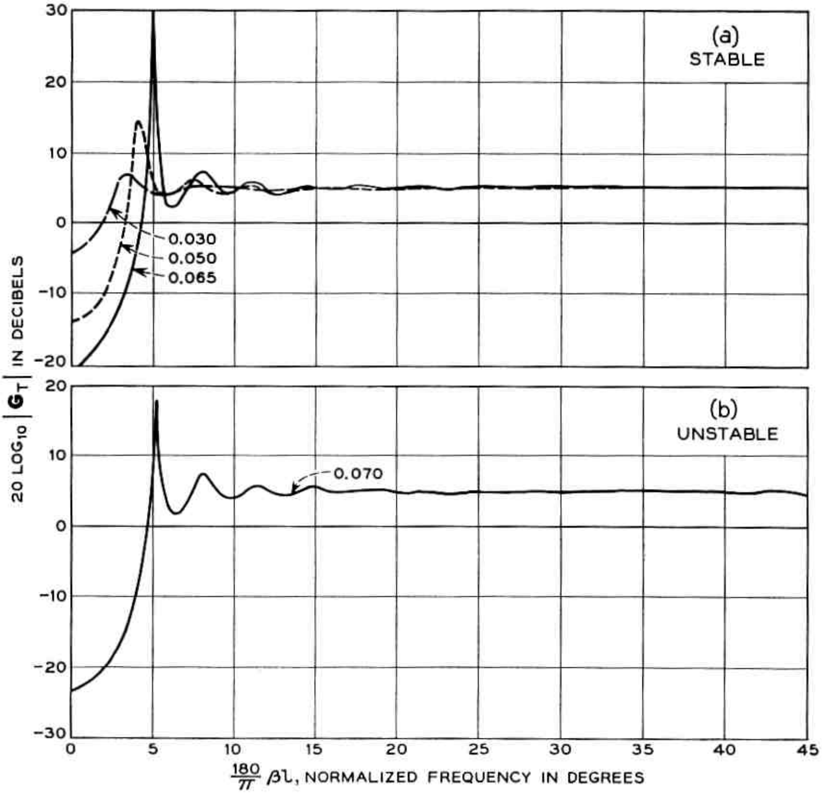


Fig. 7 — Transmission gain vs normalized frequency for one-dimensional active medium with identical, equally-spaced reflectors. $N = 50$, number of sections; $20 \log_{10} e^{\alpha l} = 0.1$ db, gain per section; total gain = 5 db; c = magnitude of reflectors, parameter indicated on curves.

by demanding that the device be at rest at $t = 0$ and examining the initial build-up of oscillation, the mathematical significance of x_{11} may be examined in the unstable case. Suppose the device is initially at rest, and a sinusoidal input is applied at $t = 0$. The total response may be divided into a steady-state response, whose envelope is constant with time, and a transient response, whose envelope ultimately grows or decays exponentially with time in the unstable and stable cases respectively. The steady-state response is given by x_{11} in both cases. In the stable case, since the transients ultimately decay with time, only the steady-state response remains. In the unstable case the steady-state response retains the same mathematical meaning, but since the tran-

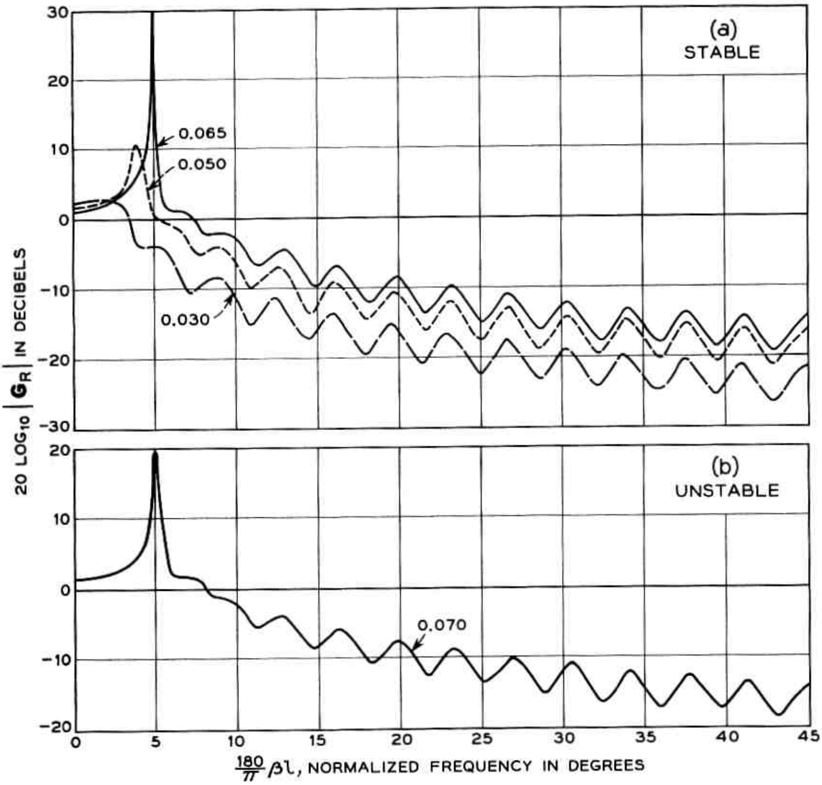


Fig. 8 — Reflection gain vs normalized frequency for one-dimensional active medium with identical, equally-spaced reflectors. $N = 50$, number of sections; $20 \log_{10} e^{\alpha l} = 0.1$ db, gain per section; total gain $= 5$ db; $c =$ magnitude of reflectors, parameter indicated on curves.

sient response grows exponentially with time, the steady-state response loses much of its physical significance.

III. RANDOM REFLECTORS

In the present section we consider active devices with reflectors having random position and/or magnitude; different reflectors are assumed statistically independent. Since the imperfections are random, the loss (or gain) is also a random variable, and we seek various statistics of the loss-frequency curve. The loss $\mathbf{L}_T$ is determined from (12)–(15); we study the average loss and the second-order statistics of the fluctuations about the average, i.e., the variance and covariance of the loss fluctuations. The form of (12)–(15) requires us to study the loss statistics rather than the gain statistics, which are of more direct interest. However, if the loss fluctuations about the average are small, then the loss and gain fluctuations will be almost identical (except for a change in sign), and their statistics will thus also be approximately identical.

As discussed above, (12)–(15) yield the transmission loss $\mathbf{L}_T$ only if the device is stable. If the device is unstable so that oscillation occurs, then the steady-state response $\mathbf{L}_T$ given by (12)–(15) loses much of its physical significance, as discussed in the previous section. The statistics of $\mathbf{L}_T$ computed below are effectively averaged over all cases, so that these results will not be meaningful unless the probability of oscillation is so small that for practical purposes it may be ignored. Thus the results below are valid in the limit of very small reflections, in analogy to the perturbation case of the previous section. In a companion paper[1] useful sufficient conditions guaranteeing stability are obtained; these stability conditions extend the range of validity of the present calculations to finite reflections.

Three different statistical models of an active device with random reflectors are considered in the present paper:

(*i*) random magnitude and spacing
(*ii*) equal magnitude, random spacing
(*iii*) random magnitude, equal spacing.

Thus for case (*i*) in (12)–(15), c_k and l_k will be random variables with appropriate distributions; we assume that the different c_k and l_k are independent random variables. In case (*ii*) the c_k are all equal to the same constant c_0, the l_k are independent random variables. In case (*iii*) the c_k are independent random variables, the l_k equal to the same constant l_0. Case (*ii*) has been suggested by R. Kompfner as being applicable to certain optical maser amplifiers.

In cases (*i*) and (*iii*) we will assume that c_k is symmetrically distributed about 0, with a distribution narrow compared to 1.

We assume in the present paper that l_k is always a large number of wavelengths, so that

$$\beta l_k \gg 2\pi. \tag{47}$$

We further assume in cases (*i*) and (*ii*) that the distribution of l_k about its mean is very narrow with respect to the mean, but wide compared to $2\pi/\beta$. These assumptions are compatible with conditions existing in certain optical amplifiers to which these results might be applied. For certain calculations we need assume in addition only a smooth, symmetrical distribution for l_k about its mean. However, for certain other calculations we must be more specific; here we will assume a Gaussian distribution for l_k, as follows:

$$p(l_k) = \frac{1}{\sqrt{2\pi}\sigma_l} e^{-(l_k - l_0)/2\sigma_l^2}, \tag{48}$$

where l_0 is the expected value and σ_l^2 the variance of l_k,

$$\begin{aligned} l_0 &= \langle l_k \rangle, \\ \sigma_l^2 &= \langle l_k^2 \rangle - \langle l_k \rangle^2. \end{aligned} \tag{49}$$

In accord with (47) and the discussion immediately following, we assume that

$$2\pi/\beta \ll \sigma_l \ll l_0 \,; \qquad \text{cases } (i) \text{ and } (ii). \tag{50}$$

Note that in case *(iii)* $l_k = l_0$, as stated above, and $\sigma_l = 0$.

In the following work we make use of the Kronecker matrix product.[3] For convenience we define this product and summarize some of its properties.

Consider two matrices A and B with elements a_{ij} and b_{ij}. The matrices A and B need not be square, have the same dimensions, or be conformable; their dimensions are completely arbitrary, so that the ordinary matrix products AB or BA may not exist. The Kronecker product, written as $A \times B$, (as opposed to the ordinary matrix product, written as AB) is defined as follows:[3]

$$A \times B = \begin{bmatrix} a_{11}B & a_{12}B & a_{13}B & \cdots \\ a_{21}B & a_{22}B & a_{23}B & \cdots \\ \cdots & \cdots & \cdots & \cdots \end{bmatrix}. \tag{51}$$

$A \times B$ has been written in (51) in partitioned form, with each submatrix consisting of a scalar element of A, a_{ij}, multiplied by the entire matrix B.

Kronecker products have the following useful properties:[3]

$$A \times B \times C = (A \times B) \times C = A \times (B \times C) \tag{52}$$

$$(A + B) \times (C + D) = A \times C + A \times D + B \times C + B \times D \tag{53}$$

$$(A \times B)\,(C \times D) = (AC) \times (BD). \tag{54}$$

As stated above, products without $\times$'s in (52) indicate ordinary matrix products, and the two matrices to be so multiplied must be conformable. Equation (54) may be extended to yield

$$\begin{aligned} &(A_1 \times B_1)\,(A_2 \times B_2) \cdots (A_N \times B_N) \\ &\qquad = (A_1 A_2 \cdots A_N) \times (B_1 B_2 \cdots B_N). \end{aligned} \tag{55}$$

We now return to the results of Section I for the transmission of a general active device. From (13) we have (see Fig. 1)

$$\begin{bmatrix} W_0(0) \\ W_1(0) \end{bmatrix} = X_1 X_2 \cdots X_N \begin{bmatrix} W_0(L_N+) \\ W_1(L_N+) \end{bmatrix}. \tag{56}$$

The output is assumed matched [see (17)], so that

$$W_1(L_N+) = 0. \tag{57}$$

In computing the loss $\mathbf{L}_T$ of (15) we might as well set

$$W_0(L_N+) = 1, \tag{58}$$

so that by (15) $\mathbf{L}_T = W_0(0)$; (56) then becomes

$$\begin{bmatrix} \mathbf{L}_T \\ W_1(0) \end{bmatrix} = X_1 X_2 \cdots X_N \begin{bmatrix} 1 \\ 0 \end{bmatrix}. \tag{59}$$

Now, in determining the average loss and the loss fluctuations about the average we are not particularly interested in the phase variations caused by the variation in total length, which may be large compared to the optical wavelength but is small compared to the average total length. Further, the variations in gain per section will also be small compared to the average gain per section. These considerations suggest the following transformations of (59), which remove these more or less irrelevant contributions to the loss and phase variations. From Fig. 1, the total length L_N is

$$L_N = \sum_{k=1}^{N} l_k . \tag{60}$$

Next define $\mathfrak{L}_T$ and $\mathfrak{R}$ as follows:

$$\mathbf{L}_T = e^{+\Gamma L_N} \cdot \mathfrak{L}_T , \qquad \mathfrak{L}_T = e^{-\Gamma L_N} \cdot \mathbf{L}_T \tag{61}$$

$$W_1(0) = e^{+\Gamma L_N} \cdot \mathfrak{R}, \qquad \mathfrak{R} = e^{-\Gamma L_N} \cdot W_1(0). \tag{62}$$

From (12) we define a new matrix Y_k in terms of X_k as follows:

$$X_k = e^{+\Gamma l_k} \cdot Y_k , \tag{63}$$

where

$$Y_k = \frac{1}{\sqrt{1 - c_k^2}} \begin{bmatrix} 1 & -jc_k \\ +jc_k \, e^{-2\Gamma l_k} & e^{-2\Gamma l_k} \end{bmatrix}. \tag{64}$$

Then from (60)–(64), (59) may be written

$$
\begin{aligned}
e^{+\Gamma L_N}\begin{bmatrix}\mathcal{L}_T\\ \mathcal{R}\end{bmatrix} &= e^{+\Gamma l_1}\cdot Y_1 e^{+\Gamma l_2}\cdot Y_2 \cdots e^{+\Gamma l_N}\cdot Y_N\begin{bmatrix}1\\ 0\end{bmatrix}\\
&= e^{+\Gamma L_N}\cdot Y_1 Y_2 \cdots Y_N\begin{bmatrix}1\\ 0\end{bmatrix}.
\end{aligned}
\tag{65}
$$

Cancelling out the $e^{+\Gamma L_N}$ factor on both sides of (65),

$$
\begin{bmatrix}\mathcal{L}_T\\ \mathcal{R}\end{bmatrix} = Y_1 Y_2 \cdots Y_N\begin{bmatrix}1\\ 0\end{bmatrix},
\tag{66}
$$

where $\mathcal{L}_T$ is defined in (61), Y_k in (64).

Equation (66) is suitable for studying the statistics of the normalized loss $\mathcal{L}_T$, which contains the essential information regarding the loss fluctuations of the device. The quantity $\mathcal{R}$ has to do with the reflected wave at the input corresponding to a unit output wave, and will not be of further interest here. The factor $e^{+\Gamma L_N} = e^{-\alpha L_N}e^{j\beta L_N}$ removed from the unnormalized loss $\mathbf{L}_T$ in (61) is of course a random variable, but for a given amplifier it has constant magnitude and delay.

We now compute $\langle \mathcal{L}_T \rangle$, the expected value of the normalized loss $\mathcal{L}_T$. Since the c_k and l_k are assumed independent random variables, the different Y_k of (66) are independent random matrices in all three cases discussed above. Taking the expected value of both sides of (66), and noting that the different Y_k have the same distribution, we have

$$
\begin{bmatrix}\langle \mathcal{L}_T \rangle\\ \langle \mathcal{R} \rangle\end{bmatrix} = \langle Y \rangle^N \begin{bmatrix}1\\ 0\end{bmatrix},
\tag{67}
$$

where $\langle Y \rangle$ is obtained from (64) as

$$
\langle Y \rangle = \begin{bmatrix}
\left\langle \dfrac{1}{\sqrt{1-c^2}} \right\rangle & -j\left\langle \dfrac{c}{\sqrt{1-c^2}} \right\rangle \\
+j\left\langle \dfrac{c}{\sqrt{1-c^2}} \right\rangle \langle e^{-2\Gamma l} \rangle & \left\langle \dfrac{1}{\sqrt{1-c^2}} \right\rangle \langle e^{-2\Gamma l} \rangle
\end{bmatrix}.
\tag{68}
$$

Note that the independence of c_k and l_k for a given k has been used in obtaining (68); the subscript k has been omitted in the above relations, since the statistics of the different c_k's and of the different l_k's are identical. Finally, since we neglect the small variations in the gain per section, we may set

$$
\langle e^{-2\Gamma l} \rangle \approx e^{2\alpha l_0} \langle e^{-j2\beta l} \rangle,
\tag{69}
$$

where l_0 is given in (49) as the average length of the sections. Then (68) becomes

$$\langle Y\rangle = \begin{bmatrix} \left\langle \dfrac{1}{\sqrt{1-c^2}} \right\rangle & -j\left\langle \dfrac{c}{\sqrt{1-c^2}} \right\rangle \\ +j\,e^{2\alpha l_0}\left\langle \dfrac{c}{\sqrt{1-c^2}} \right\rangle \langle e^{-j2\beta l}\rangle & e^{2\alpha l_0}\left\langle \dfrac{1}{\sqrt{1-c^2}} \right\rangle \langle e^{-j2\beta l}\rangle \end{bmatrix}. \tag{70}$$

Now in cases (i) and (iii) above we have

$$\left\langle \frac{c}{\sqrt{1-c^2}} \right\rangle = 0, \tag{71}$$

since the distribution of c is assumed symmetric about 0. In cases (i) and (ii) we have

$$\langle e^{-j2\beta l}\rangle \approx 0, \tag{72}$$

in view of the assumptions about the distribution of l. Consequently (70) becomes in the three cases:

$$\langle Y\rangle = \begin{cases} \left\langle \dfrac{1}{\sqrt{1-c^2}} \right\rangle \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix}, & \text{case } (i) \\[2ex] \dfrac{1}{\sqrt{1-c_0^2}} \begin{bmatrix} 1 & -jc_0 \\ 0 & 0 \end{bmatrix}, & \text{case } (ii) \\[2ex] \left\langle \dfrac{1}{\sqrt{1-c^2}} \right\rangle \begin{bmatrix} 1 & 0 \\ 0 & e^{2\alpha l_0} e^{-j2\beta l_0} \end{bmatrix}, & \text{case } (iii)\,. \end{cases} \tag{73}$$

From (67) and (73) we have the following final results:

$$\langle \mathfrak{L}_T\rangle = \begin{cases} \left\langle \dfrac{1}{\sqrt{1-c^2}} \right\rangle^N, & \text{cases } (i) \text{ and } (iii) \\[2ex] \left(\dfrac{1}{\sqrt{1-c_0^2}} \right)^N, & \text{case } (ii). \end{cases} \tag{74}$$

The result for case (ii) in (74) may be regarded simply as a special case of the results for cases (i) and (iii). Since in cases (i) and (iii) the distribution of c is assumed narrow compared to 1, we may in some calculations make the following approximation in (74):

$$\left\langle \frac{1}{\sqrt{1-c^2}} \right\rangle \approx 1 + \tfrac{1}{2} \langle c^2 \rangle , \tag{75}$$

where $\langle c^2 \rangle$ is the mean square value of the magnitude of the reflection coefficient.

Equation (74) shows that in all three cases the presence of random reflections has increased the expected value of the loss; further, the average loss is independent of β and hence of frequency. Since $\langle \mathfrak{L}_T \rangle \neq 0$, *if* the deviations of $\mathfrak{L}_T$ from its expected value are very small (as they must be in useful amplifiers), then we will have approximately

$$| \langle \mathfrak{L}_T \rangle | \approx \langle | \mathfrak{L}_T | \rangle. \tag{76}$$

This approximate relation permits us to estimate the variance of the magnitude of the loss, as discussed below. We note that

$$| \langle \mathfrak{L}_T \rangle | \leqq \langle | \mathfrak{L}_T | \rangle. \tag{77}$$

Next consider the mean square value of the loss, $\langle | \mathfrak{L}_T |^2 \rangle = \langle \mathfrak{L}_T \mathfrak{L}_T^* \rangle$. First note from (51) that

$$\begin{bmatrix} \mathfrak{L}_T \\ \mathfrak{R} \end{bmatrix} \times \begin{bmatrix} \mathfrak{L}_T^* \\ \mathfrak{R}^* \end{bmatrix} = \begin{bmatrix} \mathfrak{L}_T \mathfrak{L}_T^* \\ \mathfrak{L}_T \mathfrak{R}^* \\ \mathfrak{R} \mathfrak{L}_T^* \\ \mathfrak{R}\mathfrak{R}^* \end{bmatrix} = \begin{bmatrix} | \mathfrak{L}_T |^2 \\ \mathfrak{L}_T \mathfrak{R}^* \\ \mathfrak{R} \mathfrak{L}_T^* \\ | \mathfrak{R} |^2 \end{bmatrix}. \tag{78}$$

From (66), (55), and (78) we have

$$\begin{bmatrix} | \mathfrak{L}_T |^2 \\ \mathfrak{L}_T \mathfrak{R}^* \\ \mathfrak{R} \mathfrak{L}_T^* \\ | \mathfrak{R} |^2 \end{bmatrix} = (Y_1 \times Y_1^*)(Y_2 \times Y_2^*) \cdots (Y_N \times Y_N^*) \begin{bmatrix} 1 \\ 0 \\ 0 \\ 0 \end{bmatrix} \tag{79}$$

where Y_k is given in (64). Taking the expected value of both sides of (79), again making use of the independence of the different Y_k matrices and the fact that they have the same distribution, we have

$$\begin{bmatrix} \langle | \mathfrak{L}_T |^2 \rangle \\ \langle \mathfrak{L}_T \mathfrak{R}^* \rangle \\ \langle \mathfrak{R} \mathfrak{L}_T^* \rangle \\ \langle | \mathfrak{R} |^2 \rangle \end{bmatrix} = \langle Y \times Y^* \rangle^N \begin{bmatrix} 1 \\ 0 \\ 0 \\ 0 \end{bmatrix}, \tag{80}$$

where $\langle Y \times Y^* \rangle$ is obtained from (64) and (51) as shown in (81).

$$\langle Y \times Y^* \rangle = \begin{bmatrix} \left\langle \dfrac{1}{1-c^2} \right\rangle & +j\left\langle \dfrac{c}{1-c^2} \right\rangle & -j\left\langle \dfrac{c}{1-c^2} \right\rangle & \left\langle \dfrac{c^2}{1-c^2} \right\rangle \\ -j\left\langle \dfrac{c}{1-c^2} \right\rangle \langle e^{-2\Gamma^* l} \rangle & \left\langle \dfrac{1}{1-c^2} \right\rangle \langle e^{-2\Gamma^* l} \rangle & -\left\langle \dfrac{c^2}{1-c^2} \right\rangle \langle e^{-2\Gamma^* l} \rangle & -j\left\langle \dfrac{c}{1-c^2} \right\rangle \langle e^{-2\Gamma^* l} \rangle \\ +j\left\langle \dfrac{c}{1-c^2} \right\rangle \langle e^{-2\Gamma l} \rangle & -\left\langle \dfrac{c^2}{1-c^2} \right\rangle \langle e^{-2\Gamma l} \rangle & \left\langle \dfrac{1}{1-c^2} \right\rangle \langle e^{-2\Gamma l} \rangle & +j\left\langle \dfrac{c}{1-c^2} \right\rangle \langle e^{-2\Gamma l} \rangle \\ \left\langle \dfrac{c^2}{1-c^2} \right\rangle \langle e^{+4\alpha l} \rangle & +j\left\langle \dfrac{c}{1-c^2} \right\rangle \langle e^{+4\alpha l} \rangle & -j\left\langle \dfrac{c}{1-c^2} \right\rangle \langle e^{+4\alpha l} \rangle & \left\langle \dfrac{1}{1-c^2} \right\rangle \langle e^{+4\alpha l} \rangle \end{bmatrix} \quad (81)$$

We again omit the subscript k in the above, since the statistics of the different c_k's and of the different l_k's are assumed identical.

We now apply the same assumptions used above to (81). As in (69), neglecting the small variations in gain per section leads to

$$\begin{aligned} \langle e^{-2\Gamma l} \rangle &\approx e^{2\alpha l_0} \langle e^{-j2\beta l} \rangle , \\ \langle e^{-2\Gamma^* l} \rangle &\approx e^{2\alpha l_0} \langle e^{+j2\beta l} \rangle , \\ \langle e^{4\alpha l} \rangle &\approx e^{4\alpha l_0} , \end{aligned} \tag{82}$$

where l_0 as before is the average length of the sections [see (49)]. Further, we make use of (71) for cases (i) and (iii), and (72) for cases (i) and (ii). The resulting forms for $\langle Y \times Y^* \rangle$ differ in the three cases, but after some simplification the final quantity of interest, $\langle \mid \mathcal{L}_T \mid^2 \rangle = \langle \mathcal{L}_T \mathcal{L}_T^* \rangle$, is given by the following single relation in all three cases:

$$\begin{bmatrix} \langle \mid \mathcal{L}_T \mid^2 \rangle \\ \\ \langle \mid \mathcal{R} \mid^2 \rangle \end{bmatrix} = \begin{bmatrix} \left\langle \dfrac{1}{1 - c^2} \right\rangle & \left\langle \dfrac{c^2}{1 - c^2} \right\rangle \\ \\ e^{4\alpha l_0} \left\langle \dfrac{c^2}{1 - c^2} \right\rangle & e^{4\alpha l_0} \left\langle \dfrac{1}{1 - c^2} \right\rangle \end{bmatrix}^N \begin{bmatrix} 1 \\ \\ 0 \end{bmatrix} . \tag{83}$$

In case (ii), we have in (83)

$$\left\langle \frac{1}{1 - c^2} \right\rangle = \frac{1}{1 - c_0^2}, \qquad \left\langle \frac{c^2}{1 - c^2} \right\rangle = \frac{c_0^2}{1 - c_0^2} . \tag{84}$$

Equation (83) gives the desired result $\langle \mid \mathcal{L}_T \mid^2 \rangle$ in terms of the nth power of a real matrix. The matrix power may of course be written out explicitly in the usual way, but for the sake of simplicity this will not be done here. Some numerical examples are worked out in the next section. The variance of the loss, denoted $\sigma_{\mathcal{L}_T}^2$, is given by

$$\begin{aligned} \sigma_{\mathcal{L}_T}^2 &\equiv \langle \mid \mathcal{L}_T - \langle \mathcal{L}_T \rangle \mid^2 \rangle \\ &= \langle \mid \mathcal{L}_T \mid^2 \rangle - \mid \langle \mathcal{L}_T \rangle \mid^2 . \end{aligned} \tag{85}$$

The variance of the *magnitude* of the loss is given by

$$\begin{aligned} \sigma_{\mid \mathcal{L}_T \mid}^2 &\equiv \langle [\mid \mathcal{L}_T \mid - \langle \mid \mathcal{L}_T \mid \rangle]^2 \rangle = \langle \mid \mathcal{L}_T \mid^2 \rangle - \langle \mid \mathcal{L}_T \mid \rangle^2 \\ &\approx \langle \mid \mathcal{L}_T \mid^2 \rangle - \mid \langle \mathcal{L}_T \rangle \mid^2 \equiv \sigma_{\mathcal{L}_T}^2 , \end{aligned} \tag{86a}$$

where the approximation of (86a) follows from (76). From (77) we have

$$\sigma_{\mid \mathcal{L}_T \mid}^2 \leqq \sigma_{\mathcal{L}_T}^2 . \tag{86b}$$

In these results $\langle \mid \mathcal{L}_T \mid^2 \rangle$ is given by (83), $\langle \mathcal{L}_T \rangle$ by (74); the approximation of (86a) should be good when $\sigma_{|\mathcal{L}_T|}/\langle\mathcal{L}_T\rangle \ll 1$. We see that for all three cases $\langle\,|\,\mathcal{L}_T\,|^2\rangle$ and $\sigma_{\mathcal{L}_T}^{\;2}$ are independent of β and hence of frequency.

Finally we study the covariance of the loss $\mathcal{L}_T$, denoted $R_{\mathcal{L}_T}(\tau)$, defined by

$$R_{\mathcal{L}_T}(\tau) = \langle\mathcal{L}_T(\beta + \tau)\mathcal{L}_T^*(\beta)\rangle = R_{\mathcal{L}_T}^*(-\tau). \tag{87}$$

It will appear below that the expected value in (87) is indeed dependent only on τ, and not on β, within the approximations of the present treatment. If we regard the loss $\mathcal{L}_T(\beta)$ as a random process, then the Fourier transform of $R_{\mathcal{L}_T}(\tau)$ yields the power spectrum of the random processes $\mathcal{L}_T(\beta)$. $R_{\mathcal{L}_T}(\tau)$ thus gives information about both the dc and ac components of $\mathcal{L}_T(\beta)$; of particular interest are the mean square magnitude and the rate of fluctuation of the ac component of the loss. The total "power" (dc plus ac) P_T of the random process $\mathcal{L}_T(\beta)$ is

$$P_T = R_{\mathcal{L}_T}(0) = \langle\,|\,\mathcal{L}_T(\beta)\,|^2\rangle. \tag{88}$$

The dc "power" P_{dc} of $\mathcal{L}_T(\beta)$ is

$$P_{\text{dc}} = R_{\mathcal{L}_T}\,(\infty) = R_{\mathcal{L}_T}(-\infty), \tag{89}$$

where the limits as $\tau \to \pm\infty$ exist. Both ac and dc "powers" are necessarily pure real, and are of course independent of β, since $R_{\mathcal{L}_T}(\tau)$ is independent of β in general. Let us define the dc component of a given $\mathcal{L}_T(\beta)$ curve as

$$\mathcal{L}_{T_{\text{dc}}} = \overline{\mathcal{L}_T(\beta)} = \lim_{M\to\infty} \frac{1}{2M}\int_{-M}^{M} \mathcal{L}_T(\beta)\,d\beta, \tag{90}$$

where the bar indicates an average over β. Then it is easy to show that the dc power of (89) is also equal to

$$P_{\text{dc}} = R_{\mathcal{L}_T}(\infty) = R_{\mathcal{L}_T}(-\infty) = \langle\,|\,\mathcal{L}_{T_{\text{dc}}}\,|^2\rangle, \tag{91}$$

where $\mathcal{L}_{T_{\text{dc}}}$ is given by (90). Let us now define the ac component of a given $\mathcal{L}_T(\beta)$ curve by

$$\mathcal{L}_{T_{\text{ac}}}(\beta) = \mathcal{L}_T(\beta) - \mathcal{L}_{T_{\text{dc}}}\,. \tag{92}$$

Then the covariance $R_{\mathcal{L}_{T_{\text{ac}}}}(\tau)$ of the ac component $\mathcal{L}_{T_{\text{ac}}}(\beta)$ and the ac "power" P_{ac} of the normalized loss $\mathcal{L}_T(\beta)$ are given as follows:

$$R_{\mathcal{L}_{T_{\text{ac}}}}(\tau) = \langle\mathcal{L}_{T_{\text{ac}}}(\beta + \tau)\mathcal{L}_T^{\text{ac}*}(\beta)\rangle = R_{\mathcal{L}_T}(\tau) - R_{\mathcal{L}_T}(\infty), \tag{93a}$$

$$P_{\text{ac}} = \langle\,\overline{|\,\mathcal{L}_{T_{\text{ac}}}(\beta)\,|^2}\rangle = R_{\mathcal{L}_T}(0) - R_{\mathcal{L}_T}(\infty) = R_{\mathcal{L}_{T_{\text{ac}}}}(0). \tag{93b}$$

For convenience we define the covariance of $\mathcal{R}(\beta)$ as an auxiliary quantity, although this quantity is not of present interest to us:

$$R_{\mathcal{R}}(\tau) = \langle \mathcal{R}(\beta + \tau)\mathcal{R}^*(\beta)\rangle. \tag{94}$$

We have

$$\left\langle \begin{bmatrix} \mathcal{L}_T(\beta + \tau) \\ \mathcal{R}(\beta + \tau) \end{bmatrix} \times \begin{bmatrix} \mathcal{L}_T^*(\beta) \\ \mathcal{R}^*(\beta) \end{bmatrix} \right\rangle = \begin{bmatrix} R_{\mathcal{L}_T}(\tau) \\ \langle \mathcal{L}_T(\beta + \tau)\mathcal{R}^*(\beta)\rangle \\ \langle \mathcal{R}(\beta + \tau)\mathcal{L}_T^*(\beta)\rangle \\ R_{\mathcal{R}}(\tau) \end{bmatrix}. \tag{95}$$

From (66), (55), and (95)

$$\begin{bmatrix} R_{\mathcal{L}_T}(\tau) \\ \langle \mathcal{L}_T(\beta + \tau)\mathcal{R}^*(\beta)\rangle \\ \langle \mathcal{R}(\beta + \tau)\mathcal{L}_T^*(\beta)\rangle \\ R_{\mathcal{R}}(\tau) \end{bmatrix} = \langle Y(\beta + \tau) \times Y^*(\beta)\rangle^N \begin{bmatrix} 1 \\ 0 \\ 0 \\ 0 \end{bmatrix} \tag{96}$$

where we again make use of the independence of the different Y_k and the fact that they have the same distribution. Using the various assumptions given above in (69), (71), (72), and (82), and making appropriate simplifications in the different cases, we obtain the following final common result for cases (*i*), (*ii*), and (*iii*):

$$\begin{bmatrix} R_{\mathcal{L}_T}(\tau) \\ \\ R_{\mathcal{R}}(\tau) \end{bmatrix} = \begin{bmatrix} \left\langle \dfrac{1}{1 - c^2} \right\rangle & \left\langle \dfrac{c^2}{1 - c^2} \right\rangle \\ e^{4\alpha l_0} \left\langle \dfrac{c^2}{1 - c^2} \right\rangle \langle e^{-j2\tau l}\rangle & e^{4\alpha l_0} \left\langle \dfrac{1}{1 - c^2} \right\rangle \langle e^{-j2\tau l}\rangle \end{bmatrix}^N \begin{bmatrix} 1 \\ \\ 0 \end{bmatrix}. \tag{97}$$

In addition to the usual approximations, we have used

$$\langle e^{-j2(\beta+\tau)l}\rangle \approx 0 \tag{98}$$

in cases (*i*) and (*ii*) in obtaining the result of (97). This approximation implies that $|\tau| \ll \beta$; i.e., we examine the covariance and hence the loss over only a relatively narrow (electrical) band. In the analysis we often use the quantity $R_{\mathcal{L}_T}(\infty)$, which gives the dc "power" [see (91)]; this is justified because the covariance computed from (97) will approach its asymptotic value $R_{\mathcal{L}_T}(\infty)$ for values of τ satisfying the requirement $|\tau| \ll \beta$. We assume the distribution of l is the Gaussian distribution of (48), and note that $\langle e^{-j2\tau l}\rangle$ is simply related to the corresponding characteristic function.[4] Thus

$$\langle e^{-j2\tau l}\rangle = e^{-j2\tau l_0} e^{-2(\tau\sigma_l)^2}.\dagger \tag{99}$$

† Note that this result justifies the approximations of (72) and (98) [subject to the condition of (50)]. A similar result for $\langle e^{\Gamma l}\rangle$, where Γ is complex, may be readily derived, and justifies the approximation of (69) and (82).

In case (*iii*) we have $\sigma_l = 0$ in (99). Thus we have as our final result:

$$\begin{bmatrix} R_{\mathcal{L}_T}(\tau) \\ R_{\mathcal{R}}(\tau) \end{bmatrix} \tag{100}$$
$$= \begin{bmatrix} \left\langle \dfrac{1}{1-c^2} \right\rangle & \left\langle \dfrac{c^2}{1-c^2} \right\rangle \\ e^{4\alpha l_0} \left\langle \dfrac{c^2}{1-c^2} \right\rangle e^{-j2\tau l_0} e^{-2(\tau\sigma_l)^2} & e^{4\alpha l_0} \left\langle \dfrac{1}{1-c^2} \right\rangle e^{-j2\tau l_0} e^{-2(\tau\sigma_l)^2} \end{bmatrix}^N \begin{bmatrix} 1 \\ 0 \end{bmatrix}$$

$$\left\langle \frac{1}{1-c^2} \right\rangle = \frac{1}{1-c_0^2}, \left\langle \frac{c^2}{1-c^2} \right\rangle = \frac{c_0^2}{1-c_0^2}; \qquad \text{case } (ii)$$
$$\sigma_l = 0; \qquad \text{case } (iii). \tag{101}$$

Certain general properties of $R_{\mathcal{L}_T}(\tau)$ are readily deduced from (100). First, $R_{\mathcal{L}_T}(\tau)$ is independent of β and dependent only on τ, as assumed above in (87). Second, for $\tau = 0$, (100) becomes identical to (83), as it must. Finally, for $\tau \to \infty$, we have in cases (*i*) and (*ii*) from (100) and (101)

$$R_{\mathcal{L}_T}(\infty) = \begin{cases} \left\langle \dfrac{1}{1-c^2} \right\rangle^N, & \text{case } (i) \\ & \text{[also case } (iii) \text{ — see below]} \\ \left(\dfrac{1}{1-c_0^2} \right)^N, & \text{case } (ii). \end{cases} \tag{102}$$

$R_{\mathcal{L}_T}(\infty)$ is real, as stated above. We recall from (91) that $R_{\mathcal{L}_T}(\infty)$ is the dc "power" of $\mathcal{L}_T(\beta)$. The ac "power" is given by (93).

Now, in case (*iii*) the covariance $R_{\mathcal{L}_T}(\tau)$ is periodic, which implies that the random process $\mathcal{L}_T(\beta)$ is periodic;[4] however, this is obvious from the original formulation of the problem. $R_{\mathcal{L}_T}(\infty)$ no longer exists in the strict sense; the dc "power" is now the average value (over τ) of $R_{\mathcal{L}_T}(\tau)$. It turns out that we may approach case (*iii*) by considering case (*i*) and allowing σ_l to approach 0 in (100). [This violates the condition imposed by (50) and used in the approximations of (72) and (98) and so the limiting process $\sigma_l \to 0$ is forbidden in some of the above results; careful examination shows that it *is* valid to allow $\sigma_l \to 0$ in (100).] Then $R_{\mathcal{L}_T}(\tau)$ does approach the limit of (102) as $\tau \to \infty$; and so we take the first result of (102) as the dc "power" in case (*iii*), as well as in case (*i*).

In general

$$\langle \mathfrak{L}_T(\beta) \rangle^2 \neq P_{\mathrm{dc}} \equiv \langle | \overline{\mathfrak{L}_T(\beta)} |^2 \rangle , \tag{103}$$

$$\sigma_{\mathfrak{L}_T}{}^2 \neq P_{\mathrm{ac}} \equiv \langle \overline{| \mathfrak{L}_{T_{\mathrm{ac}}}(\beta) |^2} \rangle . \tag{104}$$

However, in case (*ii*) only — i.e., reflectors of identical magnitude and random spacing — (103) and (104) are true with the $\neq$ replaced by $=$, as seen from (74) and (102).

The matrix power of (100) is easily written explicitly in the usual way, but the results would be rather complicated. Numerical examples are worked out in the next section.

IV. NUMERICAL EXAMPLE — RANDOM REFLECTORS

Consider an optical amplifier with random reflectors of the type given in case (*ii*) of Section III: i.e., the reflectors have identical magnitude but random spacing. Assume:

$$20 \log_{10} e^{\alpha l_0} = 1 \text{ db, nominal gain per section}$$

$$N = 30, \text{ number of sections}$$

$$20 \log_{10} e^{N\alpha l_0} = 30 \text{ db, nominal total gain.}$$

Fig. 9 shows the average normalized loss and the rms fluctuation of the normalized loss about its average value, plotted versus c_0, the magnitude of the reflectors. As seen from example (*i*), Section II, instability is possible if $| c_0 | > 0.00860$. Therefore the curves of Fig. 9 are solid for $c_0 < 0.00860$, dotted for $c_0 > 0.00860$. However, this is intended only as a symbolic reminder of the question of stability. We do not know whether or not instability can occur for $| c_0 | < 0.00860$. Even though we know that instability can occur for $| c_0 | > 0.00860$, the probability of instability might remain so small for some greater range of c_0 that these curves would provide a useful approximation. In Ref. 1, Section VI, equations (122)–(131) we show that stability is guaranteed for $| c_0 | < 0.00590$, assuming that the maximum fractional variation in spacing of the reflectors [ν in (124) of Ref. 1] is small compared to 1. This is indicated in Fig. 9.

All of the above results have been independent of the precise distribution of the l_k, the spacing between reflectors, except that the conditions of (47) and the following sentence must be satisfied. However, the covariance of the loss depends explicitly on the probability distribution of the l_k. For our present example we therefore assume that the differ-

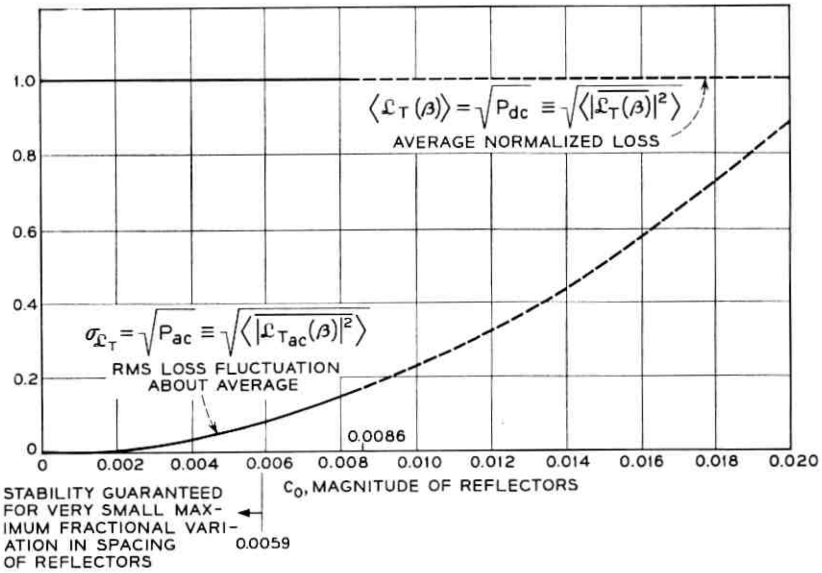


Fig. 9 — Average normalized loss and rms fluctuation about the average for one-dimensional active medium with randomly spaced reflectors of identical magnitude. $N = 30$, number of sections; $20 \log_{10} e^{\alpha l_0} = 1$ db, nominal gain per section; nominal total gain $= 30$ db.

ent l_k are independent, with the Gaussian probability density given in (48)–(50). We further assume the following numerical values:

$$(\sigma_l/l_0) = 0.01, \qquad c_0 = 0.005. \tag{105}$$

Thus, the spacing between successive reflectors is accurate to about 1 per cent, and the magnitude of the reflectors would guarantee stability in the equally spaced case of Section II. Of course a practical device would probably be built much more accurately, but the values in (105) are suitable for illustrating the general behavior. Fig. 10 shows the (complex) covariance $R_{\mathcal{L}_{T_{ac}}}(\tau)$ of the ac component $\mathcal{L}_{T_{ac}}(\beta)$ of the normalized loss for this case as a function of the normalized variable $(l_0/\pi)\tau$, for $0 < (l_0/\pi)\tau < 4$. Fig. 10(a) shows the magnitude $| R_{\mathcal{L}_{T_{ac}}}(\tau) |$ and Fig. 10(b) the phase $\angle R_{\mathcal{L}_{T_{ac}}}(\tau) + 58\, l_0\tau$; note that the linear component of phase has been removed in the plot of Fig. 10(b). The covariance is seen to be approximately a damped periodic function of τ; Fig. 11 shows a plot of the magnitude of the covariance at the points $\tau = n(\pi/l_0)$, which correspond closely to the maxima of $| R_{\mathcal{L}_{T_{ac}}}(\tau) |$.

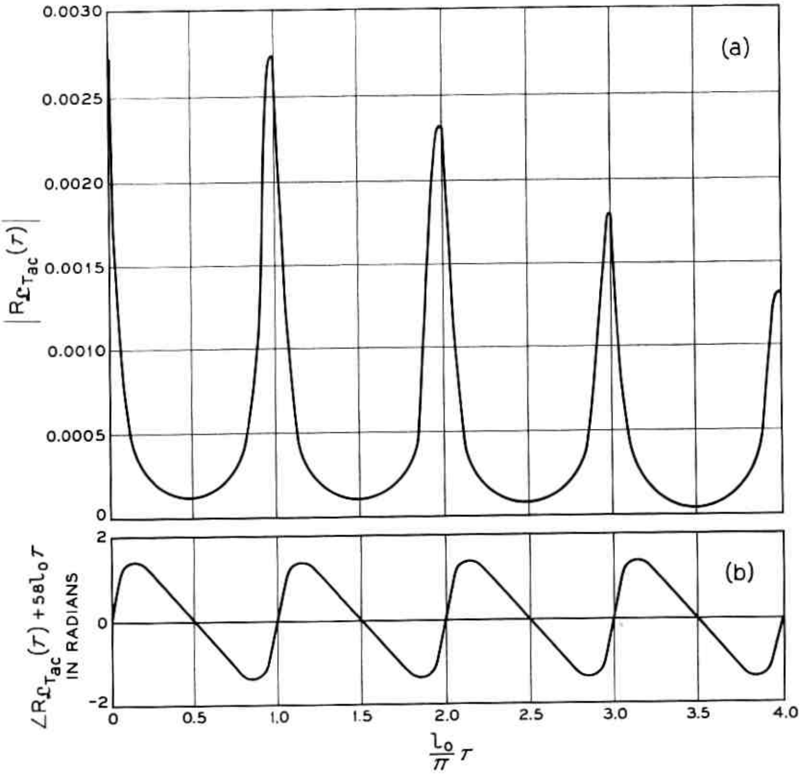


Fig. 10 — Covariance of ac component of normalized loss for one-dimensional active medium with randomly spaced reflectors. $\sigma_l/l_0 = 0.01$; $c_0 = 0.005$, magnitude of reflectors; $N = 30$, number of sections; $20 \log_{10} e^{\alpha l_0} = 1$ db, nominal gain per section; nominal total gain = 30 db.

We would expect some resemblance between the covariance of Figs. 10 and 11, for reflectors with identical magnitude but random spacing, and the (nonrandom) case of Section II for reflectors with identical magnitude and spacing. For the nonrandom case we have seen that the loss is periodic; consequently the covariance will also be periodic, and will look something like that of Figs. 10 and 11 for the random case except that it will not be damped. Note that the large linear component $-58\ l_0\tau$ that has been removed from the phase curve of Fig. 10(b) implies that the power spectrum of the random process $\mathfrak{L}_{T_{ac}}(\beta)$ is concentrated around the angular "frequency" $-58\ l_0$; this angular "frequency" corresponds to the rate of variation of the loss for two reflectors whose separation is equal to the nominal spacing of the two end reflectors in the random case.

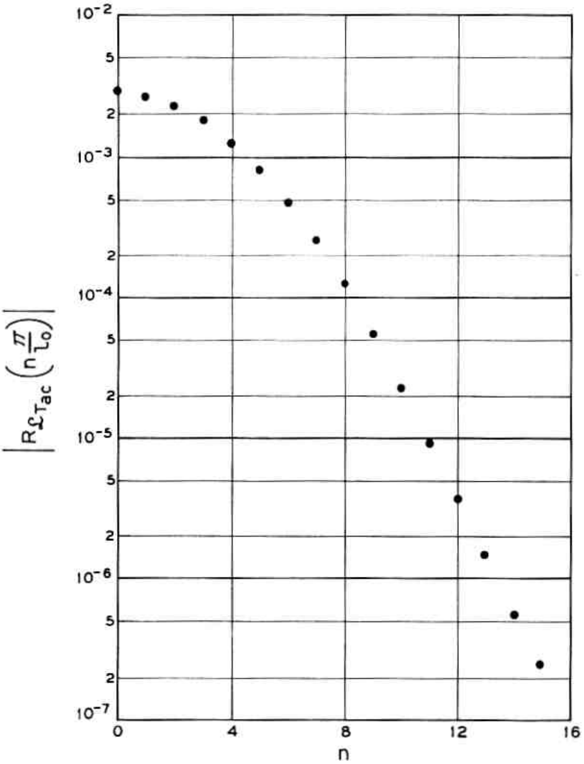


Fig. 11 — Approximate maxima of covariance of ac component of normalized loss for one-dimensional active medium with randomly spaced reflectors (see Fig. 10). $\sigma_l/l_0 = 0.01$; $c_0 = 0.005$, magnitude of reflectors; $N = 30$, number of sections; $20 \log_{10} e^{\alpha l_0} = 1$ db, nominal gain per section; nominal total gain = 30 db.

V. DISCUSSION

The question of stability has been discussed for the periodic case at the end of Section II. There it is pointed out that these calculations are valid only if the device is stable, i.e., does not oscillate. The same is true in the random case. In the periodic case we can determine by calculation the limits of stability, and this has been done in the examples of Section II. Stability in the random case is studied in Ref. 1.

Various higher-order transmission statistics may be calculated by methods similar to those used above, but the complexity of the calculations increases with the order of the statistics. In addition, statistics of the real and imaginary parts of the normalized loss $\mathcal{L}_T$ may be readily determined by similar methods.

VI. ACKNOWLEDGMENT

The author would like to thank Mrs. C. A. Lambert for programming all of the numerical calculations.

REFERENCES

1. Rowe, H. E., Stability of Active Transmission Lines with Arbitrary Imperfections, B.S.T.J., this issue, p. 293.
2. Montgomery, C. G., Dicke, R. H., and Purcell, E. M., *Principles of Microwave Circuits*, McGraw-Hill, New York, 1948.
3. Bellman, R., *Introduction to Matrix Analysis*, McGraw-Hill, New York, 1960.
4. Davenport, W. B., and Root, W. L., *Random Signals and Noise*, McGraw-Hill, New York, 1958.

Stability of Active Transmission Lines with Arbitrary Imperfections

By H. E. ROWE

(Manuscript received August 23, 1963)

Two sufficient conditions for the stability of one-dimensional active transmission lines with arbitrary imperfections (i.e., discrete or continuous reflections) are derived. The first stability condition guarantees stability for any arbitrary distribution of reflection. The second stability condition is restricted to a special case of interest that includes discrete reflectors with nominally equal magnitude and spacing; the stability condition for this restricted class is greatly improved over the general stability condition described above.

These results, aside from their own interest, provide rigorous justification for previous calculations for the gain statistics of such a device with random discrete reflectors.[1] *They may also be used to find an upper bound on the probability of instability of such a device with random reflectors.*

Certain types of optical maser amplifiers and traveling-wave tubes provide examples of practical devices with distributed gain to which these results, or similar ones, might be applied.

I. INTRODUCTION

The preceding paper[1] has considered the theory of active transmission lines with discrete imperfections. First, lines with equally-spaced identical reflectors were studied; in particular, gain-frequency curves were determined as functions of the various parameters, and the stability of the device was studied under these special conditions. It was pointed out that the mathematical expression for gain would yield a perfectly definite result for any values of the parameters, but that this mathematical result would have physical significance only if the device is stable, i.e., does not oscillate.

Next, the case of random imperfections was studied.[1] Here the statistics of the transmission were determined in terms of the statistics of the discrete reflectors, which were assumed to have random position and magnitude. Again, these results have physical significance only if the device is stable (or if the probability of instability is negligible). However, in the random case no precise information about stability was given; the computed statistics of the transmission were felt to be valid if the rms magnitude of the discrete reflectors was sufficiently small, but only intuitive feelings of what was "small enough" were available.

In the present paper we derive a sufficient condition for stability of an active transmission line with arbitrary reflectors; we further show (by one example) that this sufficient condition cannot be greatly improved (if at all) in the general case. This result gives useful information regarding the range of validity of the calculations of the preceding paper[1] for the transmission statistics of active transmission lines with random reflectors. This general bound on stability may be improved if additional information is known about the distribution of reflectors; one such case of interest is treated.

The mathematical model chosen for this problem is discussed in detail in Ref. 1. A line with N discrete reflectors is shown in Fig. 1 (which is identical to Fig. 1 of Ref. 1). The wave traveling to the right at distance z is denoted by $W_0(z)$, the wave traveling to the left by $W_1(z)$; $W_0(L_k+)$ and $W_1(L_k+)$ are the right- and left-traveling waves just to the right of the kth reflector, as indicated in this figure, while $W_0(L_k-)$ and $W_1(L_k-)$ are the right- and left-traveling waves just to the left of the kth reflector.

In the absence of reflections the forward and backward waves vary as

$$\begin{aligned} W_0(z) &\propto e^{-\Gamma z} \qquad \text{— forward wave} \\ W_1(z) &\propto e^{+\Gamma z} \qquad \text{— backward wave} \end{aligned} \tag{1}$$

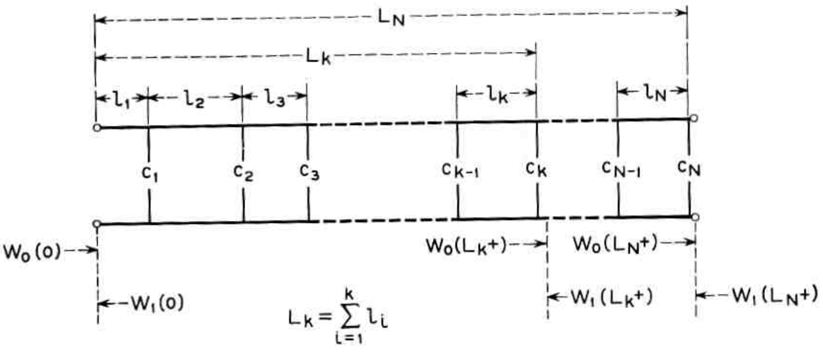


Fig. 1 — Line with N discrete reflectors.

where

$$\Gamma = -\alpha + j\beta, \qquad \alpha > 0. \tag{2}$$

The line has gain, so that $\alpha > 0$. From (12) of Ref. 1, the wave matrix for the cascade connection of the kth line section of length l_k and the kth reflector is

$$X_k = \frac{1}{\sqrt{1 - c_k^2}} \begin{bmatrix} e^{+\Gamma l_k} & -jc_k e^{+\Gamma l_k} \\ +jc_k e^{-\Gamma l_k} & e^{-\Gamma l_k} \end{bmatrix}, \qquad |\, c_k \,| \leqq 1, \tag{3}$$

$$\begin{bmatrix} W_0(L_{k-1}+) \\ W_1(L_{k-1}+) \end{bmatrix} = X_k \cdot \begin{bmatrix} W_0(L_k+) \\ W_1(L_k+) \end{bmatrix} \tag{4}$$

where $|\, c_k \,|$ is the magnitude of the reflection coefficient for the kth reflector. The over-all transmission matrix for the entire line of Fig. 1, denoted by $\bar{X}$, is given by the matrix product of (13) of Ref. 1:

$$\bar{X} = \prod_{k=1}^{N} X_k , \tag{5}$$

$$\begin{bmatrix} W_0(0) \\ W_1(0) \end{bmatrix} = \bar{X} \cdot \begin{bmatrix} W_0(L_N+) \\ W_1(L_N+) \end{bmatrix}. \tag{6}$$

For convenience, denote the elements of the over-all transmission matrix $\bar{X}$ as in (14) of Ref. 1.

$$\bar{X} = \begin{bmatrix} x_{11} & x_{12} \\ x_{21} & x_{22} \end{bmatrix}. \tag{7}$$

$\bar{X}$ is given by (3) and (5). Assume the device is operated as an amplifier with matched input and output; setting $W_1(L_N+) = 0$, the complex transmission gain $\mathbf{G}_T$ is given by

$$\mathbf{G}_T = \frac{W_0(L_N+)}{W_0(0)} = \frac{1}{x_{11}}. \tag{8}$$

Now x_{11} is a function of Γ and of all of the l_k's and c_k's. We may conceptually investigate stability in the following way. Imagine that c_k is replaced by ϵc_k throughout this analysis; ϵ is a variable parameter that scales the magnitudes of all of the coupling coefficients. Let ϵ be increased from 0, and for each value of ϵ examine x_{11} [which in (8) is the reciprocal of the transmission gain, and so may be regarded as the transmission loss] as a function of frequency ω (or of the phase constant β, which is assumed proportional to frequency, since the line is distortionless)[1] over the entire range $-\infty < \omega < +\infty$. We determine in this way the minimum value of $|x_{11}|$ for each value of ϵ. As ϵ increases, this minimum value of $|x_{11}|$ will eventually just drop to zero, for a critical value of ϵ which we denote by ϵ_c. Thus, as $\epsilon \to \epsilon_c$ the gain $|\mathbf{G}_T| \to \infty$ for a particular value of ω, and the device oscillates. ϵ_c is the dividing line between stability and instability; if $\epsilon_c > 1$, the original device, with the parameters c_k and l_k, is stable.

Such calculations have actually been carried out in Ref. 1 for devices with identical, equally-spaced reflectors. In this case the gain $\mathbf{G}_T$ is a periodic function of frequency ω, so that only a finite portion of the frequency axis (i.e., one period) must be investigated. In general, however, $\mathbf{G}_T$ is not periodic; since we cannot investigate numerically the entire ω-axis, it is not obvious how to investigate stability for the general case.

In the remainder of this paper we determine a sufficient condition that guarantees the stability of a general active line with arbitrary discrete imperfections. In particular, consider such a device, illustrated in Fig. 1, characterized by (3), (5), and (6), with arbitrary α, c_k, and l_k. We show below that any such device satisfying the condition

$$\sum_{i=1}^{N} \tanh^{-1} |c_i| < 2 \sinh^{-1} \frac{e^{-\alpha L_N}}{\sqrt{2}} \tag{9}$$

must be stable. Many practical devices will have large gain, and hence must have small reflections. In such cases $e^{-\alpha L_N} \ll 1$ and $|c_i| \ll 1$; under these conditions a slightly poorer stability condition derived from (9) is useful.

$$\sum_{i=1}^{N} |c_i| \leqq \tanh\left[2 \sinh^{-1} \frac{e^{-\alpha L_N}}{\sqrt{2}}\right]. \tag{10}$$

In the high-gain case the right-hand side of (10) may be made simpler still by further degrading this stability condition. We may show, for example, that

$$\tanh\left[2 \sinh^{-1} \frac{e^{-\alpha L_N}}{\sqrt{2}}\right] \geqq 0.932 \sqrt{2} e^{-\alpha L_N}, \qquad 8.686\alpha L_N \geqq 10 \text{ db}. \tag{11}$$

Thus a slightly poorer version of (10) is

$$\sum_{i=1}^{N} |c_i| \leqq 0.932 \sqrt{2}\, e^{-\alpha L_N}, \qquad 8.686\, \alpha L_N \geqq 10 \text{ db}. \tag{12}$$

The stability condition of (12) is valid when the one-way gain of the active medium exceeds 10 db. As the lower bound on the one-way gain of the active medium increases beyond 10 db, the numerical factor 0.932 on the right-hand side of (12) increases, approaching 1 as the lower bound on the gain approaches infinity. This is readily seen from (10); as $\alpha L_N \to \infty$, $e^{-\alpha L_N} \to 0$, so that the $\sinh^{-1}$ and tanh functions in (10) may be approximately replaced by their arguments for sufficiently large αL_N. However, direct calculation with (10) is straightforward; the result of (12) (or similar equations) is intended principally to illustrate the general behavior.

Thus (9) or the successively poorer versions of (10) and (12) guarantee that the device will be stable, even for the worst possible choice of the c_k and l_k. Equations (9), (10), and (12) are each sufficient, but not necessary, conditions for stability. These results are derived in Sections II, III, and IV. In addition, a better bound is obtained for a special case in which the reflection coefficient is distributed more or less uniformly with distance z along the active line, in a certain sense to be described more precisely in Section V below; these results include many cases of interest. Finally, some numerical examples illustrating the use of these two different types of bounds are given in Section VI.

II. DIFFERENTIAL EQUATIONS EQUIVALENT TO MATRIX RELATIONS

Consider the following differential equations:

$$\begin{aligned} W_0'(z) &= -\Gamma W_0(z) + jr(z)W_1(z), \\ W_1'(z) &= -jr(z)W_0(z) + \Gamma W_1(z). \end{aligned} \tag{13}$$

These relations have the form of the coupled line equations with a general continuous coupling coefficient. In the present case, $W_0(z)$ and $W_1(z)$ are the right- and left-directed traveling-wave complex amplitudes, and $r(z)$ is the continuous reflection that couples the two waves to each other. Equation 13 is readily obtained as a limiting form of the matrix relations of (3), (5), and (6) by assuming very small, closely spaced discrete reflectors whose magnitude varies slowly with distance. Thus in the matrix relations of Section I above set

$$l_k = \Delta z. \tag{14}$$

Assume that c_k varies slowly with k. Then we set

$$c_k = r(k\Delta z) \cdot \Delta z, \tag{15}$$

where $r(z)$ is a continuous function. We now let $\Delta z \to 0$ so that the number of discrete reflectors $\to \infty$; during this process the continuous function $r(z)$ is fixed and the c_k determined by (15), so that the magnitudes of the individual reflectors → 0 as $\Delta z \to 0$. Then the matrix relations of (3), (5), and (6) will yield the continuous differential equations of (13). The analysis is straightforward and quite similar to that of Ref. 2 for a similar problem, and so will not be given here. The above discussion of (13) as an appropriate limiting continuous form of the matrix relations of Section I is given only to provide some physical motivation for considering (13), and plays no part in the mathematical analysis to follow.

The case of isolated, discrete reflectors, characterized by (3), (5), and (6), may conversely be regarded as a special case of continuous reflection in (13), in which the continuous reflection $r(z)$ becomes a sum of suitable δ-functions, one located at each discrete reflector. Thus we show that if $r(z)$ in (13) is given by

$$r(z) = \sum_{i=1}^{N} \tanh^{-1} c_i \cdot \delta(z - L_i), \tag{16}$$

where in Fig. 1 L_i is the total distance from the input of the line to the ith reflector, then the solutions to (13) at the output of the line, i.e., $W_0(L_N+)$ and $W_1(L_N+)$, are given in terms of the input conditions $W_0(0)$ and $W_1(0)$ by (3), (5), and (6).

Consider the typical kth section of line, of length l_k, followed by the kth discrete reflector, as illustrated in Fig. 1. In the line section between the $(k-1)$th and the kth reflectors $r(z) = 0$, from (16). Therefore in this region the solution to (13) has the form of (1); the forward and backward waves are uncoupled, and have the same propagation constant. We may thus write the solution between the $(k-1)$th and kth reflectors in the matrix form

$$\begin{bmatrix} W_0(L_{k-1}+) \\ W_1(L_{k-1}+) \end{bmatrix} = \begin{bmatrix} e^{+\Gamma l_k} & 0 \\ 0 & e^{-\Gamma l_k} \end{bmatrix} \cdot \begin{bmatrix} W_0(L_k-) \\ W_1(L_k-) \end{bmatrix}, \tag{17}$$

where $W(L_k-)$ indicates a wave amplitude evaluated just to the left of the kth reflector, $W(L_k+)$ just to the right.

We next evaluate the transmission matrix for the kth reflector, i.e., the kth δ-function of (16). This calculation may be performed by setting

$$r(z) = \begin{cases} \dfrac{\tanh^{-1} c_k}{\Delta}, & L_k < z < L_k + \Delta. \\ 0, & \text{otherwise.} \end{cases} \tag{18}$$

We then determine the matrix $T(\Delta)$,

$$\begin{bmatrix} W_0(L_k + \Delta) \\ W_1(L_k + \Delta) \end{bmatrix} = T(\Delta) \cdot \begin{bmatrix} W_0(L_k) \\ W_1(L_k) \end{bmatrix}. \tag{19}$$

Then as $\Delta \to 0$, $r(z) \to \tanh^{-1} c_k \cdot \delta(z - L_k)$, and $\lim_{\Delta \to 0} T(\Delta) = T(0)$ yields a matrix relating the wave amplitudes W_0 and W_1 on the two sides of the kth δ-function of $r(z)$ [see (16)]. This analysis is again similar in motivation, although different in detail, to that of Ref. 2 for a similar problem. Since $r(z)$ in (18) is constant throughout the region of interest, (13) becomes a linear differential equation with constant coefficients, and is readily solved by the usual techniques. The solution for general Δ may be written in matrix form, yielding $T(\Delta)$ of (19), as follows:

$$T(\Delta) = \frac{1}{K_+ - K_-}\begin{bmatrix} -K_- e^{\Gamma\Delta\sqrt{\ }} + K_+ e^{-\Gamma\Delta\sqrt{\ }} & e^{\Gamma\Delta\sqrt{\ }} - e^{-\Gamma\Delta\sqrt{\ }} \\ -e^{\Gamma\Delta\sqrt{\ }} + e^{-\Gamma\Delta\sqrt{\ }} & K_+ e^{\Gamma\Delta\sqrt{\ }} - K_- e^{-\Gamma\Delta\sqrt{\ }} \end{bmatrix} \tag{20}$$

$$K_\pm = -j\,\frac{1 \pm \sqrt{\ }}{\dfrac{\tanh^{-1} c_k}{\Gamma\Delta}}; \qquad K_+K_- = 1 \tag{21}$$

$$\frac{1}{K_+ - K_-} = \frac{j}{2}\,\frac{\dfrac{\tanh^{-1} c_k}{\Gamma\Delta}}{\sqrt{\ }} \tag{22}$$

$$\sqrt{\ } = \sqrt{1 + \left(\frac{\tanh^{-1} c_k}{\Gamma\Delta}\right)^2} \tag{23}$$

Taking the limit as $\Delta \to 0$, (20)–(23) yield

$$\begin{bmatrix} W_0(L_k+) \\ W_1(L_k+) \end{bmatrix} = T(0) \cdot \begin{bmatrix} W_0(L_k-) \\ W_1(L_k-) \end{bmatrix} \tag{24}$$

where

$$T(0) \equiv \lim_{\Delta \to 0} T(\Delta) = \frac{1}{\sqrt{1 - c_k^2}} \begin{bmatrix} 1 & jc_k \\ -jc_k & 1 \end{bmatrix}. \tag{25}$$

Inverting (24),

$$\begin{bmatrix} W_0(L_k-) \\ W_1(L_k-) \end{bmatrix} = T^{-1}(0) \cdot \begin{bmatrix} W_0(L_k+) \\ W_1(L_k+) \end{bmatrix} \tag{26}$$

where, from (25)

$$T^{-1}(0) = \frac{1}{\sqrt{1 - c_k^2}} \begin{bmatrix} 1 & -jc_k \\ +jc_k & 1 \end{bmatrix}. \tag{27}$$

From (17), (26), and (27) we now have

$$\begin{bmatrix} W_0(L_{k-1}+) \\ W_1(L_{k-1}+) \end{bmatrix} = X_k \cdot \begin{bmatrix} W_0(L_k+) \\ W_1(L_k+) \end{bmatrix} \tag{28}$$

where X_k is as given in (3). Equation (28) is identical to (4). Finally, the solution to (13), with $r(z)$ given by (16), is given by (3), (5), and (6).

The equivalence of (13) and (16) with (3), (5) and (6) is useful because the original matrix problem may thus be regarded as a special case of a pair of differential equations. Stability appears to be more readily studied for the more general continuous case described by the differential equations; these results may then be applied to the special discrete case of interest here.

III. SOLUTION BY SUCCESSIVE APPROXIMATIONS (PICARD'S METHOD)

We summarize the solution of (13) by successive approximation, following the same general approach as in Ref. 3 for a similar problem. First, it is convenient to make the following transformations:

$$\begin{aligned} W_0(z) &= e^{-\Gamma z} \cdot G_0(z) \\ W_1(z) &= e^{+\Gamma z} \cdot G_1(z). \end{aligned} \tag{29}$$

Substituting (29) into (13), we have

$$\begin{aligned} G'_0(z) &= jr(z)\, e^{+2\Gamma z} G_1(z) \\ G_1'(z) &= -jr(z)\, e^{-2\Gamma z} G_0(z). \end{aligned} \tag{30}$$

Assume that the device is operated as an amplifier with matched input and output. It proves convenient in the following analysis to take the input at the right-hand end of the amplifier, i.e., at $z = L_N$, where L_N is the total length, and the output at the left-hand end, i.e., $z = 0$; this is just opposite to the choice made in Ref. 1 and in Section I above [particularly in (8)]. The useful output is then the left-directed traveling wave at $z = 0$, i.e., $W_1(0)$, corresponding to an input taken to be the left-directed traveling wave at $z = L_N$, $W_1(L_N)$. Since the device is matched at both ends, $W_0(0) = 0$; $W_0(L_N) \neq 0$, since this quantity corresponds to the reflected wave at the input end (i.e., at $z = L_N$) of the amplifier.

Now assume for convenience a unit-amplitude output wave:

$$W_1(0) = 1. \tag{31}$$

As noted above, since the output is matched,

$$W_0(0) = 0. \tag{32}$$

We seek $W_1(L_N)$, the input corresponding to the output of (31); since unit output has been assumed in (31), the complex transmission gain $\mathbf{G}_T$ will be

$$\mathbf{G}_T = \frac{1}{W_1(L_N)}, \tag{33}$$

where $W_1(L_N)$ is the solution to (13) subject to the initial conditions of (31) and (32).

The transmission gain is readily stated in terms of the solutions to (30), which were obtained from (13) via the transformation of (29). Thus, consider (30) subject to the initial conditions

$$\begin{aligned} G_0(0) &= 0, \\ G_1(0) &= 1, \end{aligned} \tag{34}$$

obtained from (31) and (32) via (29). The complex transmission gain $\mathbf{G}_T$ of the amplifier is then given by

$$\mathbf{G}_T = e^{-\Gamma L_N} \cdot \frac{1}{G_1(L_N)}, \tag{35}$$

where $G_1(L_N)$ is the solution to (30) subject to the initial conditions of (34).

We now seek the solution to (30), with the initial conditions of (34), via Picard's method of successive approximations.[4,5] Assume the $(n - 1)$th approximation to the solution is available; let us denote this approximation by $G_{0(n-1)}(z)$ and $G_{1(n-1)}(z)$. Then the $(n - 1)$th approximation is substituted into the right-hand side of (30) and the right-hand side integrated to yield the nth approximation.

$$\begin{aligned} G_{0(n)}(z) &= j \int_0^z r(s)\, e^{+2\Gamma s} G_{1(n-1)}(s)\, ds. \\ G_{1(n)}(z) &= 1 - j \int_0^z r(s)\, e^{-2\Gamma s} G_{0(n-1)}(s)\, ds. \end{aligned} \tag{36}$$

We take the initial (0th) approximation as simply the initial conditions of (34):

$$\begin{aligned} G_{0(0)}(z) &= 0, \\ G_{1(0)}(z) &= 1. \end{aligned} \tag{37}$$

Writing

$$\begin{aligned} G_{0(n)}(z) - G_{0(n-1)}(z) &= g_{0(n)}(z), \\ G_{1(n)}(z) - G_{1(n-1)}(z) &= g_{1(n)}(z), \end{aligned} \tag{38}$$

we have

$$\begin{aligned} G_{0(n)}(z) &= \sum_{k=1}^{n} g_{0(k)}(z), \\ G_{1(n)}(z) &= 1 + \sum_{k=1}^{n} g_{1(k)}(z). \end{aligned} \tag{39}$$

From (36) and (38), the g's of (39) are given as follows:

$$g_{0(n)}(z) = j \int_0^z r(s)\, e^{+2\Gamma s} g_{1(n-1)}(s)\, ds, \qquad n \geqq 1. \tag{40}$$

$$g_{1(n)}(z) = -j \int_0^z r(s)\, e^{-2\Gamma s} g_{0(n-1)}(s)\, ds, \qquad n \geqq 1. \tag{41}$$

$$g_{0(0)}(z) = 0, \qquad\qquad g_{1(0)}(z) = 1. \tag{42}$$

From (40)–(42)

$$\begin{aligned} g_{0(n)}(z) &= 0, \qquad n \text{ even.} \\ g_{1(n)}(z) &= 0, \qquad n \text{ odd.} \end{aligned} \tag{43}$$

Thus only odd terms appear in the top summation of (39), and only even terms appear in the bottom summation of (39).

We next obtain bounds on the magnitudes of the terms in the series of (39), thus showing that these series converge as $n \to \infty$ for all finite z, so that the solutions to (30) subject to the initial conditions of (34) are

$$\begin{aligned} G_0(z) &= \sum_{n=0}^{\infty} g_{0(n)}(z), \\ G_1(z) &= \sum_{n=0}^{\infty} g_{1(n)}(z), \end{aligned} \tag{44}$$

with $g_{0(n)}(z)$ and $g_{1(n)}(z)$ as given by (40)–(42). The analysis is suggested by that of Ref. 3. We show that:

$$| g_{0(n)}(z) | \begin{cases} = 0, & n \text{ even.} \\ \leqq \dfrac{\left[\int_0^z | r(s) | \, ds \right]^n}{n!}, & n \text{ odd.} \end{cases} \tag{45}$$

$$| g_{1(n)}(z) | \begin{cases} \leqq e^{2\alpha z} \dfrac{\left[\int_0^z | r(s) | \, ds \right]^n}{n!}, & n \text{ even.} \\ = 0, & n \text{ odd.} \end{cases} \tag{46}$$

where from (2)

$$\Gamma = -\alpha + j\beta, \qquad \alpha = -\text{Re } \Gamma > 0. \tag{47}$$

Suppose that (46) is true for some even n. Then from (40)

$$\begin{aligned} | g_{0(n+1)}(z) | &\leqq \int_0^z | r(t) | \, e^{-2\alpha t} \, e^{+2\alpha t} \frac{\left[\int_0^t | r(s) | \, ds \right]^n}{n!} \, dt \\ &= \frac{1}{n!} \int_0^{t=z} \left[\int_0^t | r(s) | \, ds \right]^n d\left[\int_0^t | r(s) | \, ds \right] \\ &= \frac{\left[\int_0^z | r(s) | \, ds \right]^{n+1}}{(n+1)!}, \end{aligned} \tag{48}$$

in agreement with (45). Substituting this result into (41),

$$\begin{aligned} | g_{1(n+2)}(z) | &\leqq \int_0^z | r(t) | \, e^{+2\alpha t} \frac{\left[\int_0^t | r(s) | \, ds \right]^{n+1}}{(n+1)!} \, dt \\ &\leqq \frac{e^{+2\alpha z}}{(n+1)!} \int_0^{t=z} \left[\int_0^t | r(s) | \, ds \right]^{n+1} \\ &\qquad \cdot d\left[\int_0^t | r(s) | \, ds \right] \\ &= e^{+2\alpha z} \frac{\left[\int_0^z | r(s) | \, ds \right]^{n+2}}{(n+2)!}, \end{aligned} \tag{49}$$

in agreement with (46). Noting (42) and (43), the results of (45) and (46) hold for all n by induction.

The bounds of (45) and (46) guarantee the convergence of the series solutions of (44) under quite general conditions. It is readily seen that

$$\begin{aligned} | G_0(z) | &\leqq \sinh \left[\int_0^z | r(s) | \, ds \right], \\ | G_1(z) | &\leqq e^{+2\alpha z} \cosh \left[\int_0^z | r(s) | \, ds \right]. \end{aligned} \tag{50}$$

The series solutions of (44) converge for all finite z, so long as the continuous reflection coefficient is absolutely integrable,

$$\int_0^z |r(s)|\,ds < \infty. \tag{51}$$

In particular, note that $r(z)$ may contain δ-functions, as in (16), so that the above bounds may be applied directly to the discrete case of Section I.

The solutions to (30) given by (44) and (40)–(43) thus converge for all finite z in the case of interest. However these formal mathematical solutions have physical significance only when the device to which they apply is stable, i.e., does not oscillate. In the following section we use the bounds of (45) and (46) to obtain a sufficient condition guaranteeing stability in the general case.

IV. BOUNDS ON STABILITY — GENERAL CASE

Consider a general amplifier described by (13) or equivalently by (30). Assume the total length is given by L_N. We may investigate stability as indicated following (8). Replace the continuous reflection coefficient $r(z)$ by $\epsilon\cdot r(z)$, where ϵ is a numerical parameter. Let ϵ be increased from 0, and for each value of ϵ determine the maximum value of the transmission gain $|\mathbf{G}_T|$ as a function of frequency ω. From (35) the maximum value of $|\mathbf{G}_T|$ corresponds to the minimum value of $|G_1(L_N)|$. As ϵ approaches a critical value, denoted above by ϵ_c, $|\mathbf{G}_T|_{\max} \rightarrow \infty$ and $|G_1(L_N)|_{\min} \rightarrow 0$; if $\epsilon_c > 1$ the original device is stable.

From (40)–(44),

$$G_1(L_N) = 1 + \sum_{\substack{n=2\\ n \text{ even}}}^{\infty} g_{1(n)}(L_N). \tag{52}$$

Noting that $r(z)$ has been temporarily replaced by $\epsilon\cdot r(z)$, for sufficiently small ϵ a lower bound on the magnitude of $G_1(L_N)$ is given by

$$|G_1(L_N)| \geqq 1 - \sum_{\substack{n=2\\ n \text{ even}}}^{\infty} |g_{1(n)}(L_N)|\,. \tag{53}$$

Both sides of (52) and (53) are functions of frequency ω, through their dependence on the propagation constant β. Using the result of (46) in (53),

$$|G_1(L_N)| \geqq 1 - \sum_{\substack{n=2\\ n \text{ even}}}^{\infty} e^{2\alpha L_N} \frac{\left[\int_0^{L_N} |\epsilon\cdot r(s)|\,ds\right]^n}{n!}. \tag{54}$$

Since the expression on the right-hand side of (54) is independent of the propagation constant β and hence of the frequency ω, this expression is also a lower bound on $| G_1(L_N) |_{\min}$, the minimum value of $| G_1(L_N) |$ as a function of ω.

$$| G_1(L_N) |_{\min} \geqq 1 - \sum_{\substack{n=2 \\ n \text{ even}}}^{\infty} e^{2\alpha L_N} \frac{\left[\int_0^{L_N} | \epsilon \cdot r(s) | \, ds \right]^n}{n!}. \tag{55}$$

As ϵ increases from 0, the lower bound on $| G_1(L_N) |_{\min}$ given by (55) steadily decreases, and for some particular value of $\epsilon \leqq \epsilon_c$ approaches 0. Therefore if

$$\sum_{\substack{n=2 \\ n \text{ even}}}^{\infty} \frac{\left[\int_0^{L_N} | \epsilon \cdot r(s) | \, ds \right]^n}{n!} < e^{-2\alpha L_N} \tag{56}$$

stability is guaranteed. If (56) is satisfied for $\epsilon = 1$, then stability is guaranteed for the original amplifier, with reflection coefficient $r(z)$.

Consequently, a sufficient stability condition for an active transmission line with a general continuous reflection coefficient $r(z)$, described by either (13) or (30), assuming the device to be matched at both ends, is given by

$$\sum_{\substack{n=2 \\ n \text{ even}}}^{\infty} \frac{\left[\int_0^{L_N} | r(s) | \, ds \right]^n}{n!} < e^{-2\alpha L_N}. \tag{57}$$

This may be written

$$\cosh\left[\int_0^{L_N} | r(s) | \, ds \right] - 1 < e^{-2\alpha L_N} \tag{58}$$

or further

$$\sinh^2 \left[\frac{\int_0^{L_N} | r(s) | \, ds}{2} \right] < \frac{1}{2} e^{-2\alpha L_N}. \tag{59}$$

Finally, taking the square root of both sides of (59) we obtain

$$\sinh \left[\frac{\int_0^{L_N} | r(s) | \, ds}{2} \right] < \frac{e^{-\alpha L_N}}{\sqrt{2}} \tag{60}$$

or equivalently

$$\int_0^{L_N} | r(s) | \, ds < 2 \sinh^{-1} \frac{e^{-\alpha L_N}}{\sqrt{2}} \tag{61}$$

as sufficient conditions for stability for a general active transmission line with an arbitrary continuous reflection coefficient $r(z)$.

We may now apply the result of (61) to the discrete case of Section I above by making use of the results of Section II. As noted in Section II, if the continuous coupling coefficient $r(z)$ is a series of δ-functions of the form given in (16), then the solution to (13) is identical to that for the discrete case, given in (3), (5), and (6). Since the stability condition of (61) holds true in general, it may be applied to the discrete case by substituting (16) into (61), yielding

$$\sum_{i=1}^{N} \tanh^{-1} | c_i | < 2 \sinh^{-1} \frac{e^{-\alpha L_N}}{\sqrt{2}} \, . \tag{62}$$

Equation (62) is a sufficient condition for stability for a general active transmission line with arbitrary discrete reflectors, having reflection coefficients c_i located at arbitrary positions along the line. Equation (62) is the result stated in Section I as (9). This inequality is a sufficient condition for stability; if the inequality is satisfied, the device must be stable. This condition is *not* necessary for stability; many devices that violate (62) or (9) are stable.

The weaker bounds of (10) and (12) are readily obtained from the basic result of (62) or (9) by straightforward use of inequalities. From (62) or (9) we must have

$$\tanh^{-1} | c_i | < 2 \sinh^{-1} \frac{e^{-\alpha L_N}}{\sqrt{2}} \qquad i = 1, 2, \cdots N. \tag{63}$$

Since the function $y = \tanh^{-1} x$ is concave upward for $x > 0$,

$$\tanh^{-1} x < \frac{\tanh^{-1} x_m}{x_m} \cdot x, \qquad 0 < x < x_m < 1. \tag{64}$$

Therefore, from (63),

$$\tanh^{-1} | c_i | < \frac{2 \sinh^{-1} \dfrac{e^{-\alpha L_N}}{\sqrt{2}}}{\tanh \left[2 \sinh^{-1} \dfrac{e^{-\alpha L_N}}{\sqrt{2}} \right]} \cdot | c_i | \, . \tag{65}$$

Therefore if the relation

$$\sum_{i=1}^{N} | c_i | \leqq \tanh \left[2 \sinh^{-1} \frac{e^{-\alpha L_N}}{\sqrt{2}} \right] \tag{66}$$

is satisfied, then the condition of (62) must also be satisfied, so that (66) is a slightly poorer sufficient condition for stability; this result was given in (10). Finally, since the function $y = \tanh [2 \sinh^{-1} x]$ is concave downward for $x > 0$,

$$\tanh [2 \sinh^{-1} x] \geqq \frac{\tanh [2 \sinh^{-1} x_m]}{x_m} \cdot x, \qquad 0 \leqq x \leqq x_m . \tag{67}$$

As a particular instance let us choose $x_m = (1/\sqrt{20}) = 0.2236$; then (67) becomes

$$\tanh [2 \sinh^{-1} x] \geqq 1.863\, x, \qquad 0 \leqq x \leqq \frac{1}{\sqrt{20}} = 0.2236. \tag{68}$$

By using (68) to decrease the right-hand side of (66), we obtain the slightly poorer sufficient condition for stability

$$\begin{aligned} \sum_{i=1}^{N} | c_i | &\leqq 1.863 \frac{e^{-\alpha L_N}}{\sqrt{2}} \\ &= 0.932 \sqrt{2}\, e^{-\alpha L_N}, \qquad 20 \log_{10} e^{\alpha L_N} \geqq 10\, db \end{aligned} \tag{69}$$

given in (12).

V. BOUNDS ON STABILITY — SPECIAL CASE, INCLUDING REFLECTORS OF NOMINALLY EQUAL MAGNITUDE AND SPACING

The bounds on stability derived in Section IV in the general case guarantee stability for the worst possible arrangement of reflectors. Thus in many cases the sum of the magnitudes of the reflectors may far exceed the bound given by (9), (10), or (12) without causing instability.

These general bounds guarantee stability even if we have no information whatever about the distribution of reflectors. If we do have such additional information, it should be possible to make use of it to find improved bounds. As a trivial example, in the treatment of equally spaced, identical reflectors in the previous paper[1] exact stability conditions were obtained; we will see in Section VI that for this case the sum of the magnitudes of the reflectors at the boundary of instability may far exceed that given by (9), (10), or (12).

In the present section we consider a somewhat restricted special case in which the reflection coefficient is almost uniformly distributed in a certain sense. We assume that

$$\begin{gathered} R\cdot(z-f) \leqq \int_0^z |\, r(s)\,|\, ds \leqq R\cdot(z+g), \\ R > 0, \qquad f \geqq 0, \qquad g \geqq 0, \end{gathered} \tag{70}$$

where R, f, and g are constants. Equation (70) states that the indefinite integral of the absolute magnitude of the reflection coefficient is constrained to lie between two straight lines of the same slope R, separated by the horizontal distance h given by

$$h \equiv f + g, \qquad h \geqq 0. \tag{71}$$

It turns out that the final bounds of this section are better the smaller the separation h. This is to be expected, since the smaller the separation of the two straight lines given by the right- and left-hand sides of (70), the more constrained is the reflection coefficient $r(z)$.

The presence of sufficient length of perfect (i.e., reflectionless) active line at either end will needlessly increase f and hence h in (70) and (71), and hence needlessly degrade the final stability condition given below. Such a length of perfect line cannot affect the stability, but merely alters the gain of the device (assuming it is stable). Therefore for purposes of the present stability analysis sufficient lengths of perfect active line should be removed from each end so that h is minimized, and hence the best possible bound is obtained. Removal of any additional lengths of perfect active line from either end will do neither good nor harm to the final stability condition.

A few examples serve to illustrate the general nature of the restriction of (70). First suppose that $r(z)$ is equal to a (positive) constant,

$$r(z) = r_0 . \tag{72}$$

Then (70) is true with

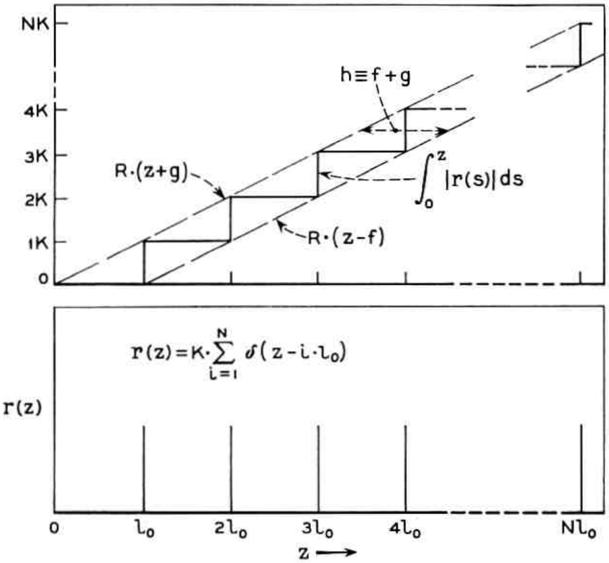
$$\begin{aligned} R &= r_0 \\ f &= 0, \qquad g = 0 \\ h &\equiv f + g = 0. \end{aligned} \tag{73}$$

The separation h [of (71)] between the straight lines of the two sides of the inequality of (70) is zero in this case. Equations (13) or (30) are readily solved exactly for the reflection coefficient of (72) by slight modification of the results of (18)–(23), in particular by first replacing

$\tanh^{-1} c_k \rightarrow r_0\Delta$ and subsequently replacing any remaining Δ's by $\Delta \rightarrow L$, where L is the total length, in these equations. From this exact solution precise stability conditions may be obtained for the case of constant (continuous) reflection coefficient; we expect the bounds of the present section to agree with this exact result when we set $f = g = 0$.

Similarly, the parameters of (73) apply to the bounds of (70) when the (continuous) reflection coefficient is a square wave of constant absolute value r_0, with arbitrary transitions between the $+r_0$ and the $-r_0$ sections.

The above two examples utilize a continuous reflection coefficient. However, our particular present interest lies in some of the discrete cases of the preceding paper.[1] First, consider the case of identical, equally-spaced reflectors of Section II, Ref. 1; the relations of (70) are illustrated for this case in Fig. 2. A less-restricted case is provided by the case of reflectors of identical magnitude but random spacing, where the fluctuation in spacing is very small compared to the average spacing, treated in Section III of Ref. 1. The relations of (70) for this case are shown in Fig. 3; the randomness in spacing has resulted in a slightly wider separa-



PARAMETERS OF EQUATION 70
$R=K/l_0 \quad f=l_0 \quad g=0$

Fig. 2 — Identical, equally spaced reflectors.

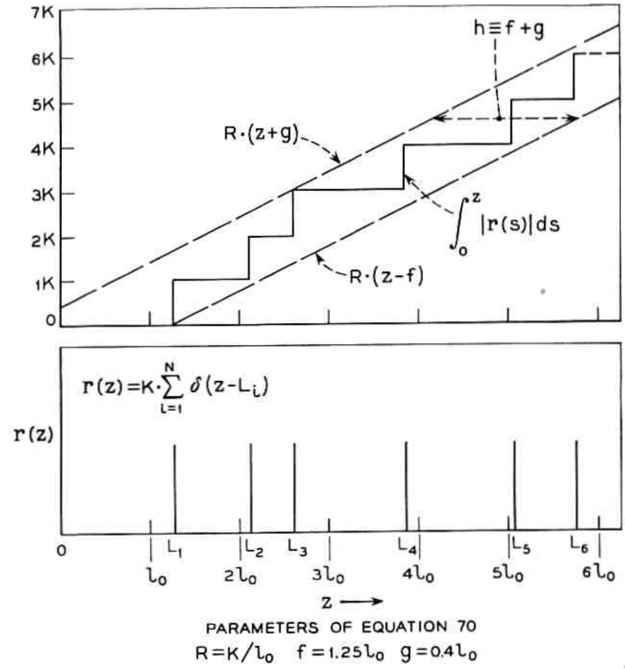


Fig. 3 — Identical, randomly spaced reflectors.

tion than in Fig. 2 between the dashed lines that enclose the staircase curve of

$$\int_0^z | r(s) | ds.$$

Since in this case the magnitudes of the reflectors are strictly constant, the "risers" of the staircase have the same size, while the "treads" vary in length. It is clear that if the magnitudes as well as the spacings of the reflectors vary slightly, both the "risers" and the "treads" of the staircase will vary slightly, but otherwise the behavior will be much the same as in Fig. 3, so that the restriction of (70) may be satisfied with small separation between the straight-line bounds.

While the discrete cases of the preceding paragraph, which have reflectors of nominally equal magnitude and spacing, are of principal interest here and supply the motivation for the analysis of the present section, discrete reflectors having quite different distributions from the above may also fall within the restriction of (70) with small separation of the bounding lines; one such case is illustrated in Fig. 4. (Note that reflectors of both signs are indicated in the lower drawing of this figure, by δ-functions with both positive and negative magnitudes.)

The above cases, which satisfy the restriction of (70), may be regarded as having the absolute magnitude of the reflection coefficient more or less constant in a certain sense, in that

$$\int_0^z | r(s) | \, ds$$

is approximately proportional to z [see (70)]. Thus we seek bounds on stability in the case of (70) that are similar to those obtained for constant reflection coefficient [see (72)].

We again use the solution by successive approximation given in Section III above. The discussion of (29)–(43) remains appropriate for our

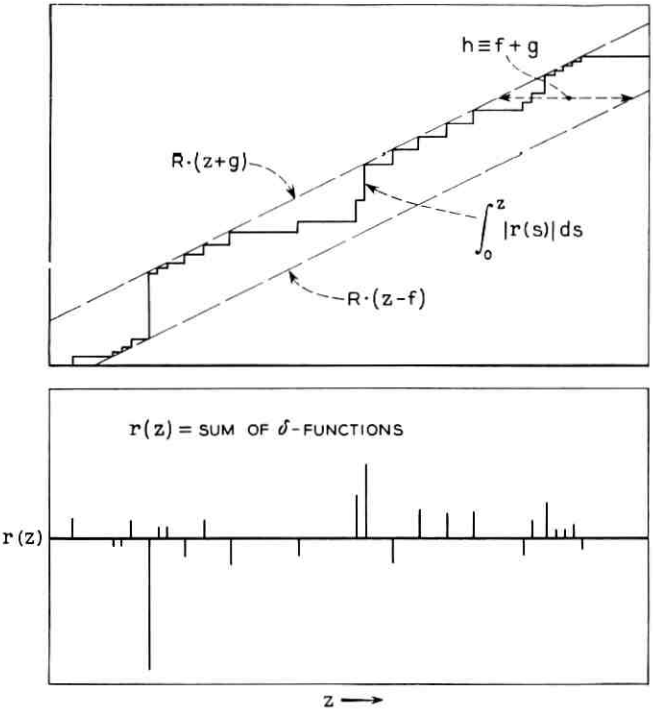


Fig. 4 — More general case satisfying the restrictions of (70).

present purposes. However, greatly improved bounds over those obtained in (44)–(51) may be obtained because of the additional restriction of (70) imposed in the present section; in contrast, the bounds of (44)–(47) of Section III hold true in general, and specifically when the restriction of (70) is not satisfied.

Consider the series solutions of (44). From (42)

$$g_{1(0)}(z) = 1, \qquad g_{0(0)}(z) = 0. \tag{74}$$

Note also (43). We show that:

$$\begin{aligned} | g_{1(n)}(z) | &< R^2 \left(\frac{1}{2\alpha} + h\right)^2 e^{2\alpha z} \\ &\quad \cdot \frac{\left\{R^2 \left(\frac{1}{2\alpha} + h\right)\left[z + \left(\frac{n}{2} - 1\right) h\right]\right\}^{(n/2)-1}}{\left(\frac{n}{2} - 1\right)!} \qquad n \text{ even}, n \geqq 2. \\ | g_{1(n)}(z) | &= 0, \qquad n \text{ odd}. \end{aligned} \tag{75}$$

$$\begin{aligned} | g_{0(n)}(z) | &= 0, \qquad n \text{ even}. \\ | g_{0(n)}(z) | &< R \left(\frac{1}{2\alpha} + h\right) \\ &\quad \cdot \frac{\left\{R^2 \left(\frac{1}{2\alpha} + h\right)\left[z + \left(\frac{n-1}{2}\right) h\right]\right\}^{(n-1)/2}}{\left(\frac{n-1}{2}\right)!}, \\ &\qquad n \text{ odd}. \end{aligned} \tag{76}$$

In (75) and (76), R and h are the parameters of (70) and (71).

First, from (40), (42) or (74), and (47),

$$\begin{aligned} | g_{0(1)}(z) | &\leqq \int_0^z e^{-2\alpha s} \, | r(s) | \, ds = \int_0^z e^{-2\alpha s} \, d\left[\int_0^s | r(t) | \, dt\right] \\ &= e^{-2\alpha z} \int_0^z | r(t) | \, dt + 2\alpha \int_0^z e^{-2\alpha s} \left[\int_0^s | r(t) | \, dt\right] ds, \end{aligned} \tag{77}$$

where we have made use of integration by parts. Using (70) in (77),

$$\begin{aligned} | g_{0(1)}(z) | &\leqq e^{-2\alpha z} \cdot R(z + g) + 2\alpha R \int_0^z e^{-2\alpha s} \cdot (s + g)\, ds \\ &= e^{-2\alpha z} \cdot R(z + g) \\ &\qquad + R \left[\frac{1 - e^{-2\alpha z}}{2\alpha} - ze^{-2\alpha z} + g(1 - e^{-2\alpha z}) \right] \\ &= \frac{R}{2\alpha}(1 - e^{-2\alpha z}) + Rg < \frac{R}{2\alpha} + R(f + g), \end{aligned} \tag{78}$$

where in the final step we have used the fact that $f \geqq 0$. Finally, substituting the definition of h from (71) into (78),

$$| g_{0(1)}(z) | < R\left(\frac{1}{2\alpha} + h\right). \tag{79}$$

Equation (79) agrees with (76) for $n = 1$.

Next, from (41), (47), and (79),

$$\begin{aligned} | g_{1(2)}(z) | &< R\left(\frac{1}{2\alpha} + h\right) \int_0^z e^{+2\alpha s} \,| r(s) |\, ds \\ &= R\left(\frac{1}{2\alpha} + h\right) \int_0^z e^{+2\alpha s}\, d\left[\int_0^s | r(t) |\, dt\right] \\ &= R\left(\frac{1}{2\alpha} + h\right) e^{+2\alpha z} \int_0^z | r(t) |\, dt \\ &\qquad - R\left(\frac{1}{2\alpha} + h\right) 2\alpha \int_0^z e^{+2\alpha s} \left[\int_0^s | r(t) |\, dt\right] ds. \end{aligned} \tag{80}$$

Using (70), (80) becomes

$$\begin{aligned} | g_{1(2)}(z) | &< R^2\left(\frac{1}{2\alpha} + h\right) e^{2\alpha z} \cdot (z + g) \\ &\qquad - R^2\left(\frac{1}{2\alpha} + h\right) 2\alpha \int_0^z e^{2\alpha s}(s - f)\, ds \\ &= R^2\left(\frac{1}{2\alpha} + h\right) e^{2\alpha z} \cdot (z + g) \\ &\qquad - R^2\left(\frac{1}{2\alpha} + h\right) \left[\frac{1 - e^{2\alpha z}}{2\alpha} + ze^{2\alpha z} + f(1 - e^{2\alpha z})\right] \\ &< R^2\left(\frac{1}{2\alpha} + h\right) e^{2\alpha z} \left[\frac{1}{2\alpha} + f + g\right]. \end{aligned} \tag{81}$$

Finally from (71), (81) becomes

$$| g_{1(2)}(z) | < R^2 \left(\frac{1}{2\alpha} + h \right)^2 e^{2\alpha z}, \tag{82}$$

which agrees with (75) for $n = 2$.

We now establish the bounds of (75) and (76) by induction. Suppose that (75) is true for some even $n \geqq 2$. Then from (40) and (47),

$$| g_{0(n+1)}(z) | < \frac{R^n \left(\frac{1}{2\alpha} + h \right)^{(n/2)+1}}{\left(\frac{n}{2} - 1 \right)!} I \tag{83}$$

where

$$I \equiv \int_0^z \left[s + \left(\frac{n}{2} - 1 \right) h \right]^{(n/2)-1} d \left[\int_0^s | r(t) | \, dt \right]. \tag{84}$$

Integrating (84) by parts,

$$\begin{aligned} I = & \left[z + \left(\frac{n}{2} - 1 \right) h \right]^{(n/2)-1} \left[\int_0^z | r(t) | \, dt \right] \\ & - \left(\frac{n}{2} - 1 \right) \int_0^z \left[s + \left(\frac{n}{2} - 1 \right) h \right]^{(n/2)-2} \\ & \cdot \left[\int_0^s | r(t) | \, dt \right] ds. \end{aligned} \tag{85}$$

Using (70) and (71), we have from (85)

$$\begin{aligned} I \leqq & \left[z + \left(\frac{n}{2} - 1 \right) h \right]^{(n/2)-1} R(z + g) \\ & - R \left(\frac{n}{2} - 1 \right) \int_0^z \left[s + \left(\frac{n}{2} - 1 \right) h \right]^{(n/2)-2} (s - f) \, ds \\ = & \left[z + \left(\frac{n}{2} - 1 \right) h \right]^{(n/2)-1} R(z + g) \\ & - R \int_0^z (s - f) \, d \left[s + \left(\frac{n}{2} - 1 \right) h \right]^{(n/2)-1} \end{aligned}$$

$$\begin{aligned}
&= \left[z + \left(\frac{n}{2} - 1\right)h\right]^{(n/2)-1} R(z + g) \\
&\qquad - R(z - f)\left[z + \left(\frac{n}{2} - 1\right)h\right]^{(n/2)-1} - Rf\left[\left(\frac{n}{2} - 1\right)h\right]^{(n/2)-1} \\
&\qquad + R\int_0^z \left[s + \left(\frac{n}{2} - 1\right)h\right]^{(n/2)-1} ds \\
&= Rh\left[z + \left(\frac{n}{2} - 1\right)h\right]^{(n/2)-1} - Rf\left[\left(\frac{n}{2} - 1\right)h\right]^{(n/2)-1} \\
&\qquad + \frac{R}{\left(\frac{n}{2}\right)}\left[z + \left(\frac{n}{2} - 1\right)h\right]^{n/2} - \frac{R}{\left(\frac{n}{2}\right)}\left[\left(\frac{n}{2} - 1\right)h\right]^{n/2} \\
&\leqq \frac{R}{\left(\frac{n}{2}\right)}\left\{\left[z + \left(\frac{n}{2} - 1\right)h\right]^{n/2} + \frac{n}{2}h\left[z + \left(\frac{n}{2} - 1\right)h\right]^{(n/2)-1}\right\},
\end{aligned} \tag{86}$$

where the last step follows from the preceding one because $n \geqq 2$ [from (75)], $f \geqq 0$ [from (70)], and $h \geqq 0$ [from (71)]. Using the inequality

$$x^k + \epsilon x^{k-1} < [x + (\epsilon/k)]^k, \qquad x \geqq 0 \quad \text{and} \quad \epsilon > 0, \tag{87}$$

(86) yields

$$I < \frac{R}{\left(\frac{n}{2}\right)}\left[z + \left(\frac{n}{2} - 1\right)h + h\right]^{n/2} = \frac{R}{\left(\frac{n}{2}\right)}[z + nh]^{n/2}. \tag{88}$$

Substituting (88) into (83),

$$| g_{0(n+1)}(z) | < R\left(\frac{1}{2\alpha} + h\right)\frac{\left[R^2\left(\frac{1}{2\alpha} + h\right)(z + nh)\right]^{n/2}}{\left(\frac{n}{2}\right)!}. \tag{89}$$

Recalling that n is some even integer $\geqq 2$ in (89), (89) agrees with (76).

Next, from (41) and (47), using the result of (89)

$$| g_{1(n+2)}(z) | < \frac{R^{n+1}\left(\frac{1}{2\alpha} + h\right)^{(n/2)+1}}{\left(\frac{n}{2}\right)!} J \tag{90}$$

where

$$J \equiv \int_0^z e^{+2\alpha s}(s + nh)^{n/2}\, d\left[\int_0^s | r(t) |\, dt\right]. \tag{91}$$

Integrating (91) by parts,

$$\begin{aligned} J = {} & e^{2\alpha z}(z + nh)^{n/2}\left[\int_0^z | r(t) |\, dt\right] \\ & - 2\alpha \int_0^z e^{+2\alpha s}(s + nh)^{n/2}\left[\int_0^s | r(t) |\, dt\right] ds \\ & - \frac{n}{2}\int_0^z e^{+2\alpha s}(s + nh)^{(n/2)-1}\left[\int_0^s | r(t) |\, dt\right] ds. \end{aligned} \tag{92}$$

Using (70) and (71), we have from (92)

$$\begin{aligned} J & \leqq e^{2\alpha z}(z + nh)^{n/2}R(z + g) \\ & \qquad - R2\alpha \int_0^z e^{+2\alpha s}(s + nh)^{n/2}\,(s - f)\, ds \\ & \qquad - R\frac{n}{|2}\int_0^z e^{+2\alpha s}(s + nh)^{(n/2)-1}(s - f)\, ds \\ & = e^{2\alpha z}(z + nh)^{n/2}R(z + g) \\ & \qquad - R\int_0^z (s - f)\, d\,[e^{+2\alpha s}(s + nh)^{n/2}] \\ & = e^{2\alpha z}(z + nh)^{n/2}R(z + g) - R(z - f)e^{2\alpha z}(z + nh)^{n/2} \\ & \qquad - Rf(nh)^{n/2} + R\int_0^z e^{+2\alpha s}(s + nh)^{n/2}\, ds \\ & = Rhe^{2\alpha z}(z + nh)^{n/2} - Rf(nh)^{n/2} \\ & \qquad + \frac{R}{2\alpha}\int_0^z (s + nh)^{n/2}\, d\,(e^{2\alpha s}) \\ & = Rhe^{2\alpha z}(z + nh)^{n/2} - Rf(nh)^{n/2} + \frac{R}{2\alpha}\, e^{2\alpha z}(z + nh)^{n/2} \\ & \qquad - \frac{R}{2\alpha}\,(nh)^{n/2} - \frac{R}{2\alpha}\,\frac{n}{2}\int_0^z e^{2\alpha s}(s + nh)^{(n/2)-1}\, ds. \end{aligned} \tag{93}$$

From (71), $h \geqq 0$, so that (93) yields

$$J < R\left(\frac{1}{2\alpha} + h\right) e^{2\alpha z}(z + nh)^{n/2}. \tag{94}$$

Substituting (94) into (90),

$$| g_{1(n+2)}(z) | < R^2 \left(\frac{1}{2\alpha} + h \right)^2 e^{2\alpha z} \frac{\left[R^2 \left(\frac{1}{2\alpha} + h \right) (z + nh) \right]^{n/2}}{\left(\frac{n}{2} \right)!}. \quad (95)$$

Recalling that n is some even integer $\geqq 2$ in (95), (95) agrees with (75). Noting (79) and (82), the results of (75) and (76) hold for all n by induction.

We now use the results of (75) together with (74) to obtain bounds on stability for those cases where the reflection coefficient $r(z)$ is restricted as in (70). This analysis is almost identical to that of Section IV, (52)–(57), for the general case, modified by replacing the relation of (46) by that of (75). Thus, making the substitution

$$\frac{\left[\int_0^z | r(s) | \, ds \right]^n}{n!} \rightarrow R^2 \left(\frac{1}{2\alpha} + h \right)^2 \cdot \frac{\left\{ R^2 \left(\frac{1}{2\alpha} + h \right) \left[z + \left(\frac{n}{2} - 1 \right) h \right] \right\}^{(n/2)-1}}{\left(\frac{n}{2} - 1 \right)!} \quad (96)$$

throughout (54)–(57), we obtain, corresponding to (57), the following sufficient condition for stability in the present case, after a minor modification of the summation index:

$$R^2 \left(\frac{1}{2\alpha} + h \right)^2 \sum_{m=0}^{\infty} \frac{\left[R^2 \left(\frac{1}{2\alpha} + h \right) (L_N + mh) \right]^m}{m!} < e^{-2\alpha L_N}. \quad (97)$$

L_N is the total length of the device. The summation of (97) is found in closed form by the analysis given in the Appendix. Using the final result of the Appendix (137), the final results of this section may be summarized as follows:

If the reflection coefficient $r(z)$ (continuous, discrete, or a combination of both) satisfies the condition

$$R \cdot (z - f) \leqq \int_0^z | r(s) | \, ds \leqq R \cdot (z + g); \qquad R > 0, f \geqq 0, g \geqq 0 \quad (98a)$$

$$h \equiv f + g; \qquad h \geqq 0.$$

then a sufficient condition for stability of the active line (with reflection) is

$$\frac{\delta\left(1+\dfrac{1}{2\alpha h}\right)}{1-\delta r_1} < \exp\left[-2\alpha L_N\left(1+\frac{\delta r_1}{2\alpha h}\right)\right] \tag{98b}$$

where

$$\delta \equiv R^2\left(\frac{1}{2\alpha}+h\right)h < \frac{1}{e} \tag{98c}$$

and r_1 is given by

$$r_1 = e^{\delta r_1}, \qquad r_1 < e. \tag{98d}$$

The results of (98) are illustrated in Fig. 5, which shows the maximum value of R for which stability is guaranteed by (98) versus the nominal total gain $20 \log_{10} e^{\alpha L_N}$, with $20 \log_{10} e^{\alpha h}$ as a parameter.

A greatly simplified but slightly poorer version of the stability condition of (98) may be obtained in the high-gain case. As one example, suppose the one-way gain of the active line exceeds 10 db,

$$e^{2\alpha L_N} \geqq 10, \qquad 8.686\ \alpha L_N \geqq 10 \text{ db}, \qquad \alpha L_N \geqq 1.151. \tag{99}$$

If δ satisfies the sufficient stability condition of (98b), it must also satisfy the weaker inequality

$$\delta < \frac{2\alpha h}{1+2\alpha h} e^{-2\alpha L_N}. \tag{100}$$

Substituting (99) into (100),

$$\delta < 0.1. \tag{101}$$

From (98d), r_1 is a monotonic increasing function of δ. Therefore

$$r_1 < 1.118. \tag{102}$$

Further, since from (98d)

$$\delta r_1 = \ln r_1, \tag{103}$$

δr_1 is a monotonic increasing function of r_1, so that

$$\delta r_1 < 0.1118. \tag{104}$$

Now writing out the right-hand side of (98b),

$$\exp\left[-2\alpha L_N\left(1+\frac{\delta r_1}{2\alpha h}\right)\right] = \exp\left(-2\alpha L_N\right)\exp\left(-\frac{L_N}{h}\,\delta r_1\right), \tag{105}$$

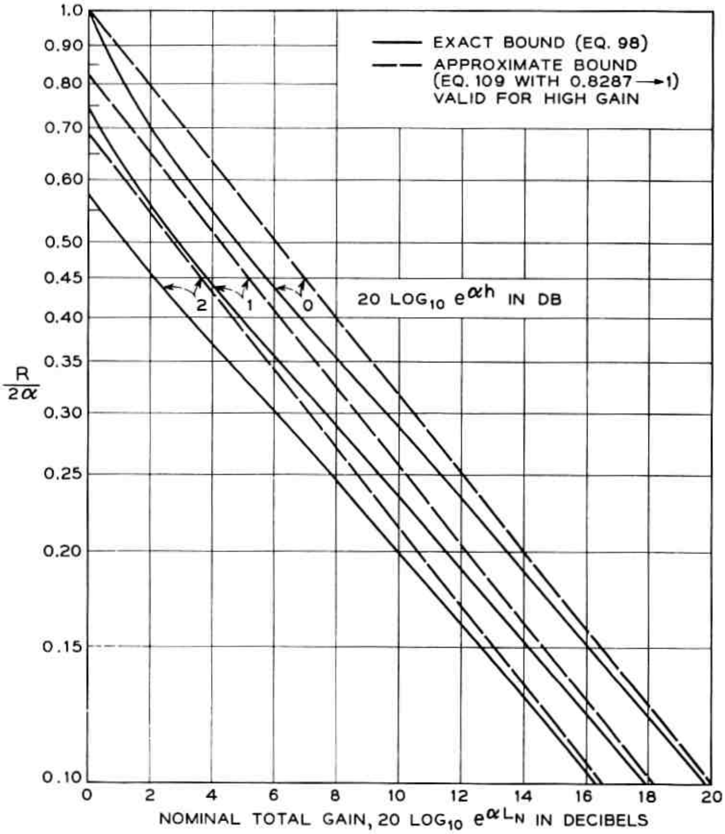


Fig. 5 — Exact and approximate bounds on R for which stability is guaranteed.

we investigate the exponent of the second factor on the right-hand side of (105). From (100),

$$\frac{L_N}{h}\,\delta r_1 < \frac{2\alpha L_N}{1 + 2\alpha h}\,e^{-2\alpha L_N}\cdot r_1 < 2\alpha L_N\,e^{-2\alpha L_N}\cdot r_1\,. \qquad (106)$$

The right-hand side of (106) is a monotonic decreasing function of $2\alpha L_N$ for $2\alpha L_N > 1$. Therefore, substituting from (99) and (102), (106) yields

$$\frac{L_N}{h}\,\delta r_1 < 0.2574. \qquad (107)$$

$$\exp\left[-\frac{L_N}{h}\,\delta r_1\right] > 0.7731. \qquad (108)$$

Finally, using (104) and (108) in (98b), we obtain the following sufficient condition for stability, subject to (98a);

$$R < 0.8287 \frac{2\alpha}{1 + 2\alpha h} e^{-\alpha L_N}; \qquad 8.686\alpha L_N \geqq 10 \; db. \tag{109}$$

The stability condition of (109) is slightly poorer than the stability condition of (98b), (98c), and (98d), from which it was derived. As the lower bound on the gain of the active line increases beyond 10 db and approaches ∞, the numerical factor 0.8287 in (109) increases and approaches 1. Equation (109) or a similar result is useful in illustrating the general behavior; however calculations using the basic result of (98) are straightforward. The result of (109), with the numerical factor 0.8287 → 1, is also shown as the dashed curves of Fig. 5, illustrating the way in which this approximate stability condition approaches the exact result of (98) in the high-gain case.

VI. EXAMPLES AND DISCUSSION

Consider first an active line with two discrete reflectors of equal magnitude c at the ends of the line, $z = 0$ and $z = \mathrm{L}_2$. c is of course real; for convenience we assume $c > 0$. In this simple case the exact stability condition is readily found, and may be compared with the two bounds derived above. From (8) of Section I, the transmission gain of this device in the stable region is

$$\mathbf{G}_T = \frac{1}{x_{11}}, \tag{110}$$

where from (1)–(7)

$$x_{11} = e^{\Gamma L_2}(1 + c^2 e^{-2\Gamma L_2}). \tag{111}$$

The condition for stability is readily found as described following (8) [this procedure is similar to that used in Section IV, (52)–(57), and Section V, (96)–(97), in obtaining bounds on stability]. Replacing c by ϵc, where ϵ is a numerical parameter greater than 0, and using (2),

$$x_{11} = e^{\Gamma L_2}[1 + (\epsilon c)^2 e^{+2\alpha L_2} e^{-j2\beta L_2}]. \tag{112}$$

For small enough ϵ the minimum value of x_{11}, and hence the maximum value of gain $\mathbf{G}_T$ of (110), occurs at

$$2\beta L_2 = \pm\pi, \pm 3\pi, \cdots . \tag{113}$$

Hence

$$|x_{11}|_{\min} = e^{-2\alpha L_2}[1 - (\epsilon c)^2 e^{+2\alpha L_2}]. \tag{114}$$

As ϵ increases from zero, instability will take place at a value of ϵ for which

$$\begin{aligned} |x_{11}|_{\min} &= 0, \\ (\epsilon c)^2 e^{2\alpha L_2} &= 1. \end{aligned} \tag{115}$$

Hence the original device (with $\epsilon = 1$) will be stable if

$$c < e^{-\alpha L_2}. \tag{116}$$

Equation (116) is an exact condition for stability for the active line described above, with two equal reflectors at the ends. We now compare this exact result with the bounds described above.

Consider first the bound of (9) or (62). This result is a sufficient condition for stability for any arbitrary distribution of discrete reflectors, and so must apply to the special case above. Setting $N = 2$, $c_1 = c_2 = c$, this general bound guarantees stability if

$$\tanh^{-1} c < \sinh^{-1} \frac{e^{-\alpha L_2}}{\sqrt{2}}. \tag{117}$$

Equation (117) yields

$$c < \frac{1}{\sqrt{2}} \frac{e^{-\alpha L_2}}{\sqrt{1 + \frac{1}{2} e^{-2\alpha L_2}}} \tag{118}$$

as a sufficient condition for stability for an active device with two equal reflectors of magnitude c at the ends. Comparing the bound of (118) with the exact stability condition of (116), we see that the general bound of (9) or (62) is conservative in the present special case; i.e., the device with two equal reflectors at the ends remains stable for the reflector magnitude c larger than that guaranteed by the general bound of (9) or (62) by a numerical factor that varies from $\sqrt{3}$ to $\sqrt{2}$ as the gain αL_2 varies from 0 to ∞. Therefore the general bound on stability given in (9) or (62) cannot be improved by a factor greater than $\sqrt{2}$ [i.e., this factor to multiply the right-hand side of (9) or (62)]; of course it may be that no improvement at all is possible, and that some distribution of reflectors can be found for which (9) is satisfied as an equality at the boundary of instability.

Next, consider the bound of Section V, (98), applied to the above special case, i.e., two discrete reflectors of identical magnitude c at the ends of the active line. In (98) we set $h = L_2$, $R = (\tanh^{-1} c)/L_2$, to yield the following (precise) bound on stability:

$$\frac{\delta r_1}{1 - \delta r_1} < \frac{2\alpha L_2}{1 + 2\alpha L_2} e^{-2\alpha L_2} \tag{119a}$$

where

$$\delta \equiv (\tanh^{-1} c)^2 \cdot \frac{1 + 2\alpha L_2}{2\alpha L_2} \tag{119b}$$

and r_1 is given by

$$r_1 = e^{\delta r_1}, \qquad r_1 < e. \tag{119c}$$

The bound on c for stability is readily determined numerically from (119) as a function of αL_2. However, when the one-way gain of the active line is large, $\alpha L_2 \gg 1$, the bound of (98) takes on the form of (109), with the numerical factor 0.8287 → 1 since $\alpha L_2 \gg 1$ (i.e., the gain is taken to be very large, not simply greater than 10 db). Thus the approximate bound on stability in the present case becomes

$$\tanh^{-1} c \lessapprox \frac{2\alpha L_2}{1 + 2\alpha L_2} e^{-\alpha L_2}; \qquad \alpha L_2 \gg 1. \tag{120}$$

The symbol $\lessapprox$ indicates that the relation of (120) is not a precise bound, but merely gives a good numerical approximation to the precise bound if αL_2 is large enough. Comparison of the (imprecise) bound of (120) with the exact stability condition of (116) shows that in the high-gain case, $\alpha L_2 \gg 1$, where $c \ll 1$, the specialized bound of Section V, (98), yields bounds on the magnitude of the reflection c in the present special case (two equal reflectors at the ends of the active line) that approach those of the exact condition for stability. Consequently the bounds of (98) cannot be further improved (in their present form).

The case of N identical, equally spaced reflectors was studied in Section II of Ref. 1, where simple expressions for stability were found in the high-gain case. If the total gain is large and the gain per section small, comparison of (109) (with the factor 0.8287 → 1) and (98a) with (43) of Ref. 1 shows again that the bound on stability of (98) cannot be further improved. It is of interest to see how close the bounds of (98) come to the exact value corresponding to instability in a few cases of interest. For this purpose we consider examples (*i*), (*ii*), and (*iii*) of

Section II, Ref. 1. In (98) we set

$$h = l, \qquad R = \frac{\tanh^{-1} c}{l}, \tag{121}$$

and compute upper bounds on $| c |$ that guarantee stability. It is also of interest to compare the general bound of (9) or (62) for this case. Table I summarizes these results. The bounds of (98) are quite good when the total gain is high, $\alpha L_N \gg 1$, and when the gain corresponding to the distance l is small, $\alpha l \ll 1$; for these conditions the stability condition of (98) gives much better results than the more general stability condition of (9), because in the former we have made use of additional information regarding the distribution of reflectors.

Table I — Identical, Equally Spaced Reflectors

N = number of reflectors
Gain (db) $= 20 \log_{10} e^{N\alpha l} \equiv 20 \log_{10} e^{\alpha L_N}$ = one-way gain of active line in db
$| c |_{\max}$ = maximum value of $| c |$ for stability, as determined in Section II, Ref. 1
Bound on $| c |$ — (98) = maximum value of $| c |$ for which stability is guaranteed by (98)
Bound on $| c |$ — (9) or (62) = maximum value of $| c |$ for which stability is guaranteed by (9) or (62).

Case (Sec. II, Ref. 1)	N	Gain, db	$\| c \|_{\max}$ (Sec. II, Ref. 1)	Bound on $\| c \|$ (98)	Bound on $\| c \|$ (9) or (62)
(*i*)	30	30	0.00860	0.00590	0.00149
(*ii*)	300	30	0.000860	0.000710	0.000149
(*iii*)	50	5	0.0650†	0.01105	0.0130

† Note that for this case in Ref. 1 the high-gain approximation given there was inappropriate, so that this result was obtained by use of a computer.

Finally, we consider the application of the above stability conditions to some of the problems involving random reflectors studied in Ref. 1. The stability of the various deterministic cases discussed above in the present section has been treated exactly here or in Ref. 1 without using the new results of the present paper; these cases have been discussed in the present section both to show that any possible improvement in these general stability conditions must be quite small, and to provide partial confirmation of these results. However, the application of (9) and (98) to cases involving random reflectors provides the principal motivation for the present analysis, since no other information whatever is available regarding stability in these cases.

Let us consider the example of the first part of Section IV, Ref. 1, in which the average normalized loss and the rms loss fluctuation were determined for an amplifier with reflections having identical magnitude but random spacing. The following parameters were chosen for this illustration:

$$\begin{aligned}
l_k &= \text{spacing between } (k-1)\text{th and } k\text{th reflectors [(3) and Fig. 1]}\\
l_0 &= \langle l_k \rangle, \text{ average value of } l_k, \text{ independent of } k\\
c_k &= \text{magnitude of } k\text{th reflection coefficient [(3) and Fig. 1]}\\
c_k &= c_0\,; \text{ all reflectors identical, } c_0 > 0\\
N &= 30, \text{ number of sections}\\
20 \log_{10} e^{N\alpha l_0} &= 30 \text{ db, nominal total gain}\\
20 \log_{10} e^{\alpha l_0} &= 1 \text{ db, nominal gain per section.}
\end{aligned} \tag{122}$$

The following assumptions were made in these calculations of Ref. 1:

(a) l_k is always a large number of wavelengths;

$$\beta l_k \gg 2\pi, \qquad \beta l_0 \gg 2\pi. \tag{123}$$

(b) The distribution of the l_k about their mean l_0 is very narrow with respect to the mean, but wide compared to $2\pi/\beta$; further, this distribution is smooth and symmetrical about l_0.

The probability density for l_k did not have to be further specified for the calculation of average loss and rms loss fluctuation in Ref. 1. (Note however that in the calculations of Ref. 1 for the covariance of the loss, the specific form of the probability density for l_k must be known, and was assumed to be Gaussian in Ref. 1.) The average loss and the rms loss fluctuation for the amplifier of (122) were given in Fig. 9 of Ref. 1 versus c_0, the magnitude of the reflections. These curves were shown dotted for $c_0 > 0.00860$, because it was known that instability is possible in this range, in particular for $l_k = l_0$, i.e., equally spaced reflectors [see Section II, Ref. 1 and case (*i*), Table I]. However it was noted that this was only a symbolic reminder of the unsolved question of stability; these results are valid for small enough c_0, but how small was not known from the results of Ref. 1.

We illustrate the utility of the results of the present paper by applying them to this problem; these results provide useful information concerning stability in this case, and of course in many similar problems. For convenience we make one further assumption in addition to those mentioned following (122):

(c) The distribution of l_k about its mean l_0 is strictly bounded; in particular

$$| l_k - l_0 | \leqq \nu l_0\,; \tag{124}$$

further, we assume for convenience that

$$\nu < 1. \tag{125}$$

ν is in (124) the upper bound on the fractional deviation in spacing from its average value; the restriction of (125) requires that $l_k \geqq 0$, and so prevents the order of the reflectors from being altered. In practical cases we will be interested in small values of ν,

$$\nu \ll 1. \tag{126}$$

We determine upper bounds on the reflector magnitude c_0 that guarantee stability, as a function of ν, the maximum fractional deviation in spacing between reflectors. For $\nu = 0$ the reflectors are equally spaced; Ref. 1 or Table I shows that stability is guaranteed if

$$c_0 < 0.00860, \qquad \nu = 0. \tag{127}$$

Next, the bound of (9) guarantees stability independently of the particular distribution of reflectors. Since however the total length may vary somewhat, we must in (9) set

$$L_N \equiv L_{30} = 30l_0(1 + \nu), \tag{128}$$

yielding

$$c_0 < 0.00149(0.03162)^{\nu} \tag{129}$$

as a sufficient stability condition.

Finally, we apply the bound of (98) to this example. We set

$$R = \frac{\tanh^{-1} c_0}{l_0}, \tag{130}$$

$$h = (1 + 60\nu)l_0 \tag{131}$$

and make use of (128) in (98) to obtain a sufficient stability condition.

The sufficient stability conditions of (127), (129), and (98) are plotted in Fig. 6; the result of (129) is identified as originating from (9), and that of (127) from Section II of Ref. 1. The curves of Fig. 6 have been plotted out to fractional spacing variations ν of 10 per cent; over this region the stability condition of (98) is superior to that of (9). However the bound of (9) [i.e., (129)] will be superior to that of (98) for large enough ν. Note that the factor $(0.03162)^{\nu}$ in (129) arises from the fact that the total length and hence the total gain is subject to statistical fluctuation [a similar factor occurs in using (98) for the problem]; in the range of probable interest, i.e., for very small fractional spacing fluctuations ν, this numerical factor will be close to 1. The fact that the limit of the bound of (98) as $\nu \to 0$ is substantially below the maximum value of c_0 given by (127) is due to the fact that the nominal gain per section in the example of (122) is 1 db, which is not too small;

as the gain per section decreases these two quantities will approach each other, as indicated above.

These results, plotted on Fig. 6, show that the range of c_0 over which the calculations of Section IV of Ref. 1 are guaranteed to be valid. If the maximum fractional variation in the spacing between reflectors is very small, then the results plotted on Fig. 9 of Ref. 1 are valid for c_0 up to approximately 0.00590.

The stability conditions of (9) and (98) may be applied to a variety of similar problems. In the above example we have found the maximum value of c_0 for which stability is guaranteed, i.e., for which the probability of oscillation is zero, as a function of the maximum departure of the spacing between reflectors from its average value. The results of (9) and (98) may also be used to determine an upper bound on the probability of oscillation in similar problems where no absolute guarantee of

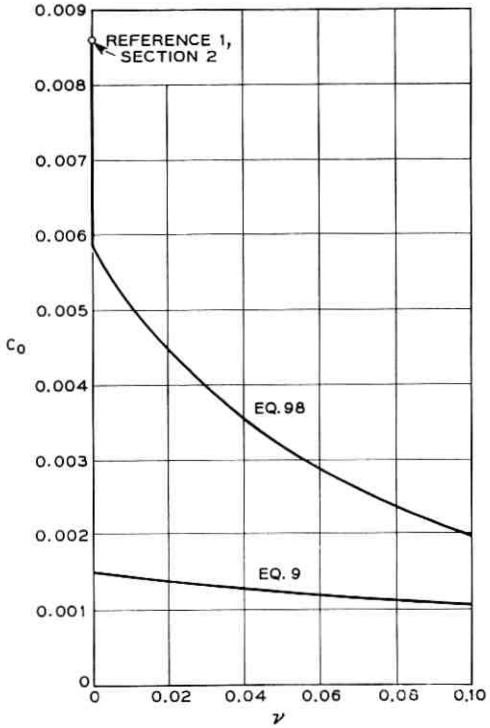


Fig. 6 — Bounds on magnitude of coupling coefficient to guarantee stability for amplifier of (122), with reflectors of identical magnitude and nominally equal spacing.

stability can be given, e.g., perhaps in cases where the probability distribution for the spacing deviations is not strictly bounded.

The main emphasis of the present paper has been on the discrete case; the continuous case was introduced only as an intermediate step leading to the desired results. However, it is clear that related problems with continuous reflection may be studied for stability using the general results derived above.

Finally, the present calculations have assumed for definiteness a rather special model; i.e., the forward and backward gains have been assumed equal and a particular form has been taken for the matrix of the discrete reflectors. These assumptions are not essential to the analysis; similar results can be derived for many related cases of interest, such as systems using isolators to partially attenuate the backward waves, etc.

VII. ACKNOWLEDGMENT

The author would like to thank S. O. Rice for first summing the series of (97), and the unknown referee for the shorter analysis of the Appendix for this series. He would also like to thank Mrs. C. L. Beattie and Miss E. Richardson for programming various numerical calculations.

APPENDIX

Summation of the Series $S = \sum_{n=0}^{\infty} \frac{(z + \delta n)^n}{n!}$

The summation of (97) was initially performed by a method suggested by S. O. Rice, employing contour integration; this method is straightforward but lengthy. A much shorter analysis presented by the unknown referee is given here. It has been shown that[6]

$$e^{ax} = 1 + \sum_{n=1}^{\infty} \frac{a(a - nb)^{n-1}}{n!} y^n \tag{132}$$

where

$$y = xe^{bx} \quad \text{and} \quad |yb| < (1/e). \tag{133}$$

Differentiate (132) with respect to y and then set $y = 1$ to obtain

$$\frac{e^{(a-b)x}}{1 + bx} = \sum_{n=0}^{\infty} \frac{[(a - b) - nb]^n}{n!} \tag{134}$$

where

$$x = e^{-bx} \quad \text{and} \quad |b| < (1/e). \tag{135}$$

Finally, set

$$a = z - \delta, \qquad b = -\delta, \qquad x = r_1 \tag{136}$$

to obtain

$$\sum_{n=0}^{\infty} \frac{(z + \delta n)^n}{n!} = \frac{e^{r_1 z}}{1 - \delta r_1}, \qquad 0 \leqq \delta < \frac{1}{e}$$

where r_1 is given by (137)

$$r_1 = e^{\delta r_1}, \qquad r_1 < e.$$

REFERENCES

1. Rowe, H. E., Imperfections in Active Transmission Lines, B.S.T.J., this issue, p. 261.
2. Rowe, H. E., and Warters, W. D., Transmission in Multimode Waveguide with Random Imperfections, B.S.T.J., **41**, May, 1962, pp. 1031–1170. See particularly Sections 2.3.2 and 2.3.3.
3. Rowe, H. E., Approximate Solutions for the Coupled Line Equations, B.S.T.J. **41**, May, 1962, pp. 1011–1029.
4. Ince, E. L., *Ordinary Differential Equations*, Dover, New York, N. Y., 1956.
5. Bellman, R., *Stability Theory of Differential Equations*, McGraw-Hill, New York, N. Y., 1953.
6. Bromwich, T. J. I'a., *An Introduction to the Theory of Infinite Series*, Macmillan, N. Y., 1955.

Contributors to This Issue

JOHN W. BALDE, B.S.E.E., 1943, Rensselaer Polytechnic Institute; Western Electric Company, 1943—. Mr. Balde's early work at Western Electric and at Bell Laboratories was in the development of airborne radar and computer systems and auxiliary test equipment. From 1957–1959, he served as a member of the teaching staff of the Western Electric Graduate Engineering Training School. Since 1959, he has been engaged in thin film process research at the Western Electric Engineering Research Center at Princeton, where he is currently a research leader in thin film evaluation. Member, Sigma Xi.

R. D. BARNARD, B.E.E., 1952, and M.E.E., 1955, Polytechnic Institute of Brooklyn; Ph.D., 1959, Case Institute of Technology; Bell Telephone Laboratories, 1959–61; faculty, Wayne State University, 1961–62; Bell Telephone Laboratories, 1962—. Presently, he is primarily concerned with theoretical problems in signal theory and control. Member, IEEE, American Physical Society, Sigma Xi, Eta Kappa Nu and Tau Beta Pi.

SIDNEY S. CHARSCHAN, B.S.M.E., 1949, Columbia University; Western Electric Company, 1951—. Mr. Charschan's early work was in plastics development, where he was associated with the first cast resin and glass reinforced plastics designs. In 1958, he transferred to the Western Electric Engineering Research Center, Princeton, N. J., where, at present, he is a research leader for a group working on the development of special vacuum systems. He is a registered professional engineer of the State of New York.

L. A. D'ASARO, B.S., 1949, and M.S., 1950, Northwestern University, Ph.D., 1955, Cornell University, Bell Telephone Laboratories, 1955—. Mr. D'Asaro's work at Bell Laboratories has been mainly concerned with exploratory development of semiconductor devices. These have included PNPN switches, the stepping transistor, Esaki diodes and gallium arsenide lasers. He is at present supervising work on high-speed diodes.

Member, American Physical Society, IEEE, Sigma Xi, and Phi Beta Kappa.

John J. Dineen, B.S.E.E., 1957, Northeastern University; Bell Telephone Laboratories, 1957–1958; Western Electric Company, 1958—. Mr. Dineen was first engaged in closed-circuit television systems studies and later in the development of a microwave radar receiver for the Nike Zeus system. He went to Western Electric's Engineering Research Center at Princeton, N. J., in 1960, where he conducted manufacturing process systems studies. More recently he has engaged in the evaluation and analysis of thin film manufacturing processing systems. He is currently attending the Western Electric Company-Lehigh University Masters Degree Program and is majoring in operations research. Member, IEEE, Tau Beta Pi and Eta Kappa Nu.

Alexander Feiner, M.S. (Electrical Engineering), 1952, Columbia University; Bell Telephone Laboratories, 1953—. He has been engaged in the application of electronic techniques to switching. He presently heads a department responsible for development of switching networks, trunks and scanners, and for transmission aspects of No. 1 ESS. Member, Sigma Xi.

Dawon Kahng, B.Sc., 1955, Seoul University (Korea); M.Sc., 1956 and Ph.D., 1959, Ohio State University; Bell Telephone Laboratories, 1959—. He has been engaged in exploratory studies of surface field effect transistors and epitaxial film doping profiles. More recently, he has been engaged in hot electron device research and development of surface barrier microwave diodes. Member, IEEE, Sigma Xi, and Pi Mu Epsilon.

Arthur C. Keller, B.S., Cooper Union, 1923; M.S., Yale University, 1925; E.E., Cooper Union, 1926, Columbia University, 1926–1930; Western Electric Company, 1917–1925; Bell Telephone Laboratories, 1925—. He is at present Director, Switching Apparatus Laboratory, having previously been Director of Component Development, Director, Switching Systems Development, and Director of Switching Apparatus Development. Mr. Keller's experience in the Bell System includes development and design of electromechanical devices, sound recording and reproducing apparatus, electronic heating and sputtering equipment, telephone switching apparatus and systems, and, during World War II, sonar equipment and systems; he holds patents in all of these fields. The division which he heads is responsible for exploratory studies of and the development, design, and preparation for manufacture of electromechanical switching apparatus for telephone systems.

Member, American Physical Society, Yale Engineering Association, SMPTE, and Society for Experimental Stress Analysis; Fellow, IEEE and Acoustical Society of America. For his contributions to sonar, he received two U.S. Navy citations. In 1962 he received the Emile Berliner Award of the Audio Engineering Society. In 1963 he was elected to the Board of Directors of the Waukesha Motor Co.

Peter Linhart, B.A., 1948, Princeton University; M.A., 1950, University of California, Berkeley; Ph.D., 1963, Columbia University; Bell Telephone Laboratories, 1956—. Mr. Linhart was first engaged in systems engineering work relating to electronic switching. He later did mathematical studies of various remote line concentrators compatible with various switching systems — e.g., a concentrator switch consisting of a random slip with common overflow group. His present work concerns patterns of test calls for a specific distributed remote line concentrator.

Samuel P. Morgan, B.S., 1943, M.S., 1944, and Ph.D., 1947, California Institute of Technology; Bell Telephone Laboratories, 1947—. A research mathematician, Mr. Morgan has been particularly concerned with the applications of electromagnetic theory to microwave and other problems. As Head, Mathematical Physics Department, he now supervises a research group in various fields of mathematical physics. Fellow, IEEE; member, American Physical Society, Sigma Xi, Tau Beta Pi and A.A.A.S.

D. J. Newman, B.A., 1951, New York University; Ph.D., 1958, Harvard University. Dr. Newman worked as an industrial mathematician from 1953 to 1957. Instructor and lecturer, Massachusetts Institute of Technology, 1957–59; Assistant Professor of Mathematics, Brown University, 1959–60; Associate Professor of Mathematics, Yeshiva University, 1960—. He has been a mathematical consultant to Bell Laboratories since January, 1961. Member, Mathematical Association of America and American Mathematical Society.

Harrison E. Rowe, B.S., 1948, M.S., 1950, and Sc.D., 1952, M.I.T.; Bell Telephone Laboratories, 1952—. He was initially associated with a group engaged in systems research. He later worked on mode conversion problems arising in multimode waveguides. Presently, he is concerned with problems relating to optical systems. Member, IEEE, Sigma Xi, Tau Beta Pi, and Eta Kappa Nu.

IRWIN W. SANDBERG, B.E.E., 1955, M.E.E., 1956, and D.E.E., 1958, Polytechnic Institute of Brooklyn; Bell Telephone Laboratories, 1958—. He has been concerned with analysis of military systems, particularly radar systems, and with synthesis and analysis of active and time-varying networks. He is currently involved in a study of the signal-theoretic properties of nonlinear systems. Member, IEEE, Society for Industrial and Applied Mathematics, Eta Kappa Nu, Sigma Xi and Tau Beta Pi.

ERLING D. SUNDE, Dipl. Ing., 1926, Technische Hochschule, Darmstadt, Germany; American Telephone and Telegraph Co., 1927–1934; Bell Telephone Laboratories, 1934—. He has made theoretical and experimental studies of inductive interference from railway and power systems, lightning protection of the telephone plant, and fundamental transmission studies in connection with the use of pulse modulation systems. He is the author of *Earth Conduction Effects in Transmission Systems*, a Bell Laboratories Series book. Fellow, IEEE; member, A.A.A.S., American Mathematical Society.

B.S.T.J. BRIEFS

Quantum Efficiency of the Green and Red Electroluminescence of GaP

By A. Pfahnl

(Manuscript received November 19, 1963)

Gallium phosphide crystals were grown from polycrystalline material in a solution of gallium contained in evacuated and sealed-off quartz tubes.[1] For the regrowth, the tube with the GaP–Ga mixture was heated to 1250°C and cooled at a rate of 1.5°C per minute. After separation of the GaP crystals from the adherent Ga, Zn was diffused into the crystals,

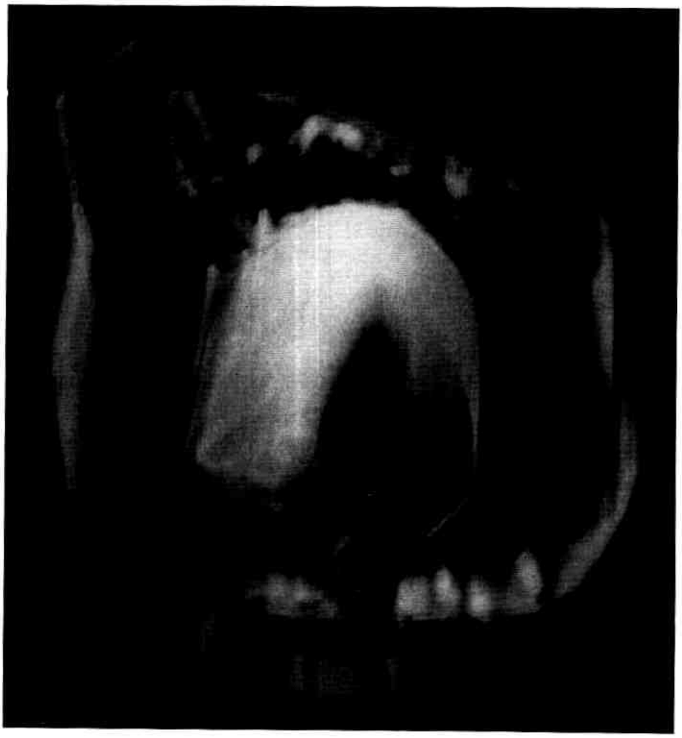


Fig. 1 — Red electroluminescent gallium phosphide crystal photographed in its own light; p-n junction prepared by diffusion of Zn at 800°C for four hours. Length of the straight side of the crystal about 1.5 mm.

leading to red (7000 Å) electroluminescent junctions. The diffusion was done in an evacuated and sealed-off quartz tube using as a source[2] a Zn + GaP mixture. The efficiency of the emission was determined with an integrating sphere and a photomultiplier with S-1 response calibrated in absolute units, and was found at room temperature to be about 1.0×10^{-3} photons per electron for the best samples. Red electroluminescence in GaP was previously reported to have efficiencies of about 10^{-4} (see Ref. 3) and $10^{-4} - 10^{-3}$ (see Ref. 4).

If silver contacts are alloyed onto the rough side of the solution-regrown GaP crystals, green electroluminescence can frequently be observed at the contact area. The efficiency of the green emission was found to be 4×10^{-5} photons (5550 Å) at 300°K observed outside the crystal per recombining electron-hole pair for the best samples. This compares with efficiencies of 3×10^{-5} measured by Gershenzon et al.[5] and efficiencies smaller than 10^{-4} as indicated by Allen et al.[3]

The figure shows one of the red electroluminescent crystals with a Zn-diffused junction photographed in its own electroluminescent light.

REFERENCES

1. Wolff, G., Keck, P. H., and Broder, J. D., Phys. Rev., **94,** 1954, pp. 753–754.
2. Foy, P. W., private communication.
3. Allen, J. W., Moncaster, M. E., and Starkiewicz, J., Solid-State Elect., **6,** 1963, pp. 95–102.
4. Gershenzon, M., and Mikulyak, R. M., Solid-State Elect., **5,** 1962, pp. 313–329.
5. Gershenzon, M., Mikulyak, R. M., Logan, R. A., and Fcy, P. W., to be published in Solid-State Elect.

Matching of Optical Modes

By H. Kogelnik

(Manuscript received November 19, 1963)

In experiments with coherent laser light it is frequently necessary to transform a given Gaussian beam[1,2] into a Gaussian beam with certain desired parameters. It is required, for example, to transform the light beam emerging from a laser oscillating in a fundamental mode in order to provide for optimum injection into a light transmission line[2,3] (consisting of a sequence of lenses), or for optimum coupling into a spherical mirror interferometer.[4] In these cases one has to "match" the incoming beam to the natural mode of the system in question. Lenses inserted in the beam perform the matching transformation. The design of a matching configuration has to take full account of the laws[1,2,3] that govern optical modes. This leads to a somewhat complex analysis.[5] The results, however, are quite simple matching formulae which are presented in this brief. A matching experiment is described for illustration.

The given beam is characterized by its minimum beam radius[1,6] (spot size) w_1 and by the location of the beam waist. The problem is to transform this beam into another with a minimum radius w_2 . The quantities w_1 and w_2 determine a characteristic "matching length" f_0 given by

$$f_0 = \pi \frac{w_1 w_2}{\lambda} \tag{1}$$

where λ is the wavelength. One beam is transformed into the other if a lens with a focal length f larger than f_0 is spaced between the two beam minima as shown in Fig. 1. The distances d_1 and d_2 between the lens and the beam minima have to satisfy the following matching conditions

$$\frac{d_1}{f} = 1 \pm \frac{w_1}{w_2} \sqrt{1 - \frac{f_0^2}{f^2}} \tag{2}$$

$$\frac{d_2}{f} = 1 \pm \frac{w_2}{w_1} \sqrt{1 - \frac{f_0^2}{f^2}} \tag{3}$$

where the same sign should be used in both equations. From (2) and (3) it follows that matching is not possible if $f < f_0$. If one chooses $f = f_0$ then $d_1 = f_0$ and $d_2 = f_0$; the beam minima are located in the two focal planes of the lens.

When one uses more than one lens to achieve the desired beam transformation, the above matching formulae are still applicable. Then f is the focal length of the lens combination, and d_1 and d_2 are measured from the principal planes. If the modes of two given optical systems are to be matched, one need not evaluate the beam parameters w_1 and w_2 , which are functions[1,6] of λ and the system parameters: the matching parame-

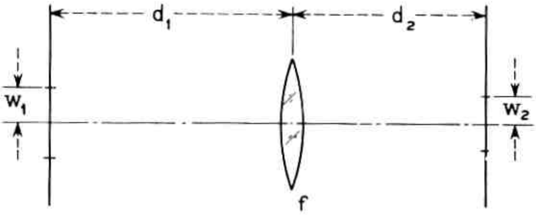


Fig. 1 — Matching configuration.

ters f_0, $\frac{w_1}{w_2}$, d_1, and d_2 are independent of λ and can be expressed in terms of the system parameters alone.

In our experimental study the light beam was taken from a He-Ne gas laser oscillating in a fundamental mode at $\lambda = 0.63$ micron. The laser cavity consisted of a concave mirror of 1 meter focal length and a flat output mirror. The mirror spacing was 1.7 meters. The (minimum) beam radius at the flat is computed[1] as $w_1 = 0.37$ mm. This beam was passed through a matching lens and then injected through a slit into a mirror system formed by two concave mirrors of 12.5 meters focal length spaced 50 centimeters apart. The injection angle was so chosen that the beam was reflected back and forth between the mirrors many times before it was finally intercepted, with the points of beam impact on each mirror forming a circular pattern. Such a beam configuration was described and analyzed in Ref. 7. As the beam passes back and forth between the mirrors its radius is changed in the same way as for transmission through a sequence of lenses[2,3,8] with corresponding parameters. The minimum beam radius of a fundamental mode of this sequence is computed as $w_2 = 0.7$ mm.

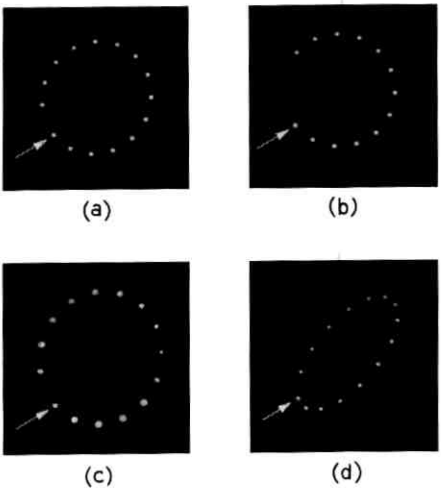


Fig. 2 — Photographs of beam spots on mirror.

From the above data one obtains a matching length of $f_0 = 1.3$ meters. A lens of a focal length of $f = 1.3$ meters was available and was used as matching lens. Therefore, spacings $d_1 = d_2 = f_0 = f = 1.3$ meters were required for matching.

A mirror of the multiple-pass system was slightly transparent and Fig. 2 shows photographs of the beam-impact points taken through this mirror. In Fig. 2(a) the arrow marks the point where the injected beam strikes the mirror first. After one return trip the point of impact is the neighboring point to the right. Subsequent impact points after a corresponding number of return trips appear counterclockwise on a circle. The beam was intercepted after 14 return trips. For illustration we show Fig. 2(b), where the beam was intercepted after 12 return trips. In both cases mode-matching conditions were fulfilled and all beam radii at impact are seen to be the same. In Fig. 2(c) one can see how the beam radii at the mirror vary periodically[9] if some mismatch is introduced: the spacing d_1 was misadjusted by about 25 cm. Fig. 2(d) shows the elliptical pattern obtained for another injection angle. Here the modes were matched again and all beam spots are of equal size.

REFERENCES

1. Boyd, G. D., and Gordon, J. P., Confocal Multimode Resonator for Millimeter Through Optical Wavelength Masers, B.S.T.J. **40,** March, 1961, pp. 489–508.
2. Goubau, G., and Schwering, F., On the Guided Propagation of Electromagnetic Wave Beams, I.R.E. Trans., **AP-9,** 1961, pp. 248–56.
3. Pierce, J. R., Modes in Sequences of Lenses, Proc. Nat. Acad. Sci. **47,** 1961, pp. 1808–31.
4. Fork, R. L., Gordon, E. I., Herriott, D. R., Kogelnik, H., and Loofbourrow, J. W., Scanning Fabry-Perot Observation of Optical Maser Output, Bull. Am. Phys. Soc. II, **8,** 1963, p. 380.
5. Kogelnik, H., Imaging of Optical Modes and Resonators with Internal Lenses, to be published.
6. Yariv, A., and Gordon, J. P., The Laser, Proc. IEEE, **51,** 1963, pp. 4–29.,
7. Herriott, D. R., Kogelnik, H., and Kompfner, R., Off-Axis Path in Spherical Mirror Interferometers, to be published.
8. Off-axis transmission of modes is discussed in a forthcoming publication by H. E. Rowe.
9. Pierce, J. R., *Theory and Design of Electron Beams*, Princeton, D. van Nostrand, 1954, p. 195.

leading to red (7000 Å) electroluminescent junctions. The diffusion was done in an evacuated and sealed-off quartz tube using as a source[2] a Zn + GaP mixture. The efficiency of the emission was determined with an integrating sphere and a photomultiplier with S-1 response calibrated in absolute units, and was found at room temperature to be about 1.0×10^{-3} photons per electron for the best samples. Red electroluminescence in GaP was previously reported to have efficiencies of about 10^{-4} (see Ref. 3) and $10^{-4} - 10^{-3}$ (see Ref. 4).

If silver contacts are alloyed onto the rough side of the solution-regrown GaP crystals, green electroluminescence can frequently be observed at the contact area. The efficiency of the green emission was found to be 4×10^{-5} photons (5550 Å) at 300°K observed outside the crystal per recombining electron-hole pair for the best samples. This compares with efficiencies of 3×10^{-3} measured by Gershenzon et al.[5] and efficiencies smaller than 10^{-4} as indicated by Allen et al.[3]

The figure shows one of the red electroluminescent crystals with a Zn-diffused junction photographed in its own electroluminescent light.

REFERENCES

1. Wolff, G., Keck, P. H., and Broder, J. D., Phys. Rev., **94**, 1954, pp. 753–754.
2. Foy, P. W., private communication.
3. Allen, J. W., Moncaster, M. E., and Starkiewicz, J., Solid-State Elect., **6**, 1963, pp. 95–102.
4. Gershenzon, M., and Mikulyak, R. M., Solid-State Elect., **5**, 1962, pp. 313–329.
5. Gershenzon, M., Mikulyak, R. M., Logan, R. A., and Foy, P. W., to be published in Solid-State Elect.

Matching of Optical Modes

By H. Kogelnik

(Manuscript received November 19, 1963)

In experiments with coherent laser light it is frequently necessary to transform a given Gaussian beam[1,2] into a Gaussian beam with certain desired parameters. It is required, for example, to transform the light beam emerging from a laser oscillating in a fundamental mode in order to provide for optimum injection into a light transmission line[2,3] (consisting of a sequence of lenses), or for optimum coupling into a spherical mirror interferometer.[4] In these cases one has to "match" the incoming beam to the natural mode of the system in question. Lenses inserted in the beam perform the matching transformation. The design of a matching configuration has to take full account of the laws[1,2,3] that govern optical modes. This leads to a somewhat complex analysis.[4] The results, however, are quite simple matching formulae which are presented in this brief. A matching experiment is described for illustration.

The given beam is characterized by its minimum beam radius[1,5] (spot size) w_1 and by the location of the beam waist. The problem is to transform this beam into another with a minimum radius w_2 . The quantities w_1 and w_2 determine a characteristic "matching length" f_0 given by

$$f_0 = \pi \frac{w_1 w_2}{\lambda} \tag{1}$$

where λ is the wavelength. One beam is transformed into the other if a lens with a focal length f larger than f_0 is spaced between the two beam minima as shown in Fig. 1. The distances d_1 and d_2 between the lens and the beam minima have to satisfy the following matching conditions

$$\frac{d_1}{f} = 1 \pm \frac{w_1}{w_2} \sqrt{1 - \frac{f_0^2}{f^2}} \tag{2}$$

$$\frac{d_2}{f} = 1 \pm \frac{w_2}{w_1} \sqrt{1 - \frac{f_0^2}{f^2}} \tag{3}$$

where the same sign should be used in both equations. From (2) and (3) it follows that matching is not possible if $f < f_0$. If one chooses $f = f_0$ then $d_1 = f_0$ and $d_2 = f_0$; the beam minima are located in the two focal planes of the lens.

When one uses more than one lens to achieve the desired beam transformation, the above matching formulae are still applicable. Then f is the focal length of the lens combination, and d_1 and d_2 are measured from the principal planes. If the modes of two given optical systems are to be matched, one need not evaluate the beam parameters w_1 and w_2 , which are functions[1,4] of λ and the system parameters: the matching parame-

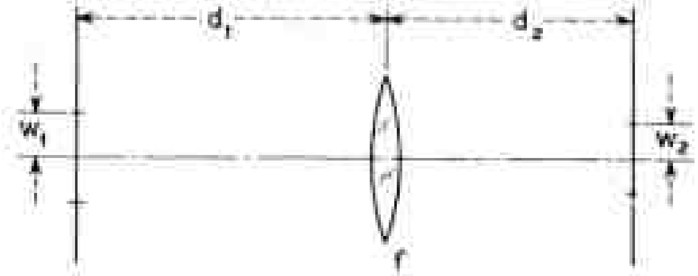


Fig. 1 — Matching configuration.

ters f_0, $\frac{w_1}{w_2}$, d_1, and d_2 are independent of λ and can be expressed in terms of the system parameters alone.

In our experimental study the light beam was taken from a He-Ne gas laser oscillating in a fundamental mode at $\lambda = 0.63$ micron. The laser cavity consisted of a concave mirror of 1 meter focal length and a flat output mirror. The mirror spacing was 1.7 meters. The (minimum) beam radius at the flat is computed[1] as $w_1 = 0.37$ mm. This beam was passed through a matching lens and then injected through a slit into a mirror system formed by two concave mirrors of 12.5 meters focal length spaced 50 centimeters apart. The injection angle was so chosen that the beam was reflected back and forth between the mirrors many times before it was finally intercepted, with the points of beam impact on each mirror forming a circular pattern. Such a beam configuration was described and analyzed in Ref. 7. As the beam passes back and forth between the mirrors its radius is changed in the same way as for transmission through a sequence of lenses[2,3,8] with corresponding parameters. The minimum beam radius of a fundamental mode of this sequence is computed as $w_2 = 0.7$ mm.

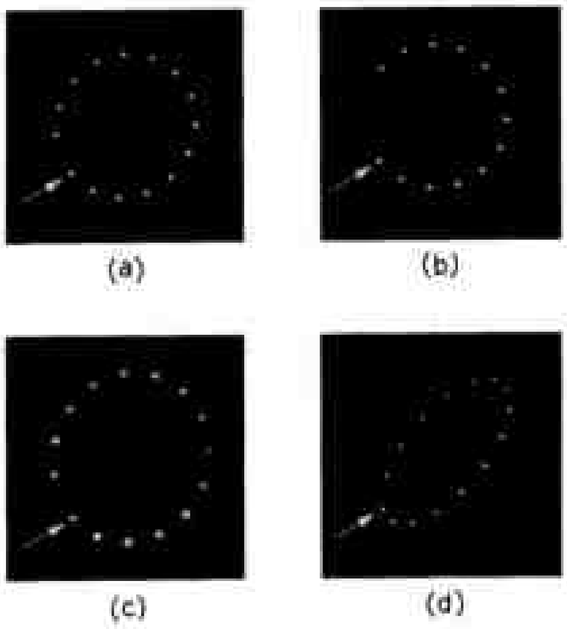


Fig. 2 — Photographs of beam spots on mirror.

From the above data one obtains a matching length of $f_0 = 1.3$ meters. A lens of a focal length of $f = 1.3$ meters was available and was used as matching lens. Therefore, spacings $d_1 = d_2 = f_0 = f = 1.3$ meters were required for matching.

A mirror of the multiple-pass system was slightly transparent and Fig. 2 shows photographs of the beam-impact points taken through this mirror. In Fig. 2(a) the arrow marks the point where the injected beam strikes the mirror first. After one return trip the point of impact is the neighboring point to the right. Subsequent impact points after a corresponding number of return trips appear counterclockwise on a circle. The beam was intercepted after 14 return trips. For illustration we show Fig. 2(b), where the beam was intercepted after 12 return trips. In both cases mode-matching conditions were fulfilled and all beam radii at impact are seen to be the same. In Fig. 2(c) one can see how the beam radii at the mirror vary periodically[9] if some mismatch is introduced: the spacing d_1 was misadjusted by about 25 cm. Fig. 2(d) shows the elliptical pattern obtained for another injection angle. Here the modes were matched again and all beam spots are of equal size.

REFERENCES

1. Boyd, G. D., and Gordon, J. P., Confocal Multimode Resonator for Millimeter Through Optical Wavelength Masers, B.S.T.J. **40**, March, 1961, pp. 489–508.
2. Goubau, G., and Schwering, F., On the Guided Propagation of Electromagnetic Wave Beams, I.R.E. Trans., **AP-9**, 1961, pp. 248–56.
3. Pierce, J. R., Modes in Sequences of Lenses, Proc. Nat. Acad. Sci. **47**, 1961, pp. 1808–31.
4. Fork, R. L., Gordon, E. I., Herriott, D. R., Kogelnik, H., and Loofbourrow, J. W., Scanning Fabry-Perot Observation of Optical Maser Output, Bull. Am. Phys. Soc. II, **8**, 1963, p. 380.
5. Kogelnik, H., Imaging of Optical Modes and Resonators with Internal Lenses, to be published.
6. Yariv, A., and Gordon, J. P., The Laser, Proc. IEEE, **51**, 1963, pp. 4–29.,
7. Herriott, D. R., Kogelnik, H., and Kompfner, R., Off-Axis Path in Spherical Mirror Interferometers, to be published.
8. Off-axis transmission of modes is discussed in a forthcoming publication by H. E. Rowe.
9. Pierce, J. R., *Theory and Design of Electron Beams*, Princeton, D. van Nostrand, 1954, p. 195.

THE BELL SYSTEM
TECHNICAL JOURNAL

VOLUME XLIII JANUARY 1964 NUMBER 1, PART 2

Demodulation of Wideband, Low-Power FM Signals*

By SIDNEY DARLINGTON

(Manuscript received October 3, 1963)

Some theoretical aspects of the demodulation of wideband, low-power FM signals are discussed. It is assumed that a band-limited, continuous, analog signal is supplied to the modulator and is recovered to a fidelity suitable for television, telephone, or carrier telephone. Much of the paper assumes that the baseband signal is sampled and clamped before it is applied to the frequency modulator. The combination has been called PAM-FM and is characterized by a piecewise constant transmitted frequency.

PAM-FM can be demodulated by spectrum analysis means not suitable for continuously varying frequencies. It is shown that a spectrum generator can be derived from the techniques of radar pulse compression, and is equivalent to an infinite set of correlators or matched filters plus means for scanning their terminals.

The spectrum analysis circuit forms are compared with demodulators using frequency detectors, with and without FM feedback, in regard to theoretical noise sensitivities. The theoretical sensitivities are quite similar for spectrum analysis and FMFB under conditions assumed. The comparisons disclose that frequency detectors (followed by filters) enjoy a disguised but efficient use of a differential phase coherence which is a characteristic of FM signals. A combination of spectrum analysis and frequency detection is described which has some of the theoretical advantages of both.

I. INTRODUCTION

This paper discusses some theoretical aspects of the demodulation of wideband, low-power frequency modulated signals. A wide trans-

* Parts of the material of this article were discussed by the author in lectures at the University of California at Berkeley during May, 1963.

mitted bandwidth permits a saving in power. Frequency modulation implies a constant power level, which makes peak power identical with average power. It is advantageous, for example, when the practical restrictions on peak power determine system power levels rather than restrictions on average power.

More specifically, the paper is concerned with FM systems subject to the following external requirements: A band-limited, continuous, analog signal is supplied to the input of a coder or modulator, which produces the transmitted signal. A demodulator reproduces the original baseband signal to a fidelity suitable for a television channel, a telephone channel, or a carrier system combining a number of telephone channels. For such purposes, for example, the average errors in the output must be more than 40 db below the baseband signal. It is assumed that a large FM index is used, to conserve signal power. These conditions are implicit in many of the conclusions. They will be referred to collectively as "the conditions assumed here."

Several different techniques and circuit forms are compared. The comparisons are concerned primarily, but not exclusively, with sensitivities to noise. Conventional FM receivers and circuits using FM feedback (FMFB) are included. However, more attention is paid to techniques which are closer to (but significantly different from) so-called frequency shift keying (FSK), a well-known method of data transmission.[1] Thus banks of correlators or matched filters appear in some of the proposed circuits, somewhat (but not exactly) as in FSK systems. Alternatively, the correlators or matched filters can be replaced by circuits resembling the pulse compressors of so-called Chirp radars,[2] and one (but not the only) purpose of the paper is to note how it can be done.

Circuits of different kinds are compared not only among themselves but also with theoretical bounds derived from general information theory. Thus the paper draws on four major disciplines within the general field of communication theory and practice, namely: conventional FM and FMFB, discrete data transmission, pulse compression radars, and information theory.

An expert in any one of the four disciplines may find some of the discussion quite familiar, and perhaps superfluous. However, it is unlikely that many readers will be thoroughly familiar with the pertinent parts of all the disciplines. Hence a somewhat tutorial approach has been adopted. However, some of the relations between disciplines and some of the circuit forms appear to be novel.

The purpose of the paper is to describe and compare the various techniques and circuit forms in simple terms. Mathematical proofs are outside the intended scope. Except in the Appendix, only the simplest formulas are stated explicitly, and circuits are represented only by simple block diagrams. A complete analysis is long, tedious, and mathematically uninteresting; a good deal of it differs only in detail from established applications to other problems. Some of the circuit forms have not actually been built; the block diagrams can be filled in with circuit details in many different ways, and best ways have not all been determined. The Appendix outlines very briefly some analytical and circuit details, which may be needed for an appreciation of some of the conclusions.

II. DEMODULATION BY SPECTRUM ANALYSIS

Much of the paper concerns systems in which the analog baseband signal is sampled, as part of the initial modulation, but is *not* quantized. Fig. 1 is a corresponding block diagram. Each sample is clamped during the sample interval, and is supplied to a frequency modulator. Then the transmitted frequency is constant over each sample interval, but changes from interval to interval. Curve B of Fig. 2 illustrates the variation in frequency with time. It differs from frequency shift keying in the following way: The transmitted frequency may be anywhere in a continuum of frequencies; it is not restricted to a finite number of discrete frequencies. The distinction has important repercussions throughout the paper.

If the sample interval is no greater than the Nyquist interval of the baseband bandwidth, the sampling destroys no information (at least in principle). It is assumed here that the sample interval equals the Nyquist interval.

Referring again to Fig. 1, the sequence of clamped samples at the input of the frequency modulator may be called a pulse amplitude modulation, or PAM representation of the original signal (with no gaps between the pulses). The corresponding output of the frequency modulator has been called PAM-FM.[3] It is a known means of adapting time

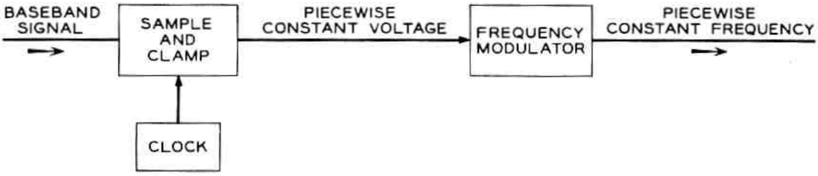


Fig. 1 — Block diagram of a PAM-FM modulator.

division multiplex to frequency modulation.* (For multiplexing, signal samples are clamped for only fractions of sample intervals and are interleaved with the samples from other channels ahead of the frequency modulator.) We are concerned here with a quite different feature of PAM-FM. The piecewise constant transmitted frequency can be demodulated by means of circuitry which cannot handle the continuously varying frequency of the more usual FM signal.

It is assumed that the demodulator is synchronized to the constant frequency intervals, as received. Some synchronization means are suggested in the Appendix (Section A.9). Then either correlators or matched filters may be used to estimate the piecewise constant frequency, sample-by-sample. The block diagram in Fig. 3 illustrates the concept, without filling in circuit details. A set of correlators or filters, tuned to a sequence of closely spaced frequencies, furnishes a spectrum analysis of the signal plus noise received over each sample interval. The signal is estimated by finding the frequency at which the spectrum is largest.

The operation is complicated by the fact that the true frequency is anywhere in a continuum, and must be estimated to closer than 1 per cent of the bandwidth of the continuum. This implies something like 100 correlators or filters, or else means for interpolation which compare the outputs of adjacent units.

2.1 *A Spectrum Generator*

The set of correlators or filters furnishes an analog representation of the desired spectrum, in which positions along a sequence of output terminals correspond to discrete values of frequency. The techniques of radar pulse compression can be used to represent the same spectrum, with time as the analog of frequency, at a single output terminal. Externally, the circuit is equivalent to an infinity of correlators or filters, with scanning means to convert the spacially distributed outputs into a function of time.

The spectrum generation hinges on a sequence of two operations. Fig. 4 is a block diagram. The first operation beats the received signal with a varying-frequency local oscillator, to obtain the difference frequency. Fig. 5 illustrates the frequencies of the true signal, of the local oscillator, and of the signal at the output of the mixer. The true frequency is constant over each sample interval, as before. The oscillator frequency varies periodically, in synchronism with the signal samples. In particular, it varies linearly over each sample interval. Thus, at the

* For example, in telemetry systems.

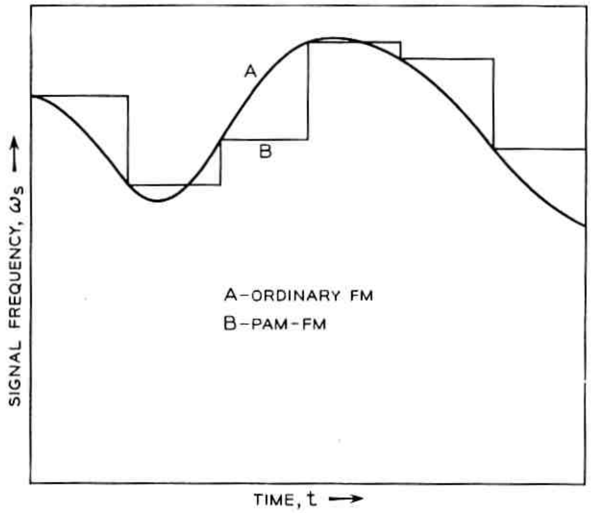


Fig. 2 — Instantaneous signal frequencies.

output of the mixer, the *variations* in frequency are the same over every sample interval, but the *average* varies from sample to sample.

The second operation transmits the modified signal through a pulse compressing (dispersive) line. The nominal delay (phase slope) varies linearly with frequency. Over any one sample interval, the instantaneous frequency varies linearly with time. Thus the nominal delay varies linearly with time. Fig. 6 illustrates the variations in delay with frequency and time.

The variations in delay are so scaled that the tail end of the signal sample just catches up with the head end. Then, on the basis of nominal delays, the entire signal sample emerges from the line in a single instant of time. Actually, of course, the nominal delay does not apply exactly to the time-varying instantaneous frequency. Thus the signal sample does not actually emerge from the line all at a single instant. However, under the conditions assumed it is squeezed into a small portion of the sample interval.

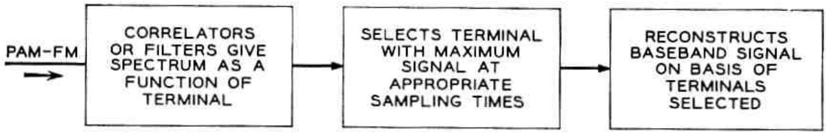

PAM-FM →
CORRELATORS OR FILTERS GIVE SPECTRUM AS A FUNCTION OF TERMINAL
SELECTS TERMINAL WITH MAXIMUM SIGNAL AT APPROPRIATE SAMPLING TIMES
RECONSTRUCTS BASEBAND SIGNAL ON BASIS OF TERMINALS SELECTED

Fig. 3 — PAM-FM demodulation by correlators or filters.

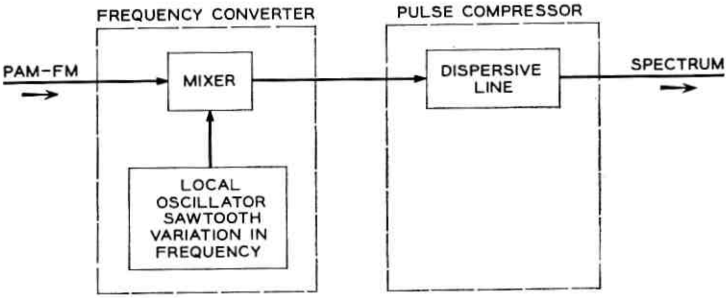


Fig. 4 — Spectrum generation.

The compression of a signal sample into a short pulse depends only on the variations in the instantaneous frequency, which are the same for each sample interval. On the other hand, the time of arrival at the output end of the line depends on the average frequency, which is the frequency of the true signal and varies from sample to sample. The baseband bandwidth and the FM index restrict the signal frequencies to a utilized RF bandwidth. With a suitable choice of circuit parameters, the corresponding variations in arrival time cover a little less than one sample interval. Then the true signal produces one pulse per sample interval, whose position in a (somewhat delayed) sample interval is a measure of the signal frequency. Fig. 7 illustrates the situation. In other terms, beating with a swept frequency and then pulse compressing converts PAM-FM into pulse position modulation, or PPM.*

It is now time to note specific formulas. For simplicity, let time t be zero at the center of a typical signal sample interval. Let the true signal, for that interval only, be

$$s(t) = \sqrt{2P_s} \cos (\omega_s t + \beta_s), \qquad -T/2 < t < T/2. \tag{1}$$

Here T is the length of the sample interval, ω_s and β_s are the frequency and phase of the true signal, and P_s the signal power. Let the corresponding output of the mixer be

$$\hat{s}(t) = \sqrt{2P_s} \cos (\omega_s t + \beta_s - \tfrac{1}{2} qt^2), \qquad - T/2 < t < T/2 \tag{2}$$

where q is an arbitrary constant. The instantaneous frequency is now $\omega_s - qt$, linear with respect to time. [Actual circuitry may introduce constant changes in amplitude, carrier frequency and phase angle, between (1) and (2), but these are trivial for present purposes.]

* In practice, the compressed pulse will have small side lobes, omitted in Fig. 7 for simplicity. See Fig. 8 below and also Section A.9 of Appendix.

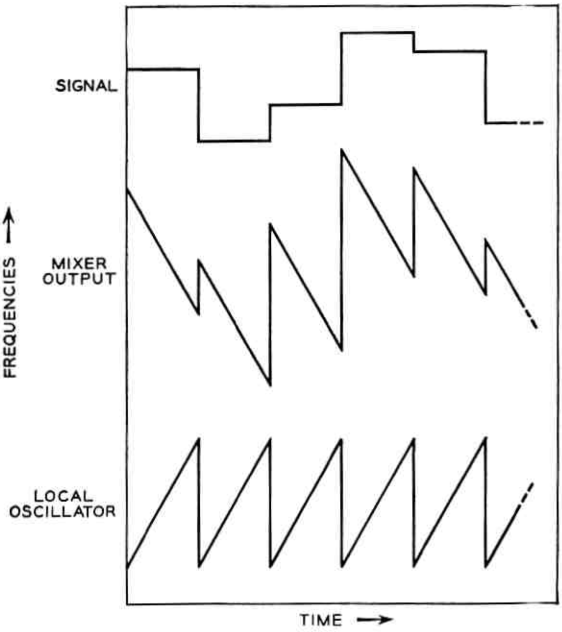


Fig. 5 — Frequency conversion.

The corresponding output of the pulse compression line is approximately

$$S(t) = \sqrt{2P_s}\, F\,(\omega_s - \omega_k) \cos\,(\omega_c t + \tfrac{1}{2}\, qt^2 + \beta_s - \beta_c)$$

$$\omega_k = \omega_c + qt \tag{3}$$

$$F(\lambda) = \frac{\sin \lambda \dfrac{T}{2}}{\lambda}.$$

The expression assumes that ω_c is large compared with $|\,\omega_s - \omega_c\,|$ and $|\,\omega_k - \omega_c\,|$. For present purposes, ω_s and ω_k lie in the utilized RF band, and ω_c is the midband, or carrier, frequency. A derivation of (3) from (2) is outlined in the Appendix (Section A.1).

The processed signal $S(t)$ may be described as a high-frequency sinusoid multiplied by an envelope function. The frequency, $\omega_c + qt$, varies with time, but it is independent of the received signal. On the other hand, the phase angle is $\beta_s - \beta_c$, in which β_c is a property of the transmission line, but β_s is the phase angle of the unprocessed signal $s(t)$. The envelope is $\sqrt{2P_s}\; F(\omega_s - \omega_k)$. It is a function of time, but

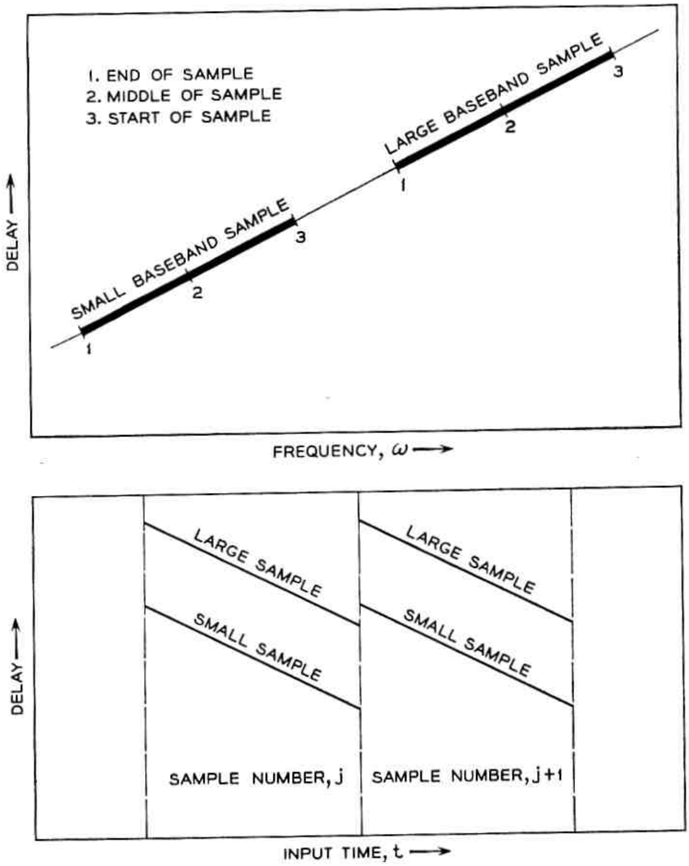


Fig. 6 — Delay vs frequency and time.

the time is an analog representation of the frequency variable ω_k. The signal frequency ω_s enters the envelope function as a parameter.

Fig. 8 is a qualitative plot of $F(\omega_s - \omega_k)$. The abscissae correspond simultaneously to time and ω_k. The largest F occurs at $\omega_k = \omega_s$. Thus the frequency ω_s may be determined by noting the time of the maximum F, and interpreting the time in terms of ω_k. The envelope F may be separated from the sinusoid by means of an envelope detector at the output of the line. Fig. 9(a) is a block diagram.

For some purposes, it is convenient to divide $S(t)$ into two components, as follows:

$$\begin{aligned} S(t) = {} & \sqrt{2P_s}\, F(\omega_s - \omega_k) \cos \beta_s \cos (\omega_c t + \tfrac{1}{2}\, qt^2 - \beta_c) \\ & - \sqrt{2P_s}\, F(\omega_s - \omega_k) \sin \beta_s \sin (\omega_c t + \tfrac{1}{2}\, qt^2 - \beta_c). \end{aligned} \tag{4}$$

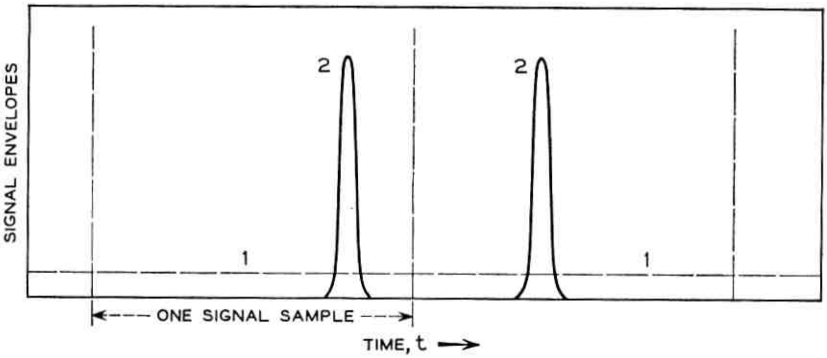


Fig. 7 — Signals at terminals of dispersive line: 1, input signal, before pulse compression; 2, output signal, after pulse compression.

The two sinusoids are independent of the signal $s(t)$. Physically, the two envelope functions can be resolved by means of phase detectors. Fig. 9(b) is a block diagram.

Consider the Fourier transform of a time function equal to $s(t)$ in the one sample interval, and zero elsewhere. More specifically, consider the transform at positive frequencies ω_k near ω_c. If the same approximations are made, as in the derivation of (3), the real and imaginary parts of the transform are the same as the two envelope functions in (4). The envelope function in (3) corresponds to the transform of the envelope of the original time function.

The same remarks apply a little more generally. Suppose the ampli-

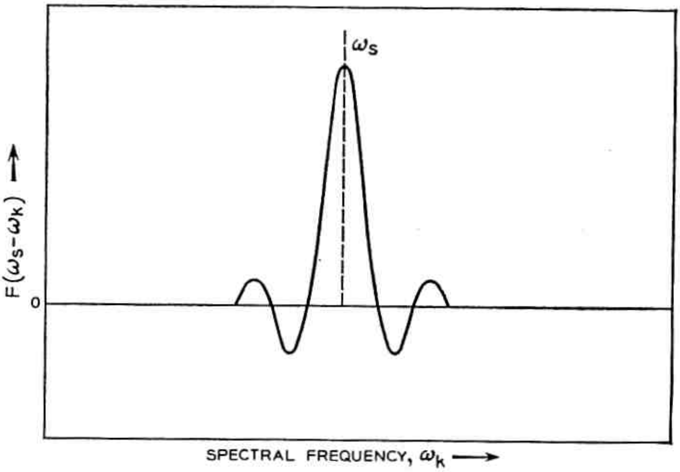


Fig. 8 — The function $F(\omega_s - \omega_k)$.

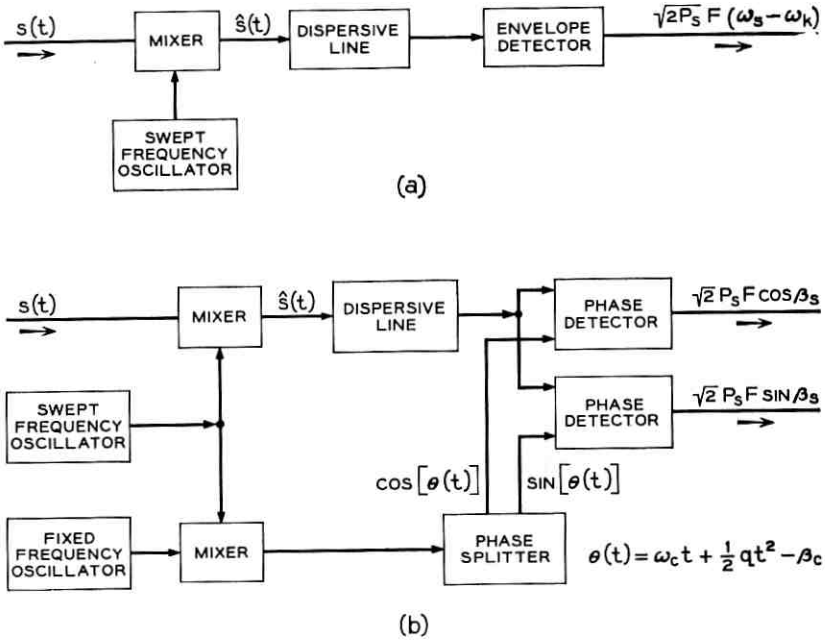


Fig. 9 — Detection of envelope and components: (a) envelope, (b) components.

tude of the received signal $s(t)$ is modified, as well as the frequency, before it reaches the pulse compressor. Then $\hat{s}(t)$ becomes

$$\hat{s}(t) = \sqrt{2P_s}\,A(t)\cos\left(\omega_2 t + \beta_2 - \tfrac{1}{2}qt^2\right), \qquad -T/2 < t < T/2. \quad (5)$$

Suppose the envelope $A(t)$ is symmetrical about the center of the sample interval. Then (3) and (4) apply except that $F(\lambda)$ is now the transform of a time function equal to the new envelope during the sample interval, and again zero elsewhere. For the analogous radar application, see Ref. 2.

The operations which convert (1) into (3) and (4) are all linear operations on the signal. If $s(t)$ is generalized to a sum of many constant-frequency sinusoids, the spectrum corresponding to a single sample interval can be generated by summing the results of the operations on the individual sinusoids. Referring again to the block diagrams, in Fig. 9(a) the output is the amplitude of the transform, and in 9(b) the two outputs are the real and imaginary parts. We will use the collection of sinusoids as a representation of the signal plus noise, received during one sample interval.

Thus pulse compression techniques generate analog representations of the transforms of signal samples. The transforms are generated as functions of time. The constant q determines the time-vs-frequency scale, and can be chosen so that the utilized RF band is scanned in less than one sample interval. The width of the peak in Fig. 8 is merely the familiar "spectrum line width" of a sinusoid of finite duration. The detailed shape (in particular the tails) can be modified to some extent through initial multiplication by an envelope function (or, alternatively, by a shaping circuit at the output of the dispersive line).

The same remarks apply to infinite sets of correlators or matched filters, except that the spectra are generated at specific instants as functions of position along arrays of output terminals. One result is: all three embodiments are equally sensitive to noise accompanying the received signal. A choice between the three must depend on practical compromises, limitations, etc., associated with the design of actual circuits. (For the external equivalence between correlators and matched filters, see, for example, Ref. 4.)

III. SENSITIVITIES TO NOISE

Demodulation by correlators, matched filters, or spectrum generators, as described in the previous section, will be referred to collectively as demodulation by spectrum analysis. This section compares the effects of noise in such circuits and in conventional FM receivers and FMFB. Between conventional FM and FMFB, some effects of noise are quite similar and some quite different. The two circuit forms will be referred to collectively as demodulation by frequency detection.

It is assumed that the noise is Gaussian and that it is added to the signal before it reaches the demodulator. It may be, for example, thermal noise associated with first stages of amplification in the receiver. In demodulation by spectrum analysis, the noise adds random processes to the spectra analyzed. These may be described as two independent Gaussian processes added to the envelope functions in (4). The independent variable in the random processes is the spectral frequency ω_k, which is also represented by time in the pulse compression embodiment. The processes are described in a little more detail in Section A.2.

It is convenient to normalize the error formulas in terms of parameters r, R, and T, defined as follows:

ω_b = baseband bandwidth (0 frequency to cutoff)

ω_r = full excursion of instantaneous signal frequency (maximum − minimum)

$r = \omega_r/\omega_b$ = bandwidth expansion ratio (6)

P_s = signal power (6) (cont.)
P_n = noise power in a frequency interval equal to one baseband
R^2 = P_s/P_n = "signal power to noise density ratio" at the input of the demodulator
T = π/ω_b = baseband Nyquist interval.

Under the conditions assumed here, the bandwidth expansion ratio, r, is fairly large — order of 10 or 20. Power thresholds (defined in the next section) set lower bounds on R, in the neighborhood of 14 to 16 db. In practical applications, practical compromises may require a somewhat larger R, and the bandwidth of the receiver must be a little greater than ω_r (whether demodulation is by spectrum analysis or frequency detection). The power spectrum of the noise is assumed to be uniform at the input of the demodulator, over the pertinent frequency interval.

3.1 *Two Different Effects of Noise*

For present purposes, one must examine two different effects of the noise, on the recovered baseband signal at the output of the demodulator. Under the conditions assumed here, the effects of the noise on the demodulated baseband signal are quite small most of the time. These may be called small noise errors, and their rms is one measure of circuit performance. On the other hand, during occasional brief intervals, peaks in the noise have a dominant effect and temporarily replace the true signal by a random false signal. This is commonly called blocking. It usually persists over intervals comparable with a baseband sample interval. The average number of blockings per second is the blocking frequency.

Fig. 10 illustrates the two effects in terms of probability densities. It is a qualitative (not quantitative) plot of the probability density of the error, due to noise, in the demodulated baseband signal at any one instant. The peak near zero is substantially Gaussian and corresponds to the small noise errors. The long tails are flat and correspond to the probability that blocking will replace the true signal by a random signal. The transitions between the Gaussian peak and the flat tails are not considered further here. They are very difficult to calculate and must be strongly dependent on design details.

The blocking frequency decreases very rapidly as the power ratio R increases. A related parameter is the power threshold. Thresholds of FM circuits (and also phase lock) have been defined in numerous ways for numerous purposes. The definition which best suits our present needs is the following: The power threshold is the signal power just sufficient

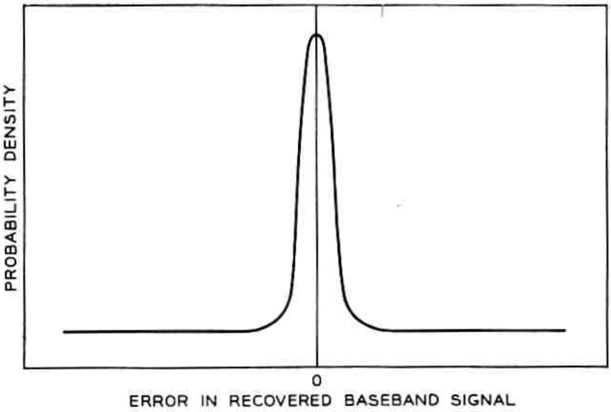


Fig. 10 — Qualitative form of error distribution.

to meet a specified limit on the portion of the samples which are blocked. It can be expressed in terms of the corresponding ratio, R, in db. Under the conditions assumed here, the specified limit on the blocking rate may be perhaps one in a thousand or one in ten thousand.

3.2 *Small Noise Errors*

Consider first the demodulation of individual signal samples by spectrum analysis. Both phase coherent and phase incoherent circuit forms are possible. More than one kind of phase coherence is of interest here. However, it will be simplest to start with the classical kind in which the phase of each constant frequency sample is independent of other samples and is determined uniquely by a rule known to the demodulator. This kind of phase coherence requires a degree of synchronization which may be impossible in practice. However, its theoretical properties bear on what follows.

Under the conditions assumed here, the corresponding small noise errors are approximately as follows:

For phase coherent demodulation:

$$\frac{\text{rms [small noise errors]}}{\text{max [true signal]}} = \frac{2\sqrt{3}}{\pi}\,\frac{1}{rR}\,. \tag{7a}$$

For phase incoherent demodulation:

$$\frac{\text{rms [small noise errors]}}{\text{max [true signal]}} = \frac{4\sqrt{3}}{\pi}\,\frac{1}{rR}\,. \tag{7b}$$

(The maximum true signal is here one half of a full signal excursion between equal + and − maxima.) Derivations are outlined in Section A.3.

According to (7), the small noise errors of phase coherent spectrum analysis are about 6 db smaller than those of phase incoherent spectrum analysis, assuming that the phases of signal samples are determined individually and uniquely by a suitable rule known to the demodulator. How do these compare with the small noise errors of demodulation by frequency detection?

Between conventional FM and FMFB, the small noise errors are approximately the same. More exactly, they are approximately the same functions of power level and bandwidths, which may themselves be quite different in practical applications of the two circuit forms. An approximate formula is

For demodulation by frequency detection:

$$\frac{\text{rms [small noise errors]}}{\text{max [true signal]}} = \frac{2}{\sqrt{3}} \frac{1}{rR} . \tag{8}$$

A well-known derivation is reviewed in Section A.4.

Superficially, conventional FM and FMFB appear to be phase incoherent. However, the (theoretical) small noise errors are almost the same as in the phase *coherent*, sample-by-sample spectrum analysis. They differ only by a voltage ratio $\pi/3$, or 0.40 db. This makes demodulation by frequency detection 5.62 db better, in regard to small noise errors, than the phase incoherent spectrum analysis.* It suggests that a more subtle form of phase coherence is at work, which perhaps can be realized also by a more subtle use of spectrum analysis.

Further evidence is as follows: Consider the usual description of noise reduction by conventional FM demodulation. (See again Section A.4.) The frequency detector, as such, produces a demodulated baseband signal plus a substantial amount of noise. However, when the FM index is large, most of the noise power is at frequencies above the baseband. Fig. 11 illustrates the usual form of the power spectrum. Then a filter which passes only the baseband eliminates most of the noise.

To approach the noise levels of phase coherent spectrum analysis, one must use an almost ideal baseband filter. But then the filter combines past outputs of the frequency detector over a "memory time" substantially longer than the baseband Nyquist interval. (Ideally it

* The 6-db difference has been noted before, with different interpretation, by, for example, Kotel'nikov.[5]

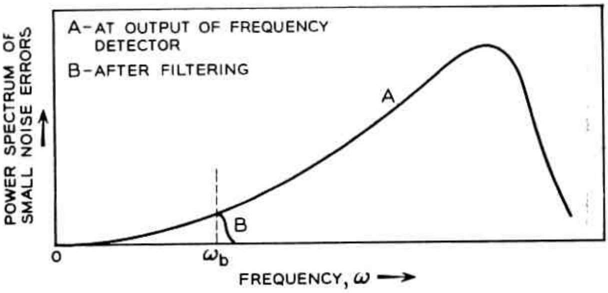


Fig. 11 — Small noise errors in frequency detection.

should be infinite.) Fig. 12(a) is a qualitative illustration of the appropriate weight function, or impulse response.

What happens if the filter is constrained to have a memory no longer than one baseband Nyquist interval? Suppose the true signal frequency is constant over that interval. Then the best weight function, within the constraint, is the parabola illustrated in Fig. 12(b).* The corresponding small noise errors turn out to be *exactly* as in phase *incoherent* spectrum analysis.

It is not at once clear how the longer memory of the ideal, unconstrained filter can reduce the (small noise) errors by anything like 5 or 6 db. The original baseband signals are substantially uncorrelated over intervals longer than one Nyquist interval. The effective correlation time of the noise process is even shorter. However, it is the *frequency* of the FM signal which has the correlation characteristics of the baseband. The *phase* is further characterized by the *continuity of phase rotations* required for a constant amplitude sinusoid of varying frequency. This may be regarded as a subtle kind of phase coherence which, in fact, is used effectively by the filter in demodulation by frequency detection.

The interpretation is clarified and supported by the following argument: Consider demodulation by spectrum analysis, and suppose the transmitted signal is generated by applying a piecewise constant control voltage to a frequency modulator. (See again Fig. 1.) Because the output of the modulator is a continuous sinusoid, the instantaneous phase rotation is continuous, even though its rate of change (which is the frequency) is discontinuous. The continuity of phase rotations, from sample to sample, has been called differential phase coherence.

* "Parabolic smoothing" is best for a finite interval, and a constant signal plus noise power proportional to ω^2. See, for example, Ref. 6.

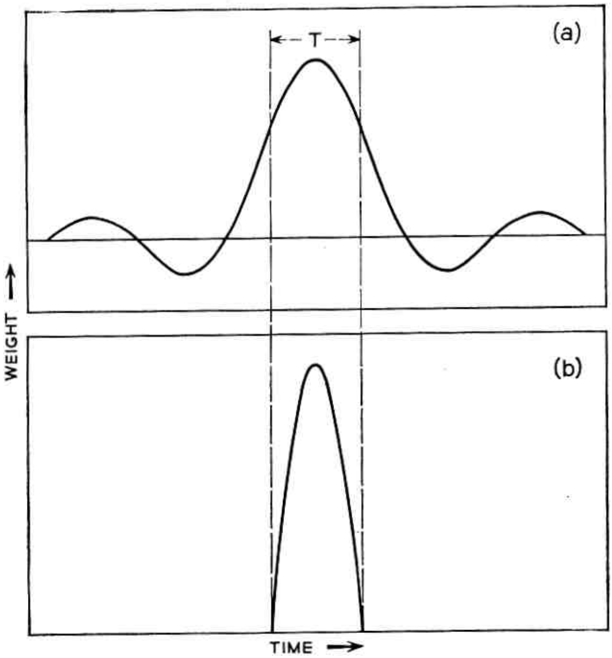


Fig. 12 — Filter weight functions: (a) ideal band-limiting filter; (b) optimum when constrained to one sample interval.

Fig. 13(a) illustrates the differentially coherent phase rotations. The slope of each straight line segment is the frequency during one signal sample interval. In contrast, if the transmitted signal is differentially phase incoherent, the phase rotations are discontinuous between samples, as in Fig. 13(b). This corresponds, for example, to forming a piecewise constant frequency signal by successive selections (or keying) from a set of phase incoherent oscillators.

Referring to Fig. 13(a), consider sample number k. The frequency can be estimated by an incoherent spectrum analysis of signal sample k by itself. [See again (7b) for the rms small noise errors.] Further information can be gleaned from spectrum analyses of samples $k - 1$ and $k + 1$. Specifically, estimates can be obtained from these samples of the phase rotations at the beginning and end of sample interval k. Only the *difference* between the two phase angles is actually needed, and hence the absolute phase reference required for the phase coherence of (7a) is no longer necessary.

The difference between the two estimated angles is the net phase rotation, modulo 2π, over sample interval k. Dividing by the duration

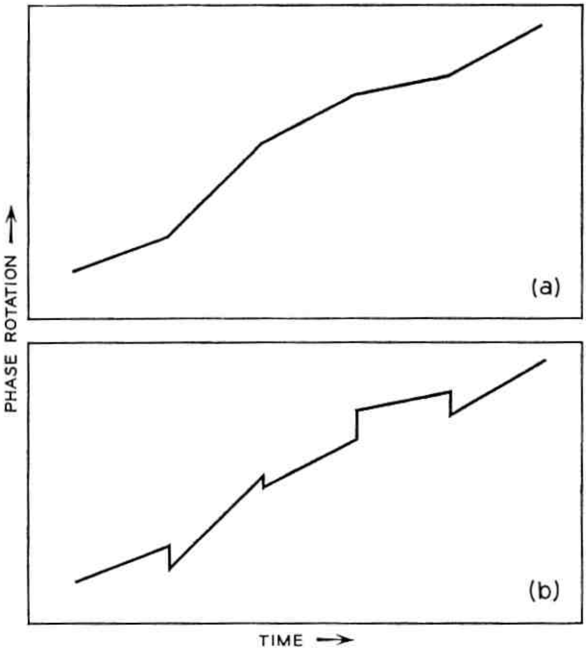


FIG. 13 — Phase rotations: (a) differential phase coherence; (b) differential phase incoherence.

T of the sample interval gives a second estimate of the frequency, but only to modulo $2\pi/T$. When the noise is small, as assumed, the first estimate is accurate enough to resolve the ambiguity. Then a weighted sum of the two estimates gives an improved estimate of the true signal frequency. (The small noise errors in the two estimates are substantially uncorrelated.) Small further improvements can be derived from frequency and phase estimates for additional sample intervals.

An optimum combination of phase and frequency measurements of all samples, $-\infty$ to $+\infty$, gives a 4.365-db theoretical improvement over sample-by-sample phase incoherent spectrum analysis. (The power ratio is $1 + \sqrt{3}$.) Of this, 3.979 db can be realized by using only samples $k - 1$, k, $k + 1$ to estimate the frequency of sample k. A derivation is described very briefly in Section A.5.

Why does one not realize the full 5.62 db apparent in conventional FM demodulation? It can be interpreted as a curious effect of the sampling of the original baseband signal, which is not part of the conventional FM system. The interpretation is supported by what follows.

Suppose the piecewise constant frequency is applied to the frequency detector in an (idealized) conventional FM receiver, and that the noise level is low enough to justify the usual small noise approximations. The output is a piecewise constant true signal, like curve A of Fig. 14, plus noise. The noise can be reduced by sampling the output of a suitable filter, as suggested by Fig. 15. Can the ideal baseband filter be used, as for an unsampled signal?

Elementary information theory includes the following: If the samples were represented by a sequence of very short impulses, like curve B of Fig. 14, the ideal filter would be as effective as for the unsampled signal. However, because they are represented, in fact, by a piecewise constant signal, like curve A, the ideal filter has two shortcomings. It produces intersample interference. It responds to the wanted sample less efficiently than to an ideal impulse.

Suppose the filter is constrained to give no intersample interference, assuming each sample to be a constant signal over its entire sample interval. The best filter within the constraint gives 4.365 db improvement over incoherent sample-by-sample spectrum analysis, which is

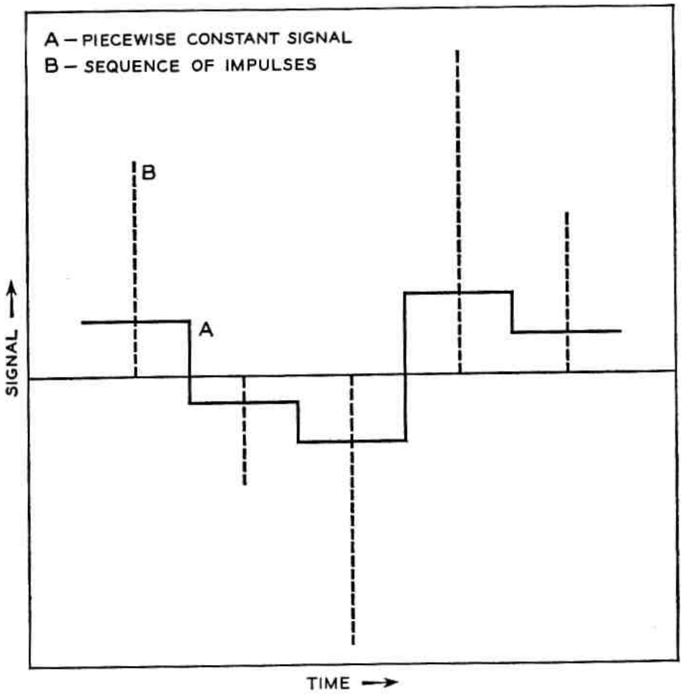


Fig. 14 — Filter inputs.

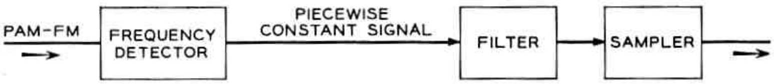


Fig. 15 — Filter and sampler after frequency detector.

exactly the same as the figure for multisample spectrum analysis using differential phase coherence. A derivation is outlined in Section A.6.

3.3 *Thresholds*

Consider first the thresholds of sample-by-sample spectrum analysis. Fig. 16(a) illustrates the spectrum of the usual signal-plus-noise sample. Fig. 16(b) illustrates the spectrum of the occasional sample which blocks. It assumes that the frequency of the spectral maximum is used as the estimate of the true frequency, as before. The blocking occurs when the spectrum of the noise sample has a peak, at a random frequency, which

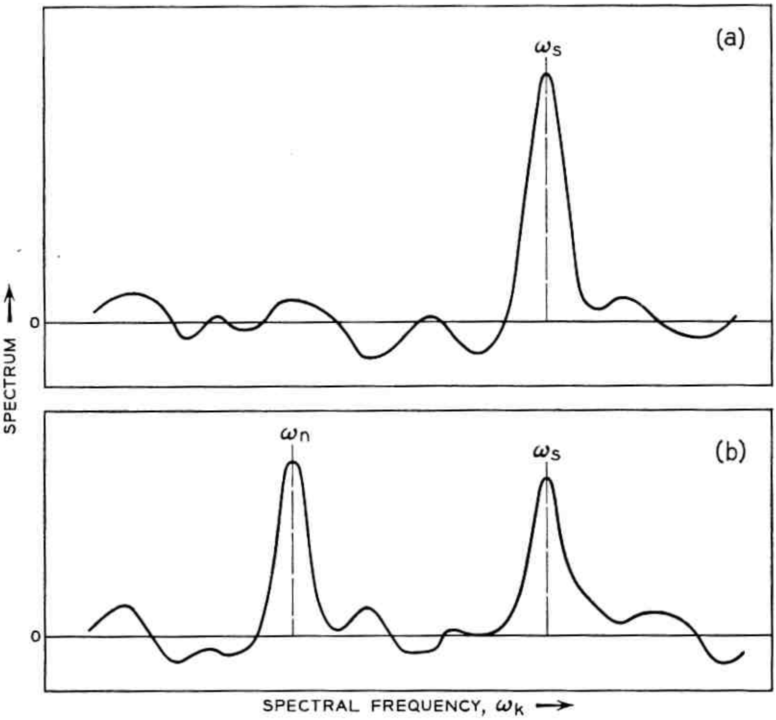


Fig. 16 — Spectrum of a single signal-plus-noise sample: (a) the usual sample; (b) the occasional sample which blocks.

exceeds the spectrum of signal-plus-noise at the true signal frequency. The remarks apply to both phase coherent and phase incoherent sample-by-sample spectrum analysis, provided the pertinent spectra are used for each.

The corresponding blocking probabilities are approximately as follows:

For phase coherent spectrum analysis:

$$P = \frac{r - 2}{2\sqrt{\pi}\,R} \exp\left(-R^2/4\right). \tag{9a}$$

For phase incoherent spectrum analysis:

$$P = \frac{r - 2}{4} \exp\left(-R^2/4\right). \tag{9b}$$

Here P is the probability that a typical sample is blocked, and blocking of different samples is uncorrelated.

Part of the derivation is the same as for the (gross) error rates of quantized frequency shift keying (FSK). However, there is an extra complication. In FSK, one is interested only in the spectrum at a finite set of discrete frequencies. The random process which is the noise spectrum is at most weakly correlated between the pertinent frequencies. Thus error rates have been approximated, for example, by assuming either zero correlation[7,8] or a manageably simple form of correlation.[9]

For our purposes, we must consider the spectra at all frequencies in a continuum, with the certainty that correlations are high across small frequency differences. An exact calculation would be extremely difficult. As an approximation, one can proceed as follows: Divide the pertinent frequency interval into, say η equal subintervals. Approximate the true spectrum in each subinterval by a constant. Assume that the constants for the η subintervals are independent random variables (over the ensemble of noise samples). Now one can estimate blocking probabilities as error rates in an η-frequency FSK system. Differences between (9) and equations in Refs. 7 and 8 reflect further approximations, appropriate under the conditions assumed here. They are described briefly in Section A.7, together with some further analytical details.

The approximation to the spectrum may be described further as follows: The covariance of the spectrum of the noise sample is approximated by perfect correlation over each subinterval and zero correlation between subintervals. The actual correlation across the (radian) frequency difference $\omega_2 - \omega_1$ is

$$\frac{\sin\,(\omega_2 - \omega_1)\,\dfrac{T}{2}}{(\omega_2 - \omega_1)\,\dfrac{T}{2}} \tag{10}$$

(see Section A.2). Equations (9) correspond to subintervals of width $2\omega_b$, which is the $|\,\omega_2 - \omega_1\,|$ at the first zeros of the true covariance function.

We have defined the threshold as the signal power required to meet a specified limit on the blocking frequency. The corresponding power ratio R, used in (9), must give the single-sample blocking probability P which corresponds to the specified blocking frequency.

Under the conditions assumed here, P is very small, say 0.001 or 0.0001. Then the exponentials in (9) are very small, and small percentage changes in R produce much larger percentage changes in P. As a result, changes in the coefficients, multiplying the exponentials, can be compensated by much smaller changes in R. For example, a two-to-one change in a coefficient is offset by something like a $\frac{1}{2}$-db change in R. Two consequences are as follows: The threshold changes only slowly with the bandwidth expansion ratio r. The threshold is rather insensitive to the size of the frequency subintervals used in the approximation described above.

Numerical examples of thresholds will be tabulated in Section IV, together with small noise errors.

Slepian[10] has derived from general information theory some important upper and lower bounds on the thresholds (as here defined) of *quantized* systems, constrained to code baseband samples individually, for transmission over channels wider than the baseband. It is interesting to compare the thresholds (9) with Slepian's bounds, even though (9) refers to *unquantized* systems. Since the bounds depend on the number of quanta, one must first decide on the appropriate quantization.

Transmission and demodulation of a quantized signal, as such, involve no counterpart of the small noise errors in unquantized systems. However, when the original baseband signal is unquantized, transmission in quantized form implies quantization or round-off errors relative to the original signal. Then, in judging system quality, one can compare the quantization errors in a quantized system with the small noise errors in an unquantized system. Thus it is interesting to compare thresholds determined by (9) with Slepian's bounds for quantized systems such that the rms quantization errors match our rms small noise errors.

Our present purposes are served by a very rough comparison, using graphical data in Slepian's paper. Under the conditions assumed here, the thresholds (9) are only very little above Slepian's lower bound. The differences are very roughly $\frac{1}{4}$ db for sample-by-sample phase coherent demodulation and one db for the phase incoherent form.

In principle, the thresholds can be reduced even a little further by combining phase and frequency estimates derived from more than one sample interval. We have seen that a second estimate of the frequency of sample k can be derived from the phases of samples $k - 1$ and $k + 1$. The same is true of the phase of sample k. This permits the phase coherent threshold to be approximated with only differential phase coherence. More complicated operations yield a further improvement. Referring again to Fig. 16(b), blocking occurs when a noise peak exceeds the signal peak, in the spectrum of the signal plus noise, and is chosen in its place. The additional phase information can be used to improve the choice between the two peaks. However, the 2π phase redundancy severely limits the improvement. For the conditions assumed here, a rough estimate is a ten-to-one reduction in the blocking frequency, or something like a one-db reduction in the threshold at the old rate (relative to phase incoherent spectrum analysis). A few further details are noted in Section A.8.

The improved threshold may be slightly below Slepian's lower bound. This is not improper, since it is obtained by violating Slepian's assumption of sample-by-sample coding and decoding.

Now consider the thresholds of conventional FM demodulators and FMFB. Fig. 17 compares simplified block diagrams of the two circuit

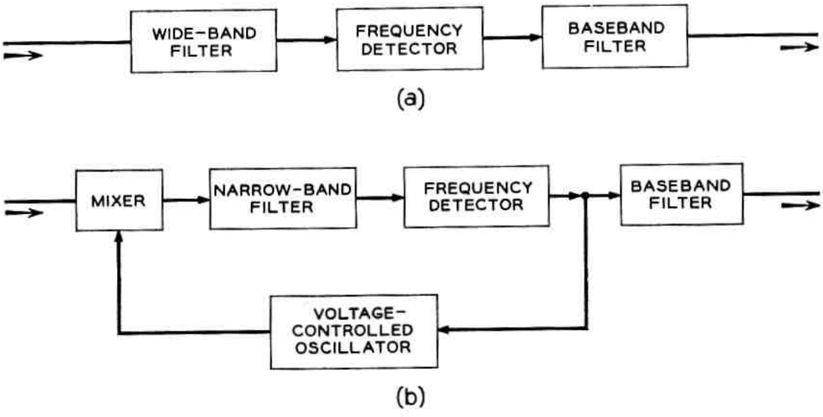


Fig. 17 — Demodulators using frequency detection: (a) conventional FM demodulator; (b) demodulator using FM feedback.

forms. The blocking phenomenon is a well-known characteristic of these circuits. Under the conditions assumed here, the thresholds are significantly lower (permit lower signal power) in FMFB circuits than in conventional FM receivers. The advantage derives from the relative bandwidths of the filters just ahead of the frequency detectors, and thereby depends on a fairly large bandwidth expansion ratio, r (which is here 10 or 20). This is, of course, the reason why FMFB is of current interest, for example for satellite communication systems.[11]

Because of the nonlinear feedback loop, it is extremely difficult to calculate for FMFB the quantitative thresholds required for specific blocking rates. However, important parameters have been identified and studied, for example by Enloe.[12] Good circuits have been built and demonstrated for voice and television channels, with thresholds which are not far above the theoretical lower bounds. Since the quantitative blocking rates have not been determined, the margins above the bounds are not known exactly.

3.4 *Comparisons with Other Methods*

At noise levels and blocking rates appropriate for television, telephone, and carrier telephone, FMFB and spectrum analysis of PAM-FM have lower theoretical thresholds than binary PCM. The binary symbols are less sensitive to noise than, say, PAM-FM samples received at the same rate. If this were the whole story, binary phase modulation would have the smaller threshold by a power ratio of about two.* Actually, of course, the symbol rate must be greater than the baseband sample rate by a factor, say ρ, equal to the number of binary symbols per sample. This, in itself, raises the power threshold by factor ρ. Thus, if there are more than two symbols per sample, the theoretical threshold for binary phase modulation is *larger*, by a power ratio of about $\rho/2$.

The threshold ratios are about the same if one compares the binary PCM with the following FSK system: A set of, say, 10 discrete frequencies is used, spaced orthogonally in the usual signal theory sense. One frequency from the set is transmitted during each baseband sample interval. But this system has only 10 quantum levels. To obtain, say, 100 quantum levels one must either transmit two symbol intervals per sample (which raises the threshold 3 db), increase the channel bandwidth by a factor of 10, or pack the frequencies much more closely than the orthogonal spacing. With close spacing, errors of one quantum level

* Binary phase modulation requires less power than binary frequency modulation. See, for example, Sunde.[13]

are more probable than larger errors, and there comes a point where they are more like the small noise errors of the analog systems.

In principle (but not likely in practice) thresholds can be *reduced* by using systems with *fewer* symbols or samples per second than the baseband sample frequency. For example, two baseband samples can be transmitted as a single analog sample provided the signal-to-noise ratio can be doubled (>80 db instead of >40 db). Transmission at the reduced sample rate yields a small reduction in threshold. It is paid for by an enormous increase in the channel bandwidth, which is required for the higher signal-to-noise ratio.

If more and more samples are combined, Shannon's fundamental channel capacity is undoubtedly approached. Turin[14] and Golay[15] have demonstrated that two closely related systems do, in fact, approach the theoretical capacity.*

Our formulas for demodulation by spectrum analysis assume that the true signal is estimated by finding the maximum point in the pertinent spectrum. The same is true of the analysis of FSK error rates in Refs. 7, 8 and 9. A well-known substitute for the determination of a maximum uses a circuit whose output is zero except when a signal-plus-noise (in this case the spectrum) exceeds a preset threshold. The threshold is set so that, most of the time, the peak due to the true signal and only that peak gets through.

Under the conditions assumed here, the threshold circuit form increases the theoretical power threshold by very roughly 3 db. More exactly, the blocking probability is dominated by an exponential factor $\exp(-R^2/8)$ as opposed to $\exp(-R^2/4)$ in equations (9).

IV. CONCLUSIONS

The techniques of radar pulse compression can be used to generate spectra of signal samples as analog functions of time. It can be done in real time in the sense that the spectrum of each signal sample is scanned in a time no greater than the sample interval. The spectra are the same as would be generated by infinite sets of correlators or matched filters. Spectrum generation of this sort may be useful for various purposes, particularly where the parameter ranges are suitable for the sort of hardware which has been developed for radar pulse compression.

Demodulation by frequency detection (with or without feedback) reduces the *small noise errors* by a disguised but efficient use of differen-

* The increase in channel bandwidth as Shannon's limit is approached is merely a property of these specific modulation schemes. In principle, it is necessary only to increase the length of the pieces of the signal which are coded as units.

tial phase coherence, which is a characteristic of FM signals. Demodulation by spectrum analysis can also take advantage of the differential phase coherence, although the pertinent operations are fairly complicated. The piecewise constant signal frequency, needed for the spectrum analysis, reduces the effectiveness by 1.24 db in the theoretical small noise errors (which can be offset by a 15 per cent increase in the FM index).

Under the conditions assumed, and for thresholds as defined here, the theoretical *power thresholds* of the spectrum analysis are very close to Slepian's lower bound. The power threshold of FMFB appears to be quite close, but just how close has not been determined.

Thus, under conditions appropriate for television, telephone, and carrier telephone systems, the theoretical noise sensitivities are very little different in FMFB and in PAM-FM with demodulation by spectrum analysis. Both techniques pose numerous practical problems, relating to, for example, stability requirements, switching time requirements, synchronization to signal samples, over-all complexity, nonlinearity in response to true signal, etc. FMFB has the advantage that it has already been used, although under somewhat special conditions.

Some theoretical thresholds and small noise errors are collected in Tables I and II, for various blocking probabilities P and bandwidth ratios r. They were calculated by (7) and (9) and refer to demodulation of PAM-FM by phase coherent and incoherent, sample-by-sample spectrum analysis. A few remarks on circuit problems are collected in Section A.9.

The noise figures obtainable with practical circuits are of course somewhat poorer. The degradations may be due to rather different practical compromises in circuits using spectrum analysis and in FMFB. Comparisons between practical noise figures may be different for different applications.

Under some conditions, a combination of spectrum analysis and frequency detection may be preferable to either alone. Fig. 18 is a block diagram of one out of many possible arranagements. A spectrum analyzer furnishes a first estimate of the frequency of a PAM-FM signal, using phase incoherent, sample-by-sample spectrum analysis. The estimated frequency variations are generated locally by a voltage-controlled oscillator. A mixer subtracts the oscillator frequency from the frequency of the received signal. (The block labeled "delay" allows for the operation time of the spectrum analysis.) Then the output of the mixer is very low index FM, corresponding to the errors in the first frequency estimate, plus noise.

TABLE I — THRESHOLDS AND SIGNAL-TO-NOISE RATIO FOR PHASE COHERENT SPECTRUM ANALYSIS

Probability of Blocking P	Bandwidth Ratio $r = \omega_r/\omega_b$	Threshold Ratio P_s/P_n	$\left[\frac{\text{Max. Demod. Signal}}{\text{rms Small Errors}}\right]$*
		(db)	(db)
0.01	10	12.1	31.3
0.005	10	12.7	31.9
0.002	10	13.4	32.6
0.001	10	13.9	33.0
0.0005	10	14.3	33.5
0.0002	10	14.8	33.9
0.0001	10	15.2	34.5
0.01	20	12.8	38.0
0.005	20	13.4	38.5
0.002	20	14.0	39.1
0.001	20	14.4	39.5
0.0005	20	14.8	39.9
0.0002	20	15.3	40.4
0.0001	20	15.6	40.7
0.01	40	13.4	44.6
0.005	40	13.9	45.0
0.002	40	14.4	45.6
0.001	40	14.8	46.0
0.0005	40	15.2	46.3
0.0002	40	15.6	46.8
0.0001	40	15.9	47.1

* At threshold signal power.

Because of the low index, it is now appropriate to use a narrow-band filter (passing something over two baseband bandwidths) followed by a frequency detector and a low-pass filter. The sampled output of the filter furnishes a correction to the first frequency estimate. The theoretical threshold of the combination is the same as for phase incoherent spectrum analysis. The theoretical small noise errors are the same as for demodulation of PAM-FM by frequency detection. The theoretical improvement over the small noise errors of the first frequency estimate is 4.365 db.

If the spectrum analysis is accomplished by correlators or matched filters, a moderate number may be sufficient even though the over-all errors must be >40 db below the true signal. The error determination by frequency detection can correct for a fairly coarse quantization of the first estimate at the same time that it reduces the errors due to noise.

The over-all circuit may be described as open-loop tuning to the passband of the narrow-band filter, as opposed to closed-loop tuning in FMFB.

TABLE II — THRESHOLDS AND SIGNAL-TO-NOISE RATIOS FOR PHASE INCOHERENT SPECTRUM ANALYSIS

Probability of Blocking P	Bandwidth Ratio $r = \omega_r/\omega_b$	Threshold Ratio P_s/P_n	$\left[\frac{\text{Max. Demod. Signal}}{\text{rms Small Errors}}\right]^*$
		(db)	(db)
0.01	10	13.3	26.5
0.005	10	13.8	27.1
0.002	10	14.4	27.6
0.001	10	14.8	27.9
0.0005	10	15.2	28.4
0.0002	10	15.7	28.8
0.0001	10	16.0	29.3
0.01	20	13.9	33.1
0.005	20	14.3	33.5
0.002	20	14.9	34.0
0.001	20	15.3	34.4
0.0005	20	15.6	34.7
0.0002	20	16.0	35.1
0.0001	20	16.3	35.4
0.01	40	14.4	39.6
0.005	40	14.8	40.0
0.002	40	15.3	40.5
0.001	40	15.6	40.8
0.0055	40	16.0	41.1
0.0002	40	16.3	41.5
0.0001	40	16.6	41.8

* At threshold signal power.

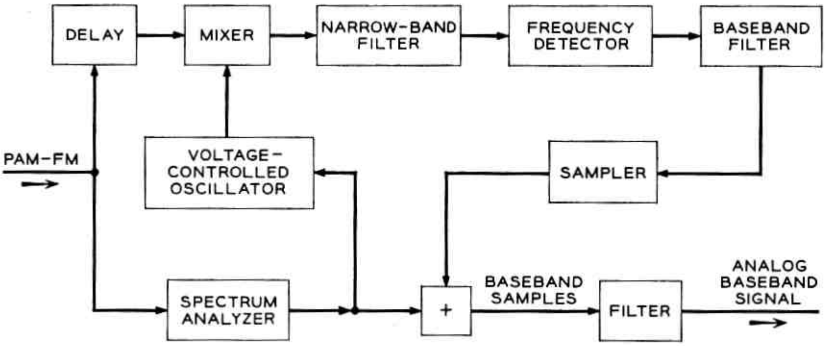


Fig. 18 — A combination of spectrum analysis and frequency detection.

V. ACKNOWLEDGMENTS

The author's work, reported here, has been strongly influenced by conversations with numerous colleagues in the data transmission, FM and FMFB, information theory, and radar pulse compression fields. These included V. E. Beneš, W. R. Bennett, L. H. Enloe, E. N. Gilbert, W. M. Goodall, C. W. Hoover, Jr., I. Jacobs, R. A. Kaenel, D. Mitchell, S. O. Rice, C. L. Ruthroff, and D. Slepian at Bell Telephone Laboratories, and T. Kailath, R. Price, and G. L. Turin while visiting the University of California at Berkeley. Several of the references were supplied by reviewers.

APPENDIX

A.1 Spectrum Generation by Pulse Compression

For a signal sample, modified by the local oscillator, assume:

$$\begin{aligned} \hat{s}(t) &= \sqrt{2P_s}\, E(t) \cos\left(\omega_s t - \tfrac{1}{2}\, qt^2 + \beta_s\right) \\ E(t) &= 0 \text{ outside of interval } -T/2 \leqq t \leqq +T/2 \\ E(-t) &= E(t). \end{aligned}$$

For the impulse response of the pulse compression line, assume:

$$w(t) = \cos\left(\omega_c t + \tfrac{1}{2}\, qt^2 - \beta_c\right).$$

When $|\omega - \omega_c| \ll \omega_c$, the frequency function is

$$Y(i\omega) = (\quad)\exp\left[-i\,\frac{(\omega - \omega_c)^2}{2q}\right].$$

The output of the line is $\hat{s} * w$. Integrate only over $E(t) \neq 0$:

$$S(t) = \sqrt{2P_s} \int_{\tau=-T/2}^{+T/2} E(\tau)\cos\left(\omega_s\tau - \tfrac{1}{2}\, q\tau^2 + \beta_s\right) \cdot \cos\left[\omega_c(t-\tau) + \tfrac{1}{2}\, q(t-\tau)^2 - \beta_c\right] d\tau.$$

Express the integrand as a sum of cosines. Neglect the high-frequency term. Then:

$$S(t) = \sqrt{2P_s} \int_{\tau=-T/2}^{+T/2} \tfrac{1}{2} E(\tau)\cos\left[\omega t + \tfrac{1}{2}\, qt^2 + \beta_s - \beta_c + (\omega_s - \omega_c - qt)\tau\right] d\tau.$$

Resolve into components per sin, cos $[(\omega_s - \omega_c - qt)\,\tau]$.

Recall that $E(\tau)$ is even. Then $E(\tau)\sin[(\omega_s - \omega_c - qt)\,\tau]$ is odd in τ.

$$S(t) = \sqrt{2P_s}\,F(\omega_s - \omega_c - qt)\cos(\omega t + \tfrac{1}{2}qt^2 + \beta_s - \beta_c)$$

$$F(\lambda) = \int_{\tau=-T/2}^{+T/2} \tfrac{1}{2}E(\tau)\cos(\lambda\tau)\,d\tau.$$

When $E(\tau) = 1,\ -T/2 \leqq \tau \leqq +T/2,\ F(\lambda) = \dfrac{\sin\lambda\dfrac{T}{2}}{\lambda}$.

A.2 Noise Contributions to Observed Spectrum

Following Rice,[16] but sacrificing some details of rigor to brevity, let the noise at the demodulator input be:

$$n(t) = \int_{\omega_1}^{\omega_2} x(\omega)\cos(\omega t + \beta_n)\,d\omega + \int_{\omega_1}^{\omega_2} y(\omega)\sin(\omega t + \beta_n)\,d\omega.$$

The interval ω_1 to ω_2 includes all signal frequencies ω_s.

Phase β_n = an arbitrary parameter in noise representation.

$x(\omega),\ y(\omega)$ = uncorrelated, zero average random variables, with uniform variances, and zero autocorrelations except across infinitesimal frequency intervals.

Let Ave denote an ensemble average, or expectation.

Let $w_1(\omega)$ and $w_2(\omega)$ be arbitrary, except for the pertinent conditions of integrability.

$$\text{Ave}\left\{\int_{\omega_1}^{\omega_2} x(\omega)\,w_1(\omega)\,d\omega \int_{\omega_1}^{\omega_2} x(\omega)\,w_2(\omega)\,d\omega\right\} = \sigma^2 \int_{\omega_1}^{\omega_2} w_1(\omega)\,w_2(\omega)\,d\omega$$

$$\text{Ave}\left\{\int_{\omega_1}^{\omega_2} y(\omega)\,w_1(\omega)\,d\omega \int_{\omega_1}^{\omega_2} y(\omega)\,w_2(\omega)\,d\omega\right\} = \sigma^2 \int_{\omega_1}^{\omega_2} w_1(\omega)\,w_2(\omega)\,d\omega$$

$$\text{Ave}\left\{\int_{\omega_1}^{\omega_2} x(\omega)\,w_1(\omega)\,d\omega \int_{\omega_1}^{\omega_2} y(\omega)\,w_2(\omega)\,d\omega\right\} = 0.$$

$$P_b = \text{noise power in one base bandwidth} = \omega_b\sigma^2.$$

Let $N(\omega_k)$ = the noise part of the spectrum of one signal-plus-noise sample.

Apply Section A.1, with $\omega_s = \omega$ and $\omega_c + qt = \omega_k$, to integrands in $n(t)$.

$$N(t) = N_1(\omega_k) \cos (\omega_c t + \tfrac{1}{2} q^2 - \beta_c) + N_2(\omega_k) \sin (\omega_c t + \tfrac{1}{2} q^2 - \beta_c)$$

$$N_1(\omega_k) = \int_{\omega_1}^{\omega_2} x(\omega)\, F(\omega - \omega_k)\, d\omega,$$

$$N_2(\omega_k) = \int_{\omega_1}^{\omega_2} y(\omega)\, F(\omega - \omega_k)\, d\omega$$

$N_1(\omega_k)$, $N_2(\omega_k)$ = independent Gaussian random processes in ω_k.

Appropriate choices of w_1, w_2 in the above expectation integrals give autocovariances of N_1, N_2.

$$\text{Ave}\,[N_\gamma(\omega_k)\, N_\gamma(\omega_j)] = \sigma^2 \int_{\omega_1}^{\omega_2} F(\omega - \omega_k)\, F(\omega - \omega_j)\, d\omega, \qquad \gamma = 1, 2.$$

Approximate the integration by integrating from $-\infty$ to $+\infty$.
Refer to Section A.1 and use $E(t) = 1$, $-T/2 \leqq t \leqq +T/2$.

$$\text{Ave}\,[N_\gamma(\omega_k)\, N_\gamma(\omega_j)] = \pi\sigma^2 \frac{\sin (\omega_k - \omega_j) \dfrac{T}{2}}{(\omega_k - \omega_j)}.$$

Let $\omega_j = \omega_k$, refer to (3), and recall that $R^2 = \dfrac{P_s}{P_b} = \dfrac{P_s}{\omega_b \sigma^2}$, $T = \dfrac{\pi}{\omega_b}$.

$$\frac{\text{Max of Signal Spectrum}}{\text{rms } N_\gamma(\omega_k)} = \sqrt{\frac{P_s}{P_b}} = R, \qquad \gamma = 1, 2.$$

A.3 Small Noise Errors in Sample-by-Sample Spectrum Analysis

Refer to Fig. 4, $S(t)$ of (3), and $N(\omega_k)$ of Section A.2.
Use $\omega_k = \omega_c + qt$ and $\beta_n = \beta_s$.

$$S(t) + N(t) = [\sqrt{2P_s}\, F(\omega_s - \omega_k) + N_1(\omega_k)] \times \cos (\omega t + \tfrac{1}{2} qt^2 + \beta_s - \beta_c) + N_2(\omega_k) \sin (\omega t + \tfrac{1}{2} qt^2 + \beta_s - \beta_c).$$

Assume (for small noise errors only):

$$N_1^2, N_2^2 \ll 2P_s F^2(0), \qquad \omega_k - \omega_s = \epsilon, \qquad \epsilon^2 \ll \omega_b^2.$$

Phase Incoherent Spectrum Analysis. Neglecting N_2^2, the envelope is $\sqrt{2P_s}\, F(\omega_s - \omega_k) + N_1(\omega_k)$.
Form a power series in ϵ and solve for max with ϵ small.

$$\text{Ave } \epsilon^2 = \frac{\text{Ave}\left(\dfrac{\partial N_1}{\partial \omega_k}\right)^2}{2P_s\left(\dfrac{\partial^2 F(\epsilon)}{\partial \epsilon^2}\right)^2_{\epsilon=0}}.$$

Evaluate by (3) and Section A.2 to get (7b).

Phase Coherent Sample-by-Sample Spectrum Analysis. Refer to (1). Make the phase β_s a linear function of ω_s :

$$S(t) = \sqrt{2P_s} \cos [\omega_s t + (\omega_s - \omega_c)(T/2)].$$

Find the components of $S(t)$ and $N(t)$ in phase with a locally generated $\cos \{[\omega_c + q(T/2)]t + \frac{1}{2} qt^2 - \beta_c\}$.

Refer to (3). Let S_c be the component of S.

$$S_c(t) = \sqrt{2P_s}\, F(\omega_s - \omega_k) \cos [(\omega_s - \omega_k)(T/2)]$$
$$= \sqrt{2P_s}\, F[2(\omega_s - \omega_k].$$

It can be shown that the frequency variable is also doubled between covariances of $N_1(\omega_k)$ and its counterpart here. Hence noise is accounted for with $\frac{1}{2}$ the frequency errors ϵ.

If the frequency-dependent signal phase appears artificial, change the time scale to $\hat{t} = t + T/2$.

$$S(\hat{t}) = \sqrt{2P_s} \cos\left(\omega_s \hat{t} - \frac{T}{2}\,\omega_c\right), \qquad 0 \leqq \hat{t} \leqq T.$$

A.4 Small Noise Errors in Frequency Detection

The FM signal is now unsampled. For simplicity assume a constant signal frequency. Resolve the noise per signal phase.

$$s(t) + n(t) = [\sqrt{2P_s} + n_a(t)] \cos (\omega_s t + \beta_s)$$
$$+ n_b(t) \sin (\omega_s t + \beta_s)$$

$$s(t) + n(t) = \rho \cos [\omega_s t + \beta_s + \varphi(t)], \qquad \tan \varphi = \frac{n_b(t)}{\sqrt{2P_s} + n_a(t)}.$$

The unfiltered frequency error is $\dot{\varphi}$. Refer to Section A.2 to get:

$$\text{When } n^2 \ll 2P_s, \quad \text{Ave } \dot{\varphi}^2 = \frac{\text{Ave } \dot{n}_b^{\,2}}{2P_s} = \frac{\sigma^2}{2P_s}\int_{\omega_1}^{\omega_2} (\omega - \omega_s)^2\, d\omega.$$

The ideal baseband filter passes only $| \omega - \omega_s | \leqq \omega_b$.

$$\text{Ave (Filtered } \epsilon)^2 = \frac{\sigma^2}{2P_s} \int_{-\omega_b}^{+\omega_b} \lambda^2 d\lambda = \frac{\sigma^2 \omega_b^3}{3P_s} = \frac{\omega_b^2 P_b}{3P_s} = \frac{\omega_b^2}{3R^2} .$$

A.5 Small Noise Errors in Multisample Spectrum Analysis

Refer to (1) and Fig. 13(a). Let $(\omega_\sigma, \beta_\sigma) = \omega_s$ and the midsample phase β_s of sample σ. With differential phase coherence,

$$\beta_\sigma - \beta_{\sigma-1} = (\omega_\sigma + \omega_{\sigma-1})(T/2).$$

Let n_σ, m_σ = noise contributions to observed ω_σ, $(2/T)\ \beta_\sigma$.
Define x_σ, y_σ, z_σ and note the relation to errors:

$$x_\sigma = \omega_\sigma + n_\sigma, \qquad y_\sigma = (2/T)\beta_\sigma + m_\sigma$$

$$z_\sigma = (x_\sigma + x_{\sigma-1}) - (y_\sigma - y_{\sigma-1}) = (n_\sigma + n_{\sigma-1}) - (m_\sigma - m_{\sigma-1}).$$

Let $\omega_\sigma + \epsilon$ = the following estimate of ω_σ :

$$\omega_\sigma + \epsilon = x_\sigma - \sum_{j=-\infty}^{+\infty} Q_j z_j .$$

Let σ_n^2 = Ave n_σ^2, σ_m^2 = Ave m_σ^2

$$\text{Ave } \epsilon^2 = \left[1 - 2(Q_\sigma + Q_{\sigma+1}) + \sum_j (Q_j + Q_{j+1})^2\right] \sigma_n^2 + \left[\sum_j (Q_j - Q_{j+1})^2\right] \sigma_m^2.$$

Choose the Q_j's for min. Ave ϵ^2 by the calculus of variations.
Compare with Ave ϵ^2 for x_σ alone, which is σ_n^2.

$$\frac{\text{Min Ave } \epsilon^2 \text{ of sum}}{\text{Ave } \epsilon^2 \text{ of } x_\sigma \text{ alone}} = \frac{\sigma_m}{\sigma_n + \sigma_m} .$$

Further analysis like that of Sections A.2 and A.3 gives

$$\sigma_n^2 = 3\sigma_m^2$$

$$\frac{\text{Min Ave } \epsilon^2 \text{ of sum}}{\text{Ave } \epsilon^2 \text{ of } x_\sigma \text{ alone}} = \frac{1}{1 + \sqrt{3}} \text{ or } - 4.365 \text{ db.}$$

A.6 Small Noise Errors in Multisample Frequency Detection of PAM-FM

Refer to Section A.4 but assume only a piecewise constant signal frequency. Refer to Figs. 13(a) and (15).

Let $w(t)$ = filter weight factor, referred to the output sample time.
Assume $w(\pm\infty) = 0$.

$$\text{Filtered error} = \int_{-\infty}^{+\infty} w(t)\,\dot{\varphi}(t)dt = -\int_{-\infty}^{+\infty} \dot{w}(t)\,\varphi(t)dl.$$

Use $\varphi(t) = [n_b(t)/\sqrt{2P_s}]$ and a white noise approximation.

$$\text{Ave } \epsilon^2 = \frac{\sigma^2}{\sqrt{2P_s}} \int_{-\infty}^{+\infty} [\dot{w}(t)]^2 dt.$$

Find $w(t)$, which gives

(a) normalized response to constant frequency in sample σ,

(b) zero response to constant frequencies in samples other than σ,

(c) minimum Ave ϵ^2 within constraints a, b.

The calculus of variations makes $w(t)$ quadratic over each sample interval and continuous at the boundaries. Then Ave ϵ^2 is a quadratic sum of the boundary values. Minimizing the boundary values is like minimizing the coefficients Q_j in Section A.5 (with $\sigma_n^2 = 3\sigma_m^2$) and gives the same result.

A.7 Blocking Probability in Sample-by-Sample Spectrum Analysis

Refer to Sections 3.3 and A.3. Approximate $N_1(\omega_k)$, $N_2(\omega_k)$ by processes piecewise constant over η subintervals.

Approximate $\sqrt{2P_s}\,F(\omega_s - \omega_k)$ by $\sqrt{2P_s}\,F(0)$ over the subinterval s and zero elsewhere.

Let x_λ, y_λ = the components of the signal-plus-noise spectrum, scaled (normalized) to unit variances. The probability densities are:

$$D_s = \frac{1}{2\pi} \exp\left[-\frac{(x_s - R)^2 + y_s^2}{2}\right],$$

$$D_\lambda = \frac{1}{2\pi} \exp\left(-\frac{x_\lambda^2 + y_\lambda^2}{2}\right), \qquad \lambda \neq s.$$

Phase Coherent Sample-by-Sample Spectrum Analysis. Rotation of the x_λ, x_s axes through $\pi/4$ gives quickly

$$P\{x_\lambda > x_s \mid \lambda \neq s\} = \frac{1}{2}\left[1 - \text{Erf}\left(\frac{R}{\sqrt{2}}\right)\right],$$

$$\text{Erf}(r) = \sqrt{\frac{2}{\pi}} \int_0^r \exp\left(-\frac{u^2}{2}\right) du.$$

This is the probability of a specific $x_\lambda > x_s$, out of $\eta - 1$ x_λ's, $\lambda \neq s$.

Under the conditions assumed here, the probability of any one or more is:

$$P \approx (\eta - 1)P\{x_\lambda > x_s\} = \frac{\eta - 1}{2}\left[1 - \operatorname{Erf}\left(\frac{R}{\sqrt{2}}\right)\right]$$
$$\approx \frac{\eta - 1}{\sqrt{\pi}R} \exp\left(-\frac{R^2}{4}\right).$$

Per Section 3.3, use

$$\eta = \frac{r\omega_b}{2\omega_b} = \frac{r}{2}.$$

Phase Incoherent Sample-by-Sample Spectrum Analysis

$$P\{(x_\lambda^2 + y_\lambda^2) > (x_s^2 + y_s^2)\} = \frac{1}{2} \exp\left(-\frac{R^2}{4}\right).$$

Under the conditions assumed here, for one or more λ's, $\lambda \neq s$:

$$P \approx \frac{\eta - 1}{2} \exp\left(-\frac{R^2}{4}\right), \qquad \text{use } n = r/2 \text{ as before.}$$

The last approximation is here a simplification, not a necessity. For an exact formula (given D_s, D_λ as above) see Ref. 7 or 8.

A.8 Reduction of the Blocking Rate of Spectrum Analysis

Refer to Section A.5. Use x, y of A.5. For a second estimation of ω_s,

$$\omega_s = (1/T)(\beta_{s+1} - \beta_{s-1}) - \tfrac{1}{2}(\omega_{s+1} + \omega_{s-1})$$
$$\omega_s + \epsilon = \tfrac{1}{2}(y_{s+1} - y_{s-1}) - \tfrac{1}{2}(x_{s+1} + x_{s-1}) + (2\nu\pi/T)$$
$$\nu = \text{unknown integer due to phase ambiguities.}$$

Refer to Fig. 16(b). Find the integers ν for the best fits to frequencies of the two peaks in the signal-plus-noise spectrum.

With no weighting for the heights of peaks, the probability that the closest is the correct choice is of the order of 0.9 (under the system conditions assumed here).

The actual choice must use also the relative heights of the peaks.

Let $P_M(M_s, M_n)$ = the probability density of the maxima M_s, M_n at the peaks due to signal-plus-noise and noise only (respectively).

Let $P_\epsilon(\epsilon_s, \epsilon_n)$ = the probability density of the observed deviations ϵ_s, ϵ_n of the second ω_s from the location of the peaks, using best ν's.

Use subscripts 1, 2 for the M's and ϵ's before the identification of which peak is signal-plus-noise and which is noise only.

The best identification corresponds to the larger of

$$P_M(M_1, M_2)\, P_\epsilon(\epsilon_1, \epsilon_2) \quad \text{and} \quad P_M(M_2, M_1)\, P_\epsilon(\epsilon_2, \epsilon_1).$$

P_M gives a strong weighting except when M_2 is close to M_1.

But when $M_n > M_s$, the difference is usually small, and P_n only very rarely gives a strong weighting to a wrong choice.

$$\text{Let } u = M_s - M_n. \text{ Then } \frac{P(u)}{P(-u)} = e^{Ru}.$$

A complete calculation of the probability of a correct choice would require integration over a complicated portion of the 4-dimensional space of M_s, M_n, ϵ_s, ϵ_n.

A.9 Some Circuit Considerations

A few circuit considerations are described below in brief, purely qualitative, terms.

Synchronization of Spectrum Analysis to PAM-FM Samples. Assume the following: The spectra represent signal-plus-noise received during intervals locally selected by a precision oscillator or clock. The length T of the intervals is almost right, without synchronizing means. The problem is to synchronize the start time to the start times of the true signal samples.

Synchronizing signals might be obtained by any of several means. One uses a very narrow band transmission channel, to send synchronizing signals from the transmitter. Others derive synchronization error signals from the communication signal itself, which must fluctuate sufficiently to supply the necessary information. (When the true signal is constant from sample to sample, there is nothing to indicate the boundaries between samples.) An error in synchronization reduces the height of the peak in the signal spectrum (on the average). It also produces a discrepancy between values of ω_s obtained from the single sample spectrum and by the second method described in Section A.5. In principle at least, a synchronization error signal can be derived from either effect and can be averaged over many sample intervals to reduce the effects of noise on the synchronization.

Shape of the Signal Sample. In (3), the tails of the function F are neither small nor short. By Section A.1, they can be reduced by shaping the envelope $E(t)$ of the signal sample before forming its spectrum. A suitable filter in the output of the spectrum generator has the same effect. Since the best spectral maximum corresponds to the F of (3), a practical compromise is needed. The pulse shaping problem is an old one, but here intersample interference due to the tails is not the important problem, but rather the way the tails can increase the blocking probability (noise-plus-tails exceeding signal-plus-noise).

Channel Bandwidth. For both ordinary FM and PAM-FM the channel bandwidth must be a little wider than the full excursion, ω_r, of the instantaneous signal frequency ω_s. The so-called Carson's Rule calls for a channel width of $\omega_r + 2\omega_b$ for ordinary FM, and the appropriate rule for PAM-FM is at least not very different. FMFB and PAM-FM spectrum analysis can tolerate wider bands without significant changes in thresholds and small noise errors.

Transition Intervals. In idealized models of spectrum analysis, certain operations happen in zero time. In any actual circuits there will be nonzero switching times. Very roughly, if a fraction α of each sample interval is lost due to the switching times, the signal power must be increased by factor $1/(1 - \alpha)$. Thus 2 per cent lost time requires roughly 0.1 db more power. In a sense, switching times are spectrum analysis counterparts of feedback stability problems in FMFB, although the comparison is purely qualitative.

REFERENCES

1. Jordan, D. B., Greenberg, H., Eldredge, E. E., and Serniuk, W., Multiple Frequency Shift Teletype Systems, Proc. IRE, **43,** November, 1955, pp. 1647–1665.
2. Klauder, J. R., Price, A. C., Darlington, S., and Albersheim, W. J., The Theory and Design of Chirp Radars, B.S.T.J., **39,** July, 1960, p. 745.
3. Feldman, C. B., and Bennett, W. R., Bandwidth and Transmission Performance, B.S.T.J., **28,** July, 1949, p. 490.
4. Turin, G. L., An Introduction to Matched Filters, Trans. IRE-PGIT, IT-**6,** June, 1960, p. 311.
5. Kotel'nikov, V. A., *The Theory of Optimum Noise Immunity*, translated by R. A. Silverman, McGraw-Hill, 1959.
6. Blackman, R. B., Bode, H. W., and Shannon, C. E., Data Smoothing and Prediction in Fire-Control Systems, Summary Technical Report of Div. 7, NDRC **1,** Report Series No. 13, MGC 12/2, National Military Establishment Research and Development Board.
7. Turin, G. L., Probability of Error in NOMAC Systems, Lincoln Lab. Tech. Report No. 57, January, 1954.
8. Reiger, S., Error Rates in Data Transmission, Proc. IRE, **46,** May, 1958, pp. 919–920.
9. Nuttall, A. H., Error Probabilities for Equicorrelated *M*-ary Signals Under Phase-Coherent and Phase-Incoherent Reception, IRE Trans. on Info. Theory, IT-**8,** July, 1962, pp. 305–314.
10. Slepian, D., The Threshold Effect in Modulation Systems that Expand Bandwidth, IRE Trans. on Info. Theory, IT-**8,** No. 5, September, 1962, pp. 122–127.
11. B.S.T.J., issue devoted to the *Telstar* Experiment, **42,** July, 1963.
12. Enloe, L. H., Decreasing the Threshold in FM by Frequency Feedback, Proc. IRE, **50,** January, 1962, pp. 18–30.
13. Sunde, E. D., Ideal Binary Pulse Transmission by AM and FM, B.S.T.J., **38,** November, 1959, pp. 1357–1426.
14. Turin, G. L., The Asymptotic Behavior of Ideal *M*-ary Systems, Proc. IRE, **47,** January, 1959, pp. 93–94.
15. Golay, M. J. E., Letter to the Editor, Proc. IRE, **37,** September, 1949, p. 103.
16. Rice, S. O., Mathematical Analysis of Random Noise, B.S.T.J., **23,** July, 1944, p. 282; and **24,** January, 1945, p. 46.

Data Transmission over a Self-Contained Error Detection and Retransmission Channel

By F. E. FROEHLICH and R. R. ANDERSON

(Manuscript received December 10, 1962)

Error control of the detection and retransmission type requires an internal storage buffer when the data source cannot be stopped. With finite capacity there will be occasions when this internal buffer is overfilled. This paper investigates the relationships among the error statistics of the channel, the storage capacity of the buffer, the round-trip transmission delay and the bit rate from the source. It is shown that the process can be treated as a Markov chain. The solution algorithm is programmed for machine computation, and representative cases are solved numerically. For typical values selected from the telephone plant, it is found that buffer capacities of a few hundred bits would be adequate.

The technique described should be useful for solving other problems in queueing theory.

I. INTRODUCTION

Studies during the last few years have shown that in the transmission of digital data over telephone lines, high accuracy can be achieved when the message is encoded in an error detecting code. Correction can then be accomplished by a repeat transmission of the portion of the information containing the errors. These so-called "feedback" techniques have been shown to be very effective in controlling errors.[1,2,3,4]

For some sources of data it is inconvenient or impossible to have the source wait while previous data are being retransmitted. There are also cases where it is required that the output from the receiver be at a uniform rate. This memorandum describes a self-contained error detection and retransmission channel capable of accepting data from the source at a steady rate, or at any rate less than a specified maximum, and of delivering it to the sink at this same rate. The channel is "self-contained," meaning that the channel itself provides enough storage of information to permit the detection of errors and their correction by retransmission without the data source and sink being aware that these processes are going on. The data source merely puts data into the transmission system at its own rate, and the data sink accepts highly reliable data from the system at the same rate. The relationships among system delay time, error probability, bit rate, and storage capacity are investigated.

The use of feedback error control with a data source which cannot be interrupted was briefly discussed by Reiffen, Schmidt, and Yudkin.[5] A. B. Fontaine has simulated such a system on a computer, using error data collected on private wire circuits.[6] Our analysis has indicated that shorter blocks could well have been used in the experimental simulation, which would have reduced the required storage capacity or increased the time to overflow.

II. THE DATA CHANNEL

A block diagram of the self-contained data channel is shown in Fig. 1. The transmitter consists of a buffer store, an encoder, a modulator, a reverse channel receiver and some logic. The transmission channel itself has a forward path and a reverse path, the latter carrying very little information compared to the former. The receiver consists of a detector, a decoder, a buffer store, and a reverse channel transmitter plus logic.

The forward channel carries data (plus any necessary redundancy and starting codes); the reverse channel carries information indicating whether retransmission is required. Errors in the reverse channel will not appreciably affect the operation. The small amount of information required over this channel permits a high degree of redundancy. In addition, a "fail-safe" code can be used, so that any undetected errors on the reverse channel result in unnecessary retransmissions (subsequently eliminated at the receiver) to ensure against loss of data.

To facilitate discussion, a specific model, chosen for its relative simplicity, is described. Modifications and improvements are apparent and will be briefly discussed. The method of operation is to accept data from the source continuously at a constant rate, R_S bits per second, which is less than the maximum rate, R_L, allowed by the data transmission system. The efficiency then, without considering the error-detecting code, is

$$E = R_S/R_L . \qquad (1)$$

The data are transmitted at an effective rate of R_S until a retransmission is requested. After a retransmission request, data are sent at the higher rate, R_L, until the system is returned to normal.

The change in rate could be made by switching the transmitting speed of the data set. Another method to achieve the data rate change is continuous transmission at rate R_L with interspersed dummy or "fill-in" bits as needed. The two methods are mathematically equivalent, and we shall assume the latter for the discussion in this paper. Thus, in the transmitting buffer the data are organized into blocks of N bits each and sent to the encoder at a rate, R_L, faster than the maximum allowable input rate. In order to equalize the input and output rates of the buffer, "fill-in" bits containing no information are inserted between the blocks of message bits as shown in Fig. 2. The data then pass through the encoder, where additional redundancy is added to allow for error detection. At the encoder, one may either ignore the fill-in bits or encode them, but will probably use them to transmit additional useful information. It is of course possible to place the error control encoder before the buffer, but this increases the required buffer size without gaining any apparent advantages. The signal is then modulated for transmission over the forward path.

Each block of information is retained in the transmitting buffer until it is certain that there will be no retransmission request from the receiver. When a sufficient time interval has elapsed and no retransmission request is received, the block of data is erased from the transmitting buffer. This time interval is taken to be T_D, the maximum round-trip delay for which the system is designed. This includes the transmission time in both directions plus any additional time for logical operations at either end.

The system as described has a sort of natural block length, the number of bits emitted by the source at rate R_S in time T_D

$$N = R_S T_D . \tag{2a}$$

With this block size, it is known that a retransmission request must apply to the immediately preceding block of data bits.

It is shown later that shorter blocks have an advantage in reducing the required buffer size, and hence we let

$$N = R_S T_D / k \tag{2b}$$

where k is an integer. For these shorter blocks, the system must assume a maximum T_D or must include some provision for determining the actual round-trip delay time so that retransmission requests can be associated with the proper blocks of data.

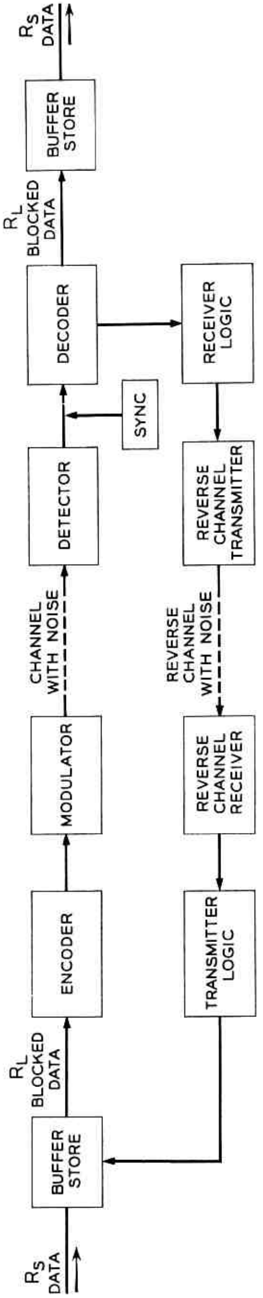


Fig. 1 — Complete self-contained error control channel.

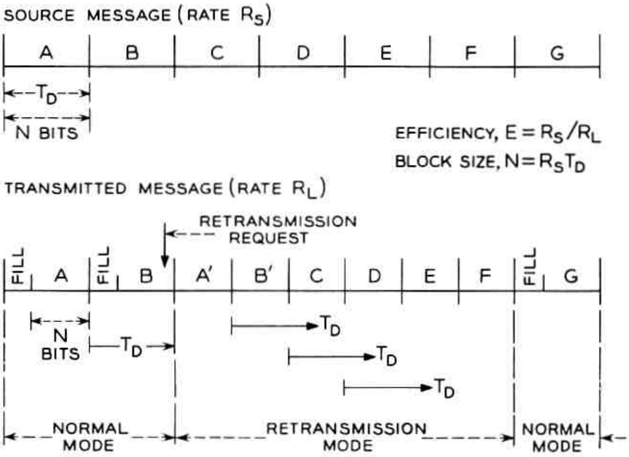


Fig. 2 — Example of time sequence at transmitter.

The number of bits, including both data and fill-in, from the buffer in the same time T_D/k is

$$N + M = R_L T_D/k. \tag{3}$$

In the receiver the demodulated signal is decoded and checked for errors. If no errors are found, the data block, with all redundancy removed, is placed in the receiving buffer. In case an error is detected in the received block of data, a retransmission request is sent to the transmitter via the reverse data channel, and no data are sent to the receiving buffer.

In the transmitter we impose the operating rule: in case a retransmission request is observed, the transmitter will complete the transmission of the current block of $N + M$ bits and then revert to the beginning of the block detected to be in error.† The transmitter then enters the retransmission mode and retransmits information starting with the block in error. During this period, the transmitting buffer continues to receive and store data from the source, thus increasing the quantity of information stored. In order to return the transmitting buffer to its normal state, the fill-in bits are now omitted between the transmitted blocks of data, so that bits will be removed faster than they arrive. This reduces the

† Another way to say this is that the transmitter takes no action on a retransmission request until the end of a full round-trip delay time, T_D , after sending the last bit of the block to be retransmitted. In this form the statement is also true when the transmitter is already in the retransmission mode. Note that in the latter case the time of decision is not necessarily at the end of a block.

information stored in the transmitting buffer and at the same time tends to refill the receiving buffer. The fill-in bits are omitted until both buffers have returned to their normal state.

The above sequence is illustrated in Fig. 2. Block A has been received in error. The retransmission request is noted by the transmitter before the completion of block B. At the conclusion of block B transmission, both A and B are retransmitted. Fill-in bits are now omitted until such time as the transmitting buffer returns to its normal state. This occurs after transmission of block F, if there are no additional retransmission requests.

We note immediately that, in case a number of nearby data blocks are found to be in error, the transmitting buffer may overflow. Similarly, the receiving buffer may empty out, so that for some time no information will be available to the data sink. The frequency of occurrence of these events depends, of course, on the error statistics of the channel, the storage capacity of the buffers, the round-trip transmission delay, the number of fill-in bits allowed between data blocks, and the size of the data block.

Questions to be answered about the self-contained data channel are: How often does the transmitting buffer store overflow and the receiving buffer empty completely? What delay is encountered by the information prior to delivery to the sink? What efficiency can this system achieve? What buffer store capacity is needed? In general, what are the relationships between buffer store size, block length, transmission efficiency, transmission delay, and average time between overflows, in any given message?

III. THE MARKOV PROCESS

In the following development, it will be assumed that retransmission requests are independent with probability P_r. For digital data transmission over telephone lines, individual bit errors are known to be not independent; however, for blocks which are long with respect to the bit error dependence, the retransmission requests will be nearly independent. There is some evidence that over voice telephone circuits at 1000–2000 bits per second the correlation among bit errors becomes so small after 10–15 bit intervals that the assumption of block error independence is acceptable.[1] An estimate of the probability of a retransmission is available, since the block error rate cannot be greater than the bit error rate times the block length.†

† Let λ be the bit error rate in B bits. Then λB is the number of bits in error. The number of blocks in error cannot be greater than λB. The total number of blocks is B/N so an upper limit of probability of block error is

We shall now devote ourselves to the question of the relationship between the storage capacity of the buffer and the average time between overflow of the buffer. It is evident that, since the number of data bits transmitted per unit time is not constant, an actual time calculation is inconvenient. We therefore quantize time into unequal units, such that the number of data bits transmitted per quantum is always the same.

The possible number of bits stored in the buffer form the states of a stochastic process. It will now be shown that, if these are considered only at certain moments of decision, the buffer states, y, form a finite Markov chain.

The only time a decision is made is exactly T_D seconds after the last bit of a block has been transmitted, and the decision consists of three parts:

(a) Which block shall be transmitted?

(b) Shall fill-in bits be transmitted following the data block?

(c) May the transmitting buffer erase a block of data?

The decision depends only on the state of the buffer and on whether a retransmission is requested; there are four cases:

(i) Normal — The system is not in the retransmission mode, and retransmission is not required. The buffer erases one block; the transmitter sends fill-in bits and then the next block in sequence from the source. By the time of the next decision, the buffer will have replaced the erased block with one block from the source. Thus, at the moment of the next decision, the total change in the buffer storage is zero. The time to the next decision is T_D/k.

(ii) The system is not in the retransmission mode, but a retransmission is requested. The buffer does not drop any bits. The transmitter backs up to the block at the beginning of the buffer in order to retransmit the block received in error. The transmitter shifts its mode and no fill-in bits are sent. The next decision will be made after one block has been completely transmitted plus T_D seconds, to allow time for another retransmission request to be received. During the retransmission time, EN bits come

$$\frac{\lambda B}{B/N} = \lambda N.$$

There may be multiple bit errors in a block, and some of the block errors may not be detected, so

$$P_r \leqq \lambda N. \tag{4a}$$

For the special case where bit errors are independent

$$P_r = 1 - (1 - \lambda)^N \doteqdot \lambda N. \tag{4b}$$

for λ much smaller than 1.

from the source, and during T_D, R_ST_D bits. The total increase in storage due to one retransmission is thus

$$I = R_ST_D + EN = R_ST_D(1 + E/k). \tag{5}$$

The time to the next decision is $T_D(1 + E/k)$.

(*iii*) *Off-Normal — The system has previously entered the retransmission mode and no additional retransmission is requested.* The buffer can drop the block which was received correctly. The transmitter continues with the block following the one just sent, without fill-in bits. The next decision will take place after the time required to transmit one block, in which time EN bits are added to the buffer. Since the buffer has dropped a full block, the amount of data in the buffer has decreased by

$$D = N(1 - E). \tag{6}$$

The time to the next decision is T_DE/k.

(*iv*) *The system is in the retransmission mode and another retransmission is required.* This is similar to case (*ii*), except that the transmitter shift is not required since it is already in the retransmission mode. The same number of bits will be discarded at the receiver, but, being already in the retransmission mode, none of these are fill-in bits, so the number of blocks to be retransmitted is greater by the ratio $(N + M)/N$. The transmitter remains in the retransmission mode and fill-in bits will not be sent. The increase in storage is given by (5), and the time to the next decision is $T_D(1 + E/k)$.

Let C be the total storage capacity of the transmitting buffer. When the source rate is constant, the transmitter can send the block as it is received. In this case, the smallest useful capacity, $C_{\min}$, includes the one block to which the retransmission request applies, if received, plus the data which arrive from the source during the round-trip delay preceding the request

$$C_{\min} = N + R_ST_D. \tag{7}$$

If the source rate may fluctuate and the start of transmission must be delayed, C must be larger. The worst case is that in which the source may intermittently stop so the transmitter must wait until the full block is received, in which case the minimum C is one block more. This additional block of storage to compensate for an intermittent source should probably not be charged to the error control system. The ability to provide this feature in a simple manner is, however, an advantage of the system.

There is another meaning for $C_{\min}$. In the normal mode of operation there must be just this many bits in storage at each time of decision. In setting up the Markov states below, we do not count this irreducible storage, but it is included in the final results for total storage capacity.

We have defined the state of the buffer, y, as the number of bits stored at any instant of decision. With a capacity of C bits, the range of this variable is

$$0 \leqq y \leqq C + 1. \tag{8}$$

The normal state is $y = 0$; overflow is $y = C + 1$.

We can now write down the transition probabilities, p_{ij}, of going from buffer state y_i to state y_j. Starting in the zero or normal state, the buffer stays in the normal state with probability $1 - P_r$ and increases by I with probability P_r

$$p_{0,0} = 1 - P_r, \qquad p_{0,I} = P_r. \tag{9a}$$

If the buffer is within D states of normal, at the next decision it will either return to normal or will increase by I

$$p_{y,0} = 1 - P_r, \qquad p_{y,y+I} = P_r \quad \text{for} \quad 0 \leqq y \leqq D \tag{9b}$$

If the buffer is more than D states from normal and more than I states from overflow, it will decrease by D or increase by I, but can neither return to normal nor overflow

$$p_{y,y-D} = 1 - P_r, \quad p_{y,y+I} = P_r \quad \text{for} \quad D < y \leqq C - I. \tag{9c}$$

If the buffer is within I states of overflow, the buffer will either decrease by D or go to overflow

$$p_{y,y-D} = 1 - P_r, \quad p_{y,C+1} = P_r \quad \text{for} \quad C - I < y < C + 1. \tag{9d}$$

In order to calculate the time to overflow, we force the buffer to stay in the overflow condition once it enters this state; i.e., the overflow state is made "absorbing"

$$p_{C+1,\ C+1} = 1. \tag{9e}$$

For all other transitions $p_{ij} = 0$. The transition matrix is

$$T = \{p_{ij}\}. \tag{9f}$$

In addition, we let the process start in the normal state with probability 1. The buffer state, in response to the retransmission signal, depends only on the buffer state at the previous moment of decision. This is the fundamental property for a process to be a Markov chain.

A schematic representation of the Markov chain described by equations (9) is given in Fig. 3. The over-all operation of the transmitter may be seen in Fig. 4, which shows the internal state diagram of a sequential machine which might be used to implement the transmitter. The states of the sequential machine are the same as the states of the Markov process, except that several of the latter may map into a single one of the former.

The arrow labels — A,B/C,D — are identified as follows. In all cases, a dash means the item is immaterial.

Transmitter inputs:

A — Has a retransmission request been received?

0 — no 1 — yes

B — What is the state of the buffer?

0 — empty (except for $C_{\min}$)

I — partially filled

1 — over-filled

Transmitter outputs:

C — May a block be dropped from storage?

0 — no 1 — yes

D — Which block shall be sent next?

D_{n-1} was the block which was just sent. D_n is the next block in sequence, and D_{n-m} is the mth block before.

F_1 and F_2 are fill-in bits. Note that if $F_1 = F_2$, two states may be combined.

Fig. 4 also applies to the receiver, except for reinterpretation of the labels.

Receiver inputs:

A — Has an error been detected?

0 — no 1 — yes

B — State of receiving buffer

0 — full

I — intermediate

1 — empty

Receiver outputs:

C — Shall this block be sent to output store?

0 — no 1 — yes

D — Shall a retransmission request be sent?

These will all be 0 except the two labelled D_{n-m} and D_{n-m-1}, which will be 1.

For certain relations among the quantities involved, the matrix can be partitioned into several closed sets[7] of states, such that it is not possible to make the transition from a state in any one closed set to a state in

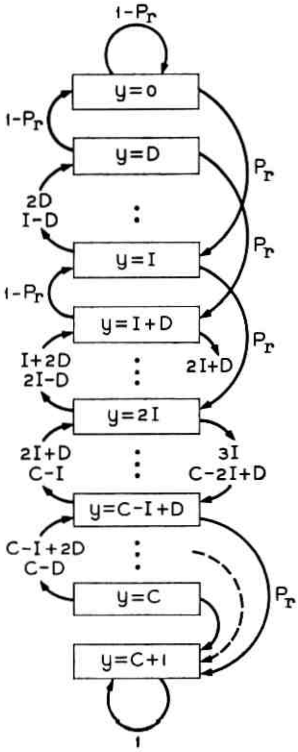


Fig. 3 — Markov state diagram.

any other such set. The states which cannot be entered from the normal state by any path may be removed from the matrix, thus reducing its size. This can be done by dividing out the greatest common factor in D, I, N, and C. A large number of the cases of interest are still included when this "normalizing factor" is made equal to D.

IV. CALCULATIONS

Following the method outlined in Kemeny and Snell,[8] we let Q be the transition matrix of all the transient states, i.e., matrix T, excluding the overflow state. Let J be the identity matrix. Then

$$G = (J - Q)^{-1} \tag{10}$$

exists and is called the fundamental matrix of the Markov process, with the following interpretation. Each element n_{ij} of G is the mean number

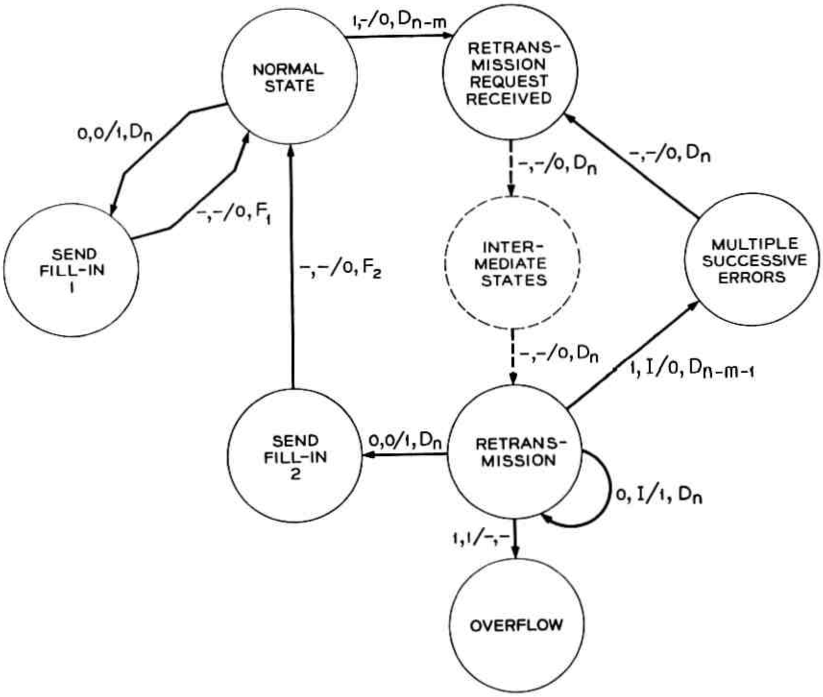


Fig. 4 — Diagram of internal states for transmitter.

of times the process is in state j, given that it started in state i. With $i = 0$ for starting in the normal state, the row sum over j is the mean number of times the process is in any of the transient states, from which we can calculate the mean time to the first overflow. Thus, the average number of decisions before overflow is

$$\langle n \rangle = \sum_{j=0}^{c} n_{0j} \, . \tag{11}$$

Higher moments, in particular the second, can be found by additional operations on the fundamental matrix.[7]

A computer program was written to do the matrix arithmetic, and a few representative cases were solved numerically. The program computes the average number of blocks transmitted before overflow and the variance about this mean. The standard deviation is usually large, nearly equal to the mean. Typical examples are: when mean number of blocks before overflow was 23, standard deviation was 19; when mean was 949, standard deviation was 943; and when mean was 4795, standard deviation was 4792. Thus the mean is a poor estimate of the actual time to overflow for any specific message, but is meaningful when a large number of transmissions are considered.

The calculations to this point have been in terms of the number of blocks, and we now convert back to actual time. Instead of a straight sum on n_{0j}, we multiply each term by the actual time taken.

There are four terms corresponding to the four cases described under the Markov process. The average time for each of the four cases is

$$(i)\ n_{00}(1 - P_r)T_D/k$$

$$(ii)\ n_{00}P_r(1 + E/k)T_D$$

$$(iii)\ \sum_{j=1}^{C} n_{0j}(1 - P_r)ET_D/k$$

$$(iv)\ \left[\sum_{j=1}^{C} n_{0j} - 1\right] P_r(1 + E/k)T_D .$$

The average time to overflow is the sum of these four:

$$\frac{t_{\text{ave}}}{T_D} = n_{00}\left(P_r + \frac{1 - P_r}{k} + \frac{EP_r}{k}\right) + \sum_{j=1}^{C} n_{0j}(P_r + E/k) - P_r(1 + E/k). \tag{12}$$

V. RESULTS

As expected, the average time before the buffer overflows will increase when the buffer capacity is increased, and when the following variables are decreased: the bit rate, the round-trip delay, the probability of retransmission, the efficiency, and the block size. The number of variables can be reduced by measuring time in units of T_D, the round-trip delay, and bits in units of R_LT_D, the number of bits from the buffer in time T_D. Since the block error rate depends on the length of the block, the probability of retransmission is modified by the block length. The variables of the system, all of which are now dimensionless, become

$$C^* = C/R_LT_D$$

$$N^* = N/R_LT_D$$

$$E$$

$$P^* = P_rR_LT_D/N$$

$$t^* = t/T_D$$

A number of curves are plotted to show the expected time between overflows as a function of the probability of retransmission. For each curve, the size of the buffer, the block size, and the efficiency are held constant. When the expected time, t/T_D, is greater than about 100 (corresponding to several seconds of transmission for reasonable values of T_D), the curves are nearly linear on log-log paper, and only this portion is plotted.

To use the curves, it is assumed that the transmission line parameters, R_L and T_D, are known. In order to facilitate interpretation of the curves, some reasonable specific values have been assigned to these parameters and the corresponding values of time, buffer size, and block length have been calculated. The assignments are as follows: Let R_L be 2000 bits per second; this could be a 2400 bps data set with an $83\frac{1}{3}$ per cent efficient error-detecting code. Let T_D be 120 ms. Then $R_LT_D = 240$ bits, the total number of bits sent in one round-trip delay time. Some other parameters are given in Table I.

Fig. 5 shows the time gained by increasing the capacity of the buffer store. For this set of curves the efficiency is 0.5 and the block length is 0.5 R_LT_D; that is, the block is as long as the maximum round-trip delay. When the efficiency is increased to 0.75 and 0.9, with the same block length (0.5 R_LT_D), the results are as shown in Figs. 6 and 7, respectively. The storage capacity required to provide a specified time to overflow at a given probability of retransmission increases markedly with efficiency. The same effect is shown in Fig. 8, where the capacity is held constant for several efficiencies. The source bit rate at $E = 0.75$ is 50 per cent greater than at $E = 0.5$, and at $E = 0.9$ the bit rate is up by 80 per cent. The cost of this increased bit rate is either the extra buffer storage or the reduced time between overflows. Some of the data from Figs. 5–8 are

TABLE I — OPERATING PARAMETERS
(Given that $R_L = 2000\ b/s$ and $T_D = 120$ msec)

E	R_S (bits/sec)	N (bits)	I (bits)	$C_{\min}$ (bits)
0.5	1000	20 120	130 180	140 240
0.75	1500	20 120 180	195 270 315	200 300 360
0.9	1800	20 120 216	234 324 410	236 336 432

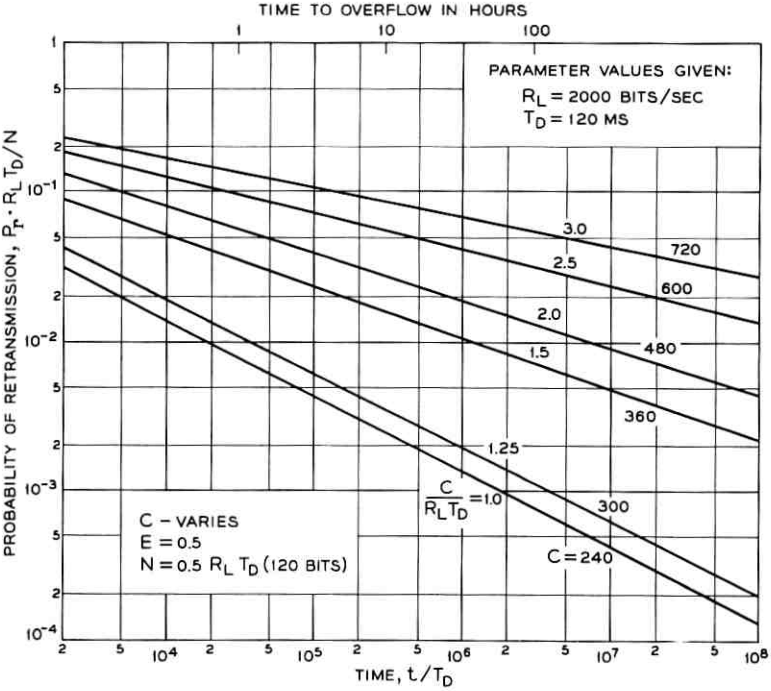


Fig. 5 — Effect of buffer size: $E = 0.5$.

shown in Table II, using the arbitrary assignments $R_L = 2000$ b/s, $T_D = 120$ ms, and $N = 120$ bits.

In all the above cases, the block lengths have been the same, $0.5\ R_LT_D$ (120 bits). Only when the efficiency is 0.5 does this represent the so-called "natural" block, i.e., the number of bits from the source in one round-trip delay time; at the increased bit rates of the higher efficiencies, the natural block length is also increased. The effect of increasing the block length in one case is shown in Fig. 9, which can be compared to Fig. 6. The required capacity for a given time to overflow has increased markedly. We therefore investigate the effects of shorter blocks.

Fig. 10 illustrates the case where each natural block is divided into three shorter blocks. A decision is made at the end of each arrow, and the fourth block back is either dropped from the buffer or is retransmitted. For example, when a retransmission is received while sending B_3, both A_1 and A_2 have been dropped and A_3 is the next block to be sent. With sufficiently inexpensive logic in the terminals, improved per-

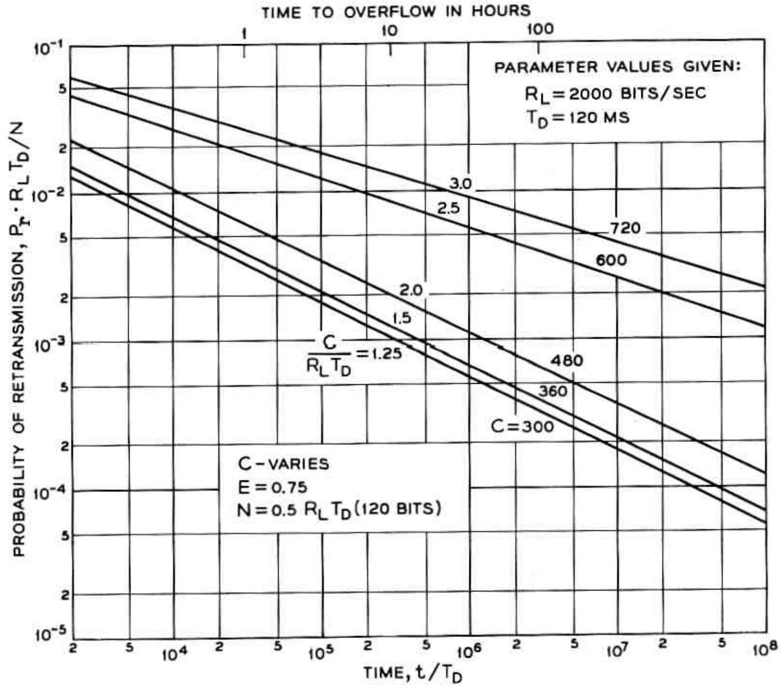


Fig. 6 — Effect of buffer size: $E = 0.75$.

formance is possible on short loops by using the actual value of T_D. In the example, we might have already dropped A_3 and therefore start the retransmission with B_1.

In Fig. 11 we show the effect of decreasing the block size, at constant capacity and efficiency. Similar results for a larger capacity and efficiencies of 0.5 and 0.75 are shown in Fig. 12.

It is somewhat difficult to visualize all of these effects when plotted separately. We attempt to summarize some of the results in Fig. 13. For these curves the normalized retransmission probability, P^*, is held constant, and buffer storage capacity is held to the minimum usable value, as given by (7); that is, the capacity is the natural block length plus the actual block length, and therefore decreases with either the block size or the efficiency. Both the latter are allowed to vary and we show the effect on the time to overflow.

There is little effect from changing the block size — except on the buffer capacity. One would therefore choose the smallest practical block.

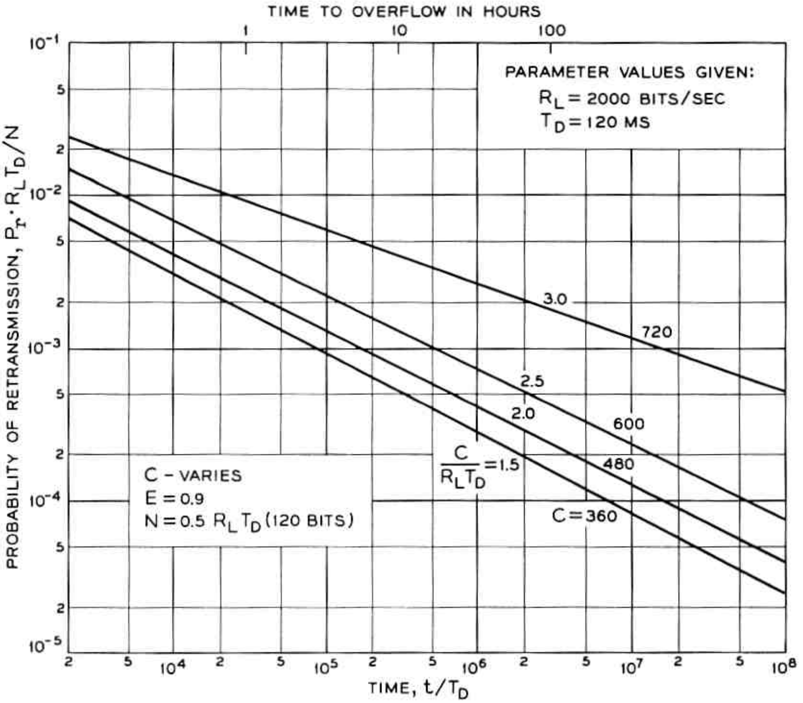


Fig. 7 — Effect of buffer size: $E = 0.90$.

However, as the efficiency is increased, the required buffer capacity is increased, although not rapidly, and the time between overflows decreases. As shown earlier (Figs. 5–7, 9) it is possible to regain this loss in time to overflow by modest increases in buffer capacity over the minimum used here. Since the increased efficiency increases the maximum source rate, this is certainly the direction to go, up to the point where the increased rate is worth less than the cost of the additional storage required.

VI. DELAY

For smooth flow to the sink the receiving buffer must have the same capacity as the transmitting buffer, and will normally be kept full. Thus the receiving buffer will introduce a delay in the message of

$$\tau = C/R_S . \tag{14}$$

This is in addition to the delay of $T_D/2$ from the transmission line.

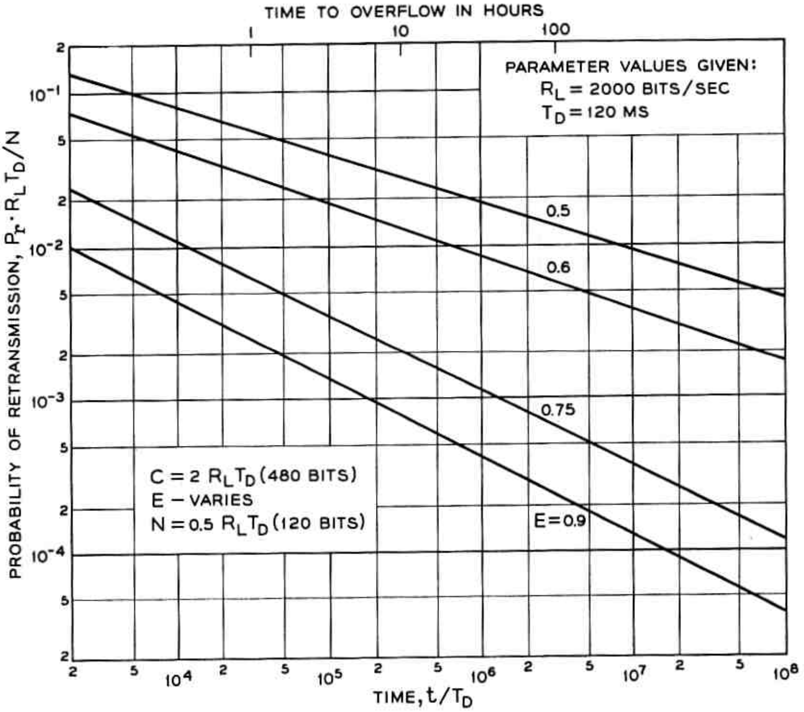


Fig. 8 — Effect of efficiency: buffer capacity fixed.

TABLE II — MEAN TIME TO OVERFLOW
(Given that R_L = 2000 b/s, T_D = 120 msec, N = 120 bits)

$E = R_S/R_L$	C (bits)	Ave. Time to Overflow (Hours)	
		$P^* = 0.01$	$P^* = 0.001$
0.5	$C_{\min}$ (240)	0.67	66.6
0.75	$C_{\min}$ (300)	0.12	11.2
0.9	$C_{\min}$ (336)	0.03	2.90
0.5	360	44.4	>1 year
0.75	360	0.15	14.9
0.9	360	0.04	3.14
0.5	480	245.3	>1 year
0.75	480	0.42	44.1
0.9	480	0.06	5.32
0.5	600	>1 year	>1 year
0.75	600	6.29	>1 year
0.9	600	0.15	17.93

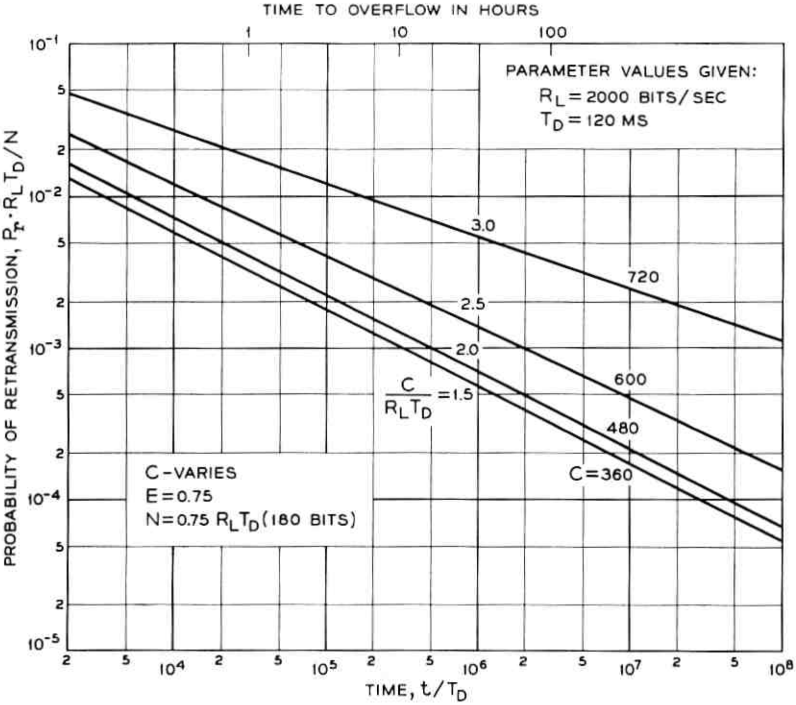


Fig. 9 — Effect of buffer size: longer block, $E = 0.75$.

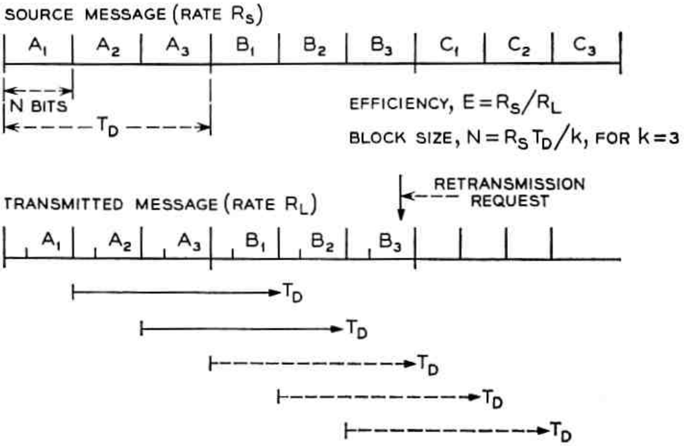

SOURCE MESSAGE (RATE R_S)
A_1 A_2 A_3 B_1 B_2 B_3 C_1 C_2 C_3
N BITS
T_D
EFFICIENCY, $E = R_S/R_L$
BLOCK SIZE, $N = R_S T_D / k$, FOR $k=3$
RETRANSMISSION REQUEST
TRANSMITTED MESSAGE (RATE R_L)
A_1 A_2 A_3 B_1 B_2 B_3
T_D
T_D
T_D
T_D
T_D

Fig. 10 — Example of time sequence with shorter blocks.

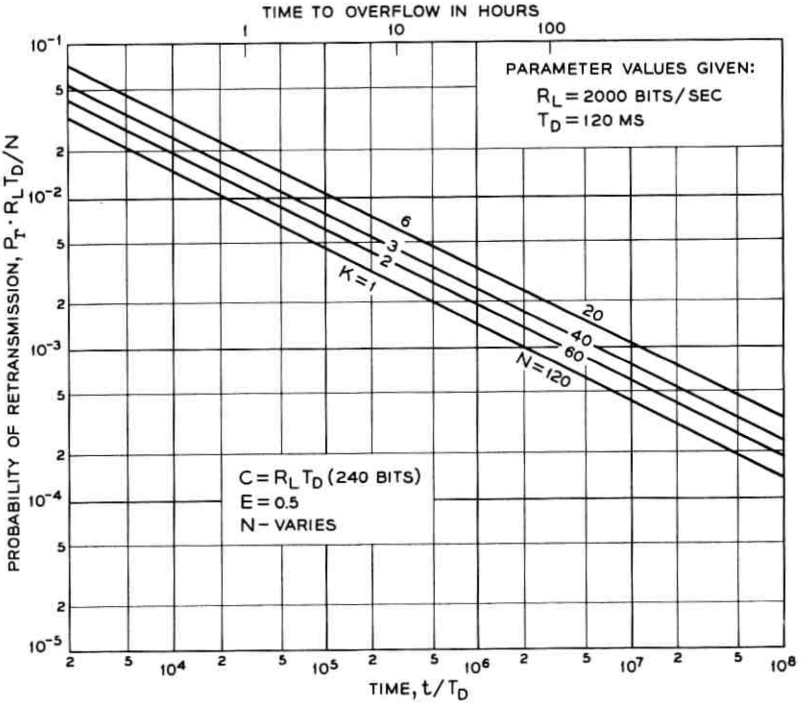


Fig. 11 — Effect of block size: fixed buffer capacity, $E = 0.5$.

There are other choices for operating the receiving buffer which will decrease the delay at the expense of irregularity of flow to the sink, which may be tolerable in many cases. If there were no receiving buffer at all, the delay would be zero except when retransmissions were required. When retransmissions are required, however, there would be additional delay until the block is received correctly, up to a maximum given by (14). The flow to the sink would not be smooth; each block would be delivered at rate R_L, followed by an interval when no data are being delivered. Various compromises between these extremes are possible. For example, buffer capacity of a single block would permit data to be delivered to the sink at the source rate with no interruptions until a retransmission is requested. Then the sink must alternately wait and accept data at the higher line rate until the process returns to normal. The delay is variable, the minimum being

$$\tau = N/R_S \tag{15}$$

with the maximum again given by (14).

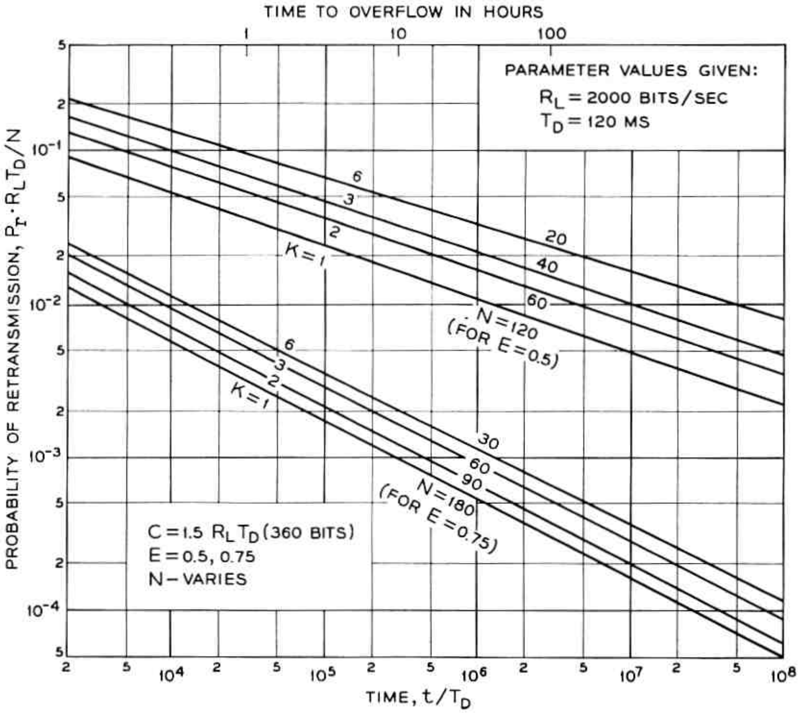


Fig. 12 — Effect of block size: fixed buffer capacity, E = 0.5 and 0.75.

This is one case where it has been possible to develop a calculable relationship between the message delay involved in error control and the resulting error rate.

VII. OTHER MODIFICATIONS

The system may be designed to take any of several actions when an overflow of the buffer occurs. The source and sink may be stopped, requiring manual resetting; they may be temporarily halted for a time sufficient for the system to clear; or, without stopping the source, the uncorrected data block may be delivered to the data sink, with or without an indication that the particular block contains errors.

One desirable modification would be to act sooner on receipt of the retransmission request. The transmitter would not continue to the end of the current block, but would immediately back up to the beginning of the block in error. This procedure could be quantized by using blocks a fraction of N in length. As indicated above, this procedure would require

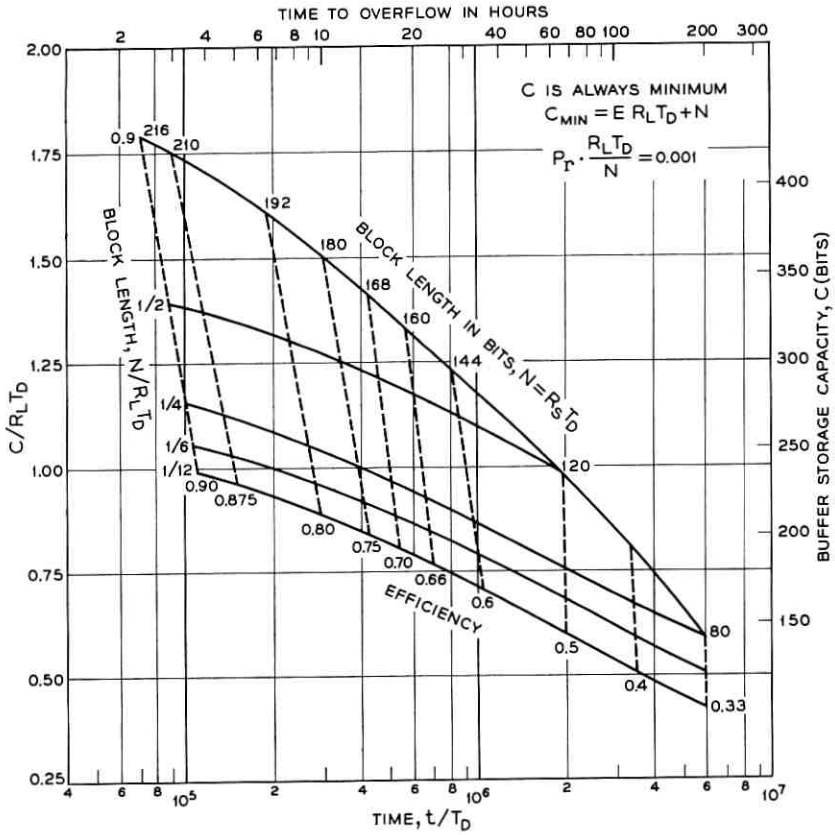


Fig. 13 — Capacity and time to overflow as functions of efficiency and block length.

either a knowledge of the actual round-trip delay or inclusion, in the retransmission requests and the retransmission, of an indication of the exact block (or fraction) involved. Another modification which would improve performance on shorter loops would be to make a preliminary measurement of the round-trip delay and adjust the operation accordingly. This could be done automatically.

Earlier, we mentioned the problem of an irregular input sequence and indicated that one additional block of storage was necessary. If this block is not counted, the performance level will be as given for a regular source, except for the possible gain arising from the probability of the intermittent source being stopped during the time when retransmissions are required. The output will be delayed an additional time corresponding to one block of data, but will be smoothed considerably — the rate will be constant except when waiting for the source.

It has been assumed that, once an error is detected, all subsequent received data are ignored until that block has been retransmitted and received correctly. With more complicated bookkeeping it would be possible to save some of these blocks, reducing the amount of retransmission required. On the other hand, since errors do occur in bursts on many transmission channels, the immediately succeeding block would have a higher-than-average error rate, and so might not be worth saving.

VIII. CONCLUSIONS

It has been shown that it is possible to calculate the performance of a self-contained error-control system by treating the system as a Markov process when the system consists of (*a*) an error-detection code, (*b*) provision for requesting and accomplishing retransmissions as necessary, and (*c*) buffer storage to allow smooth, uninterrupted flow from the source to the sink. Failure occurs when a sufficient number of retransmissions are requested in a short enough time that the total information to be stored exceeds the capacity of the buffer.

Whenever an overflow is about to occur, we could ignore the retransmission request and deliver the block as-is, in which case it appears to the sink as an error. It seems reasonable to require that this type of error should have about the same frequency of occurrence as undetected errors. For voice channels using reasonably simple codes, we might assume an undetected error rate of 10^{-8} or about one error per day.[1,3] We might also require the efficiency to be about that of the error detecting code.

With these criteria, it appears clear that one should not try to work with minimal storage, because of the relatively short time to overflow. Neither should one try to push the efficiency very high, or the required capacity grows out of bounds. A reasonable compromise for voice channels would be a buffer capacity somewhat less than 1000 bits.

We get a slightly different answer if we consider instrumentation. It is likely to be economically infeasible to build a buffer of this size with individual bit storage devices, especially since serial access is adequate. However, with bulk storage such as a circulating delay line or a magnetic tape loop, moderate increase in buffer size is not costly, and several thousand bits would be available about as cheaply as a few hundred. This would permit buffer efficiencies close to unity.

Results for any other specific cases can be easily calculated with this computer program. It is apparent that a number of modifications in the model are possible and would serve to reduce the required storage. The transition matrix would merely have to be changed to correspond to the new model; the matrix arithmetic would be the same.

The details of the chosen model and the examples were taken from a specific data transmission problem. The techniques, both the model and the method of solution, are applicable to a wider variety of problems where buffering is a consideration.

We should like to acknowledge the assistance of H. O. Burton in consultation on the mathematics of the Markov process. We appreciate the continued encouragement of G. W. Gilman, who suggested the use of feedback error control with a data source which cannot be interrupted.

REFERENCES

1. Bennett, W. R., and Froehlich, F. E., Some Results on the Effectiveness of Error-Control Procedures in Digital Data Transmission, I.R.E., Trans. Comm. Syst., **CS-9,** March, 1961, pp. 58–65.
2. Schwartz, L. S., Some Recent Developments in Digital Feedback Communication Systems, I.R.E. Trans. Comm. Syst., **CS-9,** March, 1961, pp. 51–57.
3. Cowell, W. R., and Burton, H. O., Computer Simulation of the Use of Group Codes with Retransmission on a Gilbert Burst Channel, Trans. A.I.E.E. (Comm. & Elect.), No. 58, January, 1962, pp. 577–585.
4. Brown, A. B., and Meyers, S. T., Evaluation of Some Error Correcting Methods Applicable to Digital Data Transmission, 1958 I.R.E. Natl. Conv. Record, Pt. 4, March, 1958, pp. 37–55.
5. Reiffen, B., Schmidt, W. G., and Yudkin, H. L., The Design of an Error-Free Data Transmission System for Telephone Circuits, Trans. A.I.E.E. (Comm. & Elect.), No. 55, July, 1961, pp. 224–231.
6. Fontaine, A. B., Queueing Characteristics of a Telephone Data Transmission System with Feedback, A.I.E.E. Conference Paper 62-1441, Fall General Meeting, October 9, 1962.
7. Feller, W., *An Introduction to Probability Theory and its Applications,* Vol. I, 2nd ed., John Wiley and Sons, New York, 1957.
8. Kemeny, J. G., and Snell, J. L., *Finite Markov Chains,* D. Van Nostrand Co., New York, 1960.

Intermodulation Distortion in Analog FM Troposcatter Systems

By E. D. SUNDE

(Manuscript received May 20, 1963)

In broadband transmission over troposcatter paths, selective fading will be encountered with resultant transmission impairments, depending on the modulation method. An analysis has been made in a companion paper of such selective fading, based on an idealized model of troposcatter paths. It indicated that selective fading will be accompanied by phase nonlinearity which in a first approximation can be regarded as quadratic over a narrow band. A probability distribution for such quadratic phase distortion was derived. On the premise of quadratic phase distortion, the error probability owing to selective fading was determined for digital transmission by various methods of carrier modulation.

The same idealized model and basic premise of quadratic phase distortion is used here to determine intermodulation distortion in FM for a signal with the statistical properties of random noise. An approximate expression for intermodulation noise owing to specified quadratic phase distortion has been derived, applying for any method of frequency pre-emphasis in FM. In turn, median intermodulation noise as well as the probability distribution of intermodulation noise has been determined, as related to certain basic system parameters.

A comparison is made of predicted with measured intermodulation noise in four troposcatter systems with lengths from 185 to 440 miles. The results indicate that phase nonlinearity owing to selective fading can be approximated by quadratic phase distortion, or linear delay distortion, over an appreciable part of the transmission band ordinarily considered for troposcatter systems, with a probability distribution that can be determined from certain basic parameters of troposcatter links, such as the length and antenna beam angles. However, to predict intermodulation distortion on any system, further experimental data than are now available are required on beam broadening by scatter.

The present random multipath FM distortion theory is shown to afford *a significant improvement over an equivalent single-echo theory that has been applied on an empirical basis to troposcatter systems.*

INTRODUCTION

An analysis has been made elsewhere[1] of error probabilities in high-speed digital transmission over idealized troposcatter paths, considering both random noise and intersymbol interference owing to pulse distortion caused by selective fading. The above analysis indicated that a principal cause of intersymbol interference is a quadratic component of phase distortion, or linear delay distortion. On the same basic premise an evaluation is made herein of intermodulation noise in analog transmission by frequency modulation, as now used for transmission of voice channels in frequency division multiplex. Expressions and curves are given of intermodulation noise in an idealized troposcatter channel for a signal with the properties of random noise, as related to certain basic system parameters and comparisons are made with the results of measurements on four troposcatter systems.[2,3]

In random multipath transmission the received wave can be considered the sum of a plurality of echoes, arriving over the various paths with varying amplitudes and different delays. Although this view is conceptually simple, it does not facilitate analysis of the statistical properties of the received signal and of signal distortion. In the combination of a number of time functions, such as echoes, the analysis is greatly facilitated by the use of Fourier transformation to determine the corresponding spectra. The latter can in turn be combined directly with appropriate attention to phase relations to obtain the resultant wave. For this reason it is preferable from the standpoint of analysis to regard the received wave as a multiplicity of sine wave components, rather than signal wave echoes, arriving over the plurality of transmission paths with varying amplitudes and phases. This is the method ordinarily used in the analysis of the statistical properties of narrow-band random noise, which has properties that with appropriate translation of the basic parameters are also applicable to random multipath transmission. It is the method underlying both the previous determination of error probabilities in digital transmission owing to noise and selective fading, and the present analysis of intermodulation noise in FM.

In certain radio systems the received wave can be considered the sum of a principal signal wave and a weaker echo, and comprehensive theoretical analyses have been published of intermodulation noise in FM owing to such echo distortion,[4,5,6] together with the results of simulative tests.[7] For these reasons this two-path model has been adopted as a coarse simile to multipath transmission in some interpretations of the result of measurements of intermodulation noise in troposcatter systems.[3] The limitations of this simile are recognized in the latter publication,[3] in which it is suggested that a more refined analysis is desirable. The idealized multipath model used in the analysis of troposcatter digital transmission affords a significant improvement, though it has certain predictable limitations, as shown herein.

I. TRANSMITTANCE PROPERTIES OF TROPOSCATTER LINKS

In tropospheric transmission beyond the horizon the received wave can be considered the sum of a large number of components of varying amplitudes resulting from a multiplicity of reflections within the common volume of the antennas. Owing to variations in the structure of the common volume, caused largely by winds, there will be relatively slow changes in the many reflections and thus in the amplitudes of the component waves. When a steady-state sine wave is transmitted, the received wave will thus exhibit random variations in its envelope and phase, known as fading.

In addition to such transmittance variations with time at a particular frequency, there will be transmittance variations with frequency at any given instant, as illustrated in Fig. 1. At a given instant the amplitude and phase characteristics of the transmission path may be as indicated in Fig. 1(a) and at a later instant as in Fig. 1(b).

Let $u = \omega - \omega_0$ represent the radian frequency relative to a reference frequency ω_0. When the transmission vs frequency characteristic of a troposcatter channel varies slowly with time t, it can be represented by

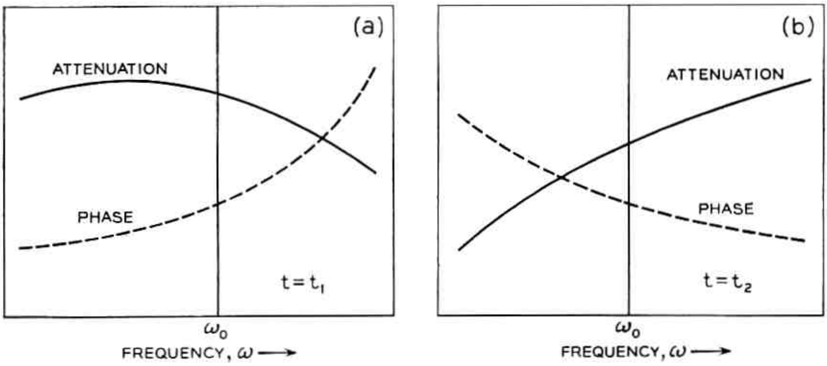


Fig. 1 — Illustrative variations in attenuation and phase characteristics with frequency at two instants t_1 and t_2.

$$T(u,t) = A(u,t)e^{-i\varphi(u,t)} \tag{1}$$

where

$A(u,t)$ = amplitude characteristic as a function of t for a fixed u, or as a function of u for a fixed time t

$\varphi(u,t)$ = phase characteristic.

If $u = u_0$ is fixed, both $A(u_0, t)$ and $\varphi(u_0, t)$ are random variables of the time t, as are the time derivatives $A'(u_0, t)$, $A''(u_0, t)$, $\varphi'(u_0, t)$, $\varphi''(u_0, t)$. The probability distributions of $A(u_0, t)$ and $\varphi(u_0, t)$ can be determined on the premise that they are the sum of a large number of randomly phased components. This results in a Rayleigh probability distribution of $A(u_0, t)$, in conformance with observations of rapid fading. To determine the probability distributions of A', A'', φ' and φ'', statistical information is required regarding the rapidity of fades. This ordinarily takes the form of the time autocorrelation functions of $A(t)$, or the related power spectrum of changes in transmittance amplitude. Such power spectra can be characterized by a certain equivalent fading bandwidth.

If the time is assumed fixed at $t = t_0$, then $A(u,t_0)$ and $\varphi(u,t_0)$ will have certain random fluctuations with the frequency u that can be characterized by probability distributions. This also applies to $\dot{A}(u,t_0)$, $\ddot{A}(u,t_0)$, $\dot{\varphi}(u,t_0)$, and $\ddot{\varphi}(u,t_0)$, where the dots indicate differentiation with respect to frequency u. The probability distributions of $\ddot{A}$, $\dot{\varphi}$, and $\dot{A}$ and $\ddot{\varphi}$ depend on the frequency autocorrelation functions, or the corresponding power spectra of variations with frequency. The latter depend on differences in transmission time over the various paths, and can be related to the maximum departure Δ from the mean transmission delay.

The amplitude and phase characteristics as a function of u at any time t_0 can in general be represented by a power series as

$$A(u,t_0) = a_0 + a_1u + a_2u^2 + a_3u^3 + \cdots \tag{2}$$

$$\varphi(u,t_0) = b_0 + b_1u + b_2u^2 + b_3u^3 + \cdots. \tag{3}$$

Certain basic relations have been developed by Carson and Fry[7] and by van der Pohl,[8] for transmission impairments in FM resulting from attenuation and phase distortion. With the aid of these relations it can be shown that intermodulation noise is caused principally by phase distortion rather than by amplitude distortion. Moreover, it can be shown that the principal contributor is quadratic phase distortion represented by b_2u^2, which corresponds to linear delay distortion $2b_2u$.

II. PROBABILITY DISTRIBUTION OF QUADRATIC PHASE DISTORTION

From (3) it follows that

$$\ddot{\varphi}(u,t_0) = 2b_2 + 6b_3u + \cdots. \tag{4}$$

For $u = 0$, i.e., at the reference or carrier frequency, the probability distribution of b_2 is the same as that of $\ddot{\varphi}(0,t)$. The latter probability distribution has been determined elsewhere[1] on the approximate premise of a linear variation in transmission delay, with maximum departures $\pm\Delta$ from the mean delay. In Fig. 2 is shown the probability that $\ddot{\varphi}$, or $2b_2$, exceeds $\Delta^2/3$ by a factor k. For example, there is a probability $p = 0.5$ that $\ddot{\varphi}$ exceeds $\Delta^2/3$ by a factor $k \approx 1.2$, and a probability $p = 0.1$ that $\ddot{\varphi}$ exceeds $\Delta^2/3$ by a factor $k \approx 19$.

Thus in general

$$\ddot{\varphi}_p = 2b_2(p) = k_p\Delta^2/3 \tag{5}$$

where $k_p\Delta^2/3$ is the value of $\ddot{\varphi}$, or $2b_2$ with a probability p of being exceeded.

Alternatively, the value of b_2 with a probability p of being exceeded is

$$b_2(p) = \frac{k_p}{6}\Delta^2. \tag{6}$$

Thus

$$b_2(0.5) \approx \frac{1.2}{6}\Delta^2 = 0.2\Delta^2 \tag{7}$$

$$b_2(0.1) \approx \frac{19}{6}\Delta^2 = 3.2\Delta^2 \tag{8}$$

$$b_2(0.01) \approx \frac{400}{6}\Delta^2 = 67\Delta^2. \tag{9}$$

Thus, when Δ is known, together with intermodulation noise for quadratic phase distortion, it is possible to determine the median value of average intermodulation noise, or the value exceeded with any other specified probability p.

III. INTERMODULATION NOISE FROM QUADRATIC PHASE DISTORTION

In a first-order evaluation of intermodulation noise, only the quadratic term b_2u^2 in (3) would be considered, since it will be the principal contributor. The ratio of nonlinear distortion power to average signal power at the frequency ω will depend on the signal properties and on the pre-

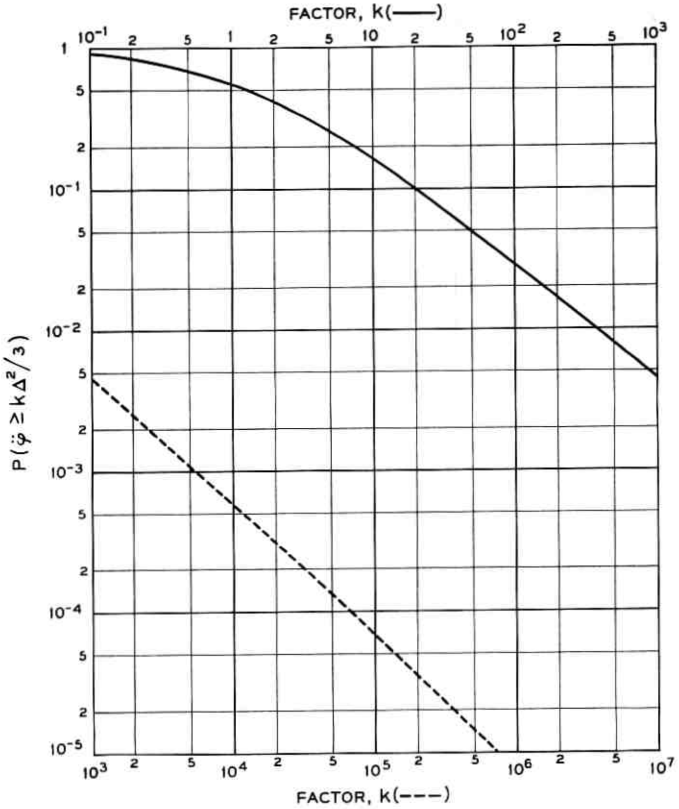


Fig. 2 — Probability that $\ddot{\varphi}$ or $2b_2$ exceeds $\Delta^2/3$ by a factor k.

emphasis used in frequency modulation. It will be assumed that the original message wave has a flat power spectrum of radian bandwidth $\Omega = 2\pi B$ and the statistical properties of random noise, and furthermore that the message wave is passed through a transmitting filter with a power transfer characteristic

$$\begin{aligned} t(\omega) &= 1 + c(\omega/\Omega)^2 \\ &= 1 + c(f/B)^2. \end{aligned} \tag{10}$$

At the receiving end a complementary filter is used to restore the message wave.

As discussed in the Appendix, exact determination of intermodulation noise from quadratic phase distortion presents formidable difficulties, except on the premise of slight phase distortion, which is not generally applicable to troposcatter systems. However, it is possible to obtain an approximate solution without the above limitation. The following relation is derived in the Appendix for the ratio $\rho(f)$ of intermodulation noise to average signal power at the frequency $f = \omega/2\pi$

$$\rho(f) = \frac{B^2}{D^2} G(c,a) H(\gamma) \tag{11}$$

where c is defined by (10)

$$\begin{aligned} a &= f/B = \omega/\Omega \\ B &= \text{bandwidth of baseband signal} = \Omega/2\pi \\ D &= \text{rms frequency deviation} = \underline{\Omega}/2\pi \end{aligned}$$

and

$$\gamma = b_2\underline{\Omega}^2 = (2\pi)^2 b_2 D^2. \tag{12}$$

The function $G(c,a)$ depends on the pre-emphasis and is given by expression (108) in the Appendix, which is

$$\begin{aligned} G(c,a) &= \frac{3a^2}{(1 + ca^2)(3 + c)} F(c,a) \\ F(c,a) &= 2 - a + \frac{2c + c^2a^2}{3} [1 + (1 - a)^3] \\ &\quad - \frac{c^2a}{2} [1 - (1 - a)^4] + \frac{c^5}{5} [1 + (1 - a)^5] \end{aligned} \tag{13}$$

This function is shown in Fig. 3 for pure FM and PM and for $c = 16$. The particular case of $c = 16$ and $a = 1$ will be considered further in the following, and for this case

$$G(16,1) = 0.192.$$

The function $H(\gamma)$ is shown in Fig. 4 and represents an approximation, as discussed in the Appendix. It will be noted that this function departs from proportionality with γ^2 for $\gamma \geqq 0.5$, reaches a certain maximum value and then diminishes.

IV. INTERMODULATION NOISE IN TROPOSCATTER PATHS

In accordance with (6), the value of b_2 with a probability p of being exceeded is $b_2(p) = k_p\Delta^2/6$. The corresponding value of γ is given by (12) as

$$\begin{aligned} \gamma_p &= \frac{k_p\Delta^2}{6} (2\pi)^2 D^2 \\ &= 6.6k_p(\Delta D)^2. \end{aligned} \tag{14}$$

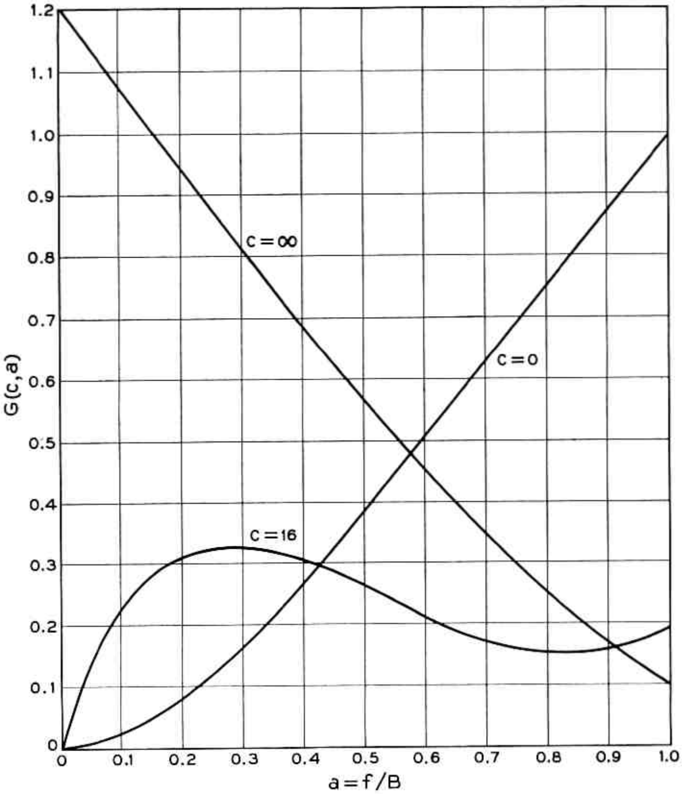


Fig. 3 — Function $G(c,a)$ for pure FM ($c = 0$), pure PM ($c = \infty$), and for pre-emphasized FM with $c = 16$.

Thus

$$\gamma_{0.5} \approx 8\ (\Delta D)^2 \tag{15}$$

$$\gamma_{0.1} \approx 125\ (\Delta D)^2 \tag{16}$$

$$\gamma_{0.01} = 2600\ (\Delta D^2). \tag{17}$$

The corresponding ratios $\rho(f)$ at $f = B$ with a probability p of being exceeded

$$\rho_p(B) = 0.192 \left(\frac{B}{D}\right)^2 H(\gamma_p) \tag{18}$$

$$\rho_{0.5}(B) = 0.192 \left(\frac{B}{D}\right)^2 H(8\Delta^2 D^2) \tag{19}$$

$$\rho_{0.1}(B) = 0.192 \left(\frac{B}{D}\right)^2 H(125\Delta^2D^2) \tag{20}$$

$$\rho_{0.01}(B) = 0.192 \left(\frac{B}{D}\right)^2 H(2600\Delta D^2). \tag{21}$$

V. DIFFERENTIAL TRANSMISSION DELAY Δ

Exact determination of the equivalent maximum departure from the mean transmission delay requires consideration of the antenna beam patterns as affected by scattering. On the approximate basis of equivalent antenna beam angles α, it follows from the geometry indicated in Fig. 5 that

$$\Delta \approx \frac{L}{v}\frac{\alpha + \beta}{2}\left(\theta + \frac{\alpha + \beta}{2}\right) \tag{22}$$

where $\beta \leqq \alpha$, v is the velocity of propagation in free space, L is the length of the link, and

$$\theta = \frac{L}{2R} = \frac{L}{2R_0K} \tag{23}$$

where R_0 is the radius of the earth and the factor K is ordinarily taken as 4/3.

The equivalent antenna beam angle α from midbeam to the 3-db loss point depends on the free-space beam angle α_0 and on the effect of scatter, which is related in a complex manner to α_0 and the length L, or alternatively θ. Narrow-beam antennas as now used in actual systems are loosely defined by $\alpha_0 \leqq 2\theta/3$. For these, $\alpha \approx \alpha_0$ on shorter links, while on longer links $\alpha > \alpha_0$ owing to beam-broadening by scatter. Analytical determination of α for longer links appears difficult, and only limited experimental data are available at present. For broad-beam antennas, $\alpha_0 \gg 2\theta/3$ and beam-broadening by scatter is in theory inappreciable.

By way of numerical example, let $L = 170$ miles and $K = 4/3$, in which case $\theta = 0.016$ radian. With $\alpha_0 = 0.004$ radian $\ll 2\theta/3$ it is permissible to take $\alpha = \alpha_0$. With $\beta = \alpha = \alpha_0$, (22) gives $\Delta = 0.08 \times 10^{-6}$ second.

The differential delay Δ in general varies with time and for narrow-beam antennas can be considered the sum of two components

$$\Delta(t) = \Delta_0 + \Delta_1(t) \tag{24}$$

where Δ_0 is a fixed component obtained from (22) by taking $\alpha = \alpha_0$, the free-space beam angles. The variable component $\Delta_1(t)$ depends on

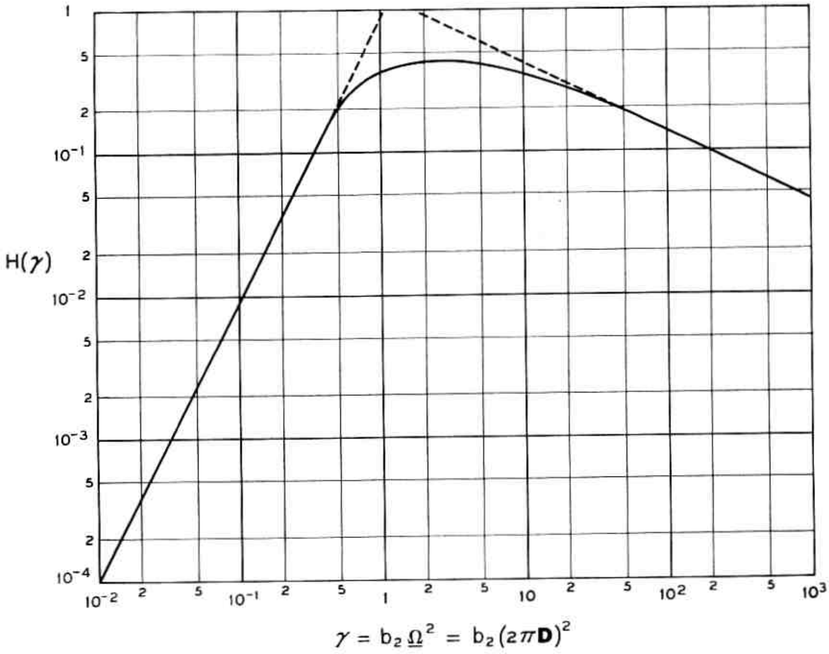


Fig. 4 — Function $H(\gamma)$. The parameter γ is the phase distortion in radians at a frequency corresponding to the rms frequency deviation $\underline{\Omega} = 2\pi D$ radians/second.

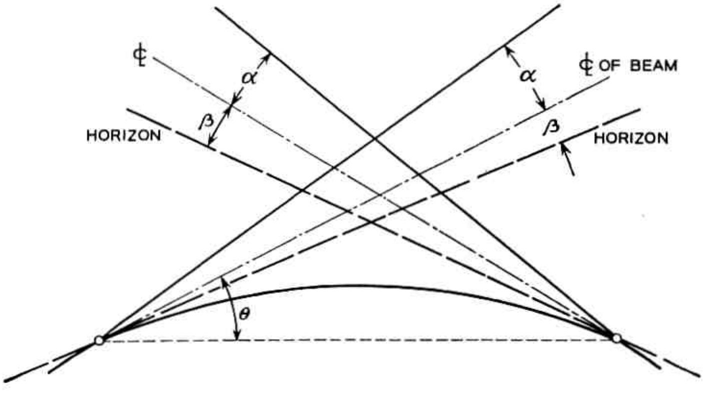


Fig. 5 — Definition of antenna beam angles α, take-off angle β and chord angle θ to midbeam. With different angles at the two ends, the mean angles are used in expressions for Δ.

scatter variation with time, as does path loss, and will have a certain correlation with path loss variations. Owing to the fixed component Δ_0 , a weaker correlation exists between $\Delta(t)$ and path loss variations.

Because of the dependence of Δ on path loss, the ratio ρ_p of intermodu lation noise to average signal power will depend somewhat on path loss However, for a given path loss ρ_p is independent of the average transmitter power and thus of the average signal power at the receiver.

VI. LIMITATIONS ON FIRST-ORDER DISTORTION THEORY

The above first-order approximation applies for sufficiently narrow signal bandwidths at the detector input such that terms in (3) of higher order than u^2 can be neglected. Results given by Rice for random variables (Section 3.4 of Ref. 10) indicate there is no correlation between $\ddot{\varphi}$ and $\dddot{\varphi}$, so that distortion owing to the term b_3u^3 will combine on a power addition basis with distortion resulting from b_2u^2. Moreover, there is a negative correlation factor between $\ddot{\varphi}$ and $\ddddot{\varphi}$, so that on the average b_4 is negative whenever b_2 is positive, and conversely. Hence distortion produced by b_4u^4 will on the average subtract directly on an amplitude basis from that resulting from b_2u^2. In the range where the function $H(\gamma)$ increases linearly with γ^2, intermodulation noise owing to the term b_2u^2 increases as $b_2^2(\Delta D)^4$. In the same range, intermodulation noise from the term b_4u^4 will vary as $b_4^2(\Delta D)^8$ and may hence have a significant effect for adequately large values of ΔD even though b_4 be much smaller than b_2 . As shown later, comparisons of measured intermodulation noise with predictions based on the above first-order theory indicate the increasing importance of the term b_4u^4 in reducing intermodulation noise as ΔD is increased.

VII. TWO-PATH VS MULTIPATH DISTORTION THEORY

The above first-order distortion theory is a mathematically derived approximation that in principle yields valid results with appropriate limitations on signal bandwidth and frequency deviation, and which retains the multipath feature that is essential to this end. By contrast, the two-path or single-echo simile mentioned in the introduction has no such basis but has been adopted principally because of the convenience of available theoretical analysis.[4,5,6] A second reason is that single-echo distortion theory yields results that in some respects are quite similar to those obtained with multipath transmission, as shown below.

It is noteworthy that, by proper choice of echo amplitude and delay, results similar to those for median quadratic phase distortion can be obtained. This is illustrated in Fig. 6, which shows the median ratio $\rho(B)$ obtained from (19) as a function of D for $B = 1$ mc/sec with $\Delta = 0.1$ and 0.5 microsecond. In the same figures are shown the ratios $\rho(B)$ obtained on the premise that the received wave consists of a main signal and an echo of equal amplitude delayed by 0.07 and 0.4 microsecond. The ratio $\rho(B)$ for the latter condition is obtained from a chart given in Fig. 9 of Ref. 3, applying for FM with virtually the same pre-

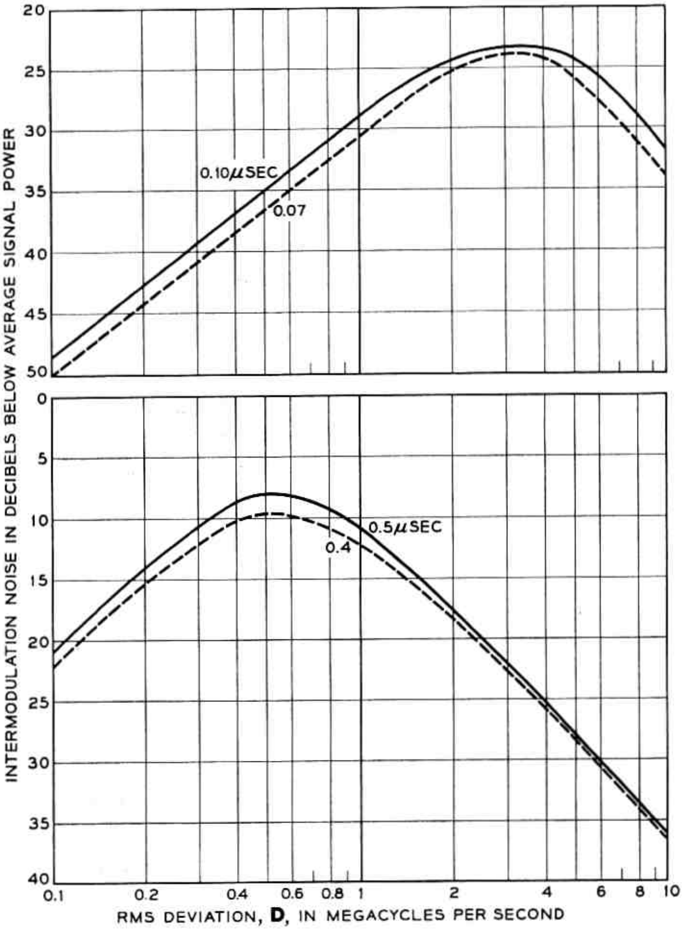


Fig. 6 — Comparison of intermodulation noise from single-echo distortion and quadratic phase distortion at $B = 1$ mc/sec: (solid lines) median intermodulation noise from quadratic phase distortion for indicated departures Δ from mean delay; (dashed lines) intermodulation noise from echo of same amplitude as signal with delays Δ_e as indicated.

emphasis as assumed herein and given by (10). The above charts are based on echo distortion theory applying for echoes that are much weaker than the signal, but this premise is ignored here in extending the theoretical results to a fictitious echo of the same amplitude as the signal. In this connection it may be noted that simulative tests[9] indicate that intermodulation noise is nearly proportional to echo amplitude, even when the latter equals the signal amplitude. With both quadratic phase distortion and single-echo distortion, intermodulation noise is virtually proportional to the second power of signal bandwidth. Hence, the relative comparisons in Fig. 6 could also apply for other bandwidths than $B = 1$ mc/sec.

The above comparisons indicate that in applying equivalent single-echo FM distortion theory to multipath transmission as in troposcatter systems, with physically tenable echo delays, certain dilemmas will be encountered. The theory could be extended beyond its validity to fictitious echoes of the same amplitude as the signal, to obtain virtually the same median intermodulation noise as for quadratic phase distortion. This would exclude the possibility of greater intermodulation noise than the median value, since the greater echo is by definition the main signal. The other procedure would be to assume an echo that is smaller than the main signal, which is physically more acceptable and does not violate the basic premise underlying echo distortion theory. In this case intermodulation noise predicted on the basis of echo distortion theory would, at least in certain cases, be much smaller than actually observed and could not be made to conform with observations, unless the echo amplitude is increased to the same amplitude as the signal.

Thus, if the ratio of echo amplitude to signal amplitude is r, intermodulation noise power based on single-echo theory will be less than for multipath transmission by a factor r^2. Hence it becomes necessary to introduce a factor $1/r^2$ to make single-echo theory applicable to multipath transmission. In Ref. 3, this factor has been determined empirically from measurements to be discussed later, and is given as 9 db.

VIII. OBSERVED MEDIAN INTERMODULATION NOISE

Measurements have been made on four troposcatter links of the median value of intermodulation noise at the frequency $f = B$. The modulating wave in these tests had a flat power spectrum, and pre-emphasis was used that closely corresponded to $c = 16$ in (10).

The basic parameters of the systems on which the measurements were made are given in Ref. 3 and are summarized in Table I. In this table α_0 is the free-space antenna beam angle from midbeam to the 3-db

Table I — Basic Parameters of Troposcatter Test Systems in Caribbean (A) and in Arctic (B,C,D)

System	A	B	C	D
Length, miles	185	194	340	440
Radio frequency, mc	725	900	900	800
Antenna/diameter, ft	60, 60	30, 60	120, 120	120, 120
α_0 (radian)	0.0115	0.017	0.0058	0.0058
θ (radian)	0.015	0.016	0.031	0.034
Δ_0 (microsecond)	0.12	0.21	0.185	0.255

loss point, which may not conform with the angle α in (22) when scatter is considered. The values of K and θ are taken from Ref. 3, and θ differs slightly from that obtained from (23) owing to differences in antenna elevations. The take-off angle β is virtually zero and has been neglected. The value Δ_0 of Δ given in the table was calculated with $\alpha = \alpha_0$, rather than the actual beam angle with scatter. Systems A, B, C and D correspond to paths 1, 2, 4 and 3 in Ref. 3.

In Figs. 7 and 8 are shown the ratios $\rho_{\frac{1}{2}}(B)$ expressed in db as a function of the rms frequency deviation D for different bandwidths B of the baseband signals.

IX. COMPARISON OF THEORETICAL WITH OBSERVED MEDIAN VALUES

In the same Figs. 7 and 8 are shown median values of intermodulation noise obtained from (19) for each case, based on values Δ_m of Δ that afford the best average approximation to the measurements. The latter values are somewhat greater than Δ_0, as indicated in Table II.

A ratio Δ_m/Δ_0 or $\alpha_m/\alpha_0 > 1$ is to be expected owing to beam-broadening by scatter, and the above ratios appear reasonable in the light of present knowledge. Thus, if the actual angles α were known so that Δ could be determined, it appears plausible that satisfactory conformance with observed intermodulation noise would be obtained.

As noted in Section V, Δ includes a component $\Delta_1(t)$ that varies with time depending on scatter conditions and which is correlated with path loss fluctuations. The ratio ρ thus depends on path loss as affected by scatter and has a certain correlation with path loss variation, as shown elsewhere.[3] Hence, if measurements had been made under different path loss conditions, the derived values Δ_m would have been somewhat different.

From Figs. 7 and 8 it will be noted that with the above choice of $\Delta = \Delta_m$ it is possible to obtain better agreement between predicted and observed intermodulation noise for small bandwidths B of the baseband

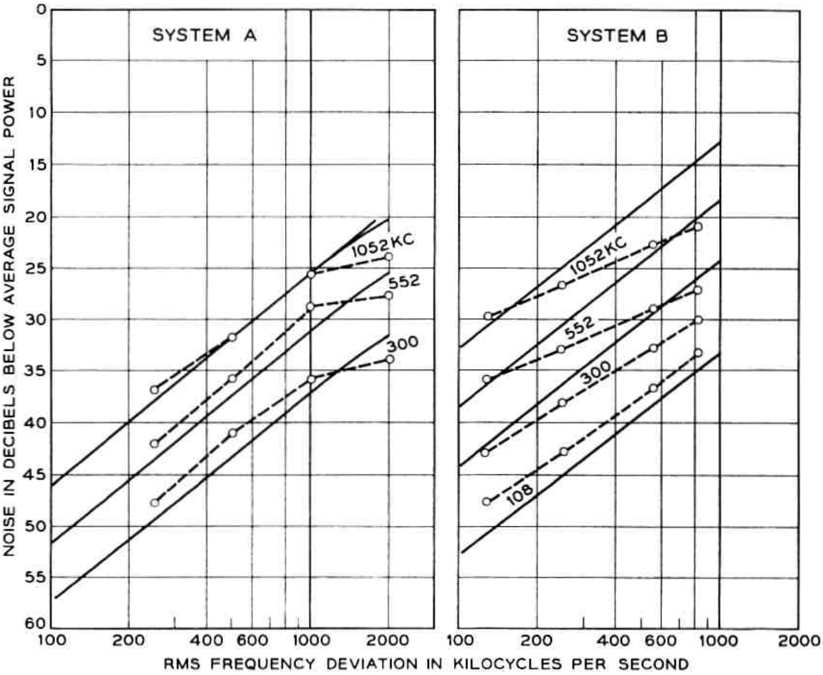


Fig. 7 — Comparison of measured and calculated median intermodulation noise: (dashed curves) measured median intermodulation noise in top channels at indicated frequencies in kc; (solid curves) calculated median intermodulation noise for idealized model with the following values of the equivalent maximum deviation Δ from the mean transmission delay: system A, $\Delta_m = 0.12$ microsecond ($\Delta_0 = 0.12$); system B, $\Delta_m = 0.25$ microsecond ($\Delta_0 = 0.12$).

signal and small deviations D than for large bandwidths and frequency deviations. This probably resides in the circumstance that the phase distortion terms of higher order than b_2u^2 have been neglected in the above first-order theory, as discussed in Section VI.

The measured median ratios given in Figs. 7 and 8 are plotted in Fig. 9 against the ratios predicted by first-order theory. It will be noted that measured intermodulation noise is less than predicted for signal-to-interference ratios less than about 30 db, owing to reduction in intermodulation noise by phase distortion of higher order than b_2u^2 that has been neglected in first-order theory. The results in Fig. 9 permit an approximate empirical correction to first-order theory.

As discussed in Section VII, with single-echo distortion theory virtually the same median intermodulation noise is obtained as with the above first-order theory, provided the echo is equal in amplitude to the

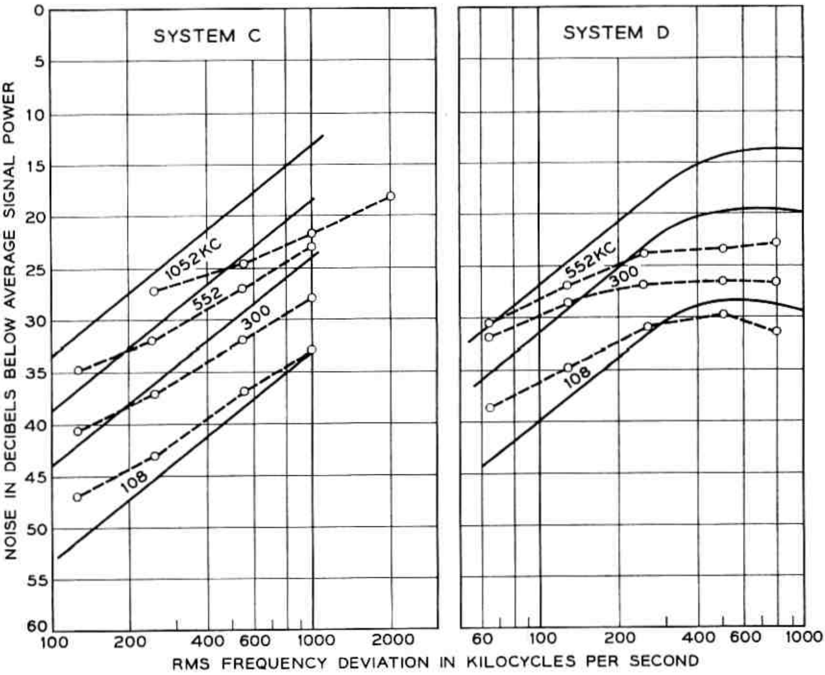


Fig. 8 — Comparison of measured and calculated median intermodulation noise: (dashed curves) measured median intermodulation noise in top channels at indicated frequencies in kc; (solid curves) calculated median intermodulation noise for idealized model with the following values of the equivalent maximum deviation Δ from the mean transmission delay: system C, $\Delta_m = 0.25$ microsecond ($\Delta_0 = 0.185$); system D, $\Delta_m = 0.55$ microsecond ($\Delta_0 = 0.255$).

mean signal. For smaller echoes, predicted intermodulation noise must be less. This conforms with results presented in Figs. 12 and 14 of Ref. 3, which show that intermodulation noise predicted from single-echo theory is significantly smaller than observed. To obtain a satisfactory average relation between predictions and observations, the predicted values must be increased by 9 db, as in Fig. 15 of Ref. 3

Table II — Ratio Δ_m/Δ_0

System	A	B	C	D
Length, miles	185	194	340	440
Δ_0 , microsecond	0.12	0.21	0.185	0.255
Δ_m , microsecond	0.12	0.25	0.25	0.55
Δ_m/Δ_0	1.0	1.2	1.35	2.15
α_m/α_0	1.0	1.1	1.35	2.15

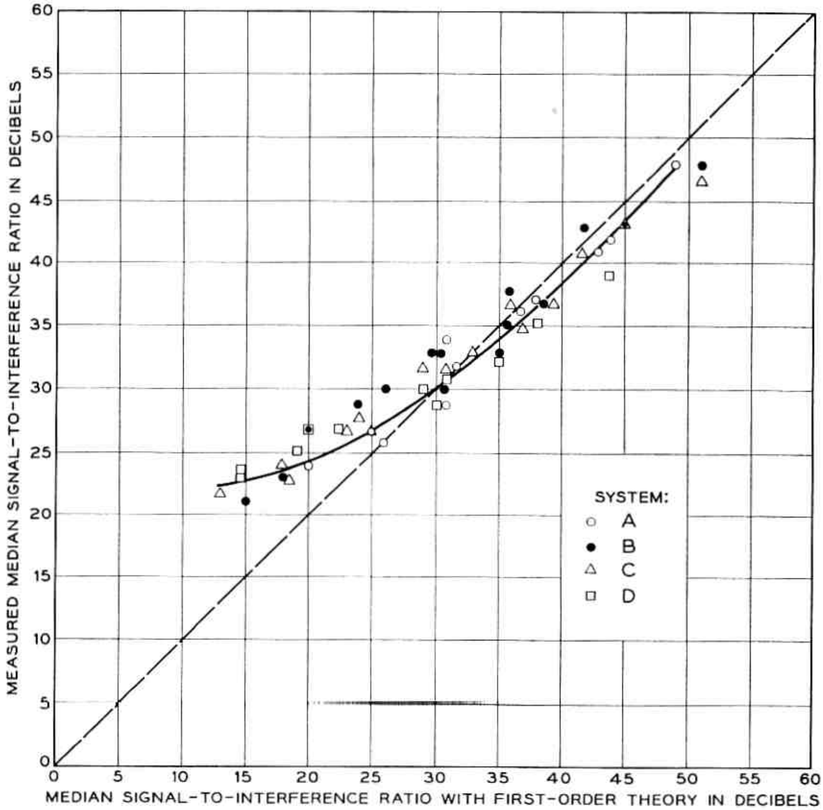


Fig. 9 — Comparison of measured median signal-to-interference ratios with median values based on first-order approximation with best choice of differential transmission delay Δ.

X. PROBABILITY DISTRIBUTION OF INTERMODULATION NOISE

From (18) it is apparent that the probability distribution of ρ is directly related to that of $H(\gamma_p)$. This function is shown in Fig. 10 as related to $(\Delta D)^2$ for $p = 0.5$, 0.1 and 0.01. It should be recognized that this function as given herein is approximate, and that the errors are likely to be greater for small values of p than for median intermodulation noise as considered previously.

From the curves in Fig. 10 it is possible to obtain approximate curves of the probability distribution of intermodulation noise, applying for various values of Δ^2D^2 as shown in Fig. 11. These curves show that the probability distributions vary markedly with the above parameter, in conformance with a few probability distributions derived from observa-

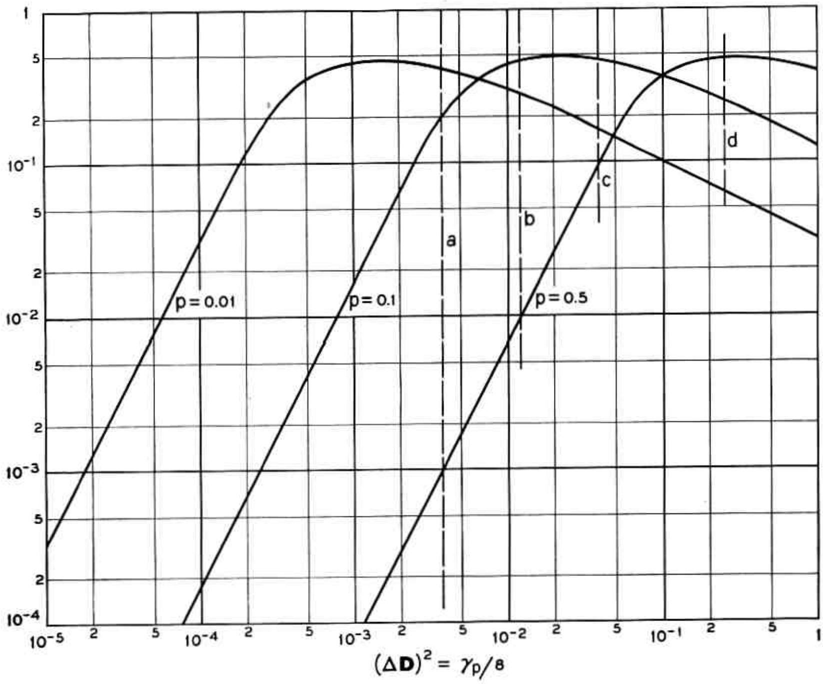


Fig. 10 — Function $H(\gamma_p)$ for various probabilities p.

tions.[2] Because of the approximations involved in the present first-order distortion theory, the above probability distribution curves should be considered illustrative and may not be accurate enough for certain engineering applications.

XI. PREDICTION OF INTERMODULATION DISTORTION

The present first-order intermodulation theory indicates that intermodulation distortion depends on the delay difference Δ, and this would apply also for an exact theory. For various troposcatter links with different angles α_0 and θ, intermodulation distortion would be the same for equal values of Δ. This is exemplified by comparison of intermodulation noise in systems B and C as shown in Figs. 7 and 8. Though these systems have different angles α_0 and θ, intermodulation noise is virtually the same since Δ is the same. Thus, if Δ could be determined, the above first-order theory, in conjunction with the above experimental data, would permit determination of intermodulation distortion for a variety

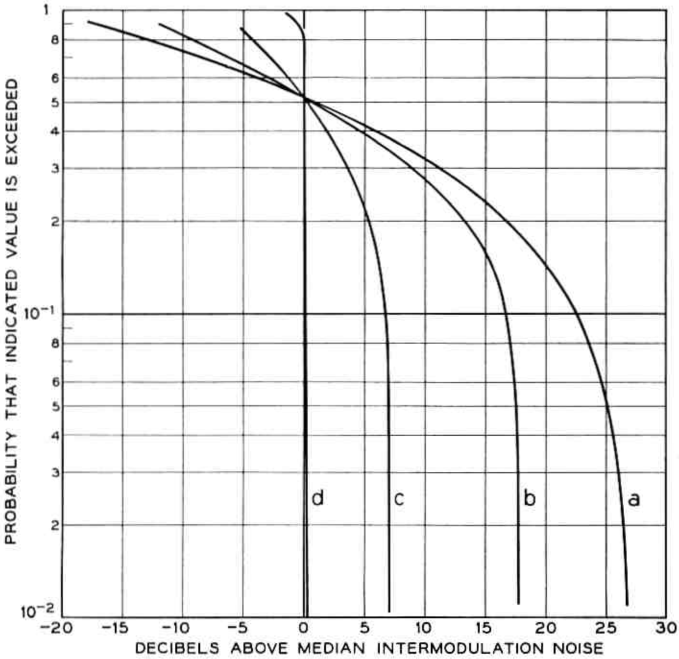


Fig. 11 — Probability distributions of intermodulation noise for various values of $(\Delta D)^2$ corresponding to dashed lines a, b, c and d in Fig. 10.

of conditions other than those in the tests. The above experimental data were confined to intermodulation noise in the top channel, i.e., for $a = 1$ in Fig. 3, and for a particular pre-emphasis, $c = 16$. The expression for $G(c,a)$, or the curves in Fig. 3, permit approximate determination of intermodulation noise at other frequencies, and also for other kinds of pre-emphasis. For example, for $a = 0.3$, intermodulation noise would be greater than for $a = 1$ by an approximate factor $0.32/0.19 \approx 1.7$. If pure FM$(c = 0)$ had been used in the tests, intermodulation noise at $a = 1$ would have been increased by an approximate factor $1/0.19 \approx 5.2$.

At present there is a principal obstacle to prediction of intermodulation distortion for other values of α_0 than in the above experimental systems. This is the lack of comprehensive experimental data on the beam angle α as affected by scatter for troposcatter links of various lengths. When and if such data become available, it will be possible to determine Δ and in turn intermodulation distortion in the manner indicated above for any kind of system.

XII. APPLICATION TO DIGITAL MULTIBAND TRANSMISSION

The distributions in Fig. 10 apply for average intermodulation noise over brief time intervals, as determined by changes in phase distortion with time. During each such interval the instantaneous amplitudes of intermodulation noise will fluctuate about the average value. For a signal with the properties of random noise, as considered here, the probability distribution of this fluctuation is approximated by the normal law. The distribution of instantaneous amplitudes or intermodulation noise is important in transmission by FM of a number of digital channels in frequency division multiplex, as discussed below.

In digital transmission over troposcatter paths, the error probability for a given signal-to-noise ratio of the receiver depends on the transmission rate, as discussed elsewhere.[1] As the transmission rate is increased, the error probability is ultimately determined by intersymbol interference owing to selective fading, and may be excessively high. The error probability can in this case be reduced, for a given total transmitter power, by transmitting at a slower rate over each of a number of narrower channels in frequency division multiplex. This could be accomplished by individual transmission over each channel, which would entail a number of independent transmitters. An alternative method would be to use a common amplifier and to transmit the combined digital signal by frequency modulation of a common carrier, as now used for transmission of voice frequency channels in frequency division multiplex. In the latter case, it is necessary to consider the possibility of additional transmission impairments owing to intermodulation noise.

With a sufficiently large number of digital channels in frequency division multiplex, the combined wave will have virtually a Gaussian amplitude distribution, like random noise. Hence the probability distribution of average intermodulation noise amplitudes would be as indicated in Fig. 11 for various conditions. The instantaneous amplitude will fluctuate with respect to the above average values, as noted in Section X.

In binary transmission it is often assumed that the error probability will not be excessive if the average noise power from all sources is about 12 db below the average signal power, or 18 db below the peak signal power in on-off binary pulse transmission. From the previous curves and expressions it appears that intermodulation noise power averaged over short intervals will be at least 10 db below the average signal power, with a small probability that it exceeds −15 db. It thus appears that intermodulation noise will not be a limiting or predominant factor even when a large number of binary channels are combined in frequency division multiplex for transmission by frequency modulation of a common carrier.

XIII. SUMMARY

In broadband transmission over troposcatter paths, selective fading will be encountered with resultant transmission impairments, depending on the modulation method. A previous analysis has been made of such selective fading, based on an idealized model of a troposcatter path. It indicated that selective fading will be accompanied by phase distortion that in a first approximation can be regarded as quadratic, and a probability distribution curve for such quadratic phase distortion was derived. On the premise of such quadratic phase distortion, the error probability owing to selective fading was determined for digital transmission by various methods of carrier modulation.

In the present study the same basic premise of quadratic phase distortion has been used in determining intermodulation distortion for a signal with the properties of random noise, based on the same idealization of a troposcatter path. An approximate relation for intermodulation noise owing to quadratic phase distortion has been derived, applying for any frequency pre-emphasis in FM. In turn, median intermodulation noise as well as the probability distribution of intermodulation noise has been determined, as related to certain basic system parameters.

Median intermodulation noise predicted on basis of free-space antenna beam angles conforms well with observations on links 185 and 194 miles in length. For links 340 and 440 miles long it is necessary to use antenna beam angles that are greater than the free-space angles by factors of about 1.35 and 2.15, respectively. On long links employing narrow-beam antennas, beam broadening is expected because of scatter. Thus if the beam angles had been determined by independent observations or by more elaborate theory, it is probable that predicted intermodulation noise would conform reasonably well with observations.

The results of intermodulation noise measurements thus appear to confirm the conclusion in a previous theoretical analysis of troposcatter transmittance, which indicated that phase distortion owing to selective fading could in a first approximation be represented by a component of quadratic phase distortion, with a probability distribution that can be determined from certain basic system parameters. This affords a simplified first-order theoretical model of selective fading in troposcatter paths that is applicable to evaluation of resultant transmission impairments in both analog and digital transmission.

It can be shown analytically, and it is confirmed by observations, that the above first-order distortion theory yields intermodulation noise that in the case of large signal bandwidths and frequency deviations will be greater than observed or obtained with a more exact distortion theory. An empirical curve presented here permits determination of the expected correction for large bandwidths and frequency deviations.

It has also been demonstrated that the first-order multipath distortion theory presented here affords a significant improvement over single-echo distortion theory applied to random multipath transmission, in that it is simpler and accounts for the probability distribution of intermodulation noise without certain contradictions that are inherent in single-echo theory. Taken in conjunction with presently available data on observed intermodulation noise on certain troposcatter links, as discussed herein, it affords a means of predicting intermodulation noise on any system when more comprehensive experimental data become available on antenna beam broadening by scatter.

APPENDIX

Intermodulation Noise from Quadratic Phase Distortion in Pre-Emphasized FM

General

To facilitate analysis of intermodulation noise in FM owing to attenuation and phase distortion, it is customary to introduce two basic approximations. One is the use of "quasistationary theory" in conjunction with the concept of instantaneous frequency, which is permissible when the signal bandwidth B is negligible in comparison with the carrier frequency, so that the frequency changes imperceptibly over a signal interval $T = 1/2B$. The other customary approximation is that distortion $\alpha(\omega) + i\beta(\omega)$ is sufficiently small to permit the approximation exp $[-\alpha(\omega) - i\beta(\omega)] \approx 1 - \alpha(\omega) - i\beta(\omega)$ over the bandwidth of the modulated carrier wave. The latter is a legitimate approximation for most transmission systems, and greatly simplifies the analysis, but may lead to appreciable errors in applications to tropospheric paths where pronounced attenuation and phase distortion can be encountered. For this reason an alternative approximate analysis is adopted herein to determine intermodulation noise from quadratic phase distortion, in which no limitation is placed on the phase distortion.

Two limiting cases are considered, from which it is possible to make an approximate determination of intermodulation noise as related to phase distortion, rms frequency deviation, and bandwidth of the baseband signal. In the first case, phase distortion is assumed adequately small, such that the maximum phase distortion in the carrier signal band is less than π radians. Under this condition it is possible by use of "quasistationary" theory to determine the power spectrum of intermodulation noise without much difficulty. In the second case, no limitation is placed on phase distortion, in which case determination of the power spectrum becomes excessively difficult or laborious. It is possible, however, to determine total intermodulation noise power at the detector output, prior to post-detection low-pass filtering. From the manner in which total intermodulation noise power behaves with increasing phase distortion, it is possible to obtain an approximate evaluation of intermodulating noise in a narrow band, such as a voice channel.

A.1 *Power Spectrum of Phase Modulation*

In FM the transmitted wave is of the general form

$$V = \cos\,[\omega_0 t + \psi(t)] \tag{25}$$

where the phase $\psi(t)$ is related to the modulating wave $m(t)$ by

$$\psi(t) = k \int_0^t m(t)\, dt \tag{26}$$

where k is a constant.

The instantaneous frequency deviation is accordingly

$$\Omega(t) = \psi'(t) = km(t). \tag{27}$$

If the original signal wave has a power spectrum $s(\omega)$ and power pre-emphasis $p(\omega)$ is used, the power spectrum of the modulating wave is

$$W_m(\omega) = s(\omega)p(\omega). \tag{28}$$

The squared rms frequency deviation $\psi'(t)$ is

$$\underline{\Omega}^2 = k^2 \int_0^\infty s(\omega)p(\omega)\, d\omega. \tag{29}$$

In accordance with (26), $\psi(t)$ is the integral of $m(t)$. Hence the power spectrum of $\psi(t)$ is given by

$$W_\psi(\omega) = k^2 s(\omega)p(\omega)/\omega^2. \tag{30}$$

From (29) and (30)

$$W_\psi(\omega) = \underline{\Omega}^2 \frac{s(\omega)p(\omega)/\omega^2}{\displaystyle\int_0^\infty s(\omega)p(\omega)\, d\omega}. \tag{31}$$

The power spectrum of $\psi'(t)$ is $\omega^2 W_\psi(\omega)$.

A.2 *Autocorrelation Function of Phase Modulation*

The autocorrelation function of $\psi(t)$ is

$$R_\psi(\tau) = \int_0^\infty W_\psi(\omega) \cos \omega\tau \, d\omega \tag{32}$$

$$= k^2 \int_0^\infty \frac{s(\omega)p(\omega)}{\omega^2} \cos \omega\tau \, d\omega. \tag{33}$$

When the constant k is determined from (29), the following relation is obtained

$$R_\psi(\tau) = \underline{\Omega}^2 \left[\int_0^\infty \frac{s(\omega)p(\omega)}{\omega^2} \cos \omega\tau \, d\omega\right] \Big/ \left[\int_0^\infty s(\omega)p(\omega) \, d\omega\right]. \tag{34}$$

When the baseband power spectrum $s(\omega)$ has a bandwidth Ω, (34) can be written

$$R_\psi(\tau) = \mu^2 \left[\Omega^2 \int_0^\Omega \frac{s(\omega)p(\omega)}{\omega^2} \cos \omega\tau \, d\omega\right] \Big/ \left[\int_0^\Omega s(\omega)p(\omega) \, d\omega\right] \tag{35}$$

where μ is the rms deviation ratio

$$\mu = \underline{\Omega}/\Omega = D/B. \tag{36}$$

In the special case of a flat power spectrum, $s(\omega) = s$ and (35) yields

$$R_\psi(\tau) = \mu^2 \left[\Omega^2 \int_0^\Omega \frac{p(\omega)}{\omega^2} \cos \omega\tau \, d\omega\right] \Big/ \left[\int_0^\Omega p(\omega) \, d\omega\right]. \tag{37}$$

With pure FM, $p(\omega) = p =$ constant and (37) reduces to

$$R_\psi(\tau) = \mu^2 \int_0^1 \frac{\cos \Omega\tau x}{x^2} \, dx \tag{38}$$

where $x = \omega/\Omega$. From (38) it follows that

$$\begin{aligned} R_\psi(0) - R_\psi(\tau) &= \mu^2 \int_0^1 \frac{1 - \cos \Omega\tau x}{x^2} \, dx \\ &= \mu^2[\Omega\tau \operatorname{Si}(\Omega\tau) + \cos \Omega\tau - 1] \\ &= \mu^2 \frac{(\Omega\tau)^2}{2}\left[1 - \frac{(\Omega\tau)^2}{36} + \cdots\right] \end{aligned} \tag{39}$$

where Si is the sine integral function.

With pure PM, $p(\omega) = \omega^2$ and (37) yields

$$\begin{aligned} R_\psi(\tau) &= 3\mu^2 \int_0^1 \cos \Omega\tau x \, dx \\ &= 3u^2 \sin \Omega\tau/\Omega\tau \end{aligned} \tag{40}$$

$$\begin{aligned} R_\psi(0) - R_\psi(\tau) &= 3\mu^2[1 - \sin \Omega\tau/\Omega\tau] \\ &= \mu^2 \frac{(\Omega\tau)^2}{2}\left[1 - \frac{(\Omega\tau)^2}{20} + \cdots\right]. \end{aligned} \tag{41}$$

A.3 *Intermodulation from Phase Distortion*

It will be assumed that the phase characteristic is of the form

$$\varphi(u) = b_0 + b_1 u + b_2 u^2 + b_3 u^3 + \cdots . \tag{42}$$

Phase distortion is then represented by the term

$$\beta(u) = b_2 u^2 + b_3 u^3 + \cdots \tag{43}$$

where $u = \omega - \omega_0$ is the frequency relative to the carrier frequency ω_0.

When the transmitted wave is of the form (25), the instantaneous frequency deviation is

$$u(t) = \psi'(t) \tag{44}$$

and the corresponding variation in phase distortion with time is

$$\beta[u(t)] = b_2[\psi'(t)]^2 + b_3[\psi'(t)]^3 + \cdots . \tag{45}$$

In the above relation $\psi'(t)$ is given by (27) and the power spectrum of $\psi'(t)$ by (28) multiplied by k^2 or

$$W_{\psi'}(u) = k^2 s(\omega) p(\omega). \tag{46}$$

In determining intermodulation distortion it must be recognized that distortion increases in the range $0 < \beta[u(t)] \leqq \pi$, diminishes in the range $\pi < \beta[u(t)] < 2\pi$, increases in the range $2\pi < \beta[u(t)] < 3\pi$, etc., as illustrated in Fig. 12.

To determine intermodulation distortion it is thus necessary to evaluate the distortion obtained when a wave with the power spectrum (46) is applied to a device with the output vs input characteristic illustrated in Fig. 12. Two limiting cases will be considered below.

A.4 *Intermodulation Spectrum for Small Quadratic Phase Distortion*

With quadratic phase distortion only, (45) becomes

$$\beta[u(t)] = b_2[\psi'(t)]^2. \tag{47}$$

It will be assumed that the probability that $\beta[u(t)]$ exceeds π is so small that it is permissible to assume $\beta[u(t)] < \pi$, and furthermore that $u(t)$ changes at a sufficiently slow rate such that $\beta'[u(t)] = 2b_2\psi''(t) \ll \pi$. For signals with the properties of random noise, these assumptions are

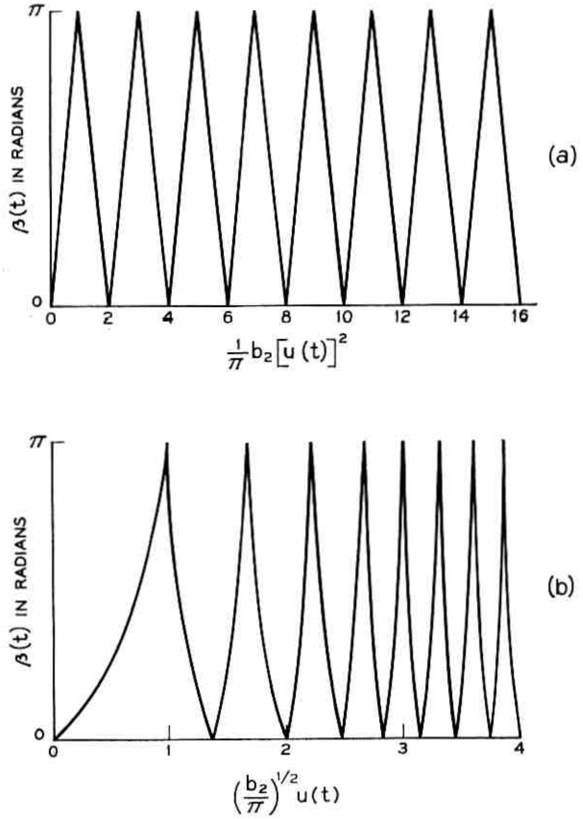


Fig. 12 — Instantaneous phase distortion $\beta(t)$ vs instantaneous frequency deviation $u(t)$ of signal.

permissible provided the rms phase error γ defined by (12) and appearing in Fig. 4 is much less than 1. With these assumptions, the autocorrelation function of the output phase distortion is the same as for a square law device and is given by (Ref. 10, Equation 4.10-1)

$$b_2^2[R_{\psi'}^2(0) + 2R_{\psi'}^2(\tau)]. \tag{48}$$

The first term can be identified with a dc component that does not give rise to noise. The power spectrum of the nonlinear output phase distortion is obtained from the second component in (48) and is given by

$$W_\psi^{(2)}(\omega) = 2b_2^2 \int_0^\infty R_{\psi'}^2(\tau) \cos \omega\tau \, d\tau. \tag{49}$$

The ratio of average intermodulation noise power at the frequency ω to the average signal power becomes

$$\rho(\omega) = \frac{W_{\psi}^{(2)}(\omega)}{W_{\psi}(\omega)} = \frac{2b_2^2 \int_0^1 R_{\psi'}^2(\tau) \cos \omega\tau \, d\tau}{k^2 p(\omega) s(\omega)/\omega^2} . \tag{50}$$

In view of (46) the following relation applies

$$R_{\psi'}(\tau) = k^2 \int_0^{\infty} s(\omega) p(\omega) \cos \omega\tau \, d\omega. \tag{51}$$

Expression (50) can be written

$$\rho(\omega) = \frac{2b_2^2 k^4 \int_0^{\infty} k^{-4} R_{\psi'}^2(\tau) \cos \omega\tau \, d\tau}{k^2 p(\omega) s(\omega)/\omega^2} \tag{52}$$

$$= \frac{2\gamma^2 a^2}{\mu^2} \frac{\int_0^{\infty} k^{-4} R_{\psi'}^2(\tau) \cos \omega\tau \, d\tau}{p(\omega) s(\omega) \int_0^{\infty} s(\omega) p(\omega) \, d\omega} \tag{53}$$

where

$$a = \omega/\Omega = f/B \tag{54}$$

$$\gamma = b_2 \mu^2 \Omega^2 = b_2 \underline{\Omega}^2 = (2\pi)^2 b_2 \, D^2. \tag{55}$$

The following relation applies*

$$\int_0^{\infty} R_{\psi'}^2(\tau) \cos \omega\tau \, d\tau = \frac{1}{2} \int_0^{\infty} W_{\psi'}(u) W_{\psi'}(\omega - u) \, du \tag{56}$$

where $W_{\psi'}(u)$ is the power spectrum given by (46).

In view of (56) and (46), expression (53) can be written

$$\rho(\omega) = \frac{a^2 \gamma^2/\mu^2}{p(\omega) s(\omega) \int_0^{\infty} p(\omega) s(\omega) d\omega} \cdot \int_{-\infty}^{\infty} s(\omega) p(\omega) s(\omega - u) p(\omega - u) du. \tag{57}$$

In the special case of a flat power spectrum $s(\omega) = s$ of bandwidth

* Ref. 10, Eq. (4C-6). In this reference the autocorrelation function is defined differently from the definition used here and has a factor 4 in integral (51), so that an additional factor $\frac{1}{4}$ appears in (56).

$\Omega = 2\pi B$, (57) becomes

$$\rho(\omega) = \frac{a^2\gamma^2/\mu^2}{p(\omega)\,\dfrac{1}{\Omega}\displaystyle\int_0^{\Omega} p(\omega)d\omega}\,\frac{1}{\Omega}\int_{\omega-\Omega}^{\Omega} p(u)p(\omega - u)du \tag{58}$$

$$= \frac{\gamma^2}{\mu^2}\,\frac{a^2}{p(a)\displaystyle\int_0^1 p(x)dx}\int_{a-1}^{1} p(x)p(a - x)dx. \tag{59}$$

When $p(x)$ is of the form

$$p(x) = 1 + c(u/\Omega)^2 = 1 + cx^2 \tag{60}$$

relation (59) becomes

$$\begin{aligned}\rho(\omega) &= \frac{3a^2\gamma^2}{\mu^2(1 + ca^2)(3 + c)}\int_{a-1}^{1} (1 + cx^2)[1 + c(a - x)^2]dx \\ &= \frac{\gamma^2 3a^2}{\mu^2(1 + ca^2)(3 + c)}\,F(c,a)\end{aligned} \tag{61}$$

where

$$\begin{aligned}F(c,a) = 2 - a + \frac{2c + c^2a^2}{3}\,&[1 + (1 - a)^3] \\ &- \frac{c^2a}{2}\,[1 - (1 - a)^4] + \frac{c^2}{5}\,[1 + (1 - a)^5].\end{aligned} \tag{62}$$

In the particular case of pure FM, $c = 0$ and $F(c,a) = 2 - a$, so that (61) yields

$$\begin{aligned}\rho(\omega) &= \frac{\gamma^2a^2}{\mu^2}\,(2 - a) \\ &= \left(\frac{B}{D}\right)^2 \gamma^2a^2(2 - a)\end{aligned} \tag{63}$$

where $a = \omega/\Omega = f/B$, $D = \Omega/2\pi$ and $\gamma = b_2\Omega^2 = b_2(2\pi D)^2$.

The above result (63) conforms with an expression derived by Rice for this limiting case (Ref. 11, Equation 5.6).

A.5 *Total Intermodulation from Quadratic Phase Distortion*

The previous analysis of the power spectrum of intermodulation noise was based on the assumption that the maximum phase distortion in the transmission band is substantially less than 180°. Without this limitation, numerical determination of the power spectrum becomes very difficult, though a formal solution may be feasible. However, it is possible to determine total intermodulation distortion without too much difficulty, without limitation on the phase distortion, as shown below.

Let x designate the instantaneous amplitude of $\psi'(t) = km(t)$, and let x have a probability density

$$p(x) = \left(\frac{2}{\pi\sigma_x^2}\right)^{\frac{1}{2}} \exp\left(-x^2/2\sigma_x^2\right). \tag{64}$$

For large instantaneous frequency deviations $\psi'(t)$ the derivative $\psi''(t)$ is on the average sufficiently small to be neglected. The total intermodulation distortion in the received signal prior to post-detection low-pass filtering is then for a nonlinear characteristic as illustrated in Fig. 12.

$$\begin{aligned} I = \int_0^{L_1} (b_2x^2)^2 p(x)dx + \int_{L_1}^{L_3} (2\pi - b_2x^2)^2 p(x)dx \\ + \int_{L_3}^{L_5} (4\pi - b_2x^2)^2 p(x)dx + \cdots + \int_{L_{2n-1}}^{L_{2m+1}} (2\pi m - b_2x^2)^2 p(x)dx \end{aligned} \tag{65}$$

where

$$L_j = (j\pi/b_2)^{\frac{1}{2}}.$$

With $b_2x^2 = u^2$, $\gamma = b_2\sigma_x^2$ and

$$p(u) = \left(\frac{2}{\pi\gamma}\right)^{\frac{1}{2}} \exp\left(-u^2/2\gamma\right) \tag{66}$$

expression (65) can be written

$$\begin{aligned} I = \int_0^{l_1} u^4 p(u)du + \int_{l_1}^{l_3} (2\pi - u^2)^2 p(u)du \\ + \int_{l_3}^{l_5} (4\pi - u^2)^2 p(u)du + \cdots \end{aligned} \tag{67}$$

where

$$l_j = (j\pi)^{\frac{1}{2}}. \tag{68}$$

Writing $2m\pi - u^2 = -\tau$, $2u\,du = d\tau$, expression (67) can be transformed into

$$\begin{aligned} I = \int_0^{\pi} \tau^{\frac{3}{2}} p(\tau)d\tau + e^{-\pi/\gamma} \int_{-\pi}^{\pi} \frac{\tau^2}{(2\pi + \tau)^{\frac{1}{2}}} p(\tau)d\tau \\ + e^{-2\pi/\gamma} \int_{-\pi}^{\pi} \frac{\tau^2}{(4\pi + \tau)^{\frac{1}{2}}} p(\tau)d\tau + \cdots \end{aligned} \tag{69}$$

where

$$p(\tau) = \left(\frac{1}{2\pi\gamma}\right)^{\frac{1}{2}} e^{-\tau/2\gamma}. \tag{70}$$

Total distortion I includes a mean or dc power component I_0 that must be subtracted from I to obtain the nonlinear component. The mean amplitude component $I_0^{\frac{1}{2}}$ ls given by

$$\begin{aligned} I_0^{\frac{1}{2}} = \int_0^{L_1} b_2x^2 p(x)dx + \int_{L_1}^{L_3} (2\pi - b_2x^2)p(x)dx \\ + \int_{L_3}^{L_5} (4\pi - b_2x^2)p(x)dx + \cdots \end{aligned} \tag{71}$$

where L_m and $p(x)$ are defined as before.

With the same notation as before, (71) can be transformed into

$$\begin{aligned} I_0^{\frac{1}{2}} = \int_0^{\pi} \tau^{\frac{1}{2}} p(\tau)d\tau + e^{-\pi/\gamma} \int_{-\pi}^{\pi} \frac{|\tau|}{(2\pi + \tau)^{\frac{1}{2}}} p(\tau)dt \\ + e^{-2\pi/\gamma} \int_{-\pi}^{\pi} \frac{|\tau|}{(4\pi + \tau)^{\frac{1}{2}}} p(\tau)d\tau + \cdots . \end{aligned} \tag{72}$$

In the above relations γ is the phase distortion corresponding to the rms frequency deviation as given by

$$\gamma = b_2\sigma_x^2 = b_2\underline{\Omega}^2 = b_2\mu^2\Omega^2. \tag{73}$$

The last relations follow from (29) since σ_x^2 is the variance of $\psi'(t)$.

The total average signal power is

$$S = R_\psi(0) = \mu^2 \left[\Omega^2 \int_0^{\Omega} \frac{p(\omega)}{\omega^2} d\omega\right] \Big/ \left[\int_0^{\Omega} p(\omega)d\omega\right] = \mu^2/C \tag{74}$$

where C is a constant depending on $p(\omega)$.

The ratio of total nonlinear intermodulation noise to total average signal power becomes

$$\begin{aligned} \rho &= \frac{I - I_0}{S} = C\,\frac{I - I_0}{\mu^2} \\ &= C(I - I_0)\left(\frac{B}{D}\right)^2. \end{aligned} \tag{75}$$

A.6 *Total Intermodulation for Small Phase Distortion*

For sufficiently small values of $\gamma = b_2\underline{\Omega}^2$, such that $\pi/\gamma \gg 1$, only the first integral in (69) needs to be considered. Hence

$$\begin{aligned} I &\approx \int_0^\pi \tau^{\frac{3}{2}} p(\tau) d\tau \\ &= 3\gamma^2 \text{ erf } (z) - 3\cdot 2^{\frac{1}{2}}\gamma^{\frac{3}{2}} \exp(-z^2) - 2^{\frac{1}{2}}\pi\gamma^{\frac{1}{2}} \exp(z^2) \end{aligned} \tag{76}$$

where

$$z^2 = \pi/2\gamma. \tag{77}$$

With a similar approximation (72) yields

$$\begin{aligned} I_0^{\frac{1}{2}} &\approx \int_0^\pi \tau^{\frac{1}{2}} p(\tau) d\tau \\ &= \gamma \text{ erf } (z) - 2^{\frac{1}{2}} \exp(-z^2). \end{aligned} \tag{78}$$

For $z \geqq 2$, or $\gamma \leqq \pi/8$:

$$I \approx 3\gamma^2 \qquad \text{and} \qquad I_0^{\frac{1}{2}} = \gamma.$$

Hence $I - I_0 = 2\gamma^2$ and (75) becomes

$$\rho = C \cdot 2 \frac{\gamma^2}{\mu^2} \tag{79}$$

where the constant C is defined through (74).

It will be noted that (79) is of the same basic form as (61) for the ratio $\rho(\omega)$ at the frequency ω. In (61) the multiplier of γ^2/μ^2 is a constant, as is the case in (79).

A.7 *Total Intermodulation for Large Phase Distortion*

When $\gamma \gg 1$, it is permissible to approximate $p(\tau)$ as given by (70) with

$$p(\tau) \approx \left(\frac{1}{2\pi\gamma}\right)^{\frac{1}{2}}. \tag{80}$$

This approximation is valid in evaluation of the various integrals in (69) and (72) provided that for the minimum value of $\tau = \pi$, $\exp(-\tau/2\gamma) \ll 1$. This is the case if

$$\pi/2\gamma \ll 1 \qquad \text{or} \qquad \gamma \gg \pi/2.$$

With (80) in (69)

$$I = \left(\frac{1}{2\pi\gamma}\right)^{\frac{1}{2}} \left[\int_0^\pi \tau^{\frac{3}{2}} d\tau + \sum_{m=1}^{\infty} e^{-m\pi/\gamma} \int_{-\pi}^{\pi} \frac{\tau^2 d\tau}{(2\pi m + \tau)^{\frac{1}{2}}}\right]. \tag{81}$$

In (81),

$$\int_{-\pi}^{\pi} \frac{\tau^2 d\tau}{(2m\pi + \tau)^{\frac{1}{2}}} = \frac{2 \cdot \pi^{\frac{5}{2}}}{15} [(2m + 1)^{\frac{3}{2}}(32m^2 - 8m + 3) - (2m + 1)^{\frac{3}{2}}(32m^2 + 8m + 3)] \tag{82}$$

$$\approx \frac{1}{m^{\frac{1}{2}}} \frac{2^{\frac{1}{2}}\pi^{\frac{5}{2}}}{3} \quad \text{for} \quad m \geqq 1 . \tag{83}$$

For $m = 1$, (82) gives about 0.5 and (83) about 0.47. Hence (83) represents a good approximation of (82).

With (83) in (81)

$$I = \left(\frac{1}{2\pi\gamma}\right)^{\frac{1}{2}} \left[\int_0^{\pi} \tau^{\frac{3}{2}}\, d\tau + \frac{2^{\frac{1}{2}}\pi^{\frac{5}{2}}}{3} \sum_{m=1}^{\infty} \frac{e^{-m\pi/\gamma}}{m^{\frac{1}{2}}}\right]. \tag{84}$$

As a first approximation the summation can be replaced by an integral, in which case

$$I \approx \left(\frac{1}{2\pi\gamma}\right)^{\frac{1}{2}} \left[\frac{2}{5}\, \pi^{\frac{5}{2}} + \frac{2^{\frac{1}{2}}\pi^{\frac{5}{2}}}{3} \int_{m=1}^{\infty} \frac{e^{-m\pi/\gamma}\, dm}{m^{\frac{1}{2}}}\right]. \tag{85}$$

With $m = u^2$

$$\begin{aligned} I &= \left(\frac{1}{2\pi\gamma}\right)^{\frac{1}{2}} \pi^{\frac{5}{2}} \left[\frac{2}{5} + \frac{2^{\frac{1}{2}} \cdot 2}{3} \int_{u=1}^{\infty} e^{-u^2\pi/\gamma}\, du\right] \\ &= \left(\frac{1}{2\pi\gamma}\right)^{\frac{1}{2}} \pi^{\frac{5}{2}} \left[\frac{2}{5} + \frac{2^{\frac{1}{2}}\gamma^{\frac{1}{2}}}{3}\, \text{erfc}\left(\sqrt{\frac{\pi}{\gamma}}\right)\right] \end{aligned} \tag{86}$$

$$= \frac{\pi^2}{3}\left[\text{erfc}\left(\sqrt{\frac{\pi}{\gamma}}\right) + \frac{6}{5 \cdot 2^{\frac{1}{2}}\gamma^{\frac{1}{2}}}\right] \tag{87}$$

$$\approx \frac{\pi^2}{3} \quad \text{for} \quad \gamma \gg 4\pi . \tag{88}$$

By a similar approximation $I_0^{\frac{1}{2}}$ as given by (72) becomes

$$I_0^{\frac{1}{2}} = \left(\frac{1}{2\pi\gamma}\right)^{\frac{1}{2}} \left[\int_0^{\pi} \tau^{\frac{1}{2}}\, d\tau + \sum_{m=1}^{\infty} e^{-m\pi/\gamma} \int_{-\pi}^{\pi} \frac{|\,\tau\,|\, d\tau}{(2m\pi + \tau)^{\frac{1}{2}}}\right] \tag{89}$$

$$= \left(\frac{1}{2\pi\gamma}\right)^{\frac{1}{2}} \left[\frac{2}{3}\, \pi^{\frac{3}{2}} + \frac{\pi\sigma}{2^{\frac{1}{2}}}\, \text{erfc}\left(\sqrt{\frac{\pi}{\gamma}}\right)\right] \tag{90}$$

$$= \frac{\sqrt{\pi}}{2}\left[\text{erfc}\left(\sqrt{\frac{\pi}{\gamma}}\right) + \frac{2^{\frac{1}{2}}}{3\gamma^{\frac{1}{2}}}\, \pi^{\frac{1}{2}}\right] \tag{91}$$

$$\approx \frac{\pi^{\frac{3}{2}}}{2} \quad \text{for} \quad \gamma \gg 4\pi . \tag{92}$$

The ratio ρ is obtained from (75) with I as given by (87) and $I_0^{\frac{1}{2}}$ by (91). In the limit of $\gamma \to \infty$ the ratio becomes

$$\begin{aligned} \rho &= C \frac{I - I_0}{\mu^2} = \frac{C}{\mu^2} \pi \left(\frac{\pi}{3} - \frac{1}{4} \right) \\ &\approx 2.5 \frac{C}{\mu^2}. \end{aligned} \tag{93}$$

A.8 *Approximation for Total Intermodulation*

The general expression for the ratio ρ of total intermodulation noise power to average signal power can be written in the form

$$\rho = \frac{2C}{\mu^2} h(\gamma). \tag{94}$$

For the limiting case of $\gamma \to 0$, the function h is in accordance with (79)

$$h = \gamma^2. \tag{95}$$

For the other limiting case in which $\gamma \to \infty$, the function h is in accordance with (93)

$$h = 2.5/2 = 1.25. \tag{96}$$

In Fig. 13 are shown the above two limiting cases, together with the function h obtained from (75) as $\eta = I - I_0$, when I and I_0 are determined from (86) and (91). The approximate function $h(\gamma)$ is obtained by drawing a transition curve between the above twɔ limiting cases, as in Fig. 13.

A.9 *Approximation for Intermodulation Spectrum*

The function $h(\gamma)$ in Fig. 13 is proportional to the total intermodulation noise power and can be related to the power spectrum $W_i(\omega)$ of intermodulation noise by

$$h(\gamma) = c_0 \int_0^\infty W_i(\omega)\, d\omega \tag{97}$$

where c_0 is a constant. Relation (94) can thus be written

$$\rho = \frac{2c_0C}{\mu^2} \int_0^\infty W_i(\omega)\, d\omega. \tag{98}$$

For $\gamma \to 0$, (98) must conform with (95), which is possible provided the power spectrum is of the general form

$$W_i^0(\omega) = c_1\gamma^2 \frac{1}{\Omega} F_0(\omega/\Omega) \tag{99}$$

where F_0 is any functional relation dependent only on the ratio $a = \omega/\Omega$.

With (99) in (98)

$$\begin{aligned} \rho &= \frac{\gamma^2}{\mu^2} 2c_0c_1C \frac{1}{\Omega} \int_0^\infty F_0(\omega/\Omega)\, d\omega \\ &= \frac{\gamma^2}{\mu^2} 2c_0c_1C \int_0^\infty F_0(u)\, du. \end{aligned} \tag{100}$$

This yields relation (95) provided

$$c_0c_1C \int_0^\infty F_0(u)\, du = 1. \tag{101}$$

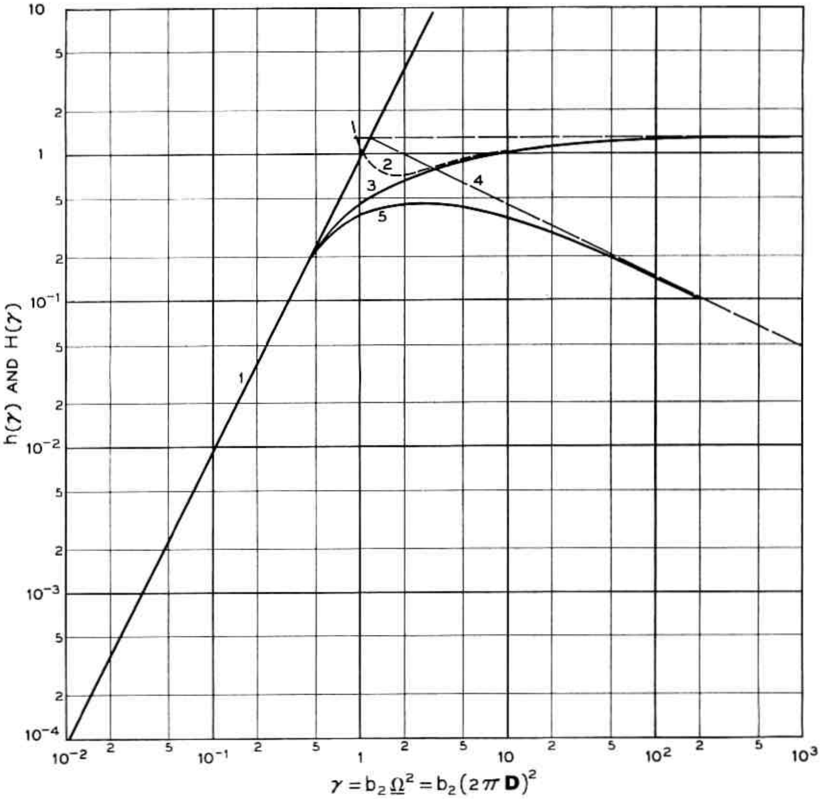


Fig. 13 — Functions $h(\gamma)$ and $H(\gamma)$: 1, functions $h(\gamma)$ and $H(\gamma)$ for $\gamma \ll 1$; 2, function $h(\gamma)$ for $\gamma \gg 1$; 3, approximate interpolated function $h(\gamma)$; 4, function $H(\gamma)$ for $\gamma \gg 1$; 5, approximate interpolated function $H(\gamma)$.

From (100) it is apparent that the ratio of intermodulation noise power to average signal power in a narrow band $d\omega$ at ω is

$$\rho(\omega) = \frac{\gamma^2}{\mu^2} 2c_0c_1C \frac{1}{\Omega} F_0(\omega/\Omega). \tag{102}$$

Comparison of (102) with (61) shows that in this case

$$2c_0c_1C \frac{1}{\Omega} F_0(\omega/\Omega) = \frac{3a^2}{(1 + ca^2)(3 + c)} F(c,a) \tag{103}$$

where $F(c,a)$ is given by (62).

In summary, for $\gamma \to 0$ the power spectrum has a fixed shape independent of γ and an amplitude proportional to γ^2.

Consider next the limiting case in which $\gamma \to \infty$. In accordance with (96) h then approaches a constant, which is possible for various power spectra of the general form

$$W_i^{(\infty)}(\omega) = \frac{c_1}{\gamma^n} F_\infty(\omega/\gamma^n) \tag{104}$$

where $F_\infty(\omega/\gamma^n)$ is any functional relation dependent only on the ratio (ω/γ^n). In this case (104) in (98) yields

$$\begin{aligned} \rho &= \frac{2c_0c_1C}{\mu^2} \frac{1}{\gamma^n} \int_0^\infty F_\infty(\omega/\gamma^n)\, d\omega \\ &= \frac{2c_0c_1C}{\mu^2} \int_0^\infty F(u)\, du \end{aligned} \tag{105}$$

where $u = \omega/\gamma^n$.

The exponent n can be determined from consideration of the input vs output characteristic shown in Fig. 12. If b_2 is increased by a factor k, the intervals between zero points are multiplied by a factor $k^{-\frac{1}{2}}$, as indicated in Fig. 14 for $k = 4$. For a given frequency deviation, the bandwidth of the power spectrum is then multiplied by a factor $k^{\frac{1}{2}}$ and the amplitude of the spectrum at each frequency multiplied by a factor $k^{-\frac{1}{2}}$. Hence in the case of quadratic phase distortion as considered here, $n = \frac{1}{2}$ in (104).

Based on the above considerations, the power spectrum at any frequency ω for the above two limiting cases would vary with γ as indicated in Fig. 13. The shape of the curves between these two limiting cases would in a first approximation be represented by the function $H(\gamma)$ shown in Fig. 13.

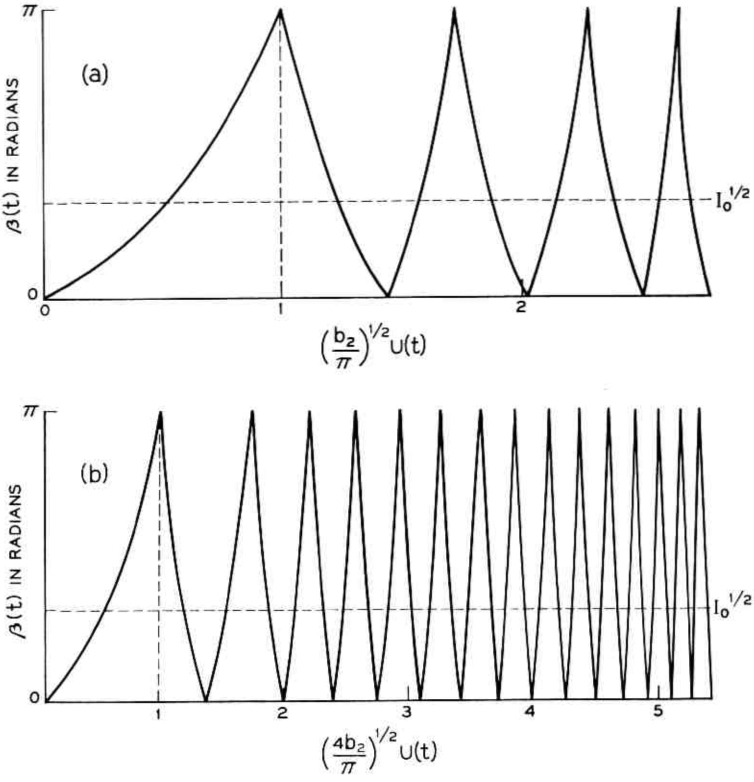


Fig. 14 — (*a*) Relation of instantaneous phase distortion $\beta(t)$ to instantaneous frequency deviation $u(t)$ for a given b_2; (*b*) relation of instantaneous phase distortion to instantaneous frequency deviation with fourfold increase in b_2.

A.10 *Approximation for* $\rho(\omega)$

The ratio $\rho(\omega)$ of intermodulation noise power in a narrow band at ω to average signal power in the same narrow band can be written

$$\rho(\omega) = \frac{2C(\omega)}{\mu^2} H(\gamma). \tag{106}$$

This relation differs from (94) in that $h(\gamma)$ as shown in Fig. 13 is replaced by $H(\gamma)$ shown in the same figure, and C is replaced by $C(\omega)$. The constant C defined through (74) depends on the frequency pre-emphasis $p(\omega)$. The function $C(\omega)$ depends both on the frequency pre-emphasis $p(\omega)$ and the frequency under consideration.

For the particular type of frequency pre-emphasis represented by

(60), expression (106) must conform with (61). This results in the following approximate relation

$$\rho(\omega) = \left(\frac{B}{D}\right)^2 G(c,a)H(\gamma) \tag{107}$$

where $H(\gamma)$ is the function shown in Fig. 13 and

$$G(c,a) = \frac{3a^2}{(1 + ca^2)(3 + c)} F(c,a) \tag{108}$$

where $F(c,a)$ is given by (62).

In the particular case in which $c = 16$ and $a = f/B = 1$

$$G(c,a) \approx 0.192 \tag{109}$$

and (107) yields

$$\rho(B) = \left(\frac{B}{D}\right)^2 \times 0.192H(\gamma) \tag{110}$$

$$= \left(\frac{B}{D}\right)^2 \times 0.192\cdot\gamma^2 \quad \text{for} \quad \gamma \ll 1. \tag{111}$$

REFERENCES

1. Sunde, E. D., Digital Troposcatter Transmission and Modulation Theory, this issue, Part 1, p. 143.
2. Clutts, C. E., Kennedy, R. N., and Trecker, J. M., Results of Bandwidth Tests on the 185-Mile Florida-Cuba Scatter Radio Systems, IRE Trans. on Communication Systems, **9**, December, 1961, p. 434.
3. Beach, C. D., and Trecker, J. M., A Method for Predicting Interchannel Modulation Due to Multipath Propagation in FM and PM Tropospheric Radio Systems, B.S.T.J., **42**, January, 1963, p. 1.
4. Bennett, W. R., Curtis, H. E., and Rice, S. O., Interchannel Interference in FM and PM Systems under Noise Loading Conditions, B.S.T.J., **34**, May, 1955, p. 601.
5. Medhurst, R. G., and Small, G. F., Distortion in Frequency-Modulation Systems Due to Small Sinusoidal Variations of Transmission Characteristics, Proc. IRE, **44**, November, 1956, p. 1608.
6. Medhurst, R. G., and Small, G. F., An Extended Analysis of Echo Distortion in FM Transmission of Frequency-Division Multiplex, Proc. IEE, **103**, Pt. *B*, March, 1956, p. 190.
7. Carson, J. R., and Fry, T. C., Variable Frequency Electric Circuit Theory with Applications to the Theory of Frequency Modulation, B.S.T.J., **16**, October, 1937, p. 513.
8. van der Pohl, B., The Fundamental Principles of Frequency Modulation, Jour. IEE, Part III, May, 1946, p. 153.
9. Albersheim, W. J., and Schafer, J. P., Echo Distortion in the FM Transmission of Frequency Division Multiplex, Proc. IRE, **40**, March, 1952, p. 316.
10. Rice. S. O., Mathematical Analysis of Random Noise-I and -II, B.S.T.J., **23**, July, 1944, p. 282, and **24**, January, 1945, p. 46.
11. Rice, S. O., Distortion Produced by a Noise Modulated Signal by Nonlinear Attenuation and Phase Shift, B.S.T.J., **36**, July, 1957, p. 879.

Cutoff Frequencies of the Dielectrically Loaded Comb Structure as Used in Traveling-Wave Masers*

By S. E. HARRIS, R. W. DEGRASSE and E. O. SCHULZ-DUBOIS

(Manuscript received June 27, 1963)

The subject of traveling-wave maser design is reviewed and a first step towards an analytical design procedure is presented. A method is derived for calculating the upper and lower cutoff frequencies of a comb-type slow-wave structure of simple geometry. It is based on the electromagnetic field pattern and the equivalent impedances which are calculated for these frequencies, both for the dielectrically loaded and the empty comb structure. The design procedure resulting from these calculations permits the prediction of a dielectric loading geometry that shifts the upper and lower cutoff frequency of the empty comb to new, lower values which can be arbitrarily specified within certain limitations. Frequencies calculated by this procedure are compared with the results of measurements, and it is found that cutoff frequencies can be predicted to better than 10 per cent.

I. INTRODUCTION

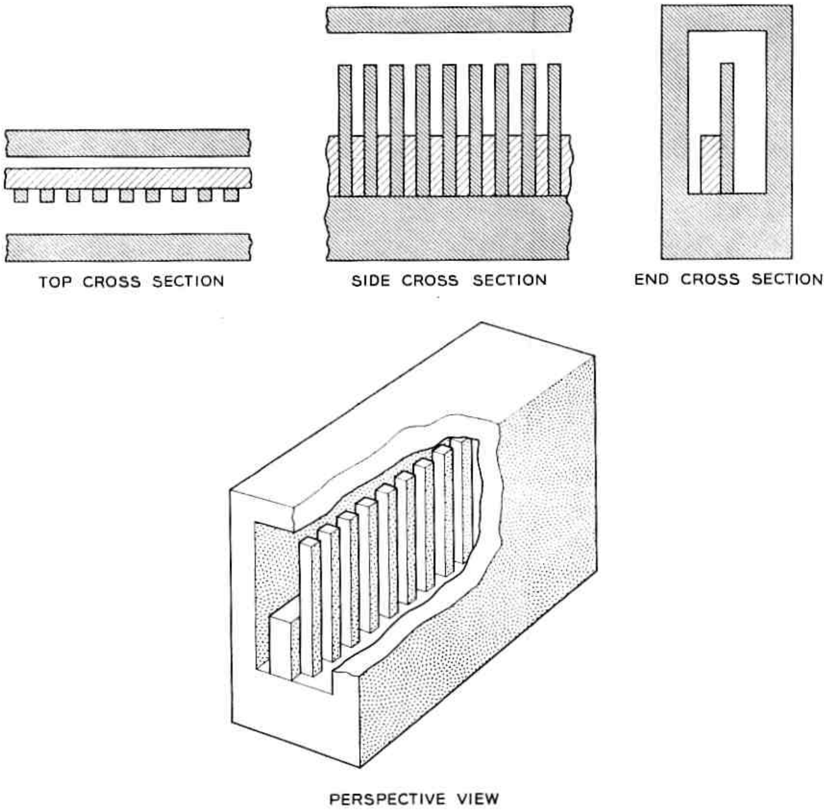
In the early development of the traveling-wave maser (TWM),[1] the design procedures used were largely empirical. Short TWM model sections were built, tested and modified in order to meet the desired performance specifications. By this cut-and-try method, a satisfactory design was finally derived which was applied in the construction of full-length TWM's.

However, a more satisfying approach is possible if the relevant theoretical aspects regarding the maser active material, the ferrimagnetic isolator and the electromagnetic behavior of the slow-wave structure are known, either rigorously or approximately. Then a TWM can be designed on the basis of analysis before actual construction. Most attractive in the analytical approach is the inherent flexibility and versa-

* This work was supported in part by the U.S. Army Signal Corps under Contract DA 36-039-sc-85357.

tility. Thus, a large number of design ideas may be explored and a near optimum configuration can be found before any hardware is built.

The present paper is a step in the direction of a more analytical approach. Using reasonably accurate approximations to the field pattern at both cutoff frequencies, the equivalent TEM line impedances, the "effective" dielectric constants and, finally, the cutoff frequencies are calculated. This results in a numerical design procedure for the TWM structure. The analysis is made for a comb having fingers of rectangular cross section and for dielectric loading with maser material in the form of one or two rectangular parallelepipeds as shown in Fig. 1. Comparison of cutoff frequencies calculated by this method with experiment shows agreement to usually better than 5 per cent.


TOP CROSS SECTION

SIDE CROSS SECTION

END CROSS SECTION

PERSPECTIVE VIEW

Fig. 1 — Typical comb structure.

1.1 *The Significance of Cutoff Frequencies in TWM Design*

Consider the TWM electronic gain formula[1]

$$G(\text{db}) = 27.3(-\chi'')fFl/v_g\,. \tag{1}$$

Here $(-\chi'')$ is the inverted susceptibility of the maser active material, f the signal frequency, F the filling factor, l the length of the maser structure and v_g the group velocity. The TWM net gain is obtained by subtracting from (1) the slow-wave structure loss (copper loss) and the ferrimagnetic isolator loss (ferrite loss).

In the development of a practical TWM, the design frequency f and the structure length l are generally determined by the application. The susceptibility $(-\chi'')$ is a property of the active material which cannot be theoretically predicted and must be experimentally determined. $(-\chi'')$ is redefined as

$$-\chi'' = I\chi_0'' \tag{2}$$

and the quantities I and χ_0'' are determined by two independent measurements. Here, I is the inversion ratio, i.e., the ratio of electronic gain from the activated maser material to electronic loss in the same material at thermal equilibrium. J. E. Geusic and W. J. Tabor have carried out inversion measurements for ruby maser material in a helix test structure, and the method and results will be described in a forthcoming paper.[2] The susceptibility at thermal equilibrium, χ_0'', is measured by standard resonance techniques[3] or may be calculated from the material composition and linewidth. In this way, $-\chi''$ can be determined to about 10 per cent, which is adequate for the present design procedure. Complications can arise in practice, however, if nominally identical crystals show variations in the active ion concentration or in the crystalline perfection.

The filling factor F may be factorized into two expressions

$$F = F_pF_v \tag{3}$$

where

$$F_p = \left[\int_M |\mu\cdot H^*|^2\,dA\right]\Big/\left[|\mu|^2\int_M |H|^2\,dA\right] \tag{4}$$

and

$$F_v = \left[\int_M |H|^2\,dA\right]\Big/\left[\int_A |H^2|\,dA\right]. \tag{5}$$

Here, μ is the magnetic dipole moment associated with the maser signal transition and H is the RF magnetic field in the TWM structure. The asterisk * denotes the conjugate complex time dependence. The integration is performed in the cross-sectional plane where M denotes the cross section of the maser material and A the total structure cross section. F_p may be called the polarization efficiency factor and F_v the volume filling factor. F_p expresses the excitation efficiency of the signal transition by the RF magnetic field present in the maser material. For example, if both μ and H are of circular polarization in the same direction, then F_p is unity. Similarly, for maser material symmetrically loaded on both sides of the comb and with a circular transition perpendicular to the finger direction, a symmetry argument shows that $F_p = \frac{1}{2}$. F_v indicates what fraction of the total magnetic field energy is contained within the maser material. F_p and F_v are functions of frequency across the passband of the comb structure. Usually, however, it is sufficient to consider F at some midband frequency where it is only a slowly varying function of frequency.

Experience suggests that it is possible to estimate F to fair accuracy from the TWM geometry and a qualitative estimate of the RF magnetic field pattern. For example, it is estimated that in TWM's designed in this laboratory for 5.6, 4.2, 2.4 and 1.4 gc the filling factor F varies over the relatively limited range from 25 to 45 per cent. Thus, from the viewpoint of the analytical design of the TWM, a detailed computation of the RF magnetic field configuration is of no great value unless the other factors entering the TWM gain formula are known with comparable accuracy.

Up to the present time, this was not the case, the factor least amenable to analytical prediction being the group velocity v_g. It is well known that a wave traveling through a slow-wave structure has field components varying like $\exp[i(\omega t - \beta z)]$, where $\omega = 2\pi f$, t is the time, β the phase propagation constant and z the length coordinate along the structure. In the comb structure, each finger is an energy storage element capable of resonant storage in the same way as a quarter-wavelength coaxial resonator. As a general rule, the phase shift between adjacent elements may assume values between 0 and $\pm\pi$ as the frequency is varied across the passband. The phase shift values 0 and $\pm\pi$ are associated with the cutoff frequencies. The comb structure is normally a forward-wave structure, where $+\pi$ is the phase shift at the upper cutoff frequency and 0 that at the lower. It is possible (although not of practical importance in TWM design) to make the comb a backward-wave structure, in which case $-\pi$ is the phase shift at the lower cutoff frequency and 0 that at the upper. In the normal forward-wave comb structure, the phase propagation constant then varies from $\beta = 0$ to $\beta = (N - 1)\pi/l$ across the passband, where N is the number of fingers and l is the structure length measured between centers of the first and last finger. The group velocity is given by

$$v_g = d\omega/d\beta. \tag{6}$$

Typical diagrams of β as a function of ω are shown in Fig. 2(a). As the curves approach the cutoff points, they assume infinite slope, corresponding to zero group velocity. There is a range at midband, however, where the group velocity is fairly constant. These graphs are typical of most of the structures studied but exceptions occasionally were found, as indicated in Fig. 2(b). These exceptions include backward-wave structures where phase and group propagation take place in opposite directions. They also include "mongrel" structures where, over part of the band, β is a double-valued function of ω; these, therefore, are forward and backward at the same time. This latter case is a very undesirable one; as discussed in Ref. 4, the existence of two propagation modes at the same frequency, one a forward wave, the other a backward wave, allows for propagation with gain in both directions despite the presence of an isolator. As a result, the maser will oscillate instead of offering stable gain. Empirically, however, this situation can be easily diagnosed and there are remedies to rectify it. Therefore, double-valued ω-β relations may be excluded from the present considerations.

With this proviso, it can be seen from Fig. 2(a) that the midband group velocity can be estimated reasonably well from a knowledge of the two cutoff frequencies alone, viz.

$$v_g = 2a\Delta fl/(N - 1) = 2a\Delta f\Delta l. \tag{7a}$$

Here Δf is the frequency width of the passband, Δl is the center-to-center spacing between comb fingers and a is a numerical factor which takes into account the detailed shape of the ω-β curve. Equation (7a) may be rearranged in terms of the group velocity slowing

$$S = \frac{c}{v_g} = \frac{1}{a}\frac{\lambda/2}{\Delta l}\frac{f}{\Delta f} \tag{7b}$$

indicating that slowing is partly a geometric effect, i.e., the compression of a half wavelength into one period of the structure, and partly the effect of compression in the frequency domain, sometimes expressed by a loaded Q. a assumes values of one for a straight line ω-β relation, 1.57 for an inverse cosine, and may in practice be as high as four for a "sagging" ω-β curve. In other words, the uncertainty in estimating the

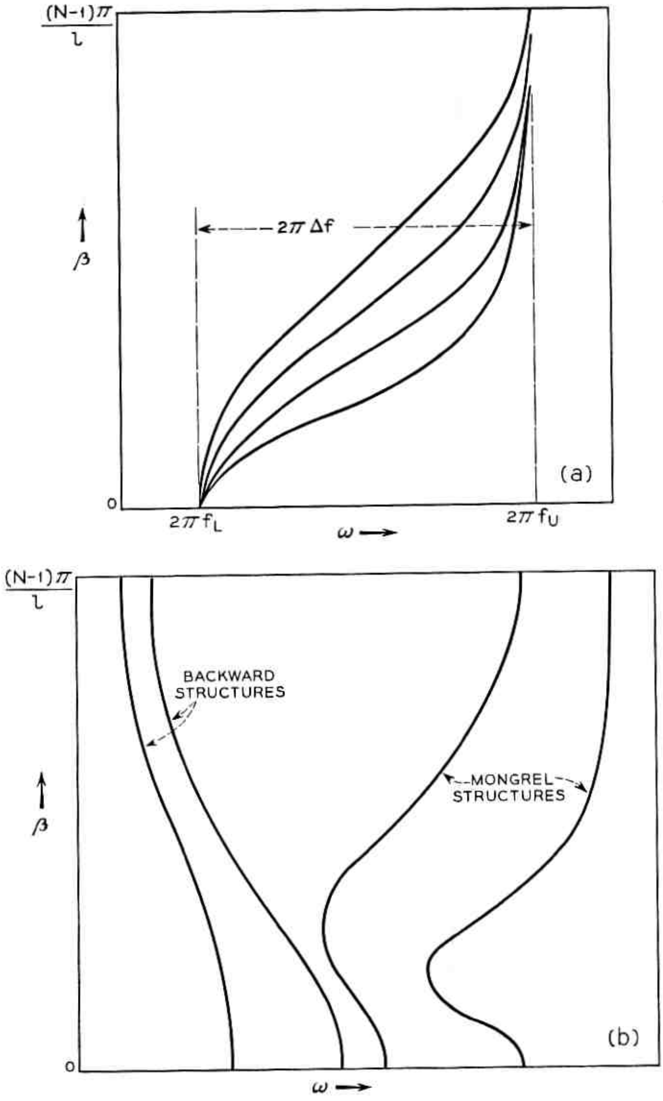


Fig. 2 — (a)Typical forward ω-β diagrams of loaded comb structures with normalized cutoff frequencies. (b) Exceptional ω-β diagrams found in comb structures with extreme dielectric loading.

group velocity or slowing from the cutoff frequencies is not very large, usually less than a factor of two.

Thus, it is clear that a method for calculating the two cutoff frequencies would be an important first step towards an analytical design procedure. Such a mathematical method should be carried out as rigorously as possible. The reason for this may be demonstrated in the following way. If fringe capacity at the finger tips and dielectric loading effects are neglected, the comb structure is electrically equivalent to the Easitron* structure. In this approximation, the comb would be a zero passband structure with identical cutoff frequencies like the Easitron. In reality, they differ only because fringe capacity and dielectric loading affect both frequencies to different degrees. Thus, the width of the passband Δf is obtained as a small difference between large numbers, the upper and lower cutoff frequencies, f_U and f_L. To obtain Δf with *fair* accuracy, f_U and f_L must be known with *good* accuracy. Similarly, a small change in the dielectric loading may change f_U and f_L each by a small percentage, but Δf by an appreciable factor. Experience has shown that comb structures with different dielectric loadings may have a passband width Δf anywhere between 1 and 50 per cent of the midband frequency. In other words, as long as the cutoff frequencies are not known, the uncertainty in an estimate of v_g may be almost two orders of magnitude. The computation of the cutoff frequencies would be very useful if it could reduce this uncertainty to about a factor of two. Besides determining the group velocity and hence, indirectly the electronic gain, the cutoff frequencies also define the center frequency and the tunable bandwidth of the TWM. Since it is impossible to match a structure right up to the cutoff frequency, the useful tunable band is well inside the structure passband Δf. An analytical design procedure that allows a reasonably accurate prediction of the cutoff frequencies would clearly be desirable, as center frequency and tunable bandwidth are among the primary TWM specifications.

1.2 *The Function of Slow-Wave Structures in Electron Beam Tubes and TWM's*

A considerable amount of work, both theoretical and experimental, has gone into the study of slow-wave structures for tubes. It would be gratifying if this knowledge could be used in TWM work. Unfortunately,

* The Easitron was analyzed by L. R. Walker, unpublished manuscript, quoted in Ref. 1. This structure consists of a rectangular waveguide with an array of uniform, identical conductors in the H plane connecting both short walls. It has zero passband, nonpropagating resonances of frequencies where the conductor length is one or more half wavelengths.

this work has only limited applicability to the TWM. This is more readily understood if slow-wave structures for electron beam tubes and for TWM's are compared.

In tubes, slowing factors between 10 and 100 are typical, while in the TWM, slowing of 50 to 1000 is used. This difference influences primarily the mechanical tolerances, which are tighter for higher slowing.

A more fundamental distinction concerns the applicable slowing concept. In an electron beam tube there must be synchronism between the electromagnetic mode propagated on the slow-wave structure and the interacting mode characterized by a charge distribution on the beam. Therefore, the analysis of tubes is concerned with the phase velocity of the slow-wave structure mode. Similarly, in a traveling-wave parametric amplifier there must be synchronism between pump, idler and signal propagation, requiring a phase velocity relation for these three frequencies. In filter circuits the condition of synchronism is usually satisfied only over a small fraction of the total structure bandwidth. By contrast, the amplification by the maser material does not depend on the existence of phase relations along the TWM structure. The maser material may be considered as an incoherent, long-time energy reservoir from which energy is withdrawn upon stimulation by an incident signal and added to the incident signal in a coherent phase preserving fashion. The function of the slowing is merely to "give the signal more time" to interact with the energy stored in the maser material, i.e., to enhance the stimulating gain interaction. Thus, the analysis of TWM's is concerned with the signal group velocity in the structure rather than phase velocity. It is not necessary that v_g be constant over the tunable band. If the gain over the tunable band is required to be constant, then the product $-\chi'' Ff/v_g$ (neglecting copper and ferrite losses) should be constant over the band. Experience has shown that this condition can be met over almost the entire passband.

Another point is the interaction mechanism between the active element and the slowing structure. An electron beam interacts with a structure mode via the RF electric field, and the interaction is conventionally represented by an interaction impedance. The interaction of the inverted spins in the maser material with the structure mode takes place via the RF magnetic field, and its strength is measured by the filling factor.

All the differences mentioned have no bearing on the question whether the knowledge of slow-wave structures accumulated in studies directed towards electron beam interaction can be applied to TWM structures. For example, the degree of slowing is not essential for a theoretical analysis, the group slowing is easily derived by differentiation from the phase propagation, and electric and magnetic interaction terms can be obtained equally well from the field analysis.

The chief difference in slow-wave structures for these two applications lies in their relation to dielectric loading. In a tube, dielectric loading is undesirable and is usually avoided as far as possible. By virtue of its dielectric constant, the glass envelope of a TWT, for example, drags away from the beam some of the electric field energy carried by the helix and thus reduces the gain interaction. In fact, most studies of slow-wave structures for beam tubes pertain to metal structures surrounded by vacuum.

Dielectric loading, being an undesirable side effect for tubes, is an essential and rather beneficial feature in maser structures. Since the gain interaction is magnetic in nature, the interaction of the electric field with dielectrics may be used to advantage without deteriorating the gain interaction. Indeed, it is being used for reducing the over-all maser size, tuning the band center frequency, adjusting the tunable bandwidth or increasing the gain by increased slowing (of course, the items mentioned are not independent). Thus, a high degree of design flexibility can be obtained, even with the identical copper comb, merely by changing the dielectric loading.

For this reason, dielectric loading must be included in any treatment of TWM structures. The present paper is a first contribution to the theoretical treatment of maser structures taking dielectric loading into account. To keep the mathematics reasonably simple, the maser comb geometry, including the dielectric loading, was chosen to be fairly simple. In the laboratory, dielectric loading techniques were developed in which the loading consists of more than one dielectric and has more complex shapes. Work to be published by F. S. Chen has generalized the analysis to take these modifications into account. It also expands the present analysis of the cutoff frequencies into a more general one which allows the prediction of the entire ω-β diagram. This will be particularly valuable in finding criteria to avoid structures having a double-valued "foldover" or "mongrel" ω-β diagram.

II. GENERAL PROBLEM AND APPROACH TO SOLUTION

The problem is to find by analysis the upper and lower cutoff frequencies of the comb-type slow-wave structure as used in a traveling-wave maser (TWM). In particular, this implies taking into account the dielectric effect resulting from loading the comb with maser material or possibly some other dielectric material and the effect of the fringe capacity at the tips of the comb finger. It was pointed out before that, in a zero-order approximation neglecting both effects, the comb is a zero passband structure.

In the course of this treatment it will be necessary to introduce a number of restrictions and approximations. These are mostly required in order to keep the mathematics manageable. Some other restrictions are introduced in order to have the geometry underlying the calculations correspond to the type of TWM geometry which is presently investigated in the laboratory. These various restrictions and approximations are labeled with lower-case roman numerals for reference in this discussion.

(i) The first restriction pertains to the cross section of the comb fingers. The treatment used here is applicable only to combs with fingers of rectangular cross section.

This means that it is not possible to apply this type of analysis to a comb having round fingers as used in the original TWM's. It may be mentioned here, however, that it is possible to treat the round-finger comb as long as certain simple frequency or impedance data are available from measurements on scale models, resistance cards or measurements in the electrolytic tank.

Besides being better suited for mathematical analysis, there is another justification for treating combs with rectangular fingers. This has to do with fabrication of combs. There is indication that it is possible to fabricate combs with rectangular fingers not only with greater ease but also with greater perfection. The subject of these fabrication techniques may be discussed at some later date.

A typical comb structure as treated here is shown in Fig. 1. The fingers shown are of square cross section and are spaced by a finger width. It should be emphasized that the general method used here is applicable to any rectangular cross section and spacing, although a great many of the computations are concerned with square fingers spaced by a finger width.

(ii) The next restriction is that maser material (or some other dielectric) is inserted into the comb in the shape of a single rectangular parallelepiped.

The restriction to parallelepipeds is rather definite. There is a possibility, however, of considering more than one slab of maser material loading the comb. No change in the general analysis is required if two identical slabs are considered which are loaded symmetrically on both sides of the comb. This is shown in Fig. 3(a). The analysis could be carried out also for the case shown in Fig. 3(b) where the maser material is in-

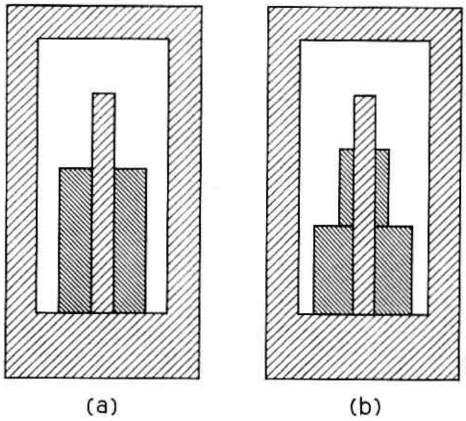


Fig. 3 — Loading geometries.

serted in the form of two pairs of identical slabs. It should be mentioned, however, that the calculation will be appreciably more cumbersome in this case. Although it will not be described in detail, it will be fairly obvious to the reader how the calculations have to be modified to take into account geometries like the one of Fig. 3(b).

(iii) A further simplifying assumption is that the dielectric loading is assumed to have an isotropic dielectric constant, at least for field components perpendicular to the finger direction.

This assumption is not too restrictive. An effective dielectric constant may be estimated in the case of a tensor dielectric constant. This estimate will usually be different for either cutoff frequency, since it depends on the electric field configuration. As the tensor components are always of the same order of magnitude, the estimated effective dielectric constant should turn out to be sufficiently accurate for most cases of practical interest.

No provisions have to be made for magnetic permeability. Outside the maser signal line, $\mu' = 1$ for the maser material. Even within the frequency range of the signal line, the deviation of μ' from unity is so small that it can be neglected for all practical purposes as a factor influencing the cutoff frequencies. A similar reasoning applies to the ferrimagnetic isolator. Even though the values of $\mu' - 1$ are larger there, they are less effective due to the very small ferrimagnetic filling factor.

The starting point for the calculation is the phase shift. At one cutoff frequency the phase shift between fingers is zero. This has the consequence that an instantaneous electric field pattern within the comb may look like Fig. 4(a). Usually, although not necessarily so, this is the case at the lower cutoff frequency, f_L. Throughout the paper this case will be referred to as the "lower cutoff," although the term "zero phase shift case" would be more appropriate. The field pattern is repetitive and shows no field lines from finger to finger since they are on the same potential. It is symmetric with respect to a cross-sectional plane in the structure which contains either the center line of a finger or the center line in the space between two fingers. Therefore the same field pattern is obtained with a single finger if the section of the comb containing this finger is enclosed by a "magnetic wall." A magnetic wall is a fictitious plane on which the electromagnetic field components obey boundary conditions such that the electric field is tangential and the magnetic field normal to the plane. These boundary conditions are opposite from those on a perfect conductor. The perfect conductor is closely approximated

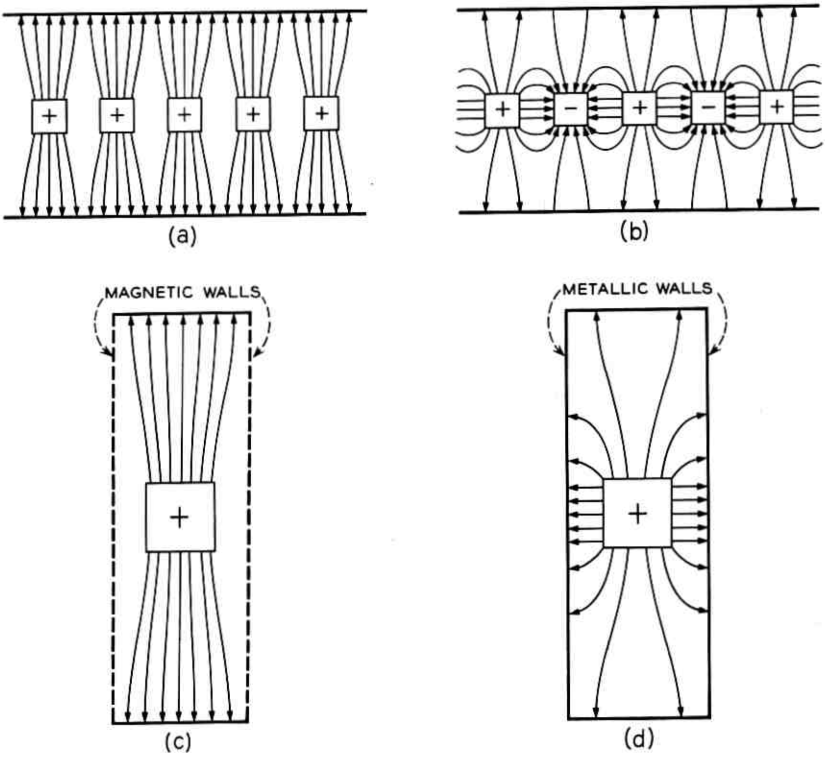


Fig. 4 — Field patterns showing phase shift conditions.

in experiments by high-conductivity metals, whereas the magnetic wall is a mathematical model only. Since the field patterns of Figs. 4(a) and 4(c) are identical, the frequencies will be the same, too. Thus the lower cutoff frequency of the comb can be found as the resonant frequency of the one-finger structure in Fig. 4(c).

A similar reasoning applies to the upper cutoff frequency f_U. Here the phase shift is π between adjacent fingers. An instantaneous field pattern will therefore look like Fig. 4(b). Since adjacent fingers are subject to opposite potential, there are strong electric field components going from finger to finger. The field pattern is symmetric with respect to a cross-sectional plane in the structure which contains the center line of a finger, but antisymmetric with respect to a cross-sectional plane which contains the center line in the space between two fingers. Thus the same field pattern can be realized on a single finger if the section of the comb containing the finger is enclosed by a perfectly conducting (or metallic) wall. This wall will take the place of the plane of antisymmetry in the comb. This is illustrated in Fig. 4(d). Again, identical field patterns require the same frequency. Thus the upper cutoff frequency of the comb can be found as the resonant frequency of the one-finger structure in Fig. 4(d).

The method of determining the resonance frequency of either one-finger model, that of Fig. 4(c) or 4(d), is suggested by Fig. 5. The finger acts essentially as a quarter-wave TEM resonator. At the comb base, this TEM line is terminated in a short. At the finger tip the TEM line is terminated by a nearly perfect "open." This is only slightly modified by fringing electric fields between the finger tip and the surrounding walls. The effect of these fields can be lumped into a fringe capacity C. In principle, C will be different for both cutoff frequencies.

Unfortunately, both capacities C_U and C_L cannot be calculated easily. Therefore, measurements have been made in an analog electrolytic tank setup. A scale model having the cross section of the one-finger lines in Figs. 4(c) and 4(d) was built. This cross section is shown in Fig. 6(a) for the upper cutoff frequency and in Fig. 6(b) for the lower. In the electrolytic tank the electric field lines of the object under study are simulated by the current lines in the tank fluid. No approximation is involved in this analogy. In particular, it is possible to simulate a magnetic wall like that of Fig. 3(c) by an insulating wall. This is done in the cross section used in the lower cutoff analog measurement shown in Fig. 6(b). In the analog measurements, the metal configuration was first lowered to the insulating bottom of the tank as indicated in Fig. 6(c). The resistance measured between electrodes in this fashion is proportional to the impedance of the corresponding TEM mode of the one-finger line; it is

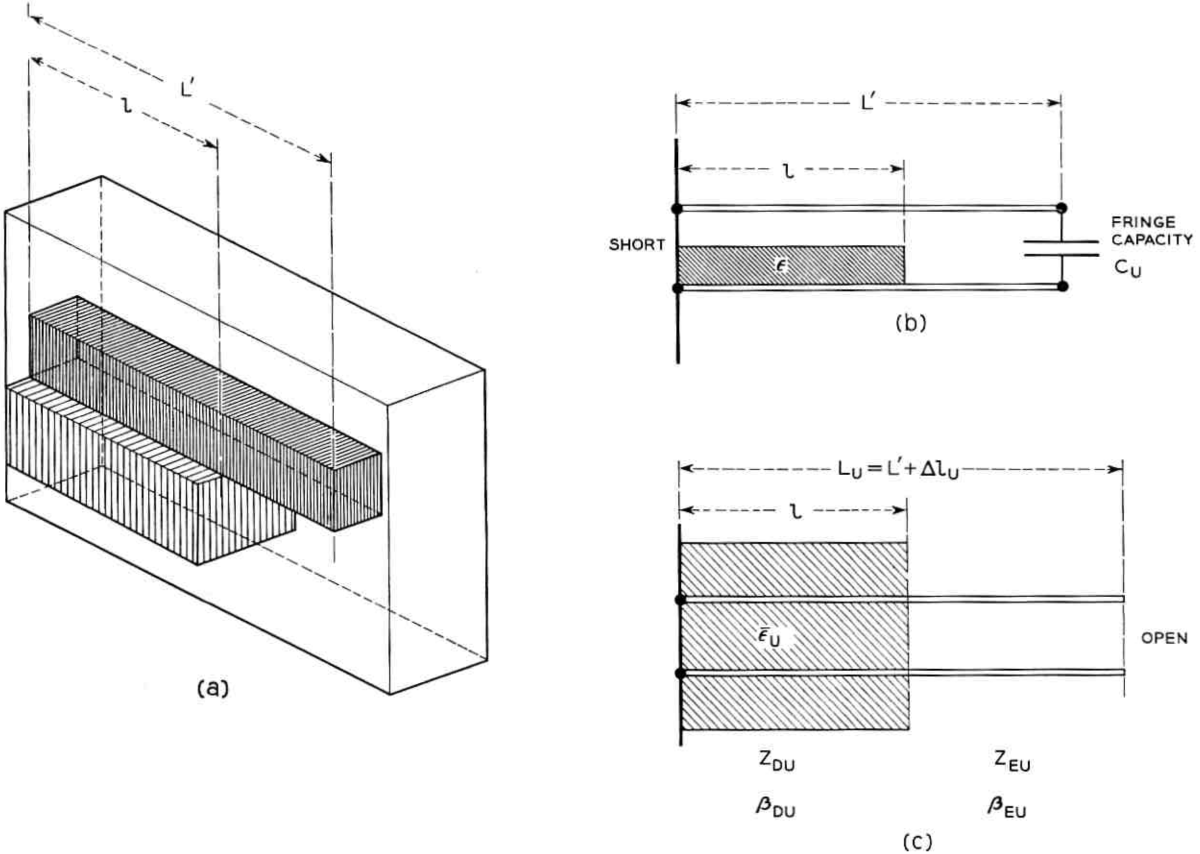


Fig. 5 — Equivalent one-finger line: (a) geometry considered, (b) two-wire line model, (c) simplified equivalent TEM mode line model; use either subscript U for upper, or L for lower, cutoff.

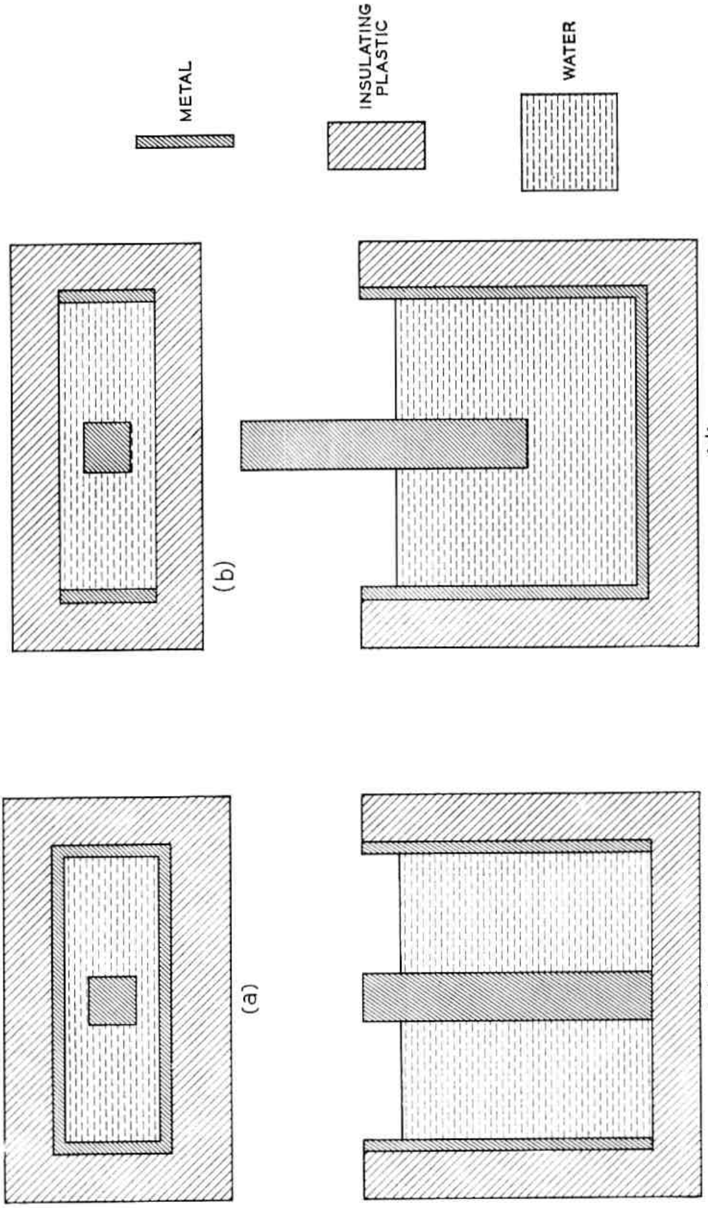


Fig. 6 — Electrolytic tank measurements.

inversely proportional to the capacitance of the line. For a second measurement, the finger was raised to the proper scaled height and a metal plate was placed on the bottom of the tank. The inverse of the resistance so measured is proportional to the capacity of the appropriate length of one-finger TEM line plus the fringe capacity arising from the diverging field pattern beyond the end of the finger. [See Fig. 6(d).] The conductivity of the tap water used was measured also. From these measurements it is possible then to evaluate the fringe capacity as well as impedance and capacitance of the TEM mode on the one-finger line.

In Fig. 7, the fringe capacitance values C_L for the lower cutoff frequency and C_U for the upper cutoff frequency obtained from the tank

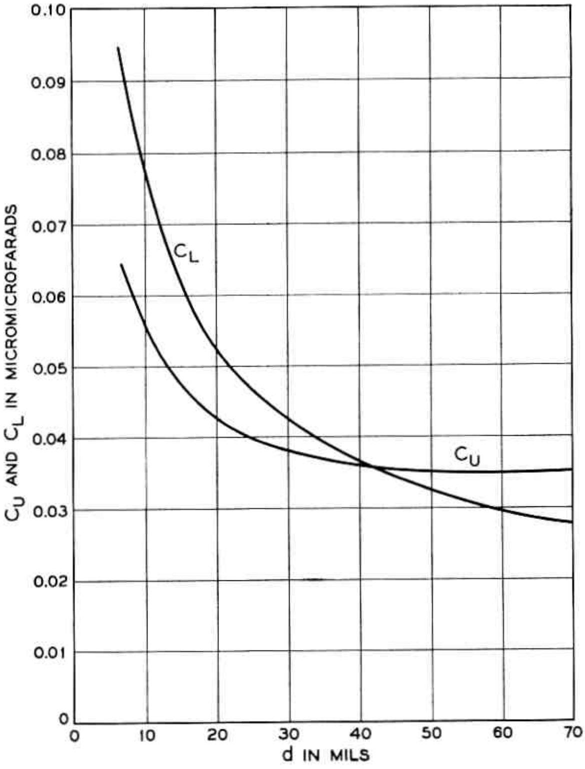


Fig. 7 — Fringing capacitance C_L for lower cutoff frequency and C_U for upper cutoff frequency. The geometry of the comb used includes fingers of cross section 0.040 x 0.040 inch, spaced 0.080 inch on center in a housing 0.240 inch wide (ratio $W_U/D_U = 1.25$). Capacitance is plotted vs spacing d between finger tips and the opposing housing wall.

measurements are shown as a function of the distance d between the finger tips and the opposite waveguide wall. The data are valid for fingers of square cross section, $D_U/2 \times D_U/2 = 0.040 \times 0.040$ inch, spaced center-to-center by $D_U = 0.080$ inch and contained in a housing of width $2W_U + D_U/2 = 0.240$ inch (aspect ratio $W_U/D_U = 1.25$). The W_U and D_U are the dimensions of the empty comb, as shown in Fig. 9 in Section III. It should be mentioned here that these data can be applied to dimensions other than those indicated if one observes two facts. First, if all linear dimensions are scaled simultaneously by some factor, the capacity is scaled by the same factor. Second, experience has shown that the fringe capacity is a very slow function of the ratio W_U/D_U ; no noticeable errors were found when these capacity values were used for W_U/D_U values ranging from 0.75 to 1.5.

Another experimental method to determine the fringe capacity is based on the availability of either a comb structure of exact size or a scale model. The upper and lower cutoff frequencies, f_{EU} and f_{EL} , of the empty, unloaded structure are measured by direct measurement. The result of the measurement can best be expressed in terms of a new effective finger length, L_U or L_L . This is based on the fact that a transmission line of physical length L', shorted at one end and terminated with a small capacity at the other, is electrically equivalent to a somewhat longer transmission line which is shorted at one end and open at the other. The effective finger lengths are different for both cutoff frequencies

$$L_U = \frac{c}{4f_{EU}} = L' + \Delta l_U \tag{8a}$$

$$L_L = \frac{c}{4f_{EL}} = L' + \Delta l_L . \tag{8b}$$

Here c is the velocity of light.

The relation between these length dimensions and the fringe capacity involves the characteristic impedance of the line. The fringe capacity follows from

$$\frac{1}{2\pi f_{EU} C_U} = Z_{EU} \tan \frac{2\pi f_{EU} L'}{c} = Z_{EU} \cot \frac{2\pi f_{EU} \Delta l_U}{c} \tag{9a}$$

and

$$\frac{1}{2\pi f_{EL} C_L} = Z_{EL} \tan \frac{2\pi f_{EL} L'}{c} = Z_{EL} \cot \frac{2\pi f_{EL} \Delta l_L}{c} . \tag{9b}$$

For the particular structure geometry investigated here in detail, the fringe capacity was determined using these equations and the characteristic impedances derived later in this section. The values of C_L and C_U obtained agree well with these from the tank measurement.

For the subsequent calculations it is assumed that the new effective lengths, L_U and L_L, are known. If, instead, the fringe capacities, C_U and C_L, are known, the new effective lengths can be calculated using the impedances Z_{EU} and Z_{EL}. Since the capacities are small, (9a) and (9b) can be approximated by

$$\Delta l_U = Z_{EU}C_U c \tag{10a}$$

$$\Delta l_L = Z_{EL}C_L c. \tag{10b}$$

In this fashion, the problem of Fig. 5(b) is reduced to that of Fig. 5(c). The cutoff frequencies, f_L and f_U, are found as resonance frequencies of a transmission line L_L or L_U long, where one end is shorted, the other open, and a length l is partially loaded with dielectric.

The field pattern in the unloaded part of the transmission line is rigorously a TEM mode. Therefore, the impedance of this line can be found by a resistance card or an electrolytic tank technique. The electrodes are shaped for the model in the same way as the conductors in the unloaded TEM line. Then the impedance of the line is simply equal to the resistance measured in the model provided the resistance per square is adjusted to or scaled to 377 ohms. In addition to this measuring technique, these impedances, Z_{EU} and Z_{EL}, will be determined analytically below. This involves a calculation with good accuracy of the electric field pattern.

The dielectrically loaded section of the transmission line would, if treated with the same rigor, require a much more involved procedure. Therefore, at this point an approximation is introduced.

(iv) The field configuration in the loaded part of the transmission line can be treated as a TEM mode.

In reality, this is not true. An exact solution of Maxwell's equations for a TEM-type transmission line having a cross section partly filled with dielectric is not a TEM mode. Instead, the process of matching boundary conditions requires the presence of longitudinal field components. It can be seen, however, that these longitudinal components will become smaller with decreasing frequency and vanish in the zero frequency limit. Thus this approximation implies the representation of a dynamic field configuration by its static analog. The accuracy of such an approximation, therefore, tends to be better the shorter the linear dimensions involved are with respect to the wavelength. In the range of dimensions used here it is expected that no appreciable loss of accuracy is incurred in this connection.

The consequences of treating the field configuration in the loaded part of the one-finger model as a TEM mode are far-reaching and very helpful for the subsequent analysis. Considering the same metal boundaries as in the unloaded part, the field configuration in the loaded part has to be the same. This follows from the fact that the TEM fields are given as a unique solution to Laplace's equation for the appropriate geometry. Thus one way to treat the loaded part of the one-finger model, consistent with a TEM mode in the same geometry, is by an effective dielectric constant. This allows for a reformulation of approximation (*iv*):

The part of the transmission line loaded partially by a high dielectric constant material can be treated as if it were loaded uniformly throughout the cross section with a material of a lower "effective" dielectric constant.

This effective dielectric constant will, of course, be different for the upper and lower cutoff frequencies. Using these effective dielectric constants, $\bar{\epsilon}_U$ and $\bar{\epsilon}_L$, the impedances and propagation constants of the loaded section are related to those of the empty section by

$$Z_{DL} = Z_{EL}/\sqrt{\bar{\epsilon}_L} \qquad Z_{DU} = Z_{EU}/\sqrt{\bar{\epsilon}_U} \tag{11}$$

$$\beta_{DL} = \sqrt{\bar{\epsilon}_L}\,\beta_{EL} \qquad \beta_{DU} = \sqrt{\bar{\epsilon}_U}\,\beta_{EU}\,. \tag{12}$$

Here the first indices E and D refer to the empty and dielectrically loaded line, the second indices L and U to the lower and upper cutoff frequencies. The propagation constants in the empty TEM line are, of course, identical to that in vacuum

$$\beta_{EL} = (2\pi f_L/c) \qquad \beta_{EU} = (2\pi f_U/c). \tag{13}$$

Assuming for the moment that the effective finger lengths, L_U and L_L, the characteristic impedances of the empty line, Z_{EU} and Z_{EL}, and the effective dielectric constants, $\bar{\epsilon}_U$ and $\bar{\epsilon}_L$, are known, the cutoff frequencies, f_U and f_L, can be calculated. The procedure is to match voltage and current at the boundary between the loaded and unloaded section of the line. This results in impedance equations

$$Z_{EU} \cot \beta_{EU}(L_U - l) = Z_{DU} \tan \beta_{DU} l \tag{14a}$$

$$Z_{EL} \cot \beta_{EL}(L_L - l) = Z_{DL} \tan \beta_{DL} l. \tag{14b}$$

These are rewritten in a more convenient form

$$\sqrt{\bar{\epsilon}_U} = \tan\left[\frac{\pi}{2}\sqrt{\bar{\epsilon}_U}\,\frac{l}{L_U}\,\frac{f_U}{f_{EU}}\right] \tan\left[\frac{\pi}{2}\left(1 - \frac{l}{L_U}\right)\frac{f_U}{f_{EU}}\right] \tag{15a}$$

$$\sqrt{\bar{\epsilon}_L} = \tan\left[\frac{\pi}{2}\sqrt{\bar{\epsilon}_L}\,\frac{l}{L_L}\,\frac{f_L}{f_{EL}}\right] \tan\left[\frac{\pi}{2}\left(1 - \frac{l}{L_L}\right)\frac{f_L}{f_{EL}}\right]. \tag{15b}$$

These equations are identical for lower and upper cutoff frequencies. They do not contain the characteristic impedances explicitly. They are solved in the following way.

$\sqrt{\bar{\epsilon}_U}$ or $\sqrt{\bar{\epsilon}_L}$ is considered a given parameter. Then the frequency ratio f_U/f_{EU} or f_L/f_{EL} is a function of l/L_L or l/L_U. This function requires the solution of transcendental equation (15a) or (15b). Numerical values were obtained by machine computations using the IBM 7090. The results are plotted in Fig. 8.

This graph can then be used to determine the upper and lower cutoff frequencies of the loaded comb structure. It is assumed here that the upper and lower cutoff frequencies of the empty comb, f_{EU} and f_{EL}, and connected with them, the effective finger lengths, L_U and L_L, are known. They are best determined by measurement, but they could also be calculated from the fringe capacity and the characteristic impedance. The quantity yet to be evaluated is the effective dielectric constant, $\bar{\epsilon}_U$ and $\bar{\epsilon}_L$, before the cutoff frequencies can be read from the graph in Fig. 8. It will be necessary, however, to work out the electric field pattern within the unloaded comb, then in the loaded section, including the respective characteristic impedances, before the effective dielectric constant can be obtained.

III. FIELD PATTERN AND CHARACTERISTIC IMPEDANCE OF UNLOADED COMB

3.1 *Upper Cutoff Frequency*

The electric field pattern of the unloaded one-finger model will look about like Fig. 9(a). This geometry is, unfortunately, too complicated for a closed analytical treatment. On the basis of the geometry and the mathematical tools at hand, the following approach may be suggested. The area available to the electric field is divided into four regions, two equivalent regions of type A and two equivalent regions of type B, as shown in Fig. 9(b). Two further approximations are then necessary.

(*v*) *The electric field in the regions A can be represented as a homogeneous, parallel plate condenser field.*

(*vi*) *The electric field in the regions B can be represented by the field produced by an infinitely thin metal fin inserted in a rectangular enclosure of corresponding dimensions.*

These approximations are illustrated in Figs. 9(c) and 9(d). Along the joints of regions A and B the field thus assumed is discontinuous. In reality, it is inhomogeneous near the boundary of region A, and it is less inhomogeneous than assumed near the boundary of region A because there is only a 90° bend, not a 180° bend as in the model used. These

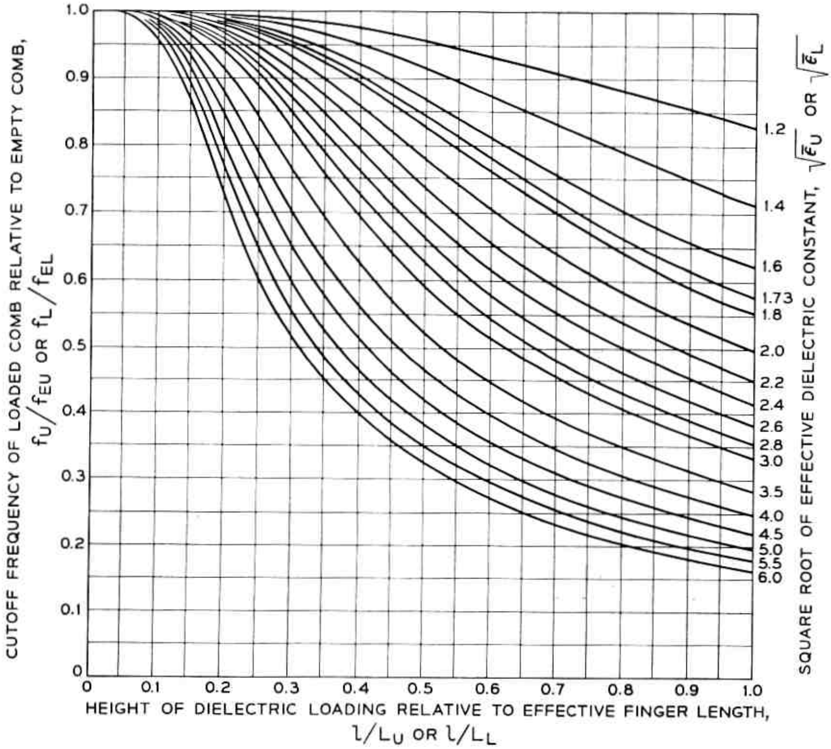


Fig. 8 — Plot of numerical values obtained by machine computations.

discrepancies of the field model from what would be expected should in reality be very small, particularly if the gap between finger and wall, the dimension $W_A = \frac{1}{2}(1 - r)D_U$ defined in Figs. 9(a) and 9(e), is small compared to other dimensions. This is so in cases of practical interest.

As far as the impedance is concerned, the two regions A and the two regions B are in parallel. The impedance of a region A is simply the ratio of its dimensions multiplied by free-space impedance. The impedance of region B is not as easily found. It is possible, however, to use a conformal transformation which maps the region B into a parallel plate geometry. This is schematically indicated in Fig. 9(e). The transformation actually utilized consists of the consecutive application of two transformations, each using elliptical functions. The procedure, including the mathematical details of the conformal transformation by elliptical functions, is outlined in the Appendix.

It is known from the theory of conformal mapping by functions of

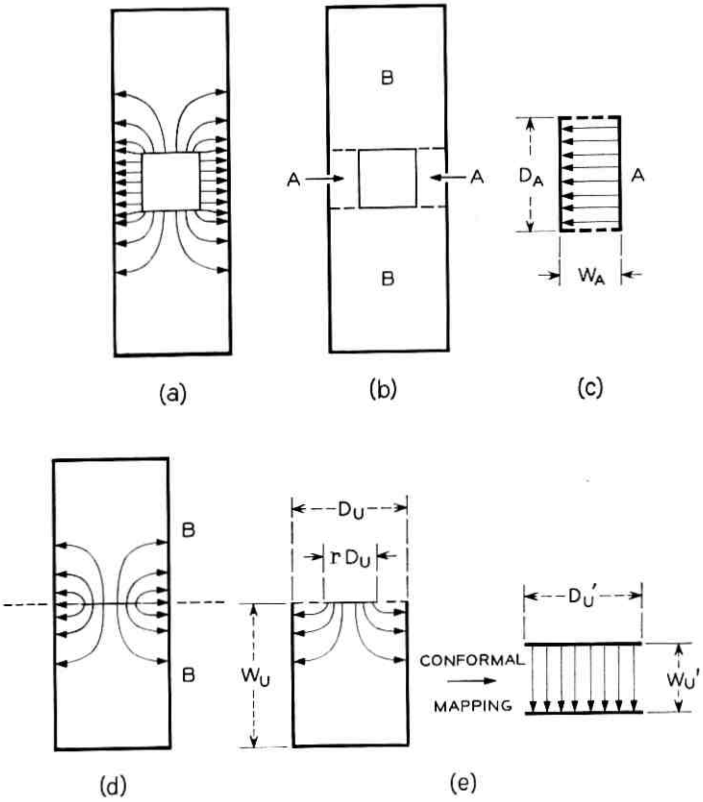


Fig. 9 — Analysis for upper cutoff frequency: (a) real field patterns, (b) regions used for analysis, (c) homogeneous field assumed in region A, (d) fin field assumed in region B, (e) fin field equivalent to homogeneous field.

complex variables that the geometry is preserved in infinitesimal regions. In particular, it is clear that infinitesimal squares with boundaries formed by field lines and equipotential lines continue to be squares. Since the impedance can be thought of as composed of the impedance of these infinitesimal squares, partly in parallel and partly in series as indicated by the over-all geometry, it follows finally that the impedance of the two transmission lines of Fig. 9(e) is the same.

The geometry before transformation is characterized by the two ratios: W_U/D_U and r. Thus W_D'/D_U' will be a function of both of these ratios. So far only combs with $r = \frac{1}{2}$ have been investigated in practice. For convenience, therefore, the subsequent calculations are carried out for this value of r. This implies a further restriction.

(*vii*) *In the numerical calculations to follow, only comb geometries with the finger width as large as the gap between fingers are considered.*

From a mathematical point of view, this restriction is somewhat arbitrary. Any other choice of r, the ratio of finger width to length of period, however, would necessitate another application of the elliptic integral conformal transformation.

With $r = \frac{1}{2}$, W_U'/D_U' is a single-valued function of W_U/D_U. This function is plotted in Fig. 10, An interesting feature of this graph is that W_U'/D_U' goes asymptotically to $\frac{1}{2}$; it reaches this value to within 2 per cent at $W_U/D_U = 0.65$. The physical interpretation of this observation is as follows. For $W_U/D_U > 0.65$, essentially all the field lines originating at the center fin in Fig. 9(d) terminate on the side wall; none reach the opposite end wall. Therefore, this wall can be moved out toward infinity with no noticeable effect on the impedance at the upper cutoff frequency.

The characteristic impedance of the empty structure at the upper cutoff frequency can now be given. It is

$$Z_{0U} = 377 \text{ ohms} \Big/ \left(2\frac{D_A}{W_A} + 2\frac{D_U'}{W_U'}\right). \qquad (16a)$$

An important special case is one where, first, W_U/D_U is greater than 0.65 so that the asymptotic value $W_U'/D_U' = \frac{1}{2}$ applies and where, second, the fingers have a square cross section so that $W_A/D_A = \frac{1}{2}$. Then the characteristic impedance is simply

$$Z_{0U} = \tfrac{1}{8} \times 377 \text{ ohms} = 47.1 \text{ ohms}. \qquad (16b)$$

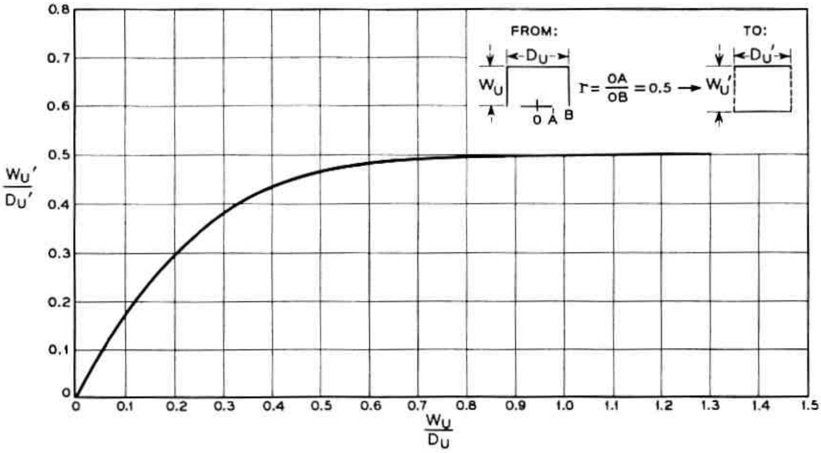


Fig. 10 — Conformal transformation for upper cutoff.

Since the partial impedances are equal, it also follows in this case that the total stored energy is equally distributed between the four regions A, A, B, B. This remark may be helpful in estimating the filling factor.

3.2 *Lower Cutoff Frequency*

The procedure here is quite similar to that in the case of the upper cutoff frequency. The field pattern is illustrated in Fig. 11(a). The cross-section area available to the electric field is divided into four regions, two electrically equivalent regions of type A and two regions of type B, as shown in Fig. 11(b). Again two approximations are required.

(*viii*) *The electric field in region A is so small that it can be neglected.*

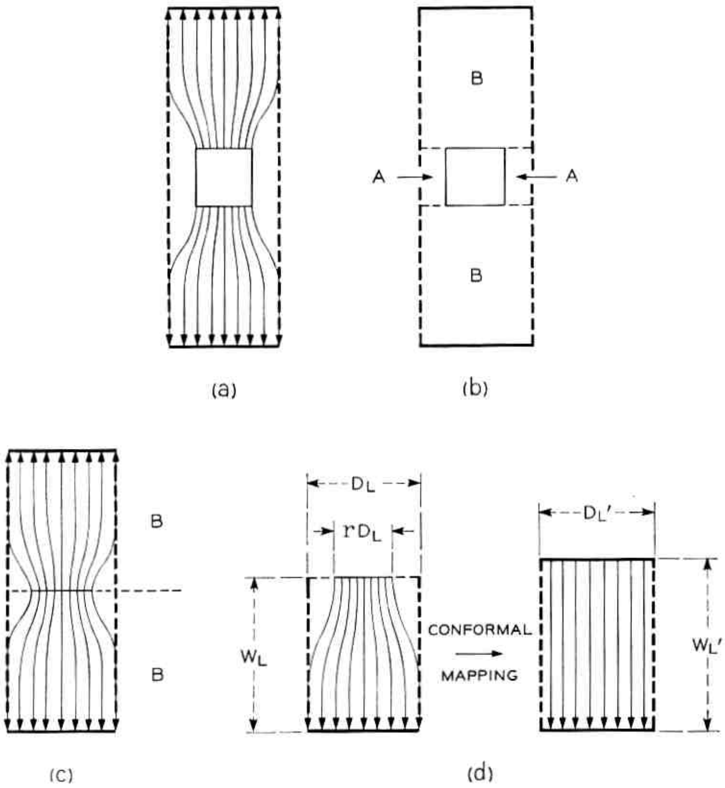


Fig. 11 — Analysis for lower cutoff frequency: (a) typical electric field pattern, (b) regions used for analysis, region A assumed field-free, (c) fin field assumed in region B, (d) fin field equivalent to homogeneous field.

(*ix*) *The electric field in region B can be represented by the field produced by an infinitely thin metal fin inserted into a rectangular enclosure with appropriate dimensions and boundary conditions.*

It is apparent that the approximation (*viii*) is justified. Only very small fringing fields will exist in region A. The implication of approximation (*ix*) is indicated in Fig. 11(c). It should also be very well justified, since there is no essential difference between the idealized field pattern and the real one. Region B can be transformed into a simple parallel plate geometry. This is indicated in Fig. 11(d). The transformation again consists of two consecutive conformal mappings by means of elliptic functions. The procedure is outlined in the Appendix. The impedance of region B is simply given by the aspect ratio W_L'/D_L' of the parallel plate geometry resulting from the transformation, multiplied by the free-space impedance. This resulting ratio W_L'/D_L' is a function of two ratios, r and W/D. For mathematical convenience and because of practical importance, only comb geometries with $r = \frac{1}{2}$ are considered in the subsequent calculations. For other ratios r, a new evaluation of the elliptical transformation is necessary. Thus restriction (*vii*) is invoked here, too.

(*vii*) *In the numerical calculations which follow, only comb geometries with the finger width equal to the gap width between fingers are considered.*

The single-valued function W_L'/D_L' of W_L/D_L with the parameter $r = \frac{1}{2}$ is shown in Fig. 12. The characteristic impedance of the empty structure at the lower cutoff frequency is then given by

$$Z_{0L} = \frac{W_L'}{2D_L'} \times 377 \text{ ohms.} \tag{17a}$$

As long as $W_L/D_L > 0.2$, it is seen from the graph that this can be approximated by

$$Z_{0L} = \frac{1}{2}\left(\frac{W_L}{D_L} + 0.11\right) 377 \text{ ohms.} \tag{17b}$$

The asymptotically linear curve in Fig. 12 and this last equation suggest an almost obvious interpretation. The electrical behavior of region B is essentially the same as that of a parallel plate geometry having the same width $D_L' = D_L$, but a slightly greater distance between plates, $W_L' > W_L$. Also, the asymptotic slope for the curve is unity. Considering a geometry with $W_L/D_L > 0.2$, this would mean the following. If W_L is increased further, the electric field pattern near the fin stays the same, while the added volume away from the finger is taken up by a homogeneous electric field.

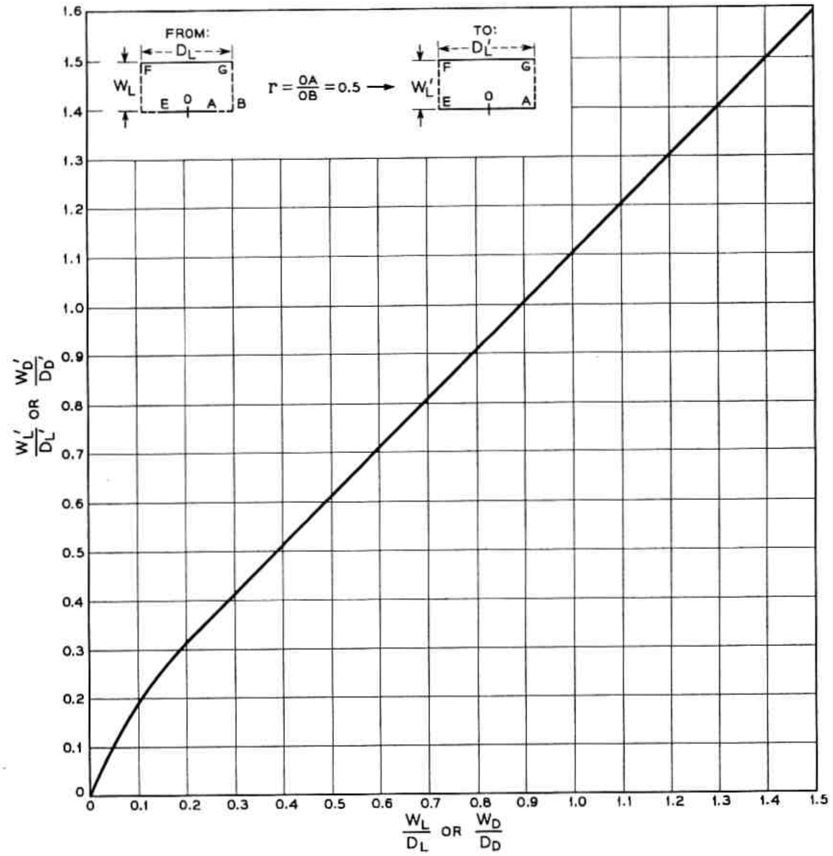


Fig. 12 — Conformal transformation for lower cutoff.

IV. CAPACITANCE AND EFFECTIVE DIELECTRIC CONSTANT OF COMB PARTIALLY FILLED WITH DIELECTRIC

It was mentioned that the electromagnetic field configuration in the comb line partially loaded with dielectric should be treated as a TEM mode. It was pointed out that this is equivalent to finding a static solution of the electric field problem. Thus the problem here is to find the static value of the capacitance per unit length of the loaded finger line. The difference in electrical behavior of the loaded line compared to the unloaded line is then fully expressed by an effective dielectric constant. This effective dielectric constant is simply the ratio of the static capacitance of the loaded line to the capacitance of the unloaded line.

4.1 *Upper Cutoff Frequency*

The field pattern in the presence of one dielectric slab is illustrated in Fig. 13. It is seen that the dielectric is present in one of the regions called B before. The usual boundary conditions for the continuity of the tangential E vector and of the normal D vector have to be observed in fitting together the electric field pattern inside and outside the dielectric.

At first sight it seems that no difficulty is incurred in this respect at the boundary of the dielectric. In the model chosen for the field configuration, the field lines run parallel to the boundary both in regions A and B. The boundary condition for tangential electric field seems to apply, with the consequence that the field pattern remains the same in the dielectric as before in the unloaded region B. Calculations are based on this assumption, and they are presumably of sufficient accuracy for present purposes.

There is a small error in this assumption. It was pointed out before that the two models chosen to represent the field in regions A and B do not match at the boundary. In the models, the field in A is homogeneous, that in B strongly inhomogeneous. The real field at the boundary of A and B should be somewhere between these two extremes. It is expected,

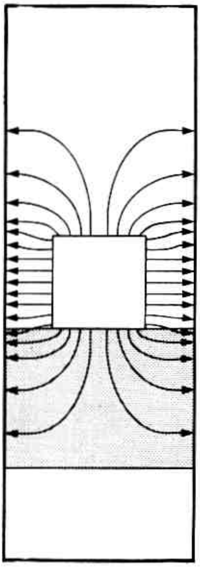


Fig. 13 — Electric field pattern at upper cutoff with dielectric loading present.

therefore, that the error in the impedance calculation of the empty comb at the upper cutoff frequency is negligible. The same is not necessarily true in the presence of dielectric. The real, inhomogeneous field in the region near the boundary of A and B will be disturbed by the insertion of dielectric. The deviations of the real field from that used in the calculations — homogeneous in A, elliptic function field in B — are now accentuated by a high dielectric constant rather than evened out as in the empty comb. This will lead to an error in the calculation of the capacitance and the effective dielectric constant. Hence it is not trivial that the approximations (*v*) and (*vi*) are still reasonably good in the presence of dielectric. Fortunately, it can be argued that the error incurred by this approximation is still negligible within the accuracy sought for here and with respect to typical structure geometries and dielectric constants considered. A formulation of the approximation follows.

(*x*) *In the presence of dielectric loading, the static electric field can still be represented by a homogeneous, parallel plate field in region A and the field of a metal fin inside a rectangular enclosure in region B filled by the dielectric.*

The next concern is the other boundary of the dielectric away from the finger. Here the field lines cross the boundary at all directions between tangential and perpendicular. It would be very difficult to apply boundary conditions to this field pattern. Therefore another restriction is introduced.

(*xi*) *The calculation is restricted to dielectric loadings thick enough so that essentially the total electric field energy of region B is contained within the dielectric.*

The numerical implication of this restriction follows directly from Fig. 10. It is assumed that the fingers are as wide as the gap between them. From the graph the following fact can be deduced. If a geometry is considered where W_U is considerably larger than D_U , then 98 per cent of the electric field energy is concentrated in a rectangle near the finger, D_U wide and 0.65 D_U deep. Restriction (*xi*) thus implies that only dielectric slabs which have a thickness of at least 0.65 times the length of a period of the comb are considered.

Fortunately, this restriction does not exclude any cases of practical interest. Since the field configuration on the finger is treated here as a TEM mode, the filling factor in the plane perpendicular to the finger is the same for the dielectric and the magnetic field energy. Thus, slabs thinner than indicated by restriction (*xi*) would also have a reduced gain interaction near the upper cutoff frequency, since not all of the magnetic field energy of region B would be contained in the maser material. Gain is still at a premium in present TWM development, and thus it does not seem to be necessary to treat cases other than those restricted by (*xi*).

It is now possible to write down the capacitance and the effective dielectric constant. By comparison with (16a), it is seen that the capacitance per unit length of the empty one-finger line is

$$c_{EU} = \epsilon_0 \left[2 \frac{D_A}{W_A} + 2 \frac{D_U'}{W_U'} \right]. \tag{18}$$

(Lower case c is used to distinguish this quantity from the fringe capacity C_U.) With dielectric loading on one side of the finger

$$c_{DU} = \epsilon_0 \left[2 \frac{D_A}{W_A} + (\epsilon + 1) \frac{D_U'}{W_U'} \right] \cdots. \tag{19a}$$

Similarly, if the dielectric is loaded on both sides of the finger

$$c_{DU} = \epsilon_0 \left[2 \frac{D_A}{W_A} + 2\epsilon \frac{D_U'}{W_U'} \right] \cdots. \tag{19b}$$

The effective dielectric constant is then simply, for loading on one side

$$\bar{\epsilon}_U = [(\epsilon + 1) + 2b]/[2 + 2b] \tag{20a}$$

and for loading on both sides

$$\bar{\epsilon}_U = (\epsilon + b)/(1 + b) \tag{20b}$$

with

$$b = D_A W_U'/W_A D_U'. \tag{21}$$

Most important perhaps for present applications is the case where, first, the fingers are square so that (16b) applies and where, second, the dielectric is ruby with an isotropic average dielectric constant of $\epsilon = 9$. In that case, for loading on one side

$$\bar{\epsilon}_U = 3 \tag{22a}$$

and for loading on both sides

$$\bar{\epsilon}_U = 5. \tag{22b}$$

4.2 *Lower Cutoff Frequency*

The field pattern in the presence of one dielectric slab is illustrated in Fig. 14. The dielectric fills part of the region called B before. For the evaluation of the capacitance it is significant that restriction (*xi*) is applied here, too. Then the following approximation can be made.

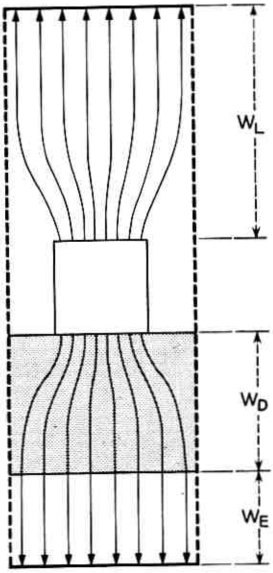


Fig. 14 — Electric field pattern at lower cutoff with dielectric loading present.

(*xii*) *In the presence of dielectric loading, the static electric field can be represented in the following way: There is zero field in region A; in the dielectric there is a field like that produced by a metal fin in a rectangular enclosure, having the dimensions of the dielectric and subject to appropriate boundary conditions. The field past the dielectric is a homogeneous parallel plate field.*

It can be argued that these approximations are well justified. There is no potential difference between fingers; hence region A should be field-free except perhaps for some very small fringe fields. In connection with 17(b) it was shown that the field has its inhomogeneities near the finger, whereas the field region near the wall is reasonably homogeneous.

The capacitance per unit length of the loaded one-finger model can now be given. For the empty line it is

$$c_{EL} = 2\epsilon_0(D_L'/W_L'). \tag{23}$$

For the loaded line, the capacitance is obtained from two contributions in parallel, one from each side of the finger. The capacitance of the loaded side comes from two contributions in series: one from the dielectric, involving an elliptical transformation using the dimensions of the dielectric, and one a parallel plate contribution from the space behind the dielectric. Thus, for dielectric loading on one side

$$c_{DL} = \epsilon_0 \left[\frac{D_L'}{W_L'} + \left(\frac{W_D'}{\epsilon D_D'} + \frac{W_E}{D_E} \right)^{-1} \right] \tag{24a}$$

and for loading on both sides

$$c_{DL} = 2\epsilon_0 \Big/ \left(\frac{W_D'}{\epsilon D_D'} + \frac{W_E}{D_E} \right). \tag{24b}$$

Here W_D and D_D are the physical dimensions of the dielectric cross section per one-finger line, W_E and D_E the dimensions of the empty space behind the dielectric. W_D'/D_D' is obtained from W_D/D_D by means of the elliptical transformation illustrated in Fig. 11.

The effective dielectric constant can now be evaluated as the ratio c_{DL}/c_{EL}. The formulas, however, turn out to be fairly long. They are given here, therefore, only for the case that the approximation in (17b) is valid both for the empty structure and the dielectric. It is further observed that

$$D_L = D_D = D_E = D$$

and

$$W_L = W_D + W_E .$$

Then the effective dielectric constant for dielectric loading on one side is

$$\bar{\epsilon}_L = \tfrac{1}{2} \frac{2\epsilon W_L - (\epsilon - 1)W_D + (\epsilon + 1)0.11D}{\epsilon W_L - (\epsilon - 1)W_D + 0.11D} \tag{25a}$$

and similarly, for loading on both sides

$$\bar{\epsilon}_L = \epsilon \frac{W_L + 0.11D}{\epsilon W_L - (\epsilon - 1)W_D + 0.11D}. \tag{25b}$$

It is seen that the effective dielectric constant is a function of ϵ, W_D/W_L and D/W_L. Once a particular structure geometry has been picked, then D/W_L is known. If a particular maser material is selected, ϵ is known. Then $\bar{\epsilon}_L$ is a unique function of the relative loading thickness, W_D/W_L. One example of such a function is given in Fig. 15. For convenience in using the graph of Fig. 8, the square root $\sqrt{\bar{\epsilon}_L}$ is given instead of $\bar{\epsilon}_L$. Curves for effective constants based on other parameters can easily be calculated using either (25a) or (25b).

V. EXAMPLE FOR DESIGN PROCEDURE

In Sections III and IV the empty and the dielectrically loaded comb structure were evaluated. Field pattern, impedance and propagation constants were obtained for both the upper and lower cutoff frequencies.

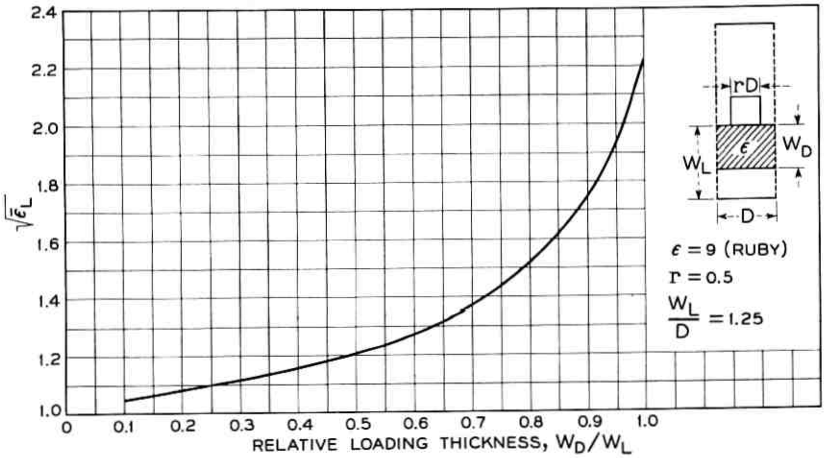


Fig. 15 — $\sqrt{\bar{\epsilon}_L}$ vs relative loading thickness, one side only, loaded at lower cutoff frequency for parameters indicated.

With this information at hand, it is now possible to arrive at a numerical design procedure. The aim is to predict the cross-sectional dimensions of a dielectric parallelepiped which will simultaneously tune the upper and lower cutoff frequencies of the comb structure to some predetermined values. Of course, it is not possible to ask for completely arbitrary design cutoff frequencies. Obviously there are limits to the amount of tuning which can be achieved by a given dielectric material within a given comb geometry. These limits can also be determined easily by the analysis.

The design procedure follows the outlines given briefly at the end of Section II. It can now be described in general terms. Perhaps it is advantageous, however, to illustrate the procedure by means of an example. The example to be described is a case of a "design on paper." That is to say, the design calculations can be made entirely on the basis of calculable values. It is not necessary to fabricate a size or scale model of the comb structure under consideration in order to determine certain values by measurement. The only empirical value required is the fringe capacity between finger tip and the structure enclosure; this may be obtained from Fig. 7.

One interesting and valuable feature of the design procedure is that of independently setting the upper and lower cutoff frequencies. This is possible because the upper cutoff frequency can be controlled by adjusting the height l of the dielectric loading alone, and because it is not dependent in any way on the dielectric thickness W_D as long as W_D exceeds a certain small minimum value. Then the dimension W_D can be used to control the lower cutoff frequency independently.

As an example for the design procedure, a comb structure is considered with the following dimensions:

(a) finger length 0.400 inch
(b) spacing between fingers 0.040 inch
(c) finger cross section 0.040 square inch
(d) wall-to-wall spacing of enclosure 0.240 inch.

As further information, the fringe capacity was measured in an electrolytic tank model and was found to be (see Fig. 7 for gap spacing greater than 70 mils):

(e) fringe capacity $C_L = 0.025\ \mu\mu F$, $C_U = 0.035\ \mu\mu F$.

The problem considered is that of finding the dimensions for a single ruby parallelepiped which brings the upper cutoff frequency to 4200 mc and the lower cutoff frequency to 3210 mc.

First step: Find effective finger length at the upper cutoff frequency of the empty comb.

Equation (10a) applies for the increase in length and (16b) applies for the impedance; thus

$$\begin{aligned}\Delta l_U &= Z_{EU} C_U c \\ &= 47.1 \times 0.025 \times 10^{-12} \times 3 \times 10^{10} \\ &= 0.035 \text{ cm.}\end{aligned}$$

The effective length for upper cutoff is then [see (8a)]

$$\begin{aligned}L_U &= L' + \Delta l_U \\ &= 2.54 \times 0.400 + 0.035 = 1.051 \text{ cm.}\end{aligned}$$

This corresponds to an upper cutoff frequency for the empty comb [see (8a)]

$$\begin{aligned}f_{EU} &= c/4L_U \\ &= 7150 \text{ mc.}\end{aligned}$$

Thus the design specification

$$f_U = 4200 \text{ mc}$$

is equivalent to specifying a ratio of

$$f_U/f_{EU} = 0.587.$$

Second step: Find in an analogous way the effective finger length at the lower cutoff frequency of the empty comb.

Equation (17b) applies for the impedance. From the dimensions given, $W_L = 0.100''$, $D_L = 0.080''$, hence

$$Z_{EL} = \frac{1}{2}\left(\frac{W_L}{D_L} + 0.11\right) 377 \text{ ohms}$$

$$= 256 \text{ ohms.}$$

The addition to length is given by (10b)

$$\Delta l_L = Z_{EL} C_{Lc}$$

$$= 256 \times 0.035 \times 10^{-12} \times 3 \times 10^{10}$$

$$= 0.268 \text{ cm.}$$

The effective length for lower cutoff becomes

$$L_L = L' + \Delta l_L$$

$$= 0.400 \times 2.54 + 0.268$$

$$= 1.284 \text{ cm}$$

corresponding to a cutoff frequency for the empty comb

$$f_{EL} = c/4L_L$$

$$= 5840 \text{ mc.}$$

The design specification of

$$f_L = 3210 \text{ mc}$$

is thus equivalent to specifying a ratio

$$f_L/f_{EL} = 0.55.$$

Third step: Satisfy the upper cutoff frequency specification by choosing an appropriate dielectric height l without regard for W_D, the thickness of the loading. This is possible because, as mentioned before, the effective dielectric constant at the upper cutoff frequency is independent of loading thickness. The effective dielectric constant, $\bar{\epsilon}_U$, for one-sided loading with ruby is 3 from (22a); thus

$$\sqrt{\bar{\epsilon}_U} = 1.73.$$

Consulting Fig. 8 for the dielectric height which makes $f_U/f_{EU} = 0.59$ with the parameter $\sqrt{\bar{\epsilon}_U}$, it is seen that

$$l/L_U = 0.96.$$

Hence, the dielectric loading height should be

$$l = 0.96 \times 1.051 = 1.010 \text{ cm}$$
$$= 0.398''$$

In other words, the dielectric loading height turns out to be very nearly the same as the finger length.

Fourth step: Satisfy the lower cutoff frequency specification by choosing an appropriate thickness W_D of the dielectric loading. This is done by the following successive measures.

From the loading height l just determined find

$$l/L_L = 1.010/1.284$$
$$= 0.79.$$

Enter the graph of Fig. 8 with $l/L_L = 0.79$ and $f_L/f_{EL} = 0.55$. The value interpolated at the point having these two coordinates is

$$\sqrt{\bar{\epsilon}_L} = 2.22.$$

The graph of Fig. 15 is valid for present calculations; entering this last value into the graph it is found that

$$W_D = W_L$$

hence

$$W_D = 0.100 \text{ inch.}$$

The final answer, then, is that the comb described initially will have the specified cutoff frequencies if a slab of ruby of height 0.398 inch and of width 0.100 inch is inserted.

An experiment was carried out to check the results of this calculation. The two cutoff frequencies of a comb as specified above were measured after inserting a single slab of polycrystalline high density alumina (dielectric constant ≈9.3) with cross-sectional dimensions of 0.400 inch and 0.100 inch. The cutoff frequencies measured were 4200 mc and 3210 mc respectively. These frequencies were then specified as design frequencies for the above example. The close agreement between the actual dimensions of the alumina slab and those calculated by the present recipe is gratifying. It may be argued, however, that the obtained agreement is somewhat fortuitous. In particular, one should expect that the fringe capacity is altered if the dielectric loading extends all the way along the fingers up to the finger tips. To investigate the accuracy of the present analysis, a series of systematic measurements was made.

For this study a number of short sections of comb structures were built and tested. They all had finger dimensions of 0.040 × 0.040 × 0.445 inch, and the fingers were spaced 0.080 inch on center. The structures were loaded symmetrically with two slabs of high-density polycrystalline alumina (dielectric constant quoted to be 9.3) of full finger height. The geometry and the result of the measurements are shown in Fig. 16. In two series of measurements, the fraction of the housing width filled by the alumina loading, W_D/W_L, was held at 0.90 and 0.95, respectively, while the gap width between the finger and the housing wall, $W_L = W_U$, was varied in the range $0.75D = 0.060$ inch, $D = 0.080$ inch, $1.25D = 0.100$ inch and $1.5D = 0.120$ inch. From the analysis, it is known that f_U should be independent of these dimensional changes. This is borne out by the experiment. Both the experimental points and the solid line for the theoretical value of f_U show the frequency independence. It is observed, however, that the experimental frequencies are 3.5 per cent higher. A somewhat greater disagreement is found for the lower cutoff frequency, which seems to indicate a systematic trend between theory and experiment. It can be said, however, that the largest deviations are

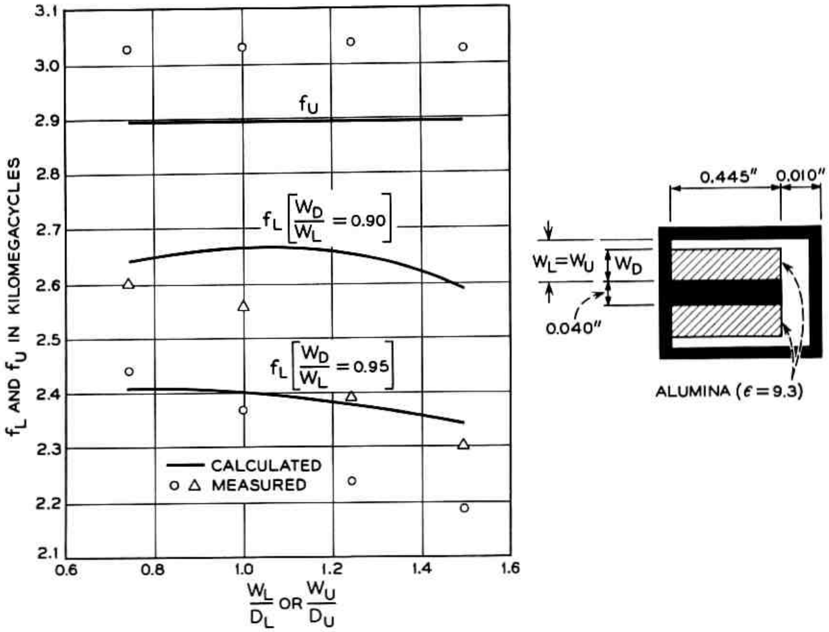


Fig. 16 — Examples of measured and calculated cutoff frequencies; the insert shows the comb geometry used.

10 per cent and that the typical discrepancy between theory and experiment is less than 5 per cent. The chance for greater systematic errors increases, of course, if comb and loading geometries are considered which comply less rigorously with the restrictions and approximations made in the text.

The numerical examples shown demonstrate that dielectric loading indeed decreases the fundamental passband frequency of the empty comb by a very appreciable factor. A one-sided loading with ruby may reduce the frequencies by a factor of 1.7, while double-sided loading may lead to a reduction by a factor 2.5. Still greater reductions may be obtained by using dielectric materials with higher dielectric constants and by modified comb geometries, in particular by changing the finger cross section from square to rectangular. It is also clear from this treatment that the shaping of the dielectric loading can be used to vary the degree of slowing within wide limits. These remarks may suffice here to illustrate the prominent role of dielectric loading techniques in the field of TWM development which was pointed out in the Introduction.

Since the original derivation of this analysis in 1960,[4] several TWM's have been developed in this laboratory. They include the TWM for the ground station receiver in the Telstar satellite communication experiment[6] and radio astronomy TWM preamplifiers for hydrogen line work at 1420 mc.[7] In these cases, the analysis has proved to be a valuable aid for arriving at a first-order design and similarly for providing guidelines in the subsequent improvements of these designs.

APPENDIX

The conformal mapping transformations are derived and evaluated, leading to the impedance transformation curves in Figs. 10 and 12. The mathematical treatment given here is not too extensive, because the type of transformation used is known from other areas of electrical engineering. Yet the description of the mathematical procedure is made reasonably complete so that it may be useful as a guide for treating other related problems: for example, traveling-wave masers where the finger width is not identical to the spacing between fingers.

A.1 *The Schwartz-Christoffel Transformation*

The particular conformal transformation used here is a special case of the more general Schwartz-Christoffel transformation. The theorem proved independently by these two mathematicians states that it is possible to find an analytical function which maps the inside of a polygon on the complex plane into the upper half of this plane. The boundary of the polygon thus is mapped into the real axis. If two transformations are considered, one of the type mentioned, the other performing the inverse function, it follows that the inside of a polygon can be mapped into the inside of any other polygon.

The general Schwartz-Christoffel transformation is illustrated in Fig. 17. For purposes of discussion, it is perhaps easier to consider first the inverse transformation of the upper half of the complex plane into a polygon. The transformation will be accomplished by a function whose derivative is given by a product of the type

$$\frac{dz}{dw} = (w - a)^{(\alpha/\pi - 1)}(w - b)^{(\beta/\pi - 1)}(w - c)^{(\gamma/\pi - 1)} \cdots. \qquad (26)$$

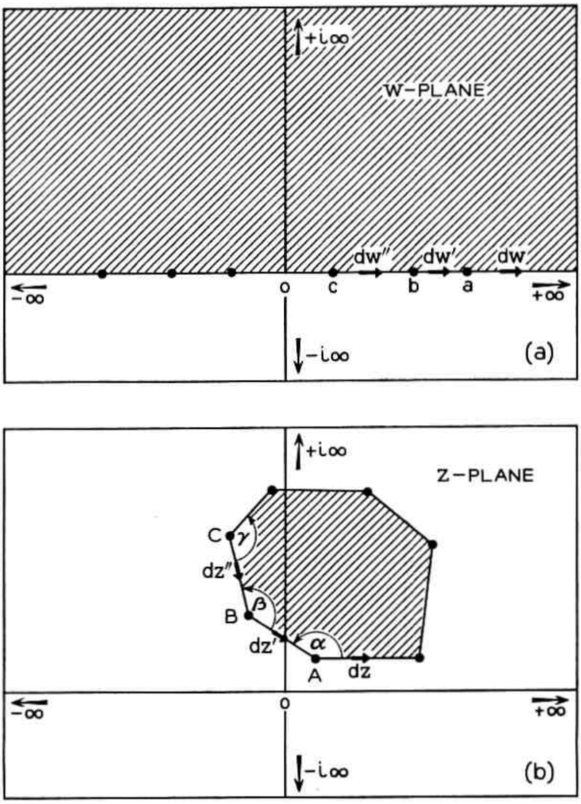


Fig. 17 — Illustration of the general mapping properties of the Schwartz-Christoffel transformation.

To demonstrate the transformation property, consider values of w and dw on the real axis. Also represent each factor in the form $r_k e^{i\phi_k}$ with a real number r_k for the magnitude and ϕ_k for the angle. It is seen then that for values w such that $w > a,b,c \cdots$ all the ϕ_k on the right-hand side of (26) vanish. Hence the angles of dz and dw are identical; that is, these line elements are parallel. Mathematically

$$\Delta dz = 0 \qquad \text{if } w \text{ and } dw \text{ are real and } w > a,b,c \cdots. \tag{27a}$$

For values $a < w < b$ the first bracket changes sign; that is, its angle is π. The angle of the first factor becomes $\alpha - \pi$

$$\Delta dz' = \alpha - \pi \qquad \text{if } w \text{ and } dw' \text{ are real and } a > w > b,c \cdots. \tag{27b}$$

That is to say, the real axis of the w plane near a is transformed in the z plane into a polygon corner at some as yet undetermined point A including an angle α. Similarly

$$\Delta dz'' = \alpha + \beta - 2\pi \qquad \text{if } w \text{ and } dw'' \text{ are real and } a,b > w > c \cdots \tag{27c}$$

indicating another polygon corner at B including an angle β and corresponding to the point b on the real axis of the w plane.

In this fashion, it is shown that the transformation (26) indeed maps the upper half of the w plane into the inside of a polygon having specified angles α, β, $\gamma \cdots$ at points in the z plane corresponding to $a,b,c \cdots$ in the w plane. While it is thus easy to satisfy conditions on the angles of the polygon, the difficulty is to find the points $A,B,C \cdots$ in the z plane which correspond to $a,b,c \cdots$ in the w plane. This requires an evaluation of the integral of (26).

Even more typical for engineering applications, and important in the present example, is the inverse situation. The corner points $A,B,C \cdots$ of the polygon are given. Then the problem is to find the real numbers $a,b,c \cdots$ which when inserted into (26) will transform this polygon into the upper half of the w plane. In most cases, this problem can only be solved numerically. The procedure would be to tabulate integrals of (26) for some range of values $a,b,c \cdots$. Numbering such tables with the given integral values $A,B,C \cdots$, the appropriate transformation parameters a,b,c could be picked.

To keep the need for tabulation down to a manageable chore, the number of significant parameters has to be restricted as much as possible. The example of importance in this connection is the mapping of a rectangle into the upper half of the complex plane. The number of significant parameters here can be reduced to one, the length ratio of two adjacent sides. Other parameters can be eliminated by trivial transformations such as scaling and rotation of the coordinate system, which is accomplished simultaneously by a complex constant factor in (26) or a shift of the coordinate origin which corresponds to the integration constant of (26).

A.2 *Mapping of a Rectangle into the Upper Half of the Complex Plane*

It is now possible to write down the transformation equation for a rectangle. The conventional notation is illustrated in Fig. 18. The corners of the rectangle in the z plane are the complex numbers K, $K + iK'$, $-K + iK'$ and $-K$. In the w plane they correspond to the points 1, $1/k$, $-1/k$ and -1 on the real axis.

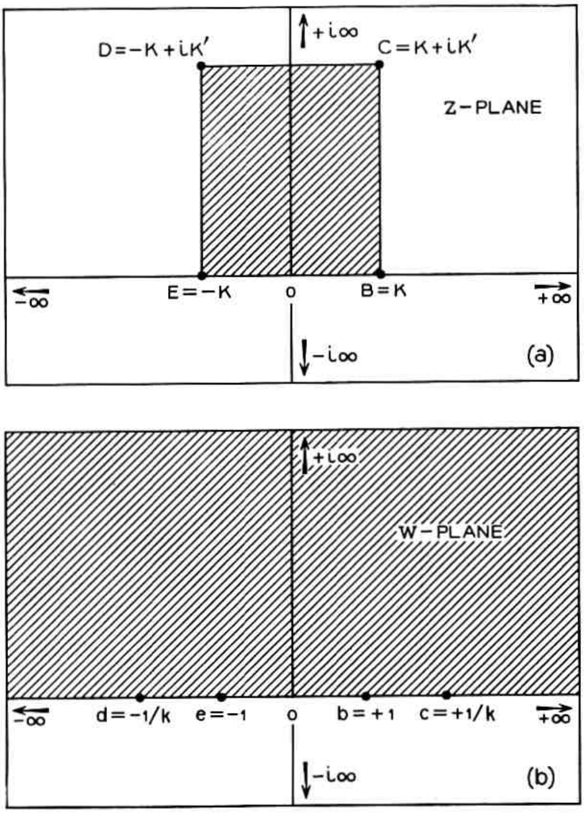


Fig. 18 — Illustration of the transformation of a rectangle in the z plane into the upper half of the w plane, introducing the conventional mathematical notation.

From (26) the transformation derivative is

$$\frac{dz}{dw} = A\left(w - \frac{1}{k}\right)^{-\frac{1}{2}} (w - 1)^{-\frac{1}{2}}(w + 1)^{-\frac{1}{2}}\left(w + \frac{1}{k}\right)^{-\frac{1}{2}}. \tag{28}$$

When the constant A is chosen appropriately ($A = -1/k$) this becomes

$$dz = \frac{dw}{(1 - w^2)^{\frac{1}{2}}(1 - k^2w^2)^{\frac{1}{2}}} \tag{29}$$

and

$$z = \int_{\omega=0}^{w} \frac{d\omega}{(1 - \omega^2)^{\frac{1}{2}}(1 - k^1\omega^2)^{\frac{1}{2}}}. \tag{30}$$

This integral is an elliptical integral of the first kind. It gives z as a function of w and k, where k is referred to as the modulus of the integral.

From the definition adapted in the figures it follows that

$$K = \int_{\omega=0}^{1} \frac{d\omega}{(1 - \omega^2)^{\frac{1}{2}}(1 - k^2\omega^2)^{\frac{1}{2}}} \tag{31}$$

and

$$iK' = \int_{\omega=1}^{1/4} \frac{d\omega}{(1 - \omega^2)^{\frac{1}{2}}(1 - k^2\omega^2)^{\frac{1}{2}}}. \tag{32}$$

K is called the complete elliptical integral. K' is the complete integral to the complementary modulus obeying the functional relationship

$$K'(k) = K(k') \tag{33}$$

where $k^2 + k'^2 = 1$ is used to define the modulus k' as complementary to k.

The definition of the elliptical integral of the first kind as given in in (30) is due to Jacobi. Many tables use also the notation of Legendre. This is obtained by setting

$$\begin{aligned} w &= \sin\phi, \qquad & dw &= \cos\phi\, d\phi \\ k &= \sin\theta, \qquad & k' &= \cos\theta. \end{aligned} \tag{34}$$

Then

$$z = \int_{\Psi=0}^{\phi} \frac{d\Psi}{(1 - \sin^2\theta \sin^2\Psi)^{\frac{1}{2}}} \tag{35}$$

$$K = \int_{\Psi=0}^{\pi/2} \frac{d\Psi}{(1 - \sin^2\theta \sin^2\Psi)^{\frac{1}{2}}} \tag{36}$$

$$K' = \int_{\Psi=0}^{\pi/2} \frac{d\Psi}{(1 - \cos^2\theta \sin^2\Psi)^{\frac{1}{2}}}. \tag{37}$$

From this discussion it is clear that the transformation of a rectangle into the upper half plane requires finding the modulus k or equivalently the modular angle θ of the elliptical integral *from the given geometry of the rectangle*. It is further clear that K and K' are not independent, but related through either (31) and (32) or (36) and (37). Therefore, it is not possible to specify both length dimensions of the rectangle of Fig. 18 but rather only their ratio. The problem thus is reduced to finding the dependence of the modulus k or θ from the aspect ratio $K'/2K$ of the rectangle.

This functional dependence was evaluated using the Smithsonian Elliptic Function Tables, in particular tables of complete elliptical integrals. The result is presented in Fig. 19.

It should be added that frequently, instead of the elliptical integral (30), its inverse is used. This inverse function is written

$$w = \text{sn } z \text{ modulo } k \tag{38}$$

which is defined to mean (30). This notation is reminiscent of the sine function, with which the sn function is indeed identical in the special case $k = 0$.

A.3 *Mapping of the Upper Cutoff Frequency Configuration*

It is now possible to carry out the mapping transformations used in the comb structure analysis. The initial geometry for the lower cutoff frequency is indicated in Fig. 20(a), where solid lines represent conducting electrodes. The final result is a parallel plate geometry like that of Fig. 20(d). This figure represents the cross section of an idealized transmission line for which the impedance is simply given by the ratio of the length dimensions times free-space impedance. The transformation makes use of two intermediate steps. The interior of the rectangle (Fig. 20a) is first mapped into the upper half of the complex plane (Fig. 20b). Then a readjustment of the scale leads to Fig. 20(c). Then the upper half plane is finally mapped into the inside of the desired rectangle (Fig. 20d) with electrodes only on opposite sides.

To keep track of these steps, the relevant points in the original geometry and their transforms are denoted by capital letters O,A,B $\cdots$. The first and second elliptical transformations are distinguished by indices 1 and 2 attached to the modulus and the complete integral values. The mapping then proceeds as follows.

(*a*) *From the z plane to the y plane.*

$$y = \text{sn } z \text{ modulo } k_1 = \sin \theta_1 \tag{39}$$

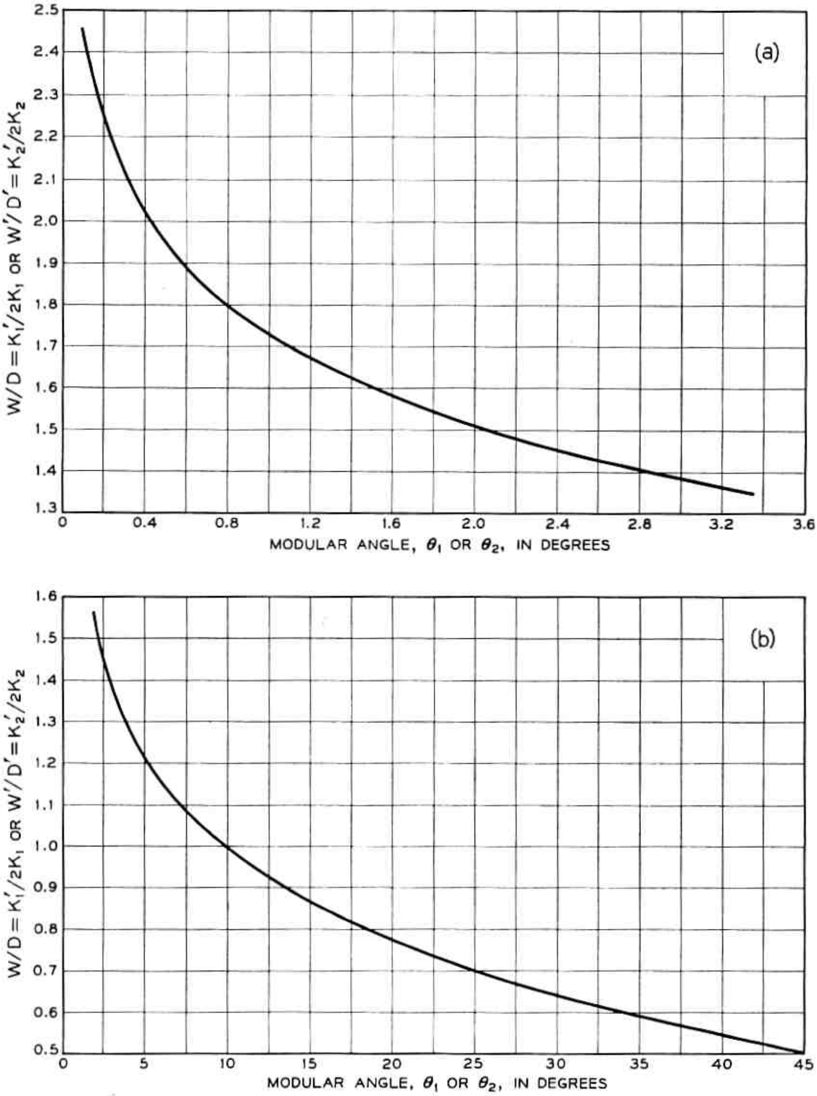


Fig. 19 — Relation between the ratio of the length dimensions of the rectangle to be transformed and the modular angle of the transforming elliptical function.

The modular angle θ_1 is found by entering the curves of Fig. 19 with the aspect ratio $W/D = K_r'/2K_1$ of the original rectangle. The corresponding coordinates in the z and y planes are given in Table I. The transformation of points O through D requires only the graph of Fig. 19. For points A and F, use has to be made of elliptic function tables. In the Smithsonian

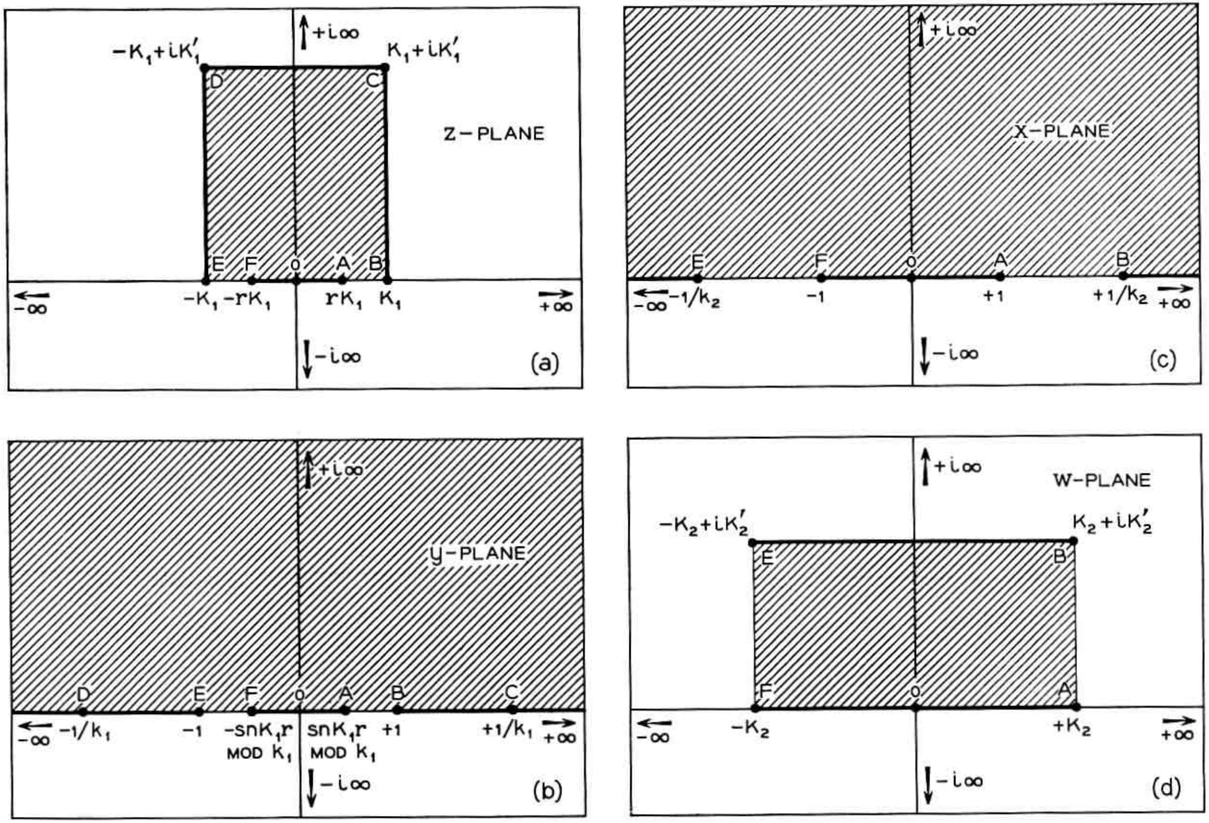


Fig. 20 — (a) Original electrode configuration of the upper cutoff frequency situation in the z plane. (b) Geometry after transformation into the upper half of the y plane. (c) Same geometry scaled in the x plane for subsequent transformation. (d) Electrode configuration transformed into a simple parallel plate geometry in the w plane.

TABLE I — SUMMARY OF TRANSFORMATIONS FOR UPPER CUTOFF FREQUENCY CASE

	z	y	x	w
O	0	0	0	0
B	K_1	1	$1/(\text{sn } K_1r \text{ mod } k_1)$	$K_2 + iK_2'$
E	$-K_1$	-1	$-1/(\text{sn } K_1r \text{ mod } k_1)$	$-K_2 + iK_2'$
C	$K_1 + iK_2'$	$1/k_1$	not of interest	
D	$-K_1 + iK_1'$	$-1/k_1$		
A	K_1r	$\text{sn } K_1r \text{ mod } k_1$	1	K_2
F	$-K_1r$	$-\text{sn } K_1r \text{ mod } k_1$	-1	$-K_2$

$k_1 = \sin \theta_1$
$k_2 = \sin \theta_2$
$\theta_2 = \phi = \sin^{-1} [\text{sn } K_1r \text{ mod } k_1 \text{ or } \theta_1]$

Tables the Legendre notation (34), (35), (36), and (37) is used. Entering these tables with $z = K_1r$ and the angle θ_r, a value of ϕ in radians is found. This value ϕ is converted to degrees and renamed θ_2.

(*b*) *From the y plane to the x plane.*

This is a change of scale and is accomplished by dividing all values by

$$\sin \phi = k_2 = \text{sn } K_1r \text{ mod } \theta_r . \tag{40}$$

After this step the arrangement of the points OBEAF on the real axis is the standard one for transformation of the upper half plane into a rectangle.

(*c*) *From the x plane to the w plane.*

This transformation finally shapes the original electrode geometry into the desired parallel plane geometry. The transformation is indicated in Table I. However, since the interest centers only on the impedance — that is, the length-dimension ratio of this final rectangle — it is not necessary to carry out this transformation in detail. This ratio $W'/D' = K_2'/2K_2$ is obtained from Fig. 19 by entering it with the modular angle $\theta_2 = \phi = \sin^{-1} k_2$.

Following these steps in the case $r = \frac{1}{2}$, the curve of Fig. 10 was obtained.

A short-cut is possible if $\theta_1 < 30°$; that is, if $W/D > 0.65$. In that case the sn function can be approximated by a sine function and $K \approx \pi/2$. Then $\phi = \theta_2 = r\pi/2$; in particular, for $r = \frac{1}{2}$, $\phi = \theta_2 = 45°$ and $W'/D' = \frac{1}{2}$.

A.4 *Mapping of the Lower Cutoff Frequency Configuration*

The procedure is quite similar to that used for the upper cutoff frequency geometry. It is summarized in Table II and Fig. 21.

TABLE II — SUMMARY OF TRANSFORMATIONS FOR LOWER CUTOFF FREQUENCY CASE

	z	y	x	w
O	0	0	0	0
B	K_1	1	not of interest	
E	$-K_1$	-1		
C	$K_1 + iK_1$	$1/k_1$	$1/(k_1 \text{ sn } K_1r \text{ mod } k_1)$	$K_2 + iK_2'$
D	$-K_1 + iK_1'$	$-1/k_1$	$-1/(k_1 \text{ sn } K_1r \text{ mod } k_1)$	$-K_2 + iK_2'$
A	rK_1	$\text{sn } K_1r \text{ mod } k_1$	1	K_2
F	$-rK_1$	$-\text{sn } K_1r \text{ mod } k_1$	-1	$-K_2$

$$\begin{aligned} k_1 &= \sin \theta_1 \\ k_2 &= \sin \theta_2 \\ \theta_2 &= \sin^{-1} [(\text{sn } K_1r \times \sin \theta_1) \text{ mod } k_1 \text{ or } \theta_1] \end{aligned}$$

(*a*) *From the z plane to the y plane.*

This step is identical to the first transformation of the upper cutoff frequency configuration.

(*b*) *From the y plane to the x plane.*

This scaling is also the same as that used before. The difference is, however, that now the points C and D are of interest, whereas before the points considered were B and E.

(*c*) *From the x plane to the w plane.*

Here the transformation differs; now a different modulus

$$k_2 = k_1 \text{ sn } K_1r \text{ mod } k_1$$

is used. The resulting complete integral values K_2 and K_2' are not to be confused with those obtained for the upper cutoff frequency case. Since the interest centers only on the impedance value $K_2'/2K_2 = W'/D'$ of the resulting rectangle, it is not necessary to evaluate this transformation in detail. The numerical evaluation is quite similar to the one of the upper cutoff situation. Using Fig. 19, one finds the first modular angle θ_1 from $K_1'/2K_1 = W/D$ of the original geometry. Entering the tables with $z = K_1r$ and θ_1, an integral value ϕ is found. This value is obtained in radians. Then form

$$\begin{aligned} k_2 = \sin \theta_2 &= \sin \theta_1 \times (\text{sn } K_1r \text{ mod } k_1) \\ &= \sin \theta_1 \times (\sin \phi \text{ mod } \theta_1). \end{aligned} \tag{41}$$

Using this formula, the angle θ_2 is evaluated in degrees. Then the graphs (Fig. 19) can be used again to obtain from θ_2 the length dimension ratio W'/D' of the transformed rectangle.

Following this procedure for the case $r = \frac{1}{2}$, the graph of Fig. 12 was obtained.

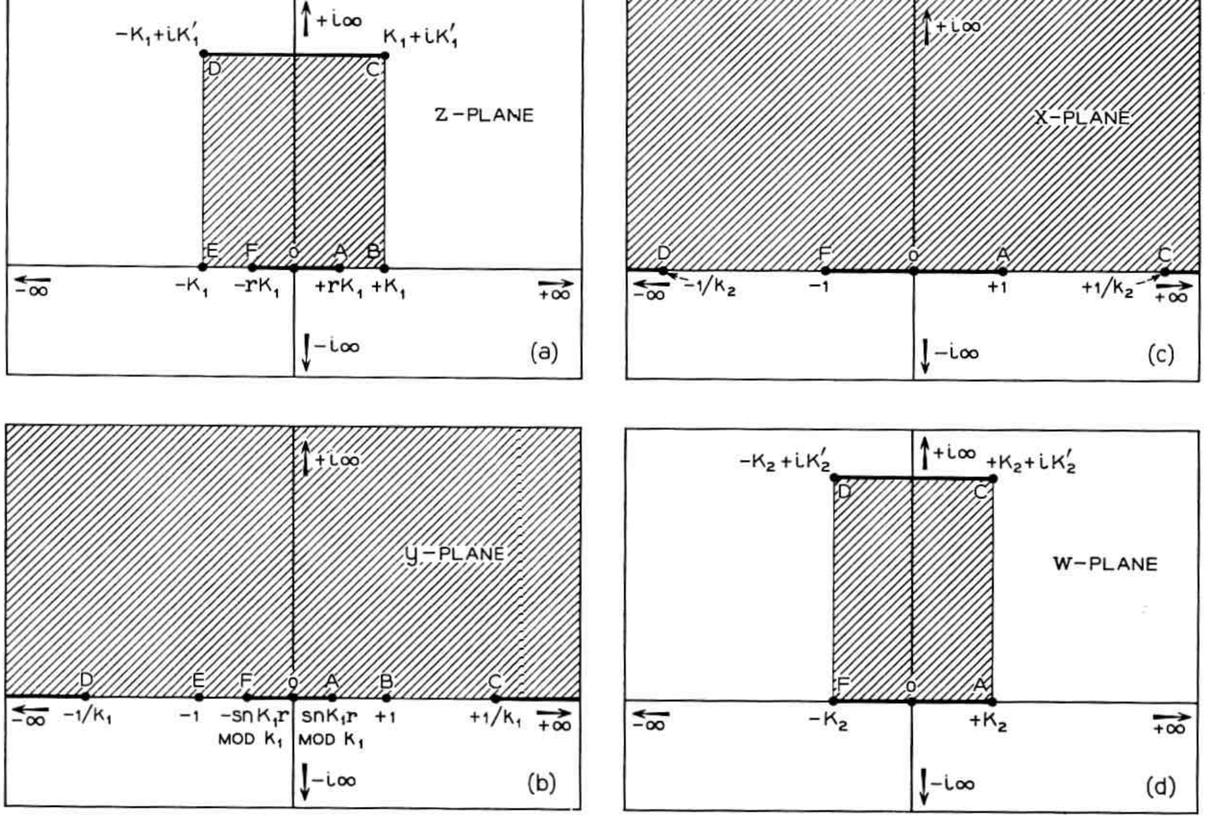


Fig. 21 — (a) Original electrode configuration corresponding to the lower cutoff frequency situation, shown in the z plane. (b) Geometry after transformation into the upper half of the y plane. (c) Same geometry scaled up in the x plane for subsequent transformation. (d) Electrode configuration transformed into a simple parallel plate geometry in the w plane.

REFERENCES

1. DeGrasse, R. W., Schulz-DuBois, E. O., and Scovil, H. E. D., Three-Level Solid-State Traveling Wave Maser, B.S.T.J., **38,** March, 1959, pp. 305–334.
2. Geusic, J. E., to be published.
3. Feher, G., Sensitivity Considerations in Microwave Paramagnetic Resonance Absorption Techniques, B.S.T.J., **36,** March, 1957, pp. 449–484.
4. DeGrasse, R. W., Kostelnick, J. J., and Scovil, H. E. D., Dual Channel 2390-mc Traveling-Wave Maser, B.S.T.J., **40,** July, 1961, pp. 1117–1128.
5. Harris, S. E., DeGrasse, R. W., and Schulz-DuBois, E. O., Solid State Maser Research, U. S. Signal Corps Report under Contract No. DA 36-039 SC-85357, First Quarterly Report, 20 September 1960. Available through ASTIA. Unpublished.
6. Tabor, W. J., and Sibilia, J. T., Masers for the *Telstar* Satellite Communications Experiment, B.S.T.J., **42,** July, 1963, pp. 1863–1886.
7. Hensel, M. L., and Treacy, E. B., to be published.

Permutation Decoding of Systematic Codes

By JESSIE MACWILLIAMS

(Manuscript received September 4, 1963)

A symmetry of a systematic code is a permutation of bit positions in each code word (the same permutation is applied to all code words) which preserves the code as a whole. Permutation decoding makes use of these symmetries to build up a decoding algorithm for the code.

It is difficult to find an appropriate set of symmetries for a code picked at random. For cyclic codes the problem is somewhat easier, and for some special cyclic codes it is solved completely in this paper. For these codes, at least, it is evident that permutation decoding is easy to implement and inexpensive compared with other decoding schemes.

Permutation decoding as a n eans of error control is evaluated for the binary symmetric channel and for the switched telephone network as represented by experimental data. It is found to be extremely effective on the binary symmetric channel and of very doubtful value on the present telephone network.

INTRODUCTION

A systematic code of block length n is a subspace of the vector space of all possible rows of n symbols chosen from a finite field. In this paper such a code will be called an alphabet,[1] and the sequences belonging to the alphabet will be called letters.

The parameters used to describe an alphabet are block length, n, number of information places, k, and error correcting capability, e: n is the number of symbols in each letter, k is the dimension of the alphabet as a vector space, and e is defined by the property that the minimum Hamming distance between two letters is either $2e + 1$ or $2e + 2$.

It is well known[1] that an alphabet with parameters n,k,e is theoretically capable of correcting all occurrences of $\leqq e$ errors in a block of length n. However for $e > 1$, the process of error correction by decoding is complicated, and likely to require expensive equipment. In this paper we describe a new decoding scheme, permutation decoding, which is conceptually simple and quite easy to implement.

The decoding procedure consists of a sequence of permutations of the received block of symbols, each of which is followed by a parity check calculation. We can thus make a rough comparison between the complexity of the equipment required for encoding and decoding. The encoder uses one parity check register, and the decoder uses r (or uses one r times), where r is the number of permutations in the decoding sequence. Real time operation with a constant time delay is possible and perhaps not too expensive.

Permutation decoding owes much to the previous work of Peter Neumann[2] and Eugene Prange.[3] It depends essentially on the symmetries of the alphabet. A symmetry of an alphabet means a permutation of digit positions which preserves the alphabet as a whole. The same permutation is applied to the digits of every letter, and each letter is changed, if at all, into another letter of the same alphabet. Very little is known about symmetries of alphabets in general, but it will be shown that even this little is enough to enable us to apply the decoding scheme to a large class of alphabets.

Permutation decoding differs from previous schemes in two important ways. First, it becomes easier as the redundancy of the alphabet increases; it is most useful for alphabets with high error correcting capabilities. Secondly, it cannot correct more than e errors in n places. A received sequence containing more than e errors either will be "corrected" wrongly or will emerge unchanged from the decoder.

It will become apparent in Section III that permutation decoding produces many more undetected errors than does error control by detection and retransmission. The simpler scheme of detection and retransmission should be used when it is at all feasible.

The plan of this paper is as follows: Section I contains a description of permutation decoding in general, without reference to the particular alphabet we wish to decode. Section II is an example; it contains a detailed account of a particular permutation group which will suffice to decode many binary cyclic alphabets. Section III describes how the probability of improper correction and of detection without correction may be estimated.

I. ERROR CONTROL AND PERMUTATION DECODING

Let V^n be the set of all possible binary sequences of length n.* The distance between two sequences is the number of places in which they

* The method will work equally well for multilevel codes. All that is needed is to find a euphonious substitute for the term "binary sequence."

differ; the distance between v_1 and v_2 is the minimum number of bits we must change in v_1 in order to convert it into v_2.

For purposes of error control, some sequences of V^n are designated as the sequences which will be transmitted. This subset of V^n is called a code. Error detection consists in finding out whether a received sequence belongs to the code. Every method of error correction consists in mapping a received sequence onto the nearest member of the code, where nearness is defined in terms of the distance function defined above. If there are several nearest members, the correction procedure chooses one in some arbitrary fashion or indicates that an uncorrectable error has been found.

The strategy for choosing a code is usually to place its members as far apart as possible in V^n. It will then take a relatively large number of errors to cause a transmitted code sequence to be received as a different code sequence. If the distance between any two code sequences is $\geqq 2e + 1$, it is theoretically possible to correct all single, double, $\cdots$ e-fold errors. This may be restated as follows: If v is a received sequence in which $\leqq e$ errors have occurred, there is a unique code sequence α at distance $\leqq e$ from v. Every other member of the code is at distance $> e$ from v.

The business of decoding is to find α, given v. To do this expeditiously we need some additional structure in the code, and from now on we restrict our choice of codes to the kind described in the next paragraph.

An alphabet* (systematic code, group code) is one in which a fixed number of fixed bit positions are designated as information places, and the other bit positions contain parity checks, which are linear combinations of the contents of the information places. For convenience, the first k bit positions are taken to be the information places. From any k-place binary sequence h we obtain a unique letter of the alphabet by adding $n - k$ parity checks. This letter will be denoted by $m(h)$.

Let $\bar{\alpha}$ stand for the first k coordinates of the n-place sequence α. α is a letter of the alphabet if and only if $\alpha = m(\bar{\alpha})$. Let π be a permutation of bit positions in V^n which preserves the alphabet; if α is a letter, so is $\alpha\pi$. The first k positions of $\alpha\pi$ are information places, and $\alpha\pi = m(\overline{\alpha\pi})$.

Let v be a received sequence containing $\leqq e$ errors. If no errors have occurred in the first k places of v, $\alpha_0 = m(\bar{v})$ is the unique letter of the alphabet at distance $\leqq e$ from v, and is the corrected version of v. On the other hand, if one or more errors have occurred in the first k places of v, the letter $m(\bar{v})$ is not the corrected version of v, since it is the same as v in the first k places. In this case $m(\bar{v})$ is at distance $> e$ from v.

* It has been shown[1] that every systematic code is a group code and that every group code is a systematic code.

The first step in the decoding procedure is to form $\alpha_0 = m(\bar{v})$, find the distance between v and α_0, and take α_0 as the corrected version of v if this distance is $\leqq e$.

Let α denote the unique letter of the alphabet at distance $\leqq e$ from v. Let π be, as before, a permutation of bit positions which preserves the alphabet. Clearly the distance between $\alpha\pi$ and $v\pi$ is the same as that between α and v.

Suppose that we can find a permutation π_i which preserves the alphabet and which moves the errors in v out of the first k positions. Then $\alpha_i = m(\overline{v\pi_i})$ is at distance $\leqq e$ from $v\pi_i$, and is the unique letter of the alphabet with this property. Consequently $\alpha = \alpha_i\pi_i^{-1}$ is the corrected version of v.

This suggests the following decoding procedure: Let I (the identity), $\pi_1, \pi_2, \cdots$ be a sequence of permutations which preserve the alphabet. Form the letters

$$\alpha_0 = m(\overline{vI}), \qquad \alpha_1 = m(\overline{v\pi_1}), \qquad \alpha_2 = m(\overline{v\pi_2}), \cdots$$

and at each step find the distance between α_i and $v\pi_i$. Continue until a letter α_i is found which is at distance $\leqq e$ from $v\pi_i$. Then $\alpha_i\pi_i^{-1}$ is the corrected version of v.

A received vector which is at distance $>e$ from all letters of the alphabet will be detected as an error but not corrected by this procedure. Some provision must be made for this eventuality. This is discussed in Section III.

It is also possible for the decoder to make an incorrect "correction." This will happen if an error pattern (of more than e errors) causes the transmitted letter α to be received as a sequence v which is at distance $\leqq e$ from a different letter α'. The probability of this occurrence is calculated in Section III.

In order for permutation decoding to work, we must be sure that one of the sequences $v\pi_i$ is correct in the first k places. If $v = \alpha + f$, f being the error sequence, the permutation π_i must move all nonzero coordinates of f out of the first k places. In order for the procedure to be practical, it must be possible to move all sets of $\leqq e$ errors out of the first k places with a fairly short sequence of permutations.

To correct all sets of $\leqq e$ errors in a block of length n, we need the following: (i) an alphabet of block length n, dimension k, and minimum distance $\geqq 2e + 1$; and (ii) a set of permutations, $\pi_1, \pi_2, \cdots$ which preserve the alphabet and at the same time move any set of $\leqq e$ errors out of the first k places.

We emphasize that the reason for insisting that the permutations π_i shall preserve the alphabet is to keep the parity check calculation always the same. The encoder (the parity check calculator) is a complicated and expensive piece of equipment; it is desirable to use only one encoder in the decoding scheme. If for any reason (such as real time operation) it is necessary to have more than one encoder, we can, to a certain extent, relax the restriction on the permutations π_i.

It may seem to be quite a trick to find at the same time both a suitable alphabet and a suitable set of permutations; really the chief difficulty is that neither alphabets nor permutation groups have been studied from this point of view. It is shown in Section II that a very simple permutation group will do for many cyclic alphabets.

We conclude this section with an example of permutation decoding applied to the Hamming alphabet with $n = 7$, $k = 4$, $e = 1$.* The alphabet is written out in Table I; it is seen to be invariant under cyclic permutation.

TABLE I — A CYCLIC ALPHABET WITH $n = 7$, $k = 4$, $e = 1$

0	0	0	0	0	0	0
0	1	1	0	1	0	0
0	0	1	1	0	1	0
0	0	0	1	1	0	1
1	0	0	0	1	1	0
0	1	0	0	0	1	1
1	0	1	0	0	0	1
1	1	0	1	0	0	0
1	1	1	1	1	1	1
1	0	0	1	0	1	1
1	1	0	0	1	0	1
1	1	1	0	0	1	0
0	1	1	1	0	0	1
1	0	1	1	1	0	0
0	1	0	1	1	1	0
0	0	1	0	1	1	1

Let T denote the cyclic permutation. Clearly at most four applications of T will move any single error out of the first four places. The decoding sequence consists of the permutations I, T, T^2, T^3, T^4.†

Let the received vector be 1110100 (the first nonzero vector of Table I with an error in the first place). The successive stages of the decoding process are shown in Table II.

* The example is chosen for simplicity. Permutation decoding is *not* the most efficient way of correcting single errors.

† E. R. Berlekamp has pointed out that the shorter decoding sequence I, T^3, T^6 is sufficient.

TABLE II — DECODING PROCEDURE FOR THE ALPHABET OF TABLE I

$V = 1110100$	
$m(\bar{V}) = 1110010$	distance = 2
$VT = 0111010$	
$m(\overline{VT}) = 0111001$	distance = 2
$VT^2 = 0011101$	
$m(\overline{VT^2}) = 0011010$	distance = 3
$VT^3 = 1001110$	
$m(\overline{VT^3}) = 1001011$	distance = 2
$VT^4 = 0100111$	
$m(\overline{VT^4}) = 0100011$	distance = 1

Thus $\alpha = 0100011$ is the unique letter at distance $\leqq 1$ from VT^4, and the corrected version of V is $\alpha T^{-4} = \alpha T^3 = 0110100$.

II. PERMUTATION DECODING OF CYCLIC ALPHABETS*

The coordinate places in V^n are labeled by the numbers $0, 1, 2, \cdots, n - 1$. This notation is convenient for describing permutations. If ω stands for one of these numbers, the cyclic permutation is

$$T: \omega \rightarrow \omega + 1 \text{ (addition mod } n).$$

The powers of the cyclic permutation are

$$T^2: \omega \rightarrow \omega + 2; \qquad T^3: \omega \rightarrow \omega + 3, \cdots, \qquad T^n: \omega \rightarrow \omega + n = \omega.$$

A cyclic alphabet in V^n is an alphabet which is invariant under T, hence also invariant under T^2, T^3, etc. We assume that we wish to decode a cyclic alphabet with parameters n,k,e.

Successive cyclic shifts will eventually bring any k consecutive bits to the first k positions, and hence will move out of the first k positions any error pattern in which there is a gap of length $\geqq k$. In particular, the sequence $I, T, \cdots, T^{n-1}$ will always correct all single errors.†

This sequence will not correct an error pattern in which there is no gap of length $\geqq k$. Suppose, for example, that $n = 23$, $k = 12$. The error pattern shown below cannot be corrected by cyclic shifts alone.

X 1 2 3 4 5 6 7 8 X 10 11 12 13 14 15 16 17 18 X 20 21 22

To deal with such cases we introduce another permutation $U: \omega \rightarrow 2\omega$ (multiplication mod n), and its powers, $U^2: \omega \rightarrow 4\omega$, $U^3: \omega \rightarrow 8\omega$, etc. If n is odd, there exists a least integer t such that $2^t \equiv 1 \bmod n$; and $U^t = I$. The choice of U is motivated by the following theorem.

* It is to be emphasized that this section is only an example. The permutations described here will not suffice to decode all cyclic alphabets of odd block length. A method of finding other permutations is given in Appendix A.

† It will also correct all double errors if $k < n/2$, and so on.

Theorem 1: Every binary cyclic alphabet of odd block length is invariant under U and the powers of U.

The proof of this theorem is given in Appendix A.

The error pattern 0, 9, 19 above is changed by U into 0, 18, 15. This pattern is moved out of the first twelve places by 21 cyclic shifts.

The permutation group on $0, 1, \cdots, n-1$ generated by T and U will be called G_n. It is easy to check that $TU = UT^2$; hence we may represent every permutation in G_n in the form U^iT^j, with $0 \leqq i \leqq t-1$, $0 \leqq j \leqq n-1$. Now every power of U leaves 0 fixed, and no power of T (except the identity) leaves 0 fixed; thus $U^iT^j = U^hT^k$ if and only if $i = h \bmod t$ and $j = k \bmod n$. It follows that the group G_n is of order nt and consists of the permutations:

$$\begin{array}{ccccc} I, & T, & T^2, & \cdots, & T^{n-1} \\ U, & UT, & UT^2, & \cdots, & UT^{n-1} \\ U^2, & U^2T, & U^2T^2, & \cdots, & U^2T^{n-1} \\ \cdot & \cdot & \cdot & \cdots, & \cdot \\ U^{t-1}, & U^{t-1}T, & U^{t-1}T^2, & \cdots, & U^{t-1}T^{n-1}. \end{array}$$

Let $0 \leqq u_1 < u_2 < \cdots < u_s \leqq n-1$ (n odd), be a set of integers; we suppose that errors occur in places $u_1, \cdots, u_s$. Let $g(u_1, \cdots, u_s, n)$ be the length of the maximum gap which can be inserted in this sequence by repeated multiplication by 2 mod n.

If $u_{\nu k}$ denotes that integer less than n which is congruent to $2^k u_\nu \bmod n$, then

$$g(u_1, \cdots, u_s, n) + 1 = \operatorname*{Max}_{i,j,k} | u_{ik} - u_{jk} |,$$

under the condition that the interval $[u_{ik}, u_{jk}]$ contain no other u_{vj}. Let $g(s,n)$ be the minimum value of this maximum gap for all possible choices of the s values $u_1, \cdots, u_s$.

$$g(s,n) = \operatorname*{Min}_{u_1, \cdots, u_s} g(u_1, \cdots, u_s, n).$$

The group G_n then contains a permutation which moves any set of s errors out of the first $g(s,n)$ places. Clearly $s' < s$ implies $g(s',n) \geqq g(s,n)$. Hence a binary cyclic n,k,e alphabet with n odd may be decoded by G_n if and only if $k \leqq g(e,n)$. The quantity $g(e,n)$ has a few obvious properties.

The numbers $0, 1, \cdots, n-1$ can be partitioned into subsets which are invariant under U.* For example, for $n = 15$, these subsets are

* The number of cyclic alphabets of block length n is determined by the number of these subsets; the dimensions of the cyclic alphabets are determined by the sizes of the invariant subsets; see, for example, Ref. 4.

$$(0),\ (1,2,4,8),\ (3,6,12,9),\ (5,10),\ (7,14,13,11).$$

The union of any number of invariant subsets is also invariant under U; from these subsets we may obtain upper bounds on $g(e,n)$ by inspection. In the example above we obtain:

$g(2,15) \leqq 9$, since the invariant set $(5,10)$ gives us a maximum gap 11,12,13,14,0,1,2,3,4 of length 9.

$g(3,15) \leqq 4$, since the invariant set $(0,5,10)$ gives us a maximum gap 1,2,3,4 of length 4.

$g(5,15) \leqq 2$, since the invariant set $(0,3,6,9,12)$ gives us a maximum gap of length 2.

These upper bounds limit the usefulness of the group G_n. However it is still sufficiently useful to be of interest.

The value of $g(e,n)$ for various choices of e,n have been computed on the IBM 7090. These are tabulated in Table III together with the parameters n,k,e for several cyclic alphabets.

Table III — Effect of Permutation U for Different Block Lengths

	Code			Gap Length		
n	k	e	$g(2)$	$g(3)$	$g(4)$	$g(5)$
47	24	5				26
31	21	2	25			
31	16	3		19		
31	11	4			12	
23	12	3		17		
21	12	2	13			
21	9	3		6		
21	5	4			6	
17	9	2	13			
15	7	2	9			
15	5	3		4		

The gap length $g(s)$ is the maximum number of consecutive error-free positions which can be inserted into an arbitrary pattern of s errors in n places by successive applications of the permutation $w \rightarrow 2w \bmod n$. If $k \leqq g(e)$, the group generated by T and U contains a sequence of permutations which will suffice to decode an n, k, e alphabet.

It is desirable, if possible, to use only part of G_n in the decoding sequence. As an example we consider the alphabet with $n = 23$, $k = 12$, $e = 3$.* In this case one of the permutations, U, U^2, U^{11} will always create a gap of length at least 12 in any set of 3 errors. The decoding sequence is:†

* This alphabet is described in detail in Table V.

† It has been shown by E. R. Berlekamp that a subsequence of length 40 is all that is really necessary.

$$\begin{array}{ccccc}
I, & T, & T^2, & \cdots, & T^{22} \\
U, & UT, & UT^2, & \cdots, & UT^{22} \\
U^2, & U^2T, & U^2T^2, & \cdots, & U^2T^{22} \\
U^{11}, & U^{11}T, & U^{11}T^2, & \cdots, & U^{11}T^{22}.
\end{array}$$

The decoding procedure for this particular alphabet has been simulated on the IBM 7090. A diagram of the logical program is given in Fig. 1, and Table IV traces a particular sequence through the decoder. In practice it is convenient to add one more permutation (in this case TU) to the decoding permutations, so that a sequence passing through the entire

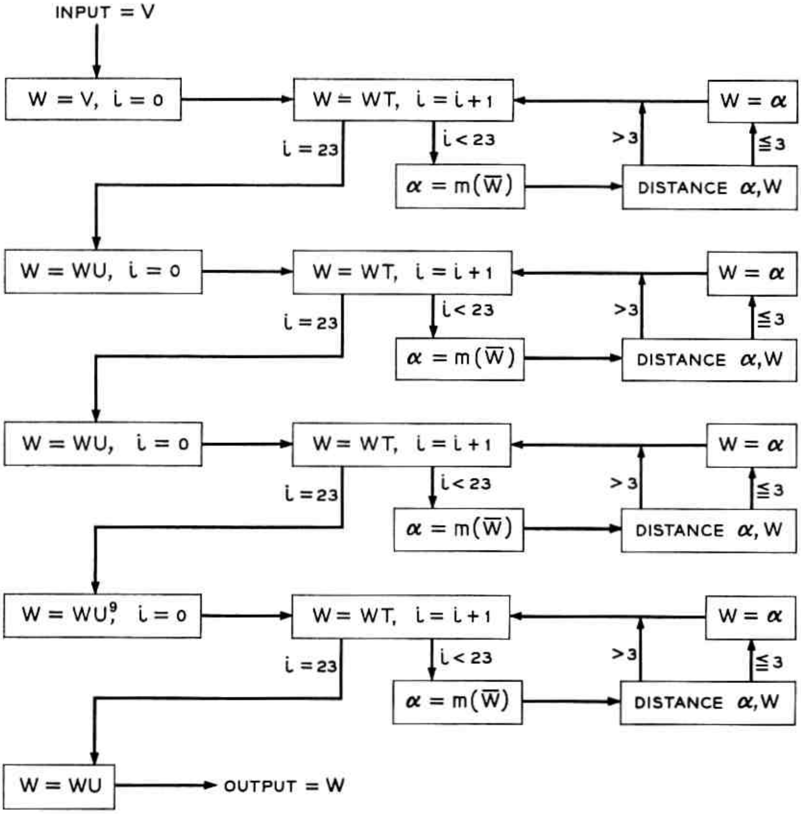


Fig. 1 — Logic of decoder for (23,12,3) alphabet. The contents of the boxes are Fortran-type instructions; for example, $i = i + 1$ is to be interpreted as "replace i by $i + 1$."

TABLE IV — DECODING PROCEDURE FOR THE (23,12,3) ALPHABET

													Parity Check												Distance
v	3	5	7	11	13	14	15	18	19	20	21	22	3	5	7	11	13	16	18	21					6
vT	0	4	6	8	12	14	15	16	19	20	21	22	0	4	6	8	16	18	19						7
vT^{22}	2	4	6	10	12	13	14	17	18	19	20	21	2	4	6	10	12	15	17	20					6
vU	3	5	6	7	10	13	14	15	17	19	21	22	3	5	6	7	10	12	15	16	17	19	20		7
vUT	0	4	6	7	8	11	14	15	16	18	20	22	0	4	6	7	8	11	12	13	14	20	21	22	6
vUT^{22}	2	4	5	6	9	12	13	14	16	18	20	21	2	4	5	6	9	13	14	17	18	19	21	22	6
vU^2	3	5	6	7	10	11	12	14	15	19	20	21	3	5	6	7	10	11	12	13	19	20	21	22	4
vU^2T	4	6	7	8	11	12	13	15	16	20	21	22	4	6	7	9	11	13	16	17	18	20	21		5
vU^2T^{12}	0	1	3	4	8	9	10	15	17	18	19	22	0	1	3	4	8	9	10	12	15	18	19		3
αU^2T^{12}	0	1	3	4	8	9	10	12	15	18	19														
αU^2T^{22}	2	5	6	9	12	14	16	18	20	21	22														
αU^{10}	0	9	10	11	13	14	17	18	19	21	22														
$\alpha U^{10}T$	0	1	10	11	12	14	15	18	19	20	21														
$\alpha U^{10}T^{22}$	8	9	10	12	13	16	17	18	20	21	22														
α	0	3	5	11	13	15	18	19	20	21	22														

The numbers indicate the positions of the ones in the 23-bit sequence. For example, the first sequence v is 00010101000101110011111.

set emerges in its original form. The final output of the permuting register is then the corrected form of the received sequence.

The operation of the 7090 program is of course sequential—it employs one subroutine to simulate the parity check calculation. It is clear from the logic diagram that it is quite convenient to split the encoder into four parallel sections, each of which contains a register capable of making a cyclic permutation and an encoder to calculate parity checks. This idea can be applied to speed up the decoding of any cyclic alphabet.

III. EVALUATION OF PERMUTATION DECODING AS A MEANS OF ERROR CONTROL

Permutation decoding of an n,k,e alphabet $\mathcal{A}$ will map a received sequence v onto the nearest letter of $\mathcal{A}$ provided that this letter is unique. This is the case if v lies at distance $\leqq e$ from some letter of $\mathcal{A}$. If v is at distance $>e$ from every letter of $\mathcal{A}$ the decoder will detect an error but will be unable to correct it.*

The decoder will also make mistakes. If f is an error sequence of more than e errors, and α the transmitted letter, the received sequence $\alpha + f$ may lie at distance $\leqq e$ from some other letter α' of $\mathcal{A}$. The decoder will then interpret $\alpha + f$ as α'. The error sequences which cause such incorrect decoding are characterized by the following theorem.

Theorem 2: The error sequences which cause the decoder to "correct" incorrectly are exactly those sequences of weight $>e$ which lie at distance $\leqq e$ from some letter of $\mathcal{A}$.

Proof: Let f be a sequence of weight $>e$ such that $f = \beta + f'$, $\beta \in \mathcal{A}$, f' of weight $\leqq e$.

For any transmitted letter α

$$\alpha + f = \alpha + \beta + f',$$

and the decoder will interpret $\alpha + f$ as $\alpha + \beta$.

Conversely, suppose that f is an error sequence such that $\alpha + f$ is decoded as $\alpha' \neq \alpha$. Then

$$\alpha + f = \alpha' + f' \qquad \alpha' \in \mathcal{A}, f' \text{ of weight } \leqq e,$$

and

$$f = \alpha' - \alpha + f'.$$

Hence f is at distance $\leqq e$ from the letter $\alpha' - \alpha$ of $\mathcal{A}$.

* One is tempted to suggest that the decoder ask for retransmission of such detected errors. This idea is of dubious value; error correction by decoding should not be used at all if error correction by detection and retransmission is a possible alternative. We must assume that retransmission is extremely awkward, if not completely infeasible.

Let $A(s), s = 0, 1, \cdots, n$ be the number of letters of $\mathcal{A}$ of weight s. Let $C(s)$, $s = 0, 1, \cdots, n$ be the number of sequences of V^n of weight s which lie at distance $\leqq e$ from letters of $\mathcal{A}$. The $C(s)$ are uniquely determined by the $A(s)$, and may be obtained from them by a simple calculation; the exact formula is given in Appendix B. The values of $A(s)$ for a number of binary cyclic alphabets are tabulated in Table V, and the values of $C(s)$ for these alphabets are given in Table VI.

For $s \leqq e$, $C(s) = \binom{n}{s}$ and is the number of sequences of weight s which are properly corrected by the decoder. For $s > e$, $C(s)$ is the number of sequences of weight s which are improperly corrected by the decoder.

Let $D(s)$, $s = e + 1, \cdots, n$ be the number of error sequences of weight s which cause the decoder to detect an error that it cannot correct. Clearly $D(s) = 0$ for $s \leqq e$, and $D(s) = \binom{n}{s} - C(s)$ for $s > e$.

P_E denotes the probability that a received sequence will be "corrected" incorrectly by the decoder; P_D denotes the probability that a received sequence will be detected as an error but not corrected. We consider first a binary symmetric memoryless channel with bit error probability p. The probability of s specific errors in a block of length n is then $p^s(1 - p)^{n-s}$, and this probability is independent of the location of the errors. Hence

$$P_E(\text{B.S.}) = \sum_{s=e+1}^{n} C(s)p^s(1 - p)^{n-s},$$

$$P_D(\text{B.S.}) = \sum_{s=e+1}^{n} D(s)p^s(1 - p)^{n-s}.$$

It is to be noted that if error correction by detection and retransmission is used, the probability of an undetected error is

$$\sum_{s=d}^{n} A(s)p^s(1 - p)^{n-s}; \qquad d = 2e + 1 \quad \text{or} \quad 2e + 2.$$

This sum starts with the first nonzero value of $A(s)$ (for $s > 0$), i.e. with $s = 2e + 1$ or $s = 2e + 2$. It is obvious that for the values of p,n currently in use in the Bell System, error correction by detection and retransmission is the preferable scheme.

The values of P_E(B.S.) and P_D(B.S.) for a number of alphabets are tabulated in Table VII; p is taken to be 3.22×10^{-5}, the over-all bit error rate on the telephone network obtained from the Alexander, Gryb and Nast task force data.[5]

Except for the first and last example, the alphabets in Tables V, VI and VII occur in pairs. The second alphabet in each pair consists of the letters of even weight in the first. Since the minimum distance of the second alphabet is $2e + 2$, its value of $C(e + 1)$ is zero. In other words, every error of weight $e + 1$ will be detected. If this is important, it is advantageous to use the second alphabet.

The error rates of Table VI are fantastically low; unfortunately the fantasy resides in the binary symmetric channel. The situation is very different on the real telephone channel.

Let $P(s,n)$ be the probability of s errors in n consecutive bits. [For the binary symmetric channel $P(s,n) = \binom{n}{s} p^s(1 - p)^{n-s}$.] Tables of $P(s,n)$ for the telephone network have been calculated from the Alexander, Gryb and Nast task force data.

The decoder will either detect or correct wrongly every error of weight $>e$. Hence

$$P_E + P_D = \sum_{s=e+1}^{n} P(s,n).$$

It is impossible to obtain exact formulae for P_E and P_D separately. Using the methods of Ref. 5 we obtain the approximate formulae

$$P_E(\text{T.N.}) = \sum_{s=e+1}^{n} C(s) \frac{P(s,n)}{\binom{n}{s}}$$

$$P_D(\text{T.N.}) = \sum_{s=e+1}^{n} D(s) \frac{P(s,n)}{\binom{n}{s}}.$$

These numbers have been computed and are tabulated in Table VIII. This shows rather clearly that on the channel described by the Alexander, Gryb and Nast data the word error rate with error correction is greater than the bit error rate with no encoding.

Of course the average error rates of Table VIII conceal something. Examination of the $P(s,n)$ tables for the individual calls shows that about half of the calls in which errors occurred would be handled successfully by permutation decoding.

Table V — Values of $A(s)$

n,k,e	47,24,5		31,21,2	31,20,2		31,16,3	31,15,3
s	$A(s)/47$	s	$A(s)/31$	$A(s)/31$	s	$A(s)/31$	$A(s)/31$
0	1/47	0	1/31	1/31	0	1/31	1/31
11	92	5	7	0	7	5	0
12	276	6	27	27	8	15	15
15	3795	7	75	0	11	168	0
16	7590	8	245	245	12	280	280
19	35420	9	655	0	15	589	0
20	49588	10	1387	1387	16	589	589
23	81720	11	2640	0	19	280	0
24	81720	12	4480	4480	20	168	168
27	49588	13	6510	0	23	15	0
28	35420	14	8310	8310	24	5	5
31	7590	15	9489	0	31	1/31	0
32	3795	16	9489	9489			
35	276	17	8310	0			
36	92	18	6510	6510			
47	1/47	19	4480	0			
		20	2640	2640			
		21	1387	0			
		22	655	655			
		23	245	0			
		24	75	75			
		25	27	0			
		26	7	7			
		31	1/31	0			

n,k,e =	23,12,3	23,11,3		(21,12,2)	(21,11,2)	(17,9,2)	(17,8,2)			(15,7,2)
s	*A*(*s*)/23	*A*(*s*)/23	*s*	*A*(*s*)	*A*(*s*)	*s*	*A*(*s*)	*A*(*s*)	*s*	*A*(*s*)
0	1/23	1/23	0	1	1	0	1	1	0	1
7	11	0	5	21	0	5	34	0	5	18
8	22	22	6	168	168	6	68	68	6	30
11	56	0	7	360	0	7	68	0	7	150
12	56	56	8	210	210	8	85	85	8	150
15	22	0	9	280	0	9	85	0	9	30
16	11	11	10	1008	1008	10	68	68	10	18
23	1/23	0	11	1008	0	11	68	0	15	1
			12	280	280	12	34	34		
			13	210	0	17	1	0		
			14	360	360					
			15	168	0					
			16	21	21					
			21	1	0					

Table VI — Values of $C(s)$

$n,k,e =$	47,24,5*	31,21,2*	31,20,2	31,16,3	31,15,3
$s = 1$	$C(s) =$ 47	31	31	31	31
2	1081	465	465	465	465
3	16215	2170	0	4495	4495
4	178365	13640	12555	5425	0
5	1533939	82274	5022	29295	26040
6	1997688	360964	339047	92225	13020
7	11700743	1276115	81685	329375	303180
8	58503719	3829585	3591040	1248525	86025
9	253516120	9788250	604655	3190675	2861455
10	1094459500	21506932	20159981	6790333	690525
11	3681363800	41087771	2569497	12963363	11812395
12	10764415000	68535730	64275400	21284445	1987720
13	28981118000	100106900	6245260	31108034	28201320
14	70307802000	128661310	120616350	40561485	3675360
15	154677160000	145890120	9085914	45969682	41569263
16	310193350000	145890120	136804210	45969682	4400419
17	565646700000	128661310	8044965	40561485	36886124
18	942103590000	100106900	93861645	31108034	2906715
19	1437947500000	68535730	4260330	21284445	19296724
20	2012287600000	41087771	38518275	12963363	1150968
21	2587226900000	21506932	1346950	6790333	6099808
22	3059047600000	9788250	9183595	3190675	329220
23	3325051800000	3829585	238545	1248525	1162500
24		1276115	1194430	329375	26195
25		360964	21917	92225	79205
26		82274	77252	29295	33255
27	+ symmetric	13640	1085	5425	5425
28	terms	2170	2170	4495	0
29		465	0	465	0
30		31	0	31	0
31		1	0	1	0

$n,k,e =$	23,12,3	23,11,3	21,12,2	21,11,2	17,9,2	17,8,2	15,7,2
$s = 1$	$C(s) =$ 23	23	21	21	17	17	15
2	253	253	210	210	136	136	105
3	1771	1771	210	0	340	0	180
4	8855	0	2625	2520	1190	1020	540
5	33649	28336	10269	1008	3910	408	1413
6	100947	14168	24024	21168	7820	6936	2355
7	245157	216568	52440	4200	11560	1428	3135
8	490314	61226	92610	85050	14450	13005	3135
9	817190	715990	131530	12810	14450	1445	2355
10	1144066	138138	161196	146748	11560	10132	1413
11	1352078	1180774	161196	14448	7820	884	540
12	1352078	171304	131530	118720	3910	3502	180
13	1144066	1005928	92610	7560	1190	170	105
14	817190	101200	52440	48240	340	340	15
15	490314	429088	24024	2856	136	0	1
16	245157	28589	10269	9261	17	0	
17	100947	86779	2625	105	1	0	
18	33649	5313	210	210			
19	8855	8855	210	0			
20	1771	0	21	0			
21	253	0	1				
22	23	0					
23	1	0					

* This table is correct to eight significant figures.

Table VII — Error Rates for the Binary Symmetric Channel

n,k,e	P_E	P_D	$P_E + P_D$
47,24,5	2.23×10^{-21}	4.56×10^{-20}	4.78×10^{-20}
31,21,2	7.23×10^{-11}	0.77×10^{-10}	1.50×10^{-10}
31,20,2	1.342×10^{-12}	1.50×10^{-10}	1.50×10^{-10}
31,16,3	5.83×10^{-15}	2.80×10^{-14}	3.38×10^{-14}
31,15,3	9.01×10^{-19}	3.38×10^{-14}	3.38×10^{-14}
23,12,3	9.59×10^{-15}	0	9.59×10^{-15}*
23,11,3	9.81×10^{-19}	9.59×10^{-15}	9.59×10^{-15}
21,12,2	7.00×10^{-12}	3.72×10^{-11}	4.42×10^{-11}
21,11,2	2.70×10^{-15}	4.42×10^{-11}	4.42×10^{-11}
17,9,2	4.54×10^{-12}	1.80×10^{-11}	2.27×10^{-11}
17,8,2	1.11×10^{-13}	2.27×10^{-11}	2.27×10^{-11}
15,7,2	6.092×10^{-12}	1.078×10^{-10}	1.078×10^{-10}

* This is a close-packed alphabet; every sequence of 23 binary bits is at distance $\leqq 3$ from some letter of the alphabet.

CONCLUSION

Permutation decoding is a simple and feasible scheme for error correction without retransmission. It is particularly suitable for use with a highly redundant alphabet. Like any such scheme it produces many more undetected errors than error correction by detection and retransmission,

Table VIII — Error Rates for the Telephone Network*

n,k,e	P_E	P_D	$P_E + P_D$
47,24,5	4.65×10^{-6}	3.51×10^{-5}	4.08×10^{-5}
31,21,2	3.17×10^{-5}	3.73×10^{-5}	6.9×10^{-5}
31,20,2	1.23×10^{-5}	5.67×10^{-5}	6.9×10^{-5}
31,16,3	7.34×10^{-6}	3.99×10^{-5}	4.72×10^{-5}
31,15,3	3.06×10^{-6}	4.41×10^{-5}	4.72×10^{-5}
23,12,3	3.69×10^{-5}	0	3.69×10^{-5}
23,11,3	1.52×10^{-5}	2.17×10^{-5}	3.69×10^{-5}
21,12,2	1.82×10^{-5}	3.13×10^{-5}	5.05×10^{-5}
21,11,2	8.72×10^{-6}	4.18×10^{-5}	5.05×10^{-5}
17,9,2	2.36×10^{-5}	1.89×10^{-5}	4.25×10^{-5}
17,8,2	0.98×10^{-5}	3.27×10^{-5}	4.25×10^{-5}
15,7,2	1.83×10^{-5}	2.0×10^{-5}	3.83×10^{-5}

* P_E and P_D are approximate values. $P_E + P_D$ is exact.

but it is quite adequate for a channel in which $P(s,n)$ decreases rapidly as s increases. It is of very doubtful value on the telephone network as described by the Alexander, Gryb and Nast data.

ACKNOWLEDGMENTS

The author would like to express her gratitude to Miss Arline Mills, who wrote most of the computer programs and performed most of the calculations used in this paper.

APPENDIX A

Idempotents and Automorphisms of Cyclic Codes

We give first a short summary of the properties of cyclic alphabets. Let V be a finite field of characteristic q. Let V^n denote the direct sum of n copies of V.

Denote by $V[y]$ the ring of polynomials in y over the field V. Let $V[x] = V[y]/(y^n - 1)$ be the residue class ring of $V[y]$ mod $y^n - 1$. $V[x]$ consists of all polynomials of degree $\leqq n - 1$ with coefficients in V. Addition of polynomials is done as usual; to multiply two polynomials we multiply in the usual way and then reduce exponents mod n.

A subset $\mathcal{A}$ of polynomials of $V[x]$ is called an ideal if

$$\begin{array}{ll} (i) & g_1, g_2 \in \mathcal{A} \Rightarrow \alpha_1 g_1 + \alpha_2 g_2 \in \mathcal{A}, \qquad \alpha_1, \alpha_2 \in V \\ (ii) & g \in \mathcal{A} \Rightarrow xg \in \mathcal{A}. \end{array}$$

A polynomial is completely determined by its coefficients; it is possible, in fact, to identify $V[x]$ and V^n. However, it is convenient to regard them as separate entities, related by the (1-1) mapping

$$\alpha_0 + \alpha_1 x + \cdots + \alpha_{n-1} x^{n-1} \rightleftarrows \alpha_0, \alpha_1, \cdots, \alpha_{n-1}.$$

An ideal in $V[x]$ is, by property (i), a linear subspace of V^n. By property (ii) it is invariant under a cyclic permutation of coordinates; hence it is a cyclic alphabet in V^n. Conversely, a cyclic alphabet in V^n is an ideal in $V[x]$. We represent the ideal and the alphabet by the same letter, $\mathcal{A}$.

The ring $V[x]$ may be regarded as the group algebra of the cyclic group $1, x, \cdots, x^{n-1}$ over V. The group algebra is semi-simple provided that q does not divide n (Ref. 6, Section 10.8). In this case it is known that every ideal contains a polynomial e with the following properties

(*i*) e is idempotent. ($e^2 = e$)
(*ii*) e is a unit for $\mathcal{A}$. ($a \in \mathcal{A} \Rightarrow ae = a$)
(*iii*) e generates $\mathcal{A}$. ($\mathcal{A}$ consists of all polynomials $fe, f \in V[x]$).

e will be called the generating idempotent of $\mathcal{A}$.

An automorphism σ of $V[x]$ is a (1-1) mapping of $V[x]$ onto itself which respects both addition and multiplication. If $v_1, v_2 \in V[x]$, then

$$(v_1 + v_2)\sigma = v_1\sigma + v_2\sigma$$

$$(v_1v_2)\sigma = (v_1\sigma)(v_2\sigma).$$

Lemma 1.1: An automorphism σ of $V[x]$ preserves an ideal $\mathcal{A}$ if and only if σ preserves the generating idempotent e of $\mathcal{A}$.

Proof: Suppose that σ preserves e. Then $\mathcal{A}\sigma = V[x] \cdot e\sigma = V[x] \cdot e = \mathcal{A}$.

Suppose that σ preserves $\mathcal{A}$. Then $e\sigma \in \mathcal{A}$, and $e\sigma e = e\sigma$ by property (*ii*).

Let b be the element of $\mathcal{A}$ such that $b\sigma = e$. Then

$$e\sigma e = e\sigma b\sigma = (eb)\sigma = b\sigma = e.$$

Hence $e = e\sigma e = e\sigma$, and σ preserves e.

Lemma 1.2: If q (the characteristic of V) is relatively prime to n (the block length of $\mathcal{A}$), then the mapping $\sigma: x^i \rightarrow x^{iq}$ is an automorphism of $V[x]$.

Proof: Clearly this mapping respects addition and multiplication in $V[x]$. We have only to show that it is 1-1.

If $x^{iq} = x^{jq}$, then $(i - j)q \equiv 0 \bmod n$. Since q is prime to n, this implies that $i - j \equiv 0 \bmod n$ or $x^i = x^j$.

This proves the lemma.

Theorem 1.3: If q is prime to n, every ideal in $V[x]$ is preserved by the mapping $\sigma: x^i \rightarrow x^{iq}$.

Proof: Let $\mathcal{A}$ be an ideal in $V[x]$ and $e = \sum_{i=0}^{n} \alpha_i x^i$, the generating idempotent of $\mathcal{A}$.

$$\sigma e = \sum_{i=0}^{n} \alpha_i x^{iq} = \left(\sum_{i=0}^{n} \alpha_i x^i\right)^q = e^q,$$

since q is the characteristic of V.

$e = e^2 = e^3 = \cdots = e^q$; hence σ preserves e, and by Lemmas 1.1 and 1.2, σ preserves $\mathcal{A}$.

Let the coordinate places of V^n be labeled $0, 1, \cdots, n - 1$; the mapping $\sigma: x^i \rightarrow x^{iq}$ in $V[x]$ corresponds to the permutation $U_q: \omega \rightarrow q\omega$ of coordinate places in V^n.

Corollary: Every binary cyclic alphabet of odd block length is preserved by the permutation $U: \omega \rightarrow 2\omega$.

This is Theorem 1 of Section II. The proof given here contains more machinery than is necessary to prove Theorem 1. This is done on purpose, in order to get Lemma 1.1, which suggests a method of finding other automorphisms of $V[x]$ which preserve a particular cyclic alphabet.

APPENDIX B

Distribution of Weights in the Cosets of a Group Code

Let $\mathcal{A}$ be an (n,k,e) binary alphabet, and let $\mathcal{B}$ be the orthogonal complement (dual alphabet) of $\mathcal{A}$ in V^n. Let $A(i)$, $B(i)$ denote the number of letters of weight i in $\mathcal{A}$, $\mathcal{B}$ respectively. The quantities $A(i)$, $B(i)$ are connected by the generating function.[7]

$$\sum_{i=0}^{n} A(i)(1+z)^{n-i}(1-z)^i = 2^k \sum_{i=0}^{n} B(i)z^i.$$

Since the $A(i)$ are known, the $B(i)$ may be calculated from this relationship.

Set

$$(1+z)^{n-i}(1-z)^i = \sum_{j=0}^{n} \psi(i,j)z^j.$$

Let $C(s,j)$ denote the number of sequences of weight s in V^n which are at distance j from some letter of $\mathcal{A}$. Then if $j \leqq e$, $C(s,j)$ and $B(i)$ are related by the generating function.[7]

$$\sum_{i=0}^{n} B(i)\psi(i,j)(1+x)^{n-i}(1-x)^i = 2^{n-k} \sum_{s=0}^{n} C(s,j)x^s.$$

Since $B(i)$ and $\psi(i,j)$ are known, $C(s,j)$ may be calculated from this relation.

Clearly

$$C(s) = \sum_{j=0}^{e} C(s,j).$$

REFERENCES

1. Slepian, D., A Class of Binary Signaling Alphabets, B.S.T.J., **35**, January, 1956, p. 634.

2. Neumann, P. G., A Note on Cyclic Permutation Error-Correcting Codes, Information and Control, **5**, March, 1962, p. 72.
3. Prange, E., The Use of Information Sets in Decoding Cyclic Codes, IRE Trans., IT-**8**, September, 1962, p. S-5–9.
4. Prange, E., An Algorithm for Factoring $x^n - 1$ over a Finite Field, Air Force Cambridge Research Center, AFCRC-TN-59-775.
5. Elliott, E. O., Estimates of Error Rates for Codes on Burst-Noise Channels, B.S.T.J., **42**, September, 1963, p. 1977.
6. Curtis, C. W., and Reiner, I., *Representation Theory of Finite Groups and Associated Algebras*, Interscience Publishers, New York, 1962.
7. MacWilliams, J., A Theorem on the Distribution of Weights in a Systematic Code, B.S.T.J., **42**, January, 1963, p. 79.

Optical Maser Oscillators and Noise

By EUGENE I. GORDON

(Manuscript received September 5, 1963)

The transmission line matrix formalism so useful for describing the transfer properties of microwave networks is extended to the electromagnetic fields associated with optical masers. The spontaneous emission noise of the optical maser is examined and shown to be amenable to a thermal description. Taking the point of view, well accepted at microwave frequencies, that a weakly nonlinear oscillator is a saturated amplifier of noise, the power and linewidth of the noise radiation emitted by the optical maser is calculated using the transmission line formalism. The significant parameters for any optical maser are shown to be the frequency, the single-pass gain of the maser medium, the effective mirror reflectivity and the population ratio. The pre-oscillation characteristics of the maser are examined and the reason for the extremely sharp oscillation threshold of the gas masers is discussed. Some observations concerning semiconductor optical masers are also made.

I. INTRODUCTION

This paper represents an attempt to describe the optical maser or laser from a microwave circuit point of view and is largely tutorial, since many of the results obtained from a circuit viewpoint are already known. The generality of the method of approach enlarges their area of validity, however.

Many of the people working on optical masers who do not have a background in microwave theory and techniques may find a fresh point of view. In particular, they may find a very modest introduction to an extremely well developed store of computational techniques which are applicable to optical masers. This may save them the trouble of inventing their own.

On the other hand, those who have previously been working in the field of microwaves may find that the analogies between optical masers and more conventional microwave devices are more cogent than they had appreciated. Finally, it is hoped that some of the distinctions between oscillating and pre-oscillating or subthreshold masers will be clarified.

II. THE CONVENTIONAL OSCILLATOR

Excluding strongly nonlinear oscillators with periodic but nonsinusoidal waveforms, it is often stated that an oscillator is a device having an internal gain which exceeds its total losses. Supposedly, noise triggers it off and it then continues to put out oscillatory power at a level determined only by saturation effects. The steady-state saturation level is defined as that for which the internal gain just equals the loss. An extensive discussion of microwave oscillators, based on this point of view, is given by Slater.[1]

Although this point of view often constitutes a good working definition of a feedback oscillator, it is incomplete in that it neglects the continuing presence of the noise. As a result, when the internal gain of the oscillator exactly equals the losses, so that the effective lifetime of a photon in the feedback loop is infinite, the noise power output of the oscillator must increase without limit. Similarly, the associated linewidth of the noise output must be zero. Since this situation is physically unrealizable, it is clear that the noise must be taken into account and that the steady-state gain could never exactly equal the loss and must always saturate at a slightly lower value. As a result, there could be no continuing, self-sustained oscillation which starts from noise.

From these considerations, it appears that a better, but still incomplete, description of an oscillator would be to say that a steady-state regenerative oscillator is a feedback amplifier driven to saturation by a noise input. Internally produced noise is usually the driving force; however, an additional source may be the noise entering through the output port. The external gain of the amplifier, that is, the gain experienced by any input signal, is usually extremely large unless the amplifier is saturated by the input signal. Since the amplification is obtained regeneratively, that is, by the use of feedback, the bandwidth of the gain is limited; the higher the gain the more limited it is. The amplified, narrow-band noise output is the output signal of the oscillator. When the large gain can be obtained without regeneration, the noise output need not be narrow-band.

This concept of the oscillator as a saturated amplifier of noise is not new and is well known in the microwave art. More recently, this concept has been employed by Gordon, Zeiger and Townes[2] in their treatment of the microwave maser oscillator, and by Wagner and Birnbaum,[3] Schawlow and Townes,[4] Shimoda,[5] Blaquier[6] and Fleck[7] in their treatments of the optical maser oscillator.

While this definition of an oscillator is somewhat more satisfying, it implies that an oscillator is merely an extremely narrow-band filter with gain. As a result, the statistical properties of the noise input should be preserved, except for the spectral narrowing. For example, a filtered Gaussian noise input[8] would remain Gaussian. Since a narrow-band Gaussian noise process shows amplitude fluctuations with a time constant approximately the inverse of its spectral range,[8] the output of a filter with gain should exhibit this property also. The fact that a true oscillator does not indicates that a correlating mechanism is operative.

Gain saturation is one mechanism that operates to eliminate fluctuations in the output intensity. The action is similar to that of a limiter. In an optical maser, gain saturation arises almost entirely from depletion of the population inversion rather than from any nonlinearity in the stimulated emission process. To a very high degree, the output waveform is sinusoidal and the saturation depends only upon the time-average power, the time average extending over a time long compared to the period of the output waveform but short compared to any relaxation mechanism or pumping rate.

Suppose now that the filtered output has a spectral range $\Delta\nu$ so that the input noise has power fluctuations with a time constant approximately $\Delta\nu^{-1}$. If the gain of the maser medium has a relaxation time $\tau_g < \Delta\nu^{-1}$, then the input fluctuations will be virtually absent in the output. On the other hand, if $\tau_g > \Delta\nu^{-1}$ the input fluctuations will appear in the output.

Thus, for a true oscillator

$$\tau_g\Delta\nu < 1 \tag{1a}$$

while for an amplifying filter

$$\tau_g\Delta\nu > 1. \tag{1b}$$

As will be seen later, for a weak optical maser $\Delta\nu$ can be quite large, while for a strong maser $\Delta\nu$ becomes vanishingly small. The cavity bandwidth $\Delta\nu_c$ represents an upper limit for $\Delta\nu$. For example, in a gas maser $\Delta\nu_c \approx 10^6$ cps and $\tau_g \approx 10^{-6}$, so that for unsaturated gains less than the loss, $\Delta\nu \approx \Delta\nu_c$, $\Delta\nu\tau_g \approx 1$, and the device acts like an amplifying filter. When the unsaturated gain exceeds the loss, $\Delta\nu$ becomes much less than $\Delta\nu_c$ and the device becomes an oscillator.

In the semiconductor maser, the lowest reported value is $\Delta\nu \approx 3 \times 10^9$

cps, corresponding to 0.1Å. Unless $\tau_g < 3 \times 10^{-10}$, it is quite probable that the device acts like an amplifying filter rather than a true oscillator.

The interested reader will find a short discussion of the effect of noise on oscillators in a book by van der Ziel.[9] His discussion indicates a method of approach to obtaining a quantitative solution to a very complicated nonlinear problem. However, the concept of the amplifying filter probably yields a good first approximation to the spectral width of the oscillator output.

In the following sections, the foregoing concepts will be exploited to exhibit the circuit formalism and to study linewidth and threshold behavior. The basic results are applicable to any uniformly pumped, single-mode maser or any multimode maser for which nonlinear mixing or coupling of modes is not significant. Some remarks concerning the lack of extremely narrow linewidths in the semiconductor maser will also be made.

III. SINGLE MODE REPRESENTATION FOR AN OPTICAL MASER

The electromagnetic fields associated with the optical maser are very close to being plane waves. To a very good approximation, each mode of the electromagnetic field can be represented by field quantities, $E(z)$ and $I(z)$. These quantities can be normalized to have the dimensions of voltage and current, respectively. The relationship between $E(z)$ and $I(z)$ is obtained by specifying the value of the function $Z(z) = E(z)/I(z)$ at some point z. In the content of this paper, a mode is one member of a complete set of transverse eigenfunctions which are appropriate for the geometry in question. No orthogonality with respect to the z coordinate is implied.

If the various modes of the electromagnetic field are uncoupled and E and I are represented as complex quantities, then a linear relationship of the form

$$\begin{vmatrix} E(z_1) \\ I(z_1) \end{vmatrix} = \begin{vmatrix} A & Z_0 B \\ Z_0^{-1} C & D \end{vmatrix} \begin{vmatrix} E(z_2) \\ I(z_2) \end{vmatrix} \tag{2a}$$

or

$$\mathbf{\Psi}(z_1) = \mathbf{T}(z_1, z_2) \cdot \mathbf{\Psi}(z_2). \tag{2b}$$

can exist for each mode.[10,11] The input side is taken at z_1 and the output side at z_2, as in Fig. 1. The quantity Z_0 is the characteristic impedance associated with the transmission medium. The directions of E and I are defined so that when the phase difference between E and I falls in

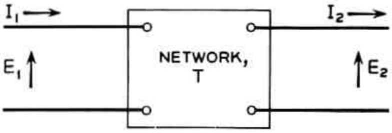


Fig. 1 — Generalized linear two-port network for single mode representation.

the first and fourth quadrant, power is flowing in the direction $z_1 \to z_2$, while for the second and third quadrant, the direction $z_2 \to z_1$. The choice of directions makes it possible to determine the result of cascading a number of sections by writing

$$\mathbf{\Psi}(z_1) = \mathbf{T}(z_1, z_2) \cdot \mathbf{T}(z_2, z_3) \cdots \mathbf{T}(z_{n-1}, z_n) \cdot \mathbf{\Psi}(z_n). \tag{3}$$

The complex quantities A, B, C and D are dimensionless and are independent of time in the steady state. In general, they are functions of frequency. For convenience, the characteristic impedance of the transmission medium will be taken as unity, i.e., $Z_0 = 1$. The transmission medium between planes z_n and z_{n+1} will be referred to as a "network." The properties of the network are described uniquely by the quantities A, B, C and D. General relations among these quantities can be determined by specifying the transfer properties of the network. These are reviewed in detail in Appendix A and are described below.† For example, a matched network which produces no reflection from side z_n when terminated by the characteristic impedance on side z_{n+1} can be written

$$\begin{vmatrix} A & B \\ B & A \end{vmatrix} \tag{4}$$

in which A and B are, in general, independent complex parameters.

A reciprocal network is characterized by the relation $AD - BC = 1$. A reactive network has the transformation

$$\begin{vmatrix} \alpha & j\beta \\ j\gamma & \delta \end{vmatrix} \tag{5}$$

in which the four independent parameters α, β, γ and δ are real. A reciprocal reactive network, in addition, has $\alpha\delta + \beta\gamma = 1$, so that only three of the parameters are independent. A matched reciprocal reactive network has $\alpha = \delta$ and $\beta = \gamma$, so that only one parameter is independent. The network of this type can be written

† The Appendix is included because there is no single convenient reference.

$$\begin{vmatrix} \cos\varphi & j\sin\varphi \\ j\sin\varphi & \cos\varphi \end{vmatrix} \tag{6}$$

in which the real parameter φ is the phase shift from z_n to z_{n+1}. A length of transmission medium is an example of a matched reciprocal reactive network. For this case, $\varphi = 2\pi\nu(z_{n+1} - z_n)/c'$, in which c' is the phase velocity of the radiation.

The most general unmatched reciprocal reactive network, which has three independent parameters, can always be characterized as

$$\begin{vmatrix} \alpha & j\beta \\ j\gamma & \delta \end{vmatrix} = \begin{vmatrix} \cos\varphi_1 & j\sin\varphi_1 \\ j\sin\varphi_1 & \cos\varphi_1 \end{vmatrix} \begin{vmatrix} N & 0 \\ 0 & N^{-1} \end{vmatrix} \begin{vmatrix} \cos\varphi_2 & j\sin\varphi_2 \\ j\sin\varphi_2 & \cos\varphi_2 \end{vmatrix}. \tag{7}$$

The network

$$\begin{vmatrix} N & 0 \\ 0 & N^{-1} \end{vmatrix} \tag{8}$$

is known as an ideal transformer of turns ratio N. Thus, the most general unmatched reactive reciprocal network, aside from phase shifts φ_1 and φ_2, is an ideal transformer.

A resistive network produces no phase shift and can be characterized as having A, B, C and D all real. A matched reciprocal resistive network has only one independent parameter and can be written

$$\begin{vmatrix} \cosh\theta & \sinh\theta \\ \sinh\theta & \cosh\theta \end{vmatrix}. \tag{9}$$

The matched reciprocal resistive network is known as an attenuator. It is shown in Appendix A that the power attenuation is given by $\exp -2\theta$. The parameter θ is referred to as the attenuator line length. The most general matched reciprocal network (resistance plus reactance) has two independent variables and can be written

$$\begin{vmatrix} \cos(\varphi - j\theta) & j\sin(\varphi - j\theta) \\ j\sin(\varphi - j\theta) & \cos(\varphi - j\theta) \end{vmatrix} = \begin{vmatrix} \cos\varphi & j\sin\varphi \\ j\sin\varphi & \cos\varphi \end{vmatrix} \begin{vmatrix} \cosh\theta & \sinh\theta \\ \sinh\theta & \cosh\theta \end{vmatrix}. \tag{10}$$

Therefore, the most general matched reciprocal network consists of an attenuator with phase shift. Note that the attenuation and phase shift commute. As a result, a matched network with distributed attenuation and phase shift can be lumped into a network with attenuation in cascade with a network having only phase shift.

The transmission factor or transmissivity of the network, denoted as L, is shown in Appendix A to have the value

$$L = \frac{4}{| A + B + C + D |^2}. \tag{11}$$

For a matched reactive reciprocal network as in (6), $L = 1$. The factor L is also known as the gain of the network. The transmission factor equals the ratio of power transmitted to power incident when the network is preceded and followed by matched terminations.

The reflection factor or reflectivity of a network is given by

$$R = \tfrac{1}{4} L \, | A + B - C - D |^2. \tag{12}$$

The reflection factor is the ratio of power reflected to power incident when the input and output terminals are matched. For a matched network, $(A = D, B = C)$, $R = 0$.

The noise generated in a network can be represented by suitably chosen current and voltage generators i and e at the input to the network as in Fig. 2.[12] The network itself can then be considered as noiseless. For example, for a series resistor r, the noise generators are appropriately $i = 0$ and $| e | = [4p(\nu)d\nu r]^{\frac{1}{2}}$ in which

$$p(\nu)\, d\nu = \frac{h\nu d\nu}{\exp h\nu/kT - 1} \tag{13}$$

is the thermally generated noise power in a frequency range $d\nu$ centered at frequency ν when the resistor is at temperature T.[13] The phase of e is a random variable. For a shunt resistor, r, the noise generators are $e = 0$ and $| i | = [4p(\nu)d\nu/r]^{\frac{1}{2}}$. For reactive networks, $e = i = 0$.

For an arbitrary network containing lossy elements, the appropriate values of e and i can be expressed in terms of the components of the transformation matrix, A, B, C and D. Since an arbitrary passive network can always be matched by suitable use of transformers and line lengths, placed on either side of the network, the arbitrary network can

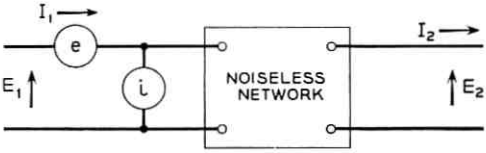


Fig. 2 — Equivalent external current and voltage generators for noiseless network representation.

always be made to appear matched and resistive. It follows that an arbitrary lossy network can always be represented as a resistive matched network imbedded in reactive networks.[14] The reactive networks on either side of the imbedded network are the inverse in reverse order of the networks required for matching the original arbitrary network.

It follows, too, that if the appropriate values of e and i can be determined for any resistive matched network, then the values can be determined for the imbedded network. These values can then be transformed through the reactive networks at the input side of the imbedded network to represent the appropriate values at the input side of the original network.

Thus, it is only necessary to have a general formula for e and i pertinent to a resistive matched network. In Appendix B it is shown that the appropriate values of e and i for any matched resistive network are given by[15]

$$\begin{aligned} |e|^2 &= |i|^2 = AB4p(\nu)d\nu \\ ei^* &= e^*i = \tfrac{1}{2}(A^2 + B^2 - 1)4p(\nu)d\nu. \end{aligned} \tag{14}$$

The values of e and i without the factor $4p(\nu)d\nu$ will be referred to as the normalized values. The quantities, $|e|^2$, $|i|^2$ and ei^* commute with a phase shift network (a length of transmission line). A value e following a transformer of turns ratio N becomes, at the input side of the transformer, Ne, while a current i becomes i/N. Thus, the procedure in finding the values of e and i preceding any given network, $\mathbf{T}$, is to find the appropriate reactive transformations of the form

$$\mathbf{T} = \mathbf{T}(\varphi_1)\cdot\mathbf{T}(N)\cdot T(\varphi_2)\cdot\mathbf{T}(r)\cdot\mathbf{T}(\varphi_3)\cdot\mathbf{T}(M)\cdot\mathbf{T}(\varphi_4) \tag{15}$$

in which $\mathbf{T}(r)$ is the imbedded matched resistive network, and find the values of e and i appropriate to $\mathbf{T}(r)$ using (14). The values of e' and i' appropriate to the input side of $\mathbf{T}$ are $e' = Ne$ and $i' = i/N$.

In Appendix B, it is shown that the noise power into a matched load following any given network at uniform temperature T (considering only the noise arising from the given network and ignoring the noise originating in the matched loads at the input and output side) is given by

$$dP = L\,|e + i|^2 p(\nu)\mathrm{d}\nu \tag{16}$$

in which L is the transmission or gain factor for the network, given by (11), and e and i are the normalized noise generators at the input to the network. For example, for an attenuator with transmission factor

L, (16) yields

$$dP = (1 - L)p(\nu)d\nu \tag{17}$$

as is shown in Appendix B.

Since the noise parameters commute with a matched phase shift network, attenuation and phase shift can be lumped with respect to noise properties as well as transfer properties.

IV. THE OPTICAL MASER

The maser medium has the property that light passing through the medium once is amplified by a factor $G_1(\nu)$. In addition, the light undergoes a phase shift, $\varphi(\nu)$. Thus, the matched maser medium can be characterized as an attenuator with a transmission factor G_1 in cascade with a matched reciprocal phase shift network.

Since the spontaneous emission from the maser medium can be considered as thermal noise,[16] the spontaneous emission power radiated into a given mode by a uniform maser medium should be given by

$$dP(\nu) = [1 - G_1(\nu)]p(\nu)d\nu \tag{18}$$

as follows from (17). Since $p(\nu) = h\nu/(\exp h\nu/kT - 1)$, the question naturally arises as to what temperature to associate with the maser medium. In particular, one wonders whether a noise formula like (18), which is valid for passive networks in thermal equilibrium, can be used when the maser medium is active, i.e., when $G_1 > 1$. Normally, the radiation temperature of the uniform maser medium is defined by the Boltzmann factor[16]

$$n_2(\nu)/n_1(\nu) = \exp -h\nu/kT \tag{19}$$

in which n_2 and n_1 are the densities of upper- and lower-state atoms, respectively. Using (19) as the definition of maser temperature, it follows that (18) is precisely correct for a uniform medium.

To illustrate, one can write for the emission power, dP, into a given mode in a frequency range $d\nu$, along the z-axis

$$\partial dP/\partial z = h\nu w_i dP(n_2 - n_1) + \tfrac{1}{2}dw_s h\nu n_2 \tag{20}$$

in which w_i is the probability per unit intensity per unit time for stimulated emission into the mode and dw_s is the probability for spontaneous emission in either direction for the same mode into frequency range $d\nu$. The population densities for the upper and lower maser levels, n_2 and

n_1, are assumed constant with z.† Solving (20) subject to the initial condition $dP = 0$ at $z = 0$ yields

$$dP(z) = \frac{\frac{1}{2}(dw_s/w_i)(1 - G_1)}{(n_1/n_2) - 1} \tag{21}$$

in which

$$G_1 = \exp h\nu w_i(n_2 - n_1)z. \tag{22}$$

Since the probability for stimulated emission is related to the probability for spontaneous emission into frequency range $d\nu$ by

$$dw_s = 2w_i h\nu d\nu \tag{23}$$

for a given mode, (21) and (18) are identical.‡ It follows that (18) correctly accounts for the spontaneous noise into a single mode, so long as one writes $p(\nu) = h\nu/(n_1/n_2 - 1)$. It also indicates the applicability of the formalism described in the preceding section to maser media.

The fact that the maser temperature T, as defined by (19), and $p(\nu)$, as defined by (13), are negative should not distract the reader from the more significant fact that (18) or (21) correctly predicts the noise power emitted by the maser medium. This quantity is never negative, and varies smoothly as T goes from positive to negative values.

The noise output from the maser can be calculated using (16), and this will be the aim of the following analysis. It is worth noting that the only significant parameters characterizing the medium, assuming that the maser medium is uniform and matched, are the total single-pass gain, G_1, the population ratio and the phase shift through the medium. The effective attenuator line length, θ, is given by the expression $G_1 = \exp -2\theta$.

V. THE REGENERATIVE OPTICAL MASER OSCILLATOR

The mirrors forming the optical cavity, being reactive (except for a small absorption loss which will be neglected) and reciprocal, can be characterized as ideal transformers. The phase shift associated with the mirrors on the cavity side can be added to the single-pass phase shift of the maser medium. The mirror phase shift external to the cavity is not significant to the problem at hand.

† This implies no saturation or uniform saturation.

‡ Equation (23) is equivalent to the statement that the total rate of spontaneous emission is related to the total rate of induced emission per unit intensity by $dw_s = w_i 8\pi h\nu^3 d\nu/c^3$, since the total number of modes per unit volume per unit frequency interval is given by $8\pi\nu^2/c^3$. (See Ref. 16.)

Since the reflection factor for an ideal transformer is, from (12) and (8)

$$R = (N^2 - 1)^2/(N^2 + 1)^2 \tag{24}$$

it follows that the equivalent turns ratio for a mirror of reflectivity R is given by

$$N^2 = (1 + R^{\frac{1}{2}})/(1 - R^{\frac{1}{2}}). \tag{25}$$

The transmission factor L_m for the maser cavity with unequal mirrors of reflectivity R and R' is obtained by combining the cascade of transformer with turns ratio N, attenuator with loss parameter $\theta = -\frac{1}{2} \ln G_1$, transmission line with phase shift $\varphi = 2\pi\nu L/c'$ + constant, in which $2L/c'$ is the equivalent round-trip time,† and transformer with turns ratio M^{-1}, $M^2 = (1 + R'^{\frac{1}{2}})/(1 - R'^{\frac{1}{2}})$, as in Fig. 3. The noise power from one end of the maser is given by (16)

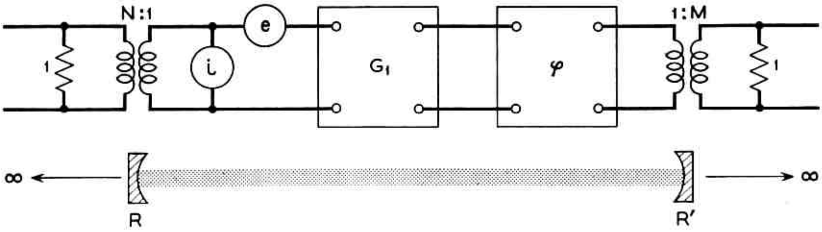


Fig. 3 — Equivalent circuit for maser cavity with mirrors of unequal reflectivity.

$$dP = L_m \mid Ne + N^{-1}i \mid^2 p(\nu)d\nu \tag{26}$$

where the normalized noise generators for the maser medium are given by

$$\begin{aligned} \mid e \mid^2 = \mid i \mid^2 &= \cosh \theta \sinh \theta \\ ei^* = e^*i &= \tfrac{1}{2}(\cosh^2 \theta + \sinh^2 \theta - 1) \end{aligned} \tag{27}$$

as follows from (14) and (9)

The cascade of mirror, maser medium and mirror takes the form

$$\begin{aligned} &\begin{vmatrix} N & 0 \\ 0 & N^{-1} \end{vmatrix} \cdot \begin{vmatrix} \cos(\varphi - j\theta) & j\sin(\varphi - j\theta) \\ j\sin(\varphi - j\theta) & \cos(\varphi - j\theta) \end{vmatrix} \cdot \begin{vmatrix} M^{-1} & 0 \\ 0 & M \end{vmatrix} \\ &= \begin{vmatrix} (N/M)\cos(\varphi - j\theta) & NMj\sin(\varphi - j\theta) \\ (NM)^{-1}j\sin(\varphi - j\theta) & (M/N)\cos(\varphi - j\theta) \end{vmatrix}. \end{aligned} \tag{28}$$

† The phase velocity c' is a function of ν and G_1 by virtue of the anomalous dispersion of the maser medium.

Using (11), the transmission factor is given by

$$\begin{aligned} L_m &= \frac{4}{\begin{aligned}|(N/M + M/N) \cos (\varphi - j\theta) \\ + j(MN + [MN]^{-1}) \sin (\varphi - j\theta)|^2\end{aligned}} . \\ &= \frac{1}{\begin{aligned}[(1 - R^{\frac{1}{2}}R'^{\frac{1}{2}}G_1)^2/G_1(1 - R)(1 - R')] \\ + [4R^{\frac{1}{2}}R'^{\frac{1}{2}}/(1 - R)(1 - R')] \sin^2 \varphi\end{aligned}} . \end{aligned} \tag{29}$$

Note that in the limit $R = R' = 0$ (no mirrors), $L_m = G_1$ as would be expected; while in the limit $R = R'$, $G_1 = 1$ (a transparent maser medium)

$$L_m = \frac{1}{1 + [4R/(1 - R)^2] \sin^2 \varphi} \tag{30}$$

which is the transmission factor for the Fabry-Perot or optical cavity surrounding the maser medium.[7,17] Equation (29), or versions of it with $R = R'$, has been derived before.[7] In these cases, however, it had been necessary to assume that the maser medium uniformly fills the region between the mirrors. No such restriction is necessary.

The noise power $| Ne + N^{-1}i |^2$ has the value, using (27)

$$\begin{aligned} | Ne + N^{-1}i |^2 &= (N^2 + N^{-2}) | e |^2 + 2ei^* \\ &= (N^2 + N^{-2}) \cosh \theta \sinh \theta \\ &\quad + (\cosh^2 \theta + \sinh^2 \theta - 1) \end{aligned} \tag{31}$$

which after some manipulation yields

$$| Ne + N^{-1}i |^2 = (1 - G_1)(1 + RG_1)/G_1(1 - R). \tag{32}$$

Combining (26), (29) and (32), the noise power in frequency range $d\nu$ leaving the maser cavity through the mirror R' can be written

$$dP = \frac{(1 + RG_1)(1 - R')(1 - G_1)p(\nu)d\nu}{(1 - R^{\frac{1}{2}}R'^{\frac{1}{2}}G_1)^2 + 4R^{\frac{1}{2}}R'^{\frac{1}{2}}G_1 \sin^2 \varphi} . \tag{33}$$

In the vicinity of a cavity resonance at frequency ν_0, the phase shift φ differs from some multiple of π by an amount $\Delta\varphi = 2\pi(\nu - \nu_0)L/c'$. The free spectral range of the cavity mode at frequency ν_0 is the range $\nu_0 - c'/4L \leqq \nu \leqq \nu_0 + c'/4L$ as illustrated in Fig. 4; the mode spacing is $c'/2L$. Thus, the total noise power leaving the cavity through R', associated with one cavity mode, is obtained by integrating dP over the free spectral range of the mode, yielding

$$P = \int_{\nu_0 - c'/4L}^{\nu_0 + c'/4L} \frac{p(\nu) C_1 \, d\nu}{C_2 + C_3 \sin^2 2\pi(\nu - \nu_0)L/c'} \, . \tag{34}$$

With the substitution $\varphi = 2\pi(\nu - \nu_0)L/c'$, (34) can be written

$$P = p(\nu_0) \frac{c'}{2\pi L} \int_{-\pi/2}^{+\pi/2} \frac{C_1 \, d\varphi}{C_2 + C_3 \sin^2 \varphi} \tag{35}$$

in which the quantities C_1, C_2 and C_3 are given by

$$C_1 = [(1 + RG_1)(1 - R')(1 - G_1)],$$

$$C_2 = [1 - R^{\frac{1}{2}} R'^{\frac{1}{2}} G_1]^2$$

and

$$C_3 = 4R^{\frac{1}{2}} R'^{\frac{1}{2}} G_1 \, .$$

Since the integrand is large only in a very small range of frequencies near ν_0, it can safely be assumed that $p(\nu)$, G_1 and R are constant with frequency and have their values at $\nu = \nu_0$.

With this approximation the integral has the value $\pi C_1/[C_2^2 + C_2 C_3]^{\frac{1}{2}}$ (see Ref. 18), yielding

$$P(R') = [(1 + RG_1)(1 - R')/(1 - RR'G_1^2)] \\ (1 - G_1)p(\nu_0)(c'/2L). \tag{36}$$

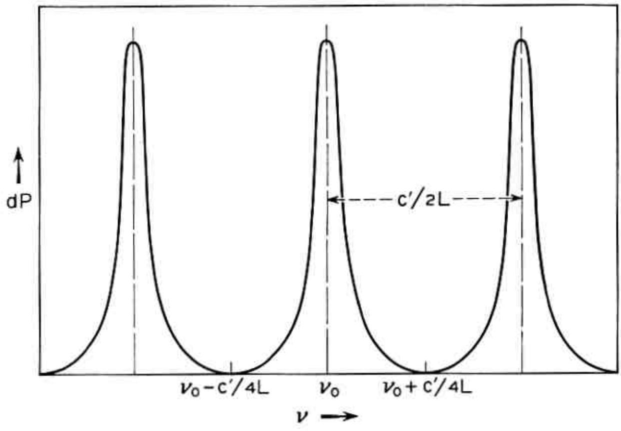


Fig. 4 — Free spectral range of cavity mode at frequency ν_0.

From (18) it may be noted that $[1 - G_1(\nu_0)]p(\nu_0)c'/2L$ is the spontaneous emission power that would be emitted by the maser medium, in the absence of the cavity, into the spectral range $c'/2L$ if the gain were constant over that range. When $G_1 > 1$, the spontaneous emission is enhanced by stimulated emission. The noise power

$$[1 - G_1(\nu_0)]p(\nu_0)c'/2L$$

will be denoted as $P_s(\nu_0)$. The noise power leaving the cavity through mirror R' is, therefore

$$P(R') = P_s(1 + RG_1)(1 - R')/(1 - RR'G_1^2). \qquad (37)$$

The power leaving the cavity through mirror R is given by (37) with R and R' interchanged. The total power leaving the cavity is

$$P_t = 2P_s[1 - RR'G_1 + \tfrac{1}{2}(G_1 - 1)(R + R')]/[1 - RR'G_1^2] \quad (38)$$

so that the cavity enhances the spontaneous emission by the factor following $2P_s$.

It should be noted that the integration of (36), which leads to (37) and (38), is valid even if $G_1(\nu)$ varies over the range $c'/2L$, so long as the frequency range over which G_1 has a significant variation is large compared to the spectral range of the noise power. This will always be so, providing the natural linewidth of the transition is large compared to the spectral range of the noise, independent of whether the transition is homogeneously or inhomogeneously broadened. Equations (37) and (38) are valid even when the gain profile is saturated, providing the maser medium is uniformly saturated with respect to the axial direction. In general, the saturation will tend to be uniform, assuming uniform pumping, since the power in the cavity tends to be uniform with length. For very high-gain maser media the latter statement is not valid.

It should be noted that (38) contains an implication which is not immediately obvious. In the limit $n_1 = n_2$ with $G_1 = 1$ and P_s finite, (38) states that $P_t = 2P_s$ independent of R or R'. Thus the total spontaneous emission noise power leaving the optical cavity is independent of its bandpass. The spectral distribution of the noise is altered, however, to correspond to the bandpass.†

The preceding results can be used to determine the linewidth of the oscillator. The spectral range or bandwidth of the noise power is obtained from (33) by determining the frequencies at which dP falls to

† In the limit $c'/2L$ very much less than the inhomogeneously broadened linewidth, (38) implies that the total spontaneous emission power into all modes is independent of the mirror reflectivity since $G_1 = 1$ over the entire line.

half its value at $\nu = \nu_0$. This occurs at frequencies $\nu_{\frac{1}{2}}$ defined by

$$4R^{\frac{1}{2}}R'^{\frac{1}{2}}G_1 \sin^2 2\pi(\nu_{\frac{1}{2}} - \nu_0)L/c' = (1 - R^{\frac{1}{2}}R'^{\frac{1}{2}}G_1)^2 \tag{39}$$

yielding a spectral width $\Delta\nu = 2(\nu_{\frac{1}{2}} - \nu_0)$ given by

$$\begin{aligned}\Delta\nu &= (c'/L\pi) \sin^{-1} \tfrac{1}{2}(1 - R^{\frac{1}{2}}R'^{\frac{1}{2}}G_1)/(RR'G_1^2)^{\frac{1}{4}} \\ &\approx (c'/2L\pi)(1 - R^{\frac{1}{2}}R'^{\frac{1}{2}}G_1)/(RR'G_1^2)^{\frac{1}{4}}.\end{aligned} \tag{40}$$

Equation (40) can be rewritten by substituting for $1 - (RR')^{\frac{1}{2}}G_1$ the value given by (38)

$$\begin{aligned}1 - (RR'G_1^2)^{\frac{1}{2}} = 2(P_s/P_t)[1 - RR'G_1 + \tfrac{1}{2}(G_1 - 1)(R + R')]/ \\ [1 + (RR'G_1^2)^{\frac{1}{2}}].\end{aligned} \tag{41}$$

The resulting expression is a function of the single-pass gain G_1. For the strong oscillator the gain saturates at a value such that $RR'G_1^2$ differs from unity by a vanishingly small amount. Thus it is expedient, in the expression for $\Delta\nu$ found by substituting (41), to write for the saturated single-pass gain

$$G_1 \equiv [1 - (1 - (RR'G_1^2)^{\frac{1}{2}})]/(RR')^{\frac{1}{2}} \tag{42}$$

and then to replace $1 - (RR'G_1^2)^{\frac{1}{2}}$ by its approximate value obtained from (41). In the reiteration, the approximate value can be written

$$1 - (RR'G_1^2)^{\frac{1}{2}} \approx \frac{(R^{\frac{1}{2}} + R'^{\frac{1}{2}})^2}{4(RR')^{\frac{1}{2}}} \frac{(1 - (RR')^{\frac{1}{2}})}{(RR')^{\frac{1}{4}}} \frac{2\pi h\nu\Delta\nu_c}{(1 - n_1/n_2)P_t} \tag{43}$$

which follows by substituting in (41) the value $G_1 = (RR')^{-\frac{1}{2}}$ and replacing P_s by its value $(1 - G_1)(c'/2L)h\nu/(n_1/n_2 - 1)$. The cavity bandwidth $\Delta\nu_c$ is given by (40) with $G_1 = 1$.

Combining (40–43) and performing the necessary algebra yields

$$\begin{aligned}\Delta\nu = 2\pi(h\nu/P_t)(\Delta\nu_c)^2(1 - n_1/n_2)^{-1}z[1 + (\pi h\nu\Delta\nu_c/P_t) \\ \cdot(1 - n_1/n_2)^{-1} \cdot (- z[x + 3] + 1 + x)/x^{\frac{1}{2}}]\end{aligned} \tag{44}$$

in which

$$z = (R^{\frac{1}{2}} + R'^{\frac{1}{2}})^2/4(RR')^{\frac{1}{2}} \tag{45}$$

is a term which is identically unity when $R = R'$ and remains close to unity even for R and R' differing by a factor of ten, and the quantity $x = (RR')^{\frac{1}{2}}$.

When $h\nu\Delta\nu_c/P_t \ll 1$ the linewidth is

$$\Delta\nu \approx 2\pi \frac{h\nu}{P_t} (\Delta\nu_c)^2 (1 - n_1/n_2)^{-1} \tag{46}$$

and differs from the well-known Schawlow-Townes[4] formula by a factor of two. The correction term $(1 - n_1/n_2)^{-1}$ approaches unity for the ideal maser ($n_1/n_2 = 0$). This term has also been found by Shimoda.[5] It should be noted that n_1/n_2 is the saturated value and in general may be very difficult to evaluate. An approximate value may be found by setting the single-pass gain given by (22) equal to the saturated gain given by (42). Extracting the appropriate value of n_1/n_2 requires a knowledge of the transition probability, the time constants for the upper and lower states and the mirror reflectivities. The degeneracy of the upper and lower levels should also be taken into account.

For the 6328 Å gas maser, typical values for a one-meter-long discharge are $P_t \approx 10^{-3}$ watts/mode, $\Delta\nu_c \approx 10^6$ cps and $n_1/n_2 = 0.98$, yielding a linewidth of 10^{-1} cps.

For the semiconductor optical maser, (46) predicts a linewidth of approximately 5×10^8 cps, taking $L = 0.06$ cm and $P_s/P_t \approx 10^{-2}$. The latter numbers are taken from the data of Quist et al.[19] The corresponding wavelength range is 10^{-2} Å. This estimate is probably a conservative lower limit because of the neglect of internal losses in the derivation and the fact that the internal losses are significant in the actual device. The lowest observed linewidths are about 10^{-1} Å. The relatively large linewidths arise from very short spontaneous emission lifetimes ($<10^{-9}$ sec). The relatively large amount of enhanced spontaneous emission available produces saturation at small values of P_t/P_s.

As has just been shown, it is possible to write formal expressions for the power output and spectral width of the noise emitted in a given mode. Since the power, and hence the spectral width, depend on the saturated single-pass gain, it is necessary to take the dynamic properties of the maser medium into account. This is illustrated graphically in Fig. 5, in which the curve of P vs G_1 as represented by (38) is shown. Also shown are a curve of $\Delta\nu$ and three dashed curves representing the dynamic properties of the maser medium. The latter curves may be found by solving the rate equations for the maser medium to obtain a relation between the power taken from the maser medium and the single-pass gain. When large power is taken from the maser medium, G_1 must approach unity, since the population inversion must approach zero. When the noise power taken from the maser medium approaches zero, the single-pass gain approaches its small signal value G_{10}. The curve representing the dynamics of the maser medium intersects the curve representing the static characteristics of the cavity at some value $G_1 < (1/RR')^{\frac{1}{2}}$. This is the operating point of the maser. The value of G_1 at the operating point determines the value of $\Delta\nu$.

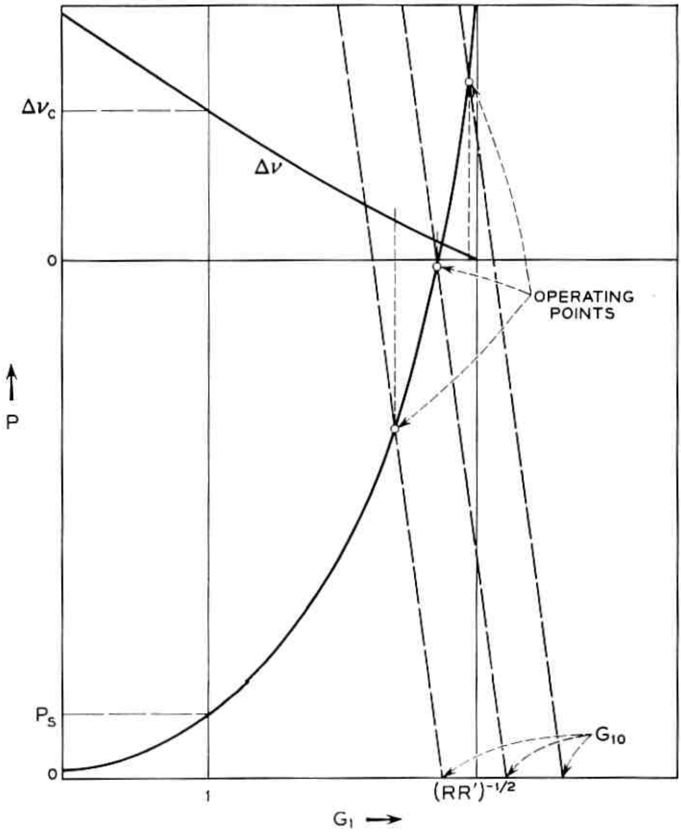


Fig. 5 — Noise power vs single-pass gain as represented by (38).

For a gas maser with small single-pass gain, the dynamic properties of the maser medium can be shown to be controlled by an equation of the form[20]

$$G_1 \approx 1 + k_0 l/(1 + \kappa P_c/P_s) \tag{47}$$

in which k_0 is the small signal gain parameter for the mode in question, l is the active length of the medium, κ/P_s is a saturation parameter dependent on the Einstein A and B coefficients for the maser levels and P_c is the power in the cavity. In most maser media $\kappa \ll 1$, so that G_1 varies quite slowly with P_c/P_s. It should be noted that P_s is essentially constant and can be evaluated for $G_1 = 1$.

The gain parameter k_0 varies with discharge current. For low currents, k_0 is proportional to current.[20] Thus, as the discharge current is increased from zero, the output power increases. From Fig. 5 it would be expected that as the discharge current approaches a value for which $1 + k_0 l$ approaches $(RR')^{-\frac{1}{2}}$, the output power would show an extremely sharp rise with current. This can be illustrated by writing for G_1 as a function of discharge current I, taking $R = R'$ for convenience,

$$G_1 \approx 1 + [(1 - R)/R](I/I_0)/(1 + \kappa P_c/P_s) \tag{48}$$

in which I_0 is the current for which $G_1 R = 1$ in the absence of saturation. From (38), taking $P_c = P_t/2(1 - R)$

$$P_c/P_s = 1/(1 - RG_1). \tag{49}$$

Solving (48) and (49) for P_c/P_s yields

$$P_c/P_s = [\Delta I/I_0 + \kappa + \sqrt{(\Delta I/I_0 + \kappa)^2 + 4\kappa}]/2\kappa \tag{50}$$

in which $\Delta I = I - I_0$. Note that when $(\Delta I/I_0) + \kappa = 0$, $P_c/P_s = \kappa^{-\frac{1}{2}}$. When $(\Delta I/I_0) + \kappa$ has the value $\kappa^{\frac{1}{2}}(f - f^{-1})$, P_c/P_s has the value $f\kappa^{-\frac{1}{2}}$. Therefore, the change in P_c/P_s by a factor f is larger than the change in $\Delta I/I_0$ of $\kappa^{\frac{1}{2}}(f - f^{-1})$ by a factor $\kappa^{-\frac{1}{2}}$. Thus, when $\kappa \ll 1$ the power output shows a very sharp threshold with current. The value of κ is of order 10^{-10} for the 6328 Å gas maser.

VI. THE NONREGENERATIVE OPTICAL MASER OSCILLATOR

Some maser media have such large single-pass gain that spontaneous emission originating at one end can be amplified sufficiently to saturate the maser medium at the opposite end. For example, the 3.39 μ transition in neon, in a helium-neon gas maser, has gains of order 50 db/meter in small-bore tubing.[21] The saturated output power is in the 1–10 mw range. Xenon at 3.5 μ has even larger gain.[22] Under these circumstances the maser can behave as a saturated oscillator without an optical cavity. The extreme line narrowing characteristic of the cavity oscillator will be lacking, but the power levels and directionality will be comparable.

This type of oscillator is characterized by a peak output always at the line center, independent of temperature and any physical dimension, line narrowing over the inhomogeneously broadened line and an extremely stable output. The power per steradian per cycle will be much greater than conventional light sources, and so these structurally simple oscillators may be very useful as frequency standards and for calibration purposes.

Some idea of the line narrowing possible with this type of oscillator can be obtained by reference to (18), which describes the noise power emanating from one end of the unsaturated maser medium in one mode with no mirrors at either end, while (33) with $R' = 0$ gives the noise power when one end has a mirror, but the other does not

$$dP = [1 + RG_1(\nu)][1 - G_1(\nu)]p(\nu)d\nu. \tag{51}$$

When $R = 0$, (51) properly yields (18); however, when $R = 1$

$$dP = [1 - G_1^2(\nu)]p(\nu)d\nu. \tag{52}$$

Since G_1^2 is the gain of a maser of twice the length of a maser of gain G_1, the perfectly reflecting mirror placed at one end serves to double the effective length of the maser medium.

Assuming no saturation, so that the gain can be written

$$G_1(\nu) = \exp k(\nu)l \tag{53}$$

in which l is the effective length of the medium, and assuming further that the gain parameter $k(\nu)$ has a Doppler profile

$$k(\nu) = k_0 \exp - [2(\nu - \nu_0)/\Delta\nu_D]^2 \ln 2 \tag{54}$$

in which ν_0 is the frequency of the line center and $\Delta\nu_D$ is the full Doppler width, the spectral width at half power of the spontaneous emission $\Delta\nu$ can be determined by writing

$$1 - \exp [k_0 l \exp - [\Delta\nu/\Delta\nu_D]^2 \ln 2] = \tfrac{1}{2}[1 - \exp k_0 l]. \tag{55}$$

In the limit $k_0 l \gg 1$, (55) can be solved for $\Delta\nu/\Delta\nu_D$ to yield

$$\Delta\nu/\Delta\nu_D = (k_0 l)^{-\frac{1}{2}}. \tag{56}$$

The 3.39 μ maser is saturated by its own spontaneous emission by gains of order 80 db, so that the maximum possible value of $k_0 l$ before saturation is about 20, corresponding to $l \approx 2$ meters. This yields $\Delta\nu \approx \Delta\nu_0/4.5$. The Doppler width at 3.39 μ is about 300 mc, yielding $\Delta\nu \approx 70$ mc. It would be possible to decrease this value somewhat by using several lengths of maser media separated by attenuators to prevent saturation. The slow dependence on l exhibited by (56) indicates the impracticality of this scheme in achieving any more than a factor of four decrease in linewidth unless k_0 can be increased drastically. Fortunately, there is evidence that this can be done, and the nonregenerative oscillator may turn out to be an extremely interesting device. In addition, it can be tuned by application of magnetic fields.

VII. CONCLUSION

The output properties of an optical maser oscillator have been derived subject to the supposition that the oscillator is a saturated amplifier of spontaneous emission noise. The most significant new results concern an expression for the linewidth of the oscillating maser which differs from the commonly accepted value. Some techniques, well-known in the microwave art and used only in a limited way in optics (stratified media), have been generalized to apply to the transmission and noise properties of optical masers.

VIII. ACKNOWLEDGMENT

The author is indebted to H. Seidel, from whom he learned most of the circuit formalism herein, and to A. D. White, J. D. Rigden, and J. E. Geusic for many stimulating discussions.

APPENDIX A

Network Representations

The basic network of interest will be a linear two-port network which will be represented schematically by Fig. 1. The major concern here will be the relationships between the quantities I_1 and E_1 at port 1 and I_2 and E_2 at port 2. The linear relationship among these quantities will be represented by the transformation matrix

$$\begin{vmatrix} E_1 \\ I_1 \end{vmatrix} = \begin{vmatrix} A & B \\ C & D \end{vmatrix} \cdot \begin{vmatrix} E_2 \\ I_2 \end{vmatrix} \tag{57a}$$

which will be abbreviated by

$$\mathbf{\Psi}_1 = \mathbf{T} \cdot \mathbf{\Psi}_2 \tag{57b}$$

in which

$$\mathbf{\Psi}_i = \begin{vmatrix} E_i \\ I_i \end{vmatrix} \qquad i = 1, 2, \cdots \tag{58}$$

and

$$\mathbf{T} = \begin{vmatrix} A & B \\ C & D \end{vmatrix}. \tag{59}$$

Note that the direction of positive current flow is defined to be into the network at port 1 and out of the network at port 2. This choice simplifies the discussion of cascaded networks. It is clear that two networks in cascade can be represented by $\mathbf{\Psi}_1 = \mathbf{T}_1 \cdot \mathbf{\Psi}_2 = \mathbf{T}_1 \cdot \mathbf{T}_2 \cdot \mathbf{\Psi}_3$. The product $\mathbf{T}_1 \cdot \mathbf{T}_2$ follows the usual rules of matrix multiplication. In general, $\mathbf{T}_1 \cdot \mathbf{T}_2 \neq \mathbf{T}_2 \cdot \mathbf{T}_1$; that is, the two networks do not commute. For a cascade of n networks

$$\mathbf{\Psi}_1 = \mathbf{T}_1 \cdot \mathbf{T}_2 \cdots \mathbf{T}_n \cdot \mathbf{\Psi}_{n+1} . \tag{60}$$

The determinant of the transformation will be a quantity of interest and will be defined as $\Delta = AD - CB$. The inverse of (57) is defined by

$$\mathbf{\Psi}_2 = \mathbf{T}^{-1} \cdot \mathbf{\Psi}_1 \tag{61}$$

in which

$$\mathbf{T}^{-1} = \frac{1}{\Delta} \begin{vmatrix} D & -B \\ -C & A \end{vmatrix} \tag{62}$$

is the inverse transformation matrix. Reference will sometimes be made to the exchange network $\underset{\sim}{\mathbf{T}}$, which is merely the network $\mathbf{T}$ with its terminals exchanged so that port 2 becomes port 1 and vice versa. The components of $\underset{\sim}{\mathbf{T}}$ are obtained by writing $\mathbf{\Psi}_2 = \mathbf{T}^{-1} \cdot \mathbf{\Psi}_1$ and noting that for the exchange network the quantities E_2, I_2 become E_1', $-I_1'$ and E_1, I_1 become E_2', $-I_2'$. As a result, the correct description is given by

$$\begin{vmatrix} E_1' \\ I_1' \end{vmatrix} = \frac{1}{\Delta} \begin{vmatrix} D & B \\ C & A \end{vmatrix} \cdot \begin{vmatrix} E_2' \\ I_2' \end{vmatrix} \tag{63}$$

and

$$\underset{\sim}{\mathbf{T}} = \frac{1}{\Delta} \begin{vmatrix} D & B \\ C & A \end{vmatrix} . \tag{64}$$

The net power *into* port 1 is given by

$$P_1 = \tfrac{1}{2}(E_1 I_1^* + E_1^* I_1) = \tfrac{1}{2} \overline{\mathbf{\Psi}}_1 \cdot \mathbf{\Sigma} \cdot \mathbf{\Psi}_1 \tag{65}$$

in which $\overline{\mathbf{\Psi}}$ is the complex transpose of $\mathbf{\Psi}$ and

$$\mathbf{\Sigma} = \begin{vmatrix} 0 & 1 \\ 1 & 0 \end{vmatrix} . \tag{66}$$

Likewise, the net power *out* of port 2 is given by

$$P_2 = \tfrac{1}{2}\overline{\boldsymbol{\Psi}}_2 \cdot \boldsymbol{\Sigma} \cdot \boldsymbol{\Psi}_2 . \tag{67}$$

Other transformations will be introduced as they are needed. In the following, certain types of networks will be characterized in terms of relationships among the components of the transformation matrix, i.e., A, B, C and D. These networks will form the basic "tools" of the analyses.

A.1 *Matched Networks*

If one of the ports of the network is terminated in a matched load and it is then observed that the input impedance of the other port is matched to the line, the network is said to be matched. First, consider the input impedance of the network in Fig. 6. The line is assumed to have unit characteristic impedance. The quantities B and C are thus normalized with respect to the characteristic impedance and admittance of the line, respectively. Writing $E_1 = AE_2 + BI_2$, $I_1 = CE_2 + DI_2$ and

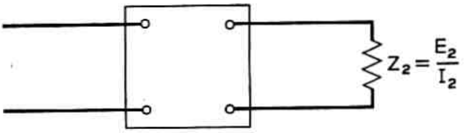


Fig. 6 — Terminating impedance for matched network.

$$Z_{\text{input}(1)} = E_1/I_1 = (AZ_2 + B)/(CZ_2 + D) \tag{68}$$

in which $Z_2 = E_2/I_2$ is the terminating impedance. The network is then matched if $Z_{\text{input}(1)} = Z_2 = 1$, yielding the requirement $A + B = C + D$. Likewise, the exchange network must also be matched, requiring $D + B = C + A$. The two requirements yield the relations $A = D$, $B = C$, and for a matched network

$$\mathbf{T} = \begin{vmatrix} A & B \\ B & A \end{vmatrix}. \tag{69}$$

A.2 *Reciprocal Networks*

A reciprocal network has the same transmission properties in either direction. It follows that for a reciprocal network $\mathbf{T}$ and $\underset{\sim}{\mathbf{T}}$ must have the same transmission factor and phase shift, which from (84) and (72) implies $\Delta = 1$. The simplest nonreciprocal network can be written

$$\mathbf{T}_a = \begin{vmatrix} \Delta^{\frac{1}{2}} & 0 \\ 0 & \Delta^{\frac{1}{2}} \end{vmatrix} \tag{70}$$

with $\Delta \neq 1$; and all nonreciprocal networks can be written

$$\mathbf{T} = \begin{vmatrix} A & B \\ C & D \end{vmatrix} = \begin{vmatrix} \Delta^{\frac{1}{2}} & 0 \\ 0 & \Delta^{\frac{1}{2}} \end{vmatrix} \cdot \begin{vmatrix} A/\Delta^{\frac{1}{2}} & B/\Delta^{\frac{1}{2}} \\ C/\Delta^{\frac{1}{2}} & D/\Delta^{\frac{1}{2}} \end{vmatrix} = \mathbf{T}_a \cdot \mathbf{T}_r \,. \tag{71}$$

The reciprocal network $\mathbf{T}_r$ is known as the reduced network; all the nonreciprocity resides in $\mathbf{T}_a$. The network $\mathbf{T}_a$, known as the abstracted network, commutes with all networks.

In general, Δ is complex and can be written $\Delta = |\Delta| \exp j\,2\varphi$. Since

$$\mathbf{T}_a = \begin{vmatrix} |\Delta|^{\frac{1}{2}} \exp j\varphi & 0 \\ 0 & |\Delta|^{\frac{1}{2}} \exp j\varphi \end{vmatrix} \tag{72}$$

and the exchange network

$$\underset{\sim}{\mathbf{T}}_a = \begin{vmatrix} |\Delta|^{-\frac{1}{2}} \exp -j\varphi & 0 \\ 0 & |\Delta|^{-\frac{1}{2}} \exp -j\varphi \end{vmatrix} \tag{73}$$

it follows that argument φ is the nonreciprocal phase shift.

A.3 *Reactive Networks*

Since a reactive network is lossless, it would be expected that $P_2 = P_1$. It follows that $\overline{\mathbf{\Psi}}_2\Sigma\mathbf{\Psi}_2 = \overline{\mathbf{\Psi}}_2\overline{\mathbf{T}}\Sigma\mathbf{T}\mathbf{\Psi}_2$, so that for a reactive network $\overline{\mathbf{T}}\Sigma\mathbf{T} = \Sigma$. Performing the indicated matrix multiplication yields

$$\begin{vmatrix} A^*C + AC^* & A^*D + C^*B \\ AD^* + CB^* & BD^* + B^*D \end{vmatrix} = \begin{vmatrix} 0 & 1 \\ 1 & 0 \end{vmatrix}. \tag{74}$$

It follows that A^*C and BD^* are imaginary and $A^*D + C^*B = 1$. Note that multiplication of $A^*D + C^*B = 1$ by CB^* yields $CB^* = (A^*C)(B^*D) + |C|^2|B|^2$, which is real. Likewise,

$$AD^* = |A|^2|D|^2 + (AC^*)(BD^*)$$

is real. If the reactive network is reciprocal so that $\Delta = AD - BC = 1 = (AC^*)(D/C^*) - D^*B(C/D^*) = (A/B^*)DB^* - (B/A^*)A^*C$, it follows that both D/C^* and A/B^* are imaginary. Since AC^* is imaginary and C^*B is real, A/B is imaginary. Thus B/B^* is real, which can only occur when B is real or imaginary. It follows that B and C are real (or imaginary) while A and D are imaginary (or real). The question of which set to choose real is decided by noting that the idemfactor (no network) is a reactive reciprocal network. Thus, the appropriate choice is A, D real and B, C imaginary, and the reactive reciprocal network can be written

$$\begin{vmatrix} \alpha & j\beta \\ j\gamma & \delta \end{vmatrix} \tag{75}$$

for which $\alpha\delta + \beta\gamma = 1$. If the reactive reciprocal network is also matched so that $\alpha = \delta$ and $\beta = \gamma$, the network can be written

$$\begin{vmatrix} \cos\varphi & j\sin\varphi \\ j\sin\varphi & \cos\varphi \end{vmatrix} \tag{76}$$

in which the parameter φ is known as the angular length or phase shift. Note that the input impedance for this network is

$$Z_{\text{input}(1)} = \frac{Z_2 + j\tan\varphi}{1 + Z_2 j\tan\varphi} \tag{77}$$

which is recognizable as the impedance transformation for a transmission line of angular length φ and unit characteristic impedance.

Next, certain basic passive circuit elements of interest will be characterized.

A.4 *Series Impedance (Fig. 7)*

Since $I_1 = I_2$ and $E_1 = I_2Z + E_2$, the transfer matrix is given by

$$\mathbf{T} = \begin{vmatrix} 1 & Z \\ 0 & 1 \end{vmatrix}. \tag{78}$$

A.5 *Shunt Admittance (Fig. 8)*

For the shunt admittance, $E_1 = E_2$ and $I_1 = E_2Y + I_2$; thus

$$\mathbf{T} = \begin{vmatrix} 1 & 0 \\ Y & 1 \end{vmatrix}. \tag{79}$$

Both networks are reciprocal but not matched.

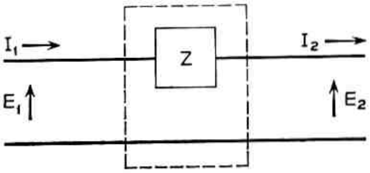


Fig. 7 — Series impedance.

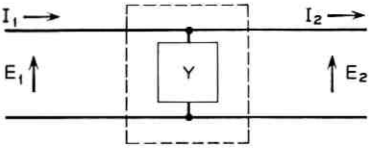


Fig. 8 — Shunt admittance.

A.6 *Ideal Transformer of Turns Ratio N:1 (Fig. 9)*

For the transformer $E_1 = NE_2$ and $I_1 = I_2/N_1$. Thus

$$T = \begin{vmatrix} N & 0 \\ 0 & N^{-1} \end{vmatrix}. \tag{80}$$

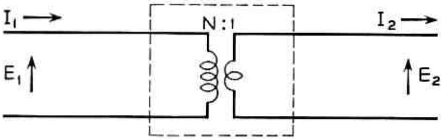


Fig. 9 — Ideal transformer.

A.7 *Network Transmissivity*

Consider a general two-port network which has a matched generator on side 1 and a matched load on side 2. The cascade of networks may be represented schematically as shown in Fig. 10, yielding

$$\begin{vmatrix} E_g \\ I_g \end{vmatrix} = \begin{vmatrix} 1 & 1 \\ 0 & 1 \end{vmatrix} \cdot \begin{vmatrix} A & B \\ C & D \end{vmatrix} \cdot \begin{vmatrix} 1 & 0 \\ 1 & 1 \end{vmatrix} \cdot \begin{vmatrix} E_2 \\ 0 \end{vmatrix}. \tag{81}$$

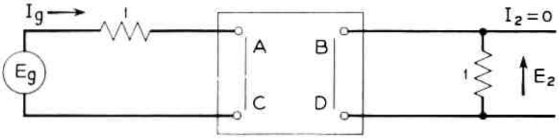


Fig. 10 — Circuit for two-port network with matched generator on side 1 and matched load on side 2.

Performing the indicated matrix multiplication yields

$$\begin{vmatrix} E_g \\ I_g \end{vmatrix} = \begin{vmatrix} A + B + C + D & B + D \\ C + D & D \end{vmatrix} \cdot \begin{vmatrix} E_2 \\ 0 \end{vmatrix}. \tag{82}$$

The network transmission factor is defined as the ratio

$$L = |E_2|^2/|E_{20}|^2 \tag{83}$$

in which E_{20} is the voltage that would be measured if the network were absent. In the latter case, the appropriate two-port transformation must be the identity matrix, so that $A = D = 1$, $B = C = 0$. Hence

$$L = \frac{|E_g|^2}{|A + B + C + D|^2} \bigg/ \frac{|E_g|^2}{4} = \frac{4}{|A + B + C + D|^2}. \tag{84}$$

A transmission factor greater than one implies gain. The network transmission factor for a matched network is

$$L = 4|2A + 2B|^{-2} = |A + B|^{-2}. \tag{85}$$

A.8 *Network Reflectivity*

When a network is followed by a matched load, the input impedance to the network is given by $Z_{\text{input}} = (A + B)/(C + D)$. The power reflectivity, R, is given by

$$\begin{aligned} R = \left|\frac{(Z_{\text{input}} - 1)}{(Z_{\text{input}} + 1)}\right|^2 &= \frac{|A + B - C - D|^2}{|A + B + C + D|^2} \\ &= \tfrac{1}{4} L |A + B - C - D|^2 \end{aligned} \tag{86}$$

in which L is the transmission factor of the network. Note that for a matched network, $A = D$, $B = C$ and $R = 0$.

A.9 *Matched Attenuator*

An attenuator is a nonreactive reciprocal device. Thus, a matched attenuator must satisfy the conditions $A = D$, $B = C$ and $AD - BC = 1$, with A, B, C and D real. This is automatically satisfied by writing

$$\mathbf{T} = \begin{vmatrix} \cosh\theta & \sinh\theta \\ \sinh\theta & \cosh\theta \end{vmatrix}. \tag{87}$$

The parameter θ is referred to as the line length of the attenuator. The attenuator commutes with a matched reciprocal reactive network (a transmission line, for example) and the resultant network

$$\begin{aligned} &\begin{vmatrix} \cos\varphi & j\sin\varphi \\ j\sin\varphi & \cos\varphi \end{vmatrix} \begin{vmatrix} \cosh\theta & \sinh\theta \\ \sinh\theta & \cosh\theta \end{vmatrix} \\ &\qquad = \begin{vmatrix} \cos(\varphi - j\theta) & j\sin(\varphi - j\theta) \\ j\sin(\varphi - j\theta) & \cos(\varphi - j\theta) \end{vmatrix} \end{aligned} \tag{88}$$

is the representation for an attenuator with phase shift or angular length φ. From the fact that a lossy transmission line would have a phase shift of the form $\exp -j(\varphi - j\theta) = \exp -\theta \exp -j\varphi$, it would be expected that the transmission factor is given by $\exp -2\theta$. For an attenuator the transmission factor is obtained by using (85) and (87)

$$L = |\cosh \theta + \sinh \theta|^{-2} = \exp -2\theta. \tag{89}$$

Thus, the attenuator matrix is specified completely by knowledge of its transmission factor.

A.10 *Isolator*

An isolator is by definition nonreciprocal although it is matched. Thus, an isolator with forward transmission unity and reverse transmission $\exp -2\theta$ can be synthesized by cascading an attenuator of line length $\theta/2$, having a transmission $\exp -\theta$, with a nonreciprocal network with transmission $\exp +\theta$ in the forward direction and transmission $\exp -\theta$ in the backward direction. In its most simple form, the nonreciprocal network (70) is given by

$$\begin{vmatrix} \exp -\theta/2 & 0 \\ 0 & \mathrm{cxp} -\theta/2 \end{vmatrix}. \tag{90}$$

Thus, the isolator can be represented by

$$\begin{aligned} \mathbf{T} &= \begin{vmatrix} \exp -\theta/2 & 0 \\ 0 & \exp -\theta/2 \end{vmatrix} \begin{vmatrix} \cosh \theta/2 & \sinh \theta/2 \\ \sinh \theta/2 & \cosh \theta/2 \end{vmatrix} \\ &= \begin{vmatrix} \exp -\theta/2 \cosh \theta/2 & \exp -\theta/2 \sinh \theta/2 \\ \exp -\theta/2 \sinh \theta/2 & \exp -\theta/2 \cosh \theta/2 \end{vmatrix}. \end{aligned} \tag{91}$$

and the transmission factor is unity in the forward direction and $\exp -2\theta$ in the backward direction. An ideal isolator is one for which the line length θ approaches infinity, yielding

$$\mathbf{T} = \begin{vmatrix} \frac{1}{2} & \frac{1}{2} \\ \frac{1}{2} & \frac{1}{2} \end{vmatrix}. \tag{92}$$

APPENDIX B

Noisy Networks

In the following, the source of noise will be considered to be spontaneous fluctuations which arise because of the thermal properties of the material. At extremely high frequencies, the thermal noise is more commonly called black-body radiation. The noise spectrum will be characterized by the following statement of Nyquist's theorem:[13] The noise power *available* per mode at frequency ν in a small frequency interval, $d\nu$, is given by

$$\begin{aligned} dP &= p(\nu)d\nu \\ p(\nu) &= h\nu(\exp h\nu/kT - 1)^{-1} \end{aligned} \tag{93}$$

for a passive circuit at temperature T. The constant k is Boltzmann's constant (1.38×10^{-23} joules/degree) and h is Planck's constant (6.6×10^{-34} joules-sec).

It is sometimes convenient to write the noise power in terms of equivalent rms voltage and current generators e and i as shown in Fig. 11. The internal impedance of the voltage generator is zero and for the current generator it is infinite, and one can write

$$|e| = (4rp(\nu)d\nu)^{\frac{1}{2}}, \qquad |i| = (4gp(\nu)d\nu)^{\frac{1}{2}}. \tag{94}$$

In the following, both e and i will have the units of $(\text{power})^{\frac{1}{2}}$ since r and g are normalized with respect to the line impedance and admittance.

A systematic method of handling the noise produced by a network will be developed next. First, the following theorem will be stated without proof: *Any passive noisy two-port network can be replaced by an equivalent noise-free network which has an added shunt current generator and series voltage generator at the input or output terminals which represent the noise contribution.*[12] Therefore, any passive noisy network can be replaced by the representation shown in Fig. 2, in which the network is now noise-free but the noise appears from equivalent voltage and current generators at one of the terminals. A proof for the theorem can be given, and although it is relatively simple, the proof is lengthy.

Next, a scheme for representing in a simple way the additional current and voltage appearing at the terminal will be described. The following technique is due to H. Seidel.[15] It is clear that one may always write

$$\begin{aligned} E_1 &= AE_2 + BI_2 + e \\ I_1 &= CE_2 + DI_2 + i. \end{aligned} \tag{95}$$

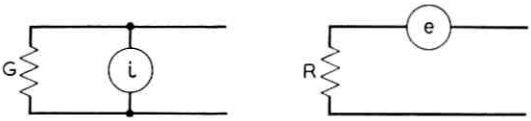


Fig. 11 — Noise-free representation of network with equivalent external voltage and current generators at one terminal.

The inclusion of the noise generators can be accomplished in an artificial but, as will be seen, highly useful way by writing

$$\begin{vmatrix} E_1 \\ I_1 \\ 1 \end{vmatrix} = \begin{vmatrix} A & B & e \\ C & D & i \\ 0 & 0 & 1 \end{vmatrix} \cdot \begin{vmatrix} E_2 \\ I_2 \\ 1 \end{vmatrix}. \tag{96}$$

This representation results from adding the trivial equation $1 = 0 + 0 + 1$ to the set in (95). No new information has been added, but the bookkeeping advantages afforded by this change will become apparent shortly.

Equation (96) will be the general representation for a passive noisy two-port network. The problem now is to learn how to characterize the noise quantities e and i in terms of the properties of the network represented by A, B, C and D. It will be seen that this can be done by comparing the network to some simple network whose properties are known.

First, it will be demonstrated that for any passive network, one may take A, B, C and D either all real or all imaginary. This is equivalent to saying that the network may always be taken to appear purely resistive. This is clear since the input impedance is $Z_{\text{in}} = (AZ_2 + B)$ $(CZ_2 + D)$. If Z_2 is real, then it would be expected that a resistive network will have Z_{in} real also. The proof follows from the fact that the input impedance can always be made real with a suitable length of transmission line. In addition, with an ideal transformer one may match the input impedance to the line. It follows, therefore, that for noise calculations, one only need consider matched networks ($A = D$, $B = C$) with the ratio A/B real.

First, consider a matched network which is cascaded with a transmission line of arbitrary length θ. The latter network has no loss and has no source of noise. In addition, the transmission line can change only the phase but not the magnitude of the noise current and voltage generators. Consequently, one expects that for a matched network

$$\begin{vmatrix} \cos\theta & j\sin\theta & 0 \\ j\sin\theta & \cos\theta & 0 \\ 0 & 0 & 1 \end{vmatrix} \begin{vmatrix} A & B & e \\ B & A & i \\ 0 & 0 & 1 \end{vmatrix} = \begin{vmatrix} X & X & e\cos\theta + ij\cos\theta \\ X & X & ej\sin\theta + i\cos\theta \\ 0 & 0 & 1 \end{vmatrix}$$

has the same noise properties independent of θ. Thus, it is required that

$$|e|^2 = |e\cos\theta + i\sin\theta|^2 = |e|^2\cos^2\theta + |i|^2\sin^2\theta$$
$$+ e(ji)^* + e^*(ji)\sin\theta\cos\theta$$

or

$$(|e|^2 - |i|^2)\sin^2\theta + j(ei^* - e^*i)\sin\theta\cos\theta = 0.$$

Choosing $\theta = \pi/2$ yields

$$|e|^2 = |i|^2 \tag{97}$$

and it follows also that

$$ei^* = e^*i. \tag{98}$$

Next, the passive matched network at temperature T will be cascaded with a shunt conductance at the same temperature and the open-circuit noise voltage at the input terminals will be determined. The network representation for the shunt conductance in the new formalism follows from Nyquist's theorem as is shown in Fig. 11

$$|i'| = (4Gp(\nu)d\nu)^{\frac{1}{2}}$$
$$\begin{vmatrix} A & B & e \\ C & D & i \\ 0 & 0 & 1 \end{vmatrix} = \begin{vmatrix} 1 & 0 & 0 \\ G & 1 & i' \\ 0 & 0 & 1 \end{vmatrix}. \tag{99}$$

Thus, the network to be studied is as shown in Fig. 12 with

$$\begin{vmatrix} E_1 \\ 0 \\ 1 \end{vmatrix} = \begin{vmatrix} A & B & e \\ B & A & i \\ 0 & 0 & 1 \end{vmatrix} \cdot \begin{vmatrix} 1 & 0 & 0 \\ G & 1 & i' \\ 0 & 0 & 1 \end{vmatrix} \cdot \begin{vmatrix} E_2 \\ 0 \\ 1 \end{vmatrix}$$
$$= \begin{vmatrix} A + BG & B & Bi' + e \\ B + AG & A & Ai' + i \\ 0 & 0 & 1 \end{vmatrix} \begin{vmatrix} E_2 \\ 0 \\ 1 \end{vmatrix} \tag{100}$$

yielding

$$\begin{aligned} E_1 &= (A + BG)E_2 + Bi' + e \\ 0 &= (B + AG)E_2 + Ai' + i \end{aligned} \tag{101}$$

so that

$$\begin{aligned} E_2 &= -\frac{(Ai' + i)}{(B + AG)} \\ E_1 &= -\frac{(A + BG)}{B + AG}(Ai' + i) + Bi' + e. \end{aligned} \tag{102}$$

At the open-circuited terminal at which the noise voltage E_1 is measured, one must have from Nyquist's theorem, (94)

$$|E_1|^2 = 4Z_{\text{input}}p(\nu)d\nu \tag{103}$$

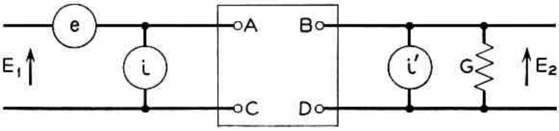


Fig. 12 — Passive matched network with shunt conductance for determination of open-circuit noise voltage.

since Z_{input} is resistive. Its value is

$$Z_{\text{input}} = \frac{A + BG}{B + AG}. \tag{104}$$

Hence, it is required that

$$\left(\frac{A + BG}{B + AG}\right) 4p(\nu)d\nu = \left| -\left(\frac{A + BG}{B + AG}\right)(Ai' + i) + Bi' + e \right|^2. \tag{105}$$

Remember also that $| e |^2 = | i |^2$ and ei^* is real. Since the noise arising from the shunt conductance G is independent of the noise produced by the network, cross terms like ei', i^*i' are zero, since their product represents a time average which must be zero. Solution of (105) yields uniquely the values[15]

$$\begin{aligned} | i |^2 = | e |^2 &= AB \\ ei^* &= \tfrac{1}{2}(A^2 + B^2 - 1) \end{aligned} \tag{106}$$

which are the desired relations measured in units of $4p(\nu)d\nu$. The veracity of (106) can be established by direct substitution into (105), which yields an identity independent of the value of G.

The equivalent values of e and i for an attenuator are given by

$$\begin{aligned} | e |^2 = | i |^2 &= AB = \cosh\theta \sinh\theta = \tfrac{1}{2}\sinh 2\theta \\ ei^* &= \tfrac{1}{2}(A^2 + B^2 - 1) = \tfrac{1}{2}(\cosh^2\theta + \sinh^2\theta - 1) = \sinh^2\theta. \end{aligned} \tag{107}$$

The phase angle between e and i^* is $\cos^{-1}\tanh\theta$. Since the phase angle is always positive, the implication is that the noise sources radiate more power toward the attenuator than away from it. This is reasonable, since the noise power leaving each end of the attenuator should be equal and the noise radiated toward the attenuator by the noise sources is attenuated before emerging from the output end. In the limit of large θ, the phase angle approaches zero.

The equivalent values for an isolator are given by

$$\begin{aligned} |e|^2 &= |i|^2 = \exp -\theta \cosh\frac{\theta}{2}\sinh\frac{\theta}{2} = \tfrac{1}{2}\exp -\theta\sinh\theta \\ ei^* &= \frac{1}{2}\left[\exp -\theta\left(\sinh^2\frac{\theta}{2} + \cosh^2\frac{\theta}{2}\right) - 1\right] = -|e|^2. \end{aligned} \qquad (108)$$

Therefore, the angle between e and i is always π, the implication being that the isolator radiates noise power away from the attenuator, i.e., only in the direction in which it attenuates.

The noise radiated into a matched resistor by any network is obtained by considering the network shown in Fig. 13

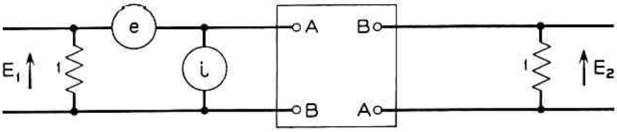


Fig. 13 — Noise radiated into matched resistor.

$$\begin{vmatrix} E_1 \\ 0 \\ 1 \end{vmatrix} = \begin{vmatrix} 1 & 0 & 0 \\ 1 & 1 & 0 \\ 0 & 0 & 1 \end{vmatrix} \begin{vmatrix} A & B & e \\ C & D & i \\ 0 & 0 & 1 \end{vmatrix} \begin{vmatrix} 1 & 0 & 0 \\ 1 & 1 & 0 \\ 0 & 0 & 1 \end{vmatrix} \begin{vmatrix} E_2 \\ 0 \\ 1 \end{vmatrix}$$

$$= \begin{vmatrix} A' & B' & e' \\ C' & D' & i' \\ 0 & 0 & 1 \end{vmatrix} \begin{vmatrix} E_2 \\ 0 \\ 1 \end{vmatrix}. \qquad (109)$$

The noise contributed by the matched input and output resistors is neglected since this is additive. The primed quantities result from matrix multiplication. The noise output dP is evaluated by noting that

$$0 = C'E_2 + i' \qquad \text{and} \qquad dP = |E_2|^2 = |i'|^2/|C'|^2.$$

Performing the matrix multiplication yields $C' = A + B + C + D$ and $i' = e + i$. Thus, the output noise power is given by

$$dP = \frac{|e + i|^2\, 4p(\nu)\, d\nu}{|A + B + C + D|^2} = L|e + i|^2\, p(\nu)\, d\nu \qquad (110)$$

in which L is the network insertion loss [see (20)]. For a matched resistive network (attenuator)

$$\begin{aligned} |e + i|^2 &= |e|^2 + |i|^2 + 2ei^* = (2AB + A^2 + B^2 - 1) \\ &= [(A + B)^2 - 1] = L^{-1} - 1 \end{aligned}$$

which follows from (9) and (21). Thus,

$$dP = (1 - L)p(\nu)d\nu \tag{111}$$

is the noise power in frequency range $d\nu$ emanating from a matched resistive network. This result is well known in network theory. In optics, it is known as one form of Kirchhoff's law.

REFERENCES

1. Slater, J. C., *Microwave Electronics*, New York, D. Van Nostrand, 1950.
2. Gordon, J. P., Zeiger, H. J., and Townes, C. H., Maser — New Type of Microwave Amplifier, Frequency Standard, and Spectrometer, Phys. Rev., **99,** August, 1955, p. 1264.
3. Wagner, W. G., and Birnbaum, G., Theory of Quantum Oscillators in Multimode Cavity, J. Appl. Phys., **32,** July, 1961, p. 1185.
4. Schawlow, A. L., and Townes, C. H., Infrared and Optical Masers, Phys. Rev., **112,** December, 1958, p. 1940, and Townes, C. H., *Advances in Quantum Electronics*, ed. J. R. Singer, Columbia University Press, 1961.
5. Shimoda, K., Theory of Masers for Higher Frequencies, Inst. Phys. and Chem. Res. (Tokyo) — Sci. Papers, **55,** March, 1961, p. 1.
6. Blaquiere, A., Largeur de raie d'un oscillateur Laser, considéré comme le siège d'une reaction en chaine, Acad. des Sciences — CR, **255,** December, 1962, p. 3141.
7. Fleck, J. A., Jr., Linewidth and Conditions for Steady Oscillation in Single and Multiple Element Lasers, J. Appl. Phys., **34,** October, 1963, p. 2997.
8. Davenport, W. B., Jr., and Root, W. L., *An Introduction to the Theory of Random Signals and Noise*, McGraw-Hill, 1958, p. 158.
9. Van der Zeil, A., *Noise*, Prentice-Hall, 1954, p. 283.
10. Guilleman, E. A., *Communication Networks*, Vol. II, John Wiley & Sons, Inc., 1935, Chap. 4.
11. Born, M., and Wolf, E., *Principles of Optics*, Pergamon Press, New York, 1959, p. 54.
12. Becking, A. G. T., Groendijk, H., and Knol, H. S., Noise Factor of Four-Terminal Networks, Philips Res. Repts., Vol. 10, No. 5, October, 1955, p. 349.
13. Van der Zeil, A., op. cit., p. 8.
14. Weissflock, A., Anwendung des Transformatorsatzes über verlustlose Vierpole auf die Hintereinanderscholtung von Vierpolen, Hochfrequenz and Elekakus, **61,** January, 1943, p. 19.
15. Seidel, H., Noise Properties of Four-Pole Networks with Application to Parametric Devices, IRE Trans., **CT-8,** December, 1961, p. 398.
16. Yariv, A., and Gordon, J. P., The Laser, Proc. IEEE, **51,** January, 1963, p. 4.
17. Born, M., and Wolf, E., op. cit., p. 324.
18. Peirce, B. O., and Foster, R. M., *A Short Table of Integrals*, 4th Ed., Ginn and Company, 1956.
19. Quist, T. M., Rediker, R. H., Keyes, R. J., Krag, W. E., Lax, B., McWhorter, A. L., and Zeigler, H. J., Semiconductor Maser of GaAs, Appl. Phys. Letters, **1,** December, 1962, p. 91.
20. White, A. D., Gordon, E. I., and Rigden, J. D., Output Power of the 6328-Å Gas Maser, Appl. Phys. Letters, **2,** March, 1963, p. 91.
21. Rigden, J. D., White, A. D., and Gordon, E. I., Visible HE-NE Maser and Some Developments, NEREM, TPM 14-4, November, 1962, p. 120.
22. Bridges, W. B., High Optical Gain at 3.5 μ in Pure Xenon, Appl. Phys. Letters, **3,** August, 1963, p. 45.

The 80 Diperiodic Groups in Three Dimensions

By ELIZABETH A. WOOD

(Manuscript received December 10, 1962)

The low-energy electron diffraction work of L. H. Germer, J. J. Lander, A. U. MacRae, J. Morrison and others is resulting in new information about surface structures. These three-dimensional structures have periodicity only in two dimensions. The 230 triperiodic space groups are not applicable to the solution of these structures. The 17 strictly two-dimensional groups do not admit the existence of a third dimension and may therefore not be appropriate for these structures which are not strictly planar. The useful space groups for these structures are the 80 diperiodic groups in three dimensions.

Nowhere in the literature have these been put into a form convenient for use, as have the other two sets of space groups. This has now been done and the tables are available on request from the Circulation Manager, Bell System Technical Journal, Bell Telephone Laboratories, Incorporated, 463 West Street, New York 14, N. Y. Sample tables are given in this paper.

I. BACKGROUND

Crystals grown under favorable conditions acquire an external shape whose symmetry has long attracted attention. Nineteenth century mineralogists systematically described the symmetry of these shapes in terms of *symmetry operations*. For example, the operation of *rotation* of a cube through 90° around an axis normal to a cube face brings the cube into a position indistinguishable from its original position. An operation that achieves this indistinguishability is called a symmetry operation. In this example the cube will present an identical appearance four times during a rotation of 360° around the axis, which is therefore called "an axis of 4-fold symmetry" or simply "a 4-fold axis." A cube has three 4-fold axes, four 3-fold axes (corner-to-corner) and six 2-fold axes (mid-edge-to-mid-edge) (Figs. 1a and b). Such axes are called *symmetry elements*. The terms "tetrad," "triad" and "diad" are also used for them.

Another type of symmetry element is a mirror plane, across which the operation of *reflection* produces an object indistinguishable from the original. Such a plane through the center of a cube parallel to two opposite faces reflects the left half into the right half and vice versa; that is, the two halves are mirror images of each other. Since there are also diagonal mirror planes in a cube there is a total of nine planes (Figs. 1c and d).

There is also a *center* of symmetry in the center of a cube which relates any feature located a given distance from it in one direction to an indistinguishable feature located the same distance away in the opposite direction. The operation is called *inversion.*

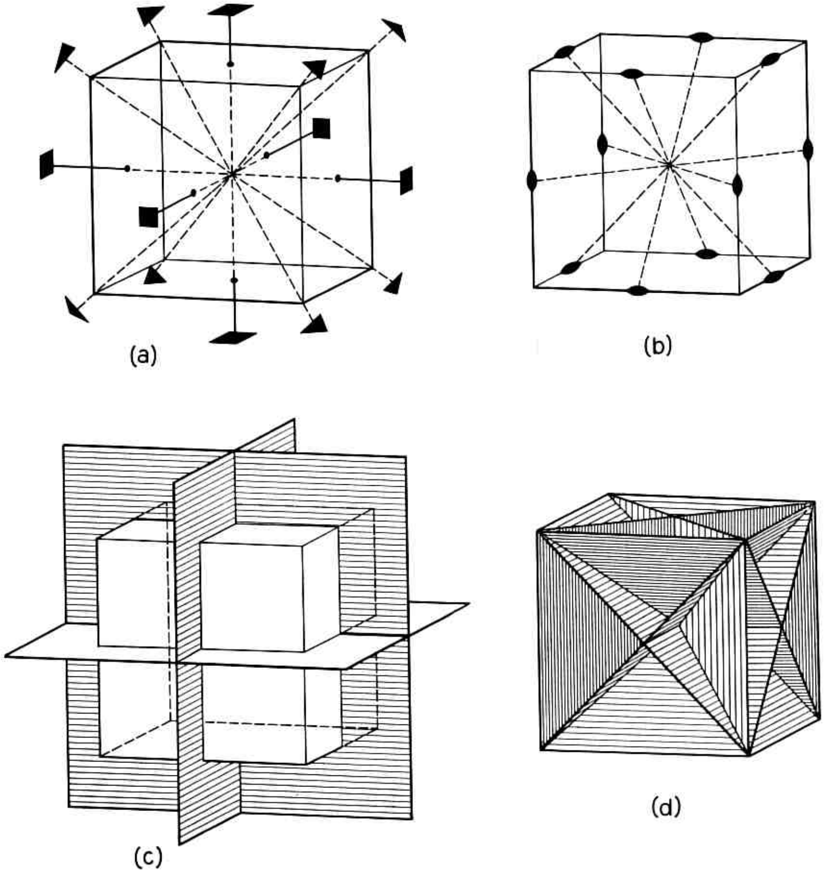


Fig. 1 — The symmetry elements of a cube: (a) the three 4-fold axes and the four 3-fold axes, (b) the six 2-fold axes, (c) three mirror planes parallel to the faces, and (d) six diagonal mirror planes.

An *inversion axis* combines the operation of rotation with that of inversion. The familiar regular tetrahedron which has neither a 4-fold axis nor a center of symmetry has a 4-fold inversion axis because after a rotation of 90° *plus* an inversion through the center point it occupies a position in space indistinguishable from its original position. A center of symmetry is equivalent to a one-fold inversion axis.

Note that during the entire group of "operations" on the cube, one point (the center of the cube) remains unmoved. Another way of saying this is to say that all of the symmetry elements pass through a single point. This group of operations or the symmetry elements which represent them therefore constitute the point group symmetry of the cube. When similar groups of operations are determined for all possible crystals, it is found that there are only 32 possible crystallographic point groups.

The symmetry of shape is the outward expression of the inner orderly atomic arrangement of the crystal. Any property of any piece of the crystal must obey the point group symmetry even though the piece be a ground sphere a few tenths of a millimeter in diameter.

When we consider in detail the *crystal structure* — that is, the positions of the atoms relative to each other — we find that the symmetry elements occur at well-defined positions in space and do not all go through the same point. This is readily illustrated by Fig. 2, the projection of the structure of calcite ($CaCO_3$) onto a plane normal to its 3-fold symmetry axis. Note that the 3-fold axis cannot be randomly placed, normal to the paper, but must pass through the black spots representing the carbon atoms, and further that there is a 3-fold axis through every carbon atom. There are also mirror planes in calcite. We could make a 3-dimensional model of the array of symmetry elements of calcite, and the operation of any symmetry element would shift every other symmetry element to a position indistinguishable from its original position.

Such a self-consistent array of symmetry elements in space is called a *space group*. Since location in space (not orientation alone) is of significance here, two other kinds of symmetry operations become meaningful: operations which combine *translation* with either rotation or reflection. The resulting symmetry elements are called, respectively, *screw axes* and *glide planes*.

As in the case of point groups, the space groups are limited in number. There are 230 possible space groups, i.e., 230 possible self-consistent arrangements in space of all the symmetry elements mentioned above.

Diagrams of these are given in the International Tables for X-ray Crystallography (edited by Henry and Lonsdale, 1952; see References).

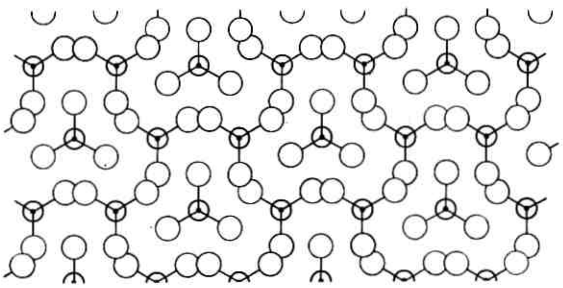


Fig. 2 — The structure of calcite projected onto a plane normal to its 3-fold symmetry axis.

The one appropriate to the structure of calcite is shown on the next page.* It is identified by the symbol $R\bar{3}c$ which states that the unit cell (the repeat unit of the structure) is rhombohedral in shape (R), that it has a 3-fold inversion axis ($\bar{3}$) with a glide plane parallel to it in which the translation is in the c direction.† Of course the 3-fold axis operating on this glide plane generates two more. Additional symmetry elements which are found to exist whenever the stated symmetry operations are performed are also shown in the calcite space group diagram.

The space group of a crystal can in many cases be uniquely determined directly from x-ray diffraction data. Since, in any given space group, the possible atom positions will be related in a well defined manner by the symmetry operations, a knowledge of the space group is a very powerful aid in determining the arrangement of atoms in the crystal, i.e., the crystal structure.

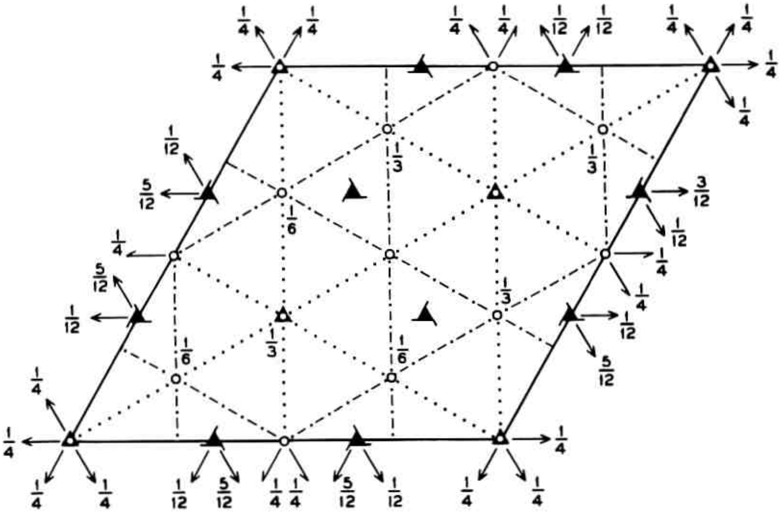
One could repetitiously extend the space-group symbols in the diagram, as we have the calcite structure in Fig. 2, by translation which would be in three dimensions if we were not limited to the printed page. (The translation vectors define the edges of the *unit cell*, the repeat unit of the three-dimensional structure.) The three-dimensional lattice of translation vectors which would represent this operation is called a *space lattice*. There are only 14 such lattices possible.

If we limit our attention strictly to two dimensions we find that, instead of 230 space groups, we have 17 *plane groups* and instead of 14 space lattices we have 5 *nets*. Here the periodicity no longer extends in three dimensions (triperiodicity) but only in two dimensions (diperiodicity).

* Fractions on the diagram refer to positions of symmetry elements along the c direction (normal to the paper). The unit is the unit length of c, i.e., the c dimension of the unit cell.

† For a very brief discussion of space group symbols see the *Crystallographic Data* section of the American Institute of Physics Handbook.

Trigonal $\bar{3}m$ $R\,\bar{3}\,2/c$ No. 167 $R\,\bar{3}\,c$ D_{3d}^{6}



Symbols of Symmetry Planes

Symbol	Symmetry plane	Graphical symbol	Nature of glide translation
		Normal to plane of projection	
c	Axial glide plane	···········	$c/2$ along z-axis; or $(a+b+c)/2$ along [111] on rhombohedral axes.
n	Diagonal glide plane (net)	–·–·–·–·–·–	$(a+b)/2$ or $(b+c)/2$ or $(c+a)/2$; or $(a+b+c)/2$ (tetragonal and cubic).

Symbols of Symmetry Axes

Symbol	Symmetry axis	Graphical symbol	Nature of right-handed screw translation along the axis	Symbol	Symmetry axis	Graphical symbol	Nature of right-handed screw translation along the axis
1	Rotation monad	None	None	2_1	Screw diad	⬮ (normal to paper)	$c/2$
$\bar{1}$	Inversion monad	○	None			→ (parallel to paper)	$a/2$ or $b/2$ or $(a+b)/2$
2	Rotation diad	⬮ (normal to paper)	None			Normal to paper	
		→ (parallel to paper)		3	Rotation triad	▲	None
				3_1	Screw triads	▲	$c/3$
				3_2		▲	$2c/3$
				$\bar{3}$	Inversion triad	△	None

It is with still a third set of groups, the 80 diperiodic groups in three dimensions, that the present paper is concerned.

II. DISCUSSION OF THE 80 DIPERIODIC GROUPS

The International Tables for X-ray Crystallography (1952) ("ITXRC") give two different sets of space groups: the familiar 230 triperiodic space groups and the 17 two-dimensional space groups in which all operations are confined strictly to two dimensions. In the latter set, any operation which admits the existence of the third dimension, such as a two-fold axis lying in the plane, is forbidden.

The existence of a set of groups which admit such operations, but still refer to arrays that are infinitely periodic in only two dimensions, was recognized by several authors at about the same time (Speiser, 1927; C. Hermann, 1928; Alexander and K. Herrmann, 1928; L. Weber, 1929; Alexander and K. Herrmann, 1929). These and subsequent authors (see references at end of this paper) have used a wide variety of nomenclature, some giving some diagrams. C. Hermann gives point positions, but in many cases chooses a different origin and in some cases a larger cell than that given in ITXRC. This work and others contain errors and omissions and none of the authors has given the groups in the form currently used in the International Tables so that they could be conveniently used for structure determination. This has now been done.

Consideration of the restrictions imposed by the loss of periodicity in the third dimension leads to the exclusion of the following symmetry elements: (*i*) screw axes normal to the plane of diperiodicity, (*ii*) glide planes with glide directions out of this plane, and (*iii*) n-fold axes not normal to this plane, with $n > 2$. Since the upper side of our diperiodic array may be like or unlike the lower side, mirror planes, glide planes, two-fold rotation and screw axes may lie in the plane.

It is possible to choose the 80 diperiodic groups in three dimensions from the pages of the existing International Tables for X-ray Crystallography by using some of the "1st setting" monoclinic groups and some of the "2nd setting" monoclinic groups as well as various orientations of the orthorhombic groups, without deletion or addition of any symmetry operations. In the diperiodic-group case we always have a unique direction in the plane-normal. Placing this direction along each of two nonequivalent directions in a single (orthorhombic) triperiodic space group gives us two nonequivalent diperiodic groups. This, of course, requires the appropriate permutation of point coordinates and indices of forbidden reflections.

Special positions of atoms with a fixed coordinate expressed as a fraction of the unit-cell length in the z-direction, other than zero, are not allowed since fractions of a period are meaningless in this nonperiodic direction.

The five nets (comparable to the 14 space lattices in three dimensions) for these diperiodic groups are the same as those for the 17 two-dimensional groups, namely, oblique ($a \neq b, \gamma \neq 90°$), primitive and centered rectangular ($a \neq b, \gamma = 90°$), square ($a = b, \gamma = 90°$), and hexagonal ($a = b, \gamma = 120°$), where γ is the angle between the a and b axes.

Alexander and Herrmann became interested in these groups because of their work with the smectic state in liquid crystals where only two-dimensional periodicity obtains. Cochran's interest in them grew out of his use of "generalized crystal-structure projections" (Cochran, 1952, b) and Holser's (1958, b) out of his investigation of the structure at the boundary between two parts of a twinned crystal (1958, a).

The interest of the writer in making these groups available in convenient form stems from cooperation with those members of Bell Laboratories who have been investigating surface structures by means of low-energy electron diffraction, in particular, L. H. Germer, J. J. Lander, A. U. MacRae and J. Morrison.* These structures are infinitely periodic in two dimensions but lack periodicity in the third dimension (normal to the surface).

Which set of diperiodic groups is appropriate for surface structures? Certainly the structures are not strictly planar: the atoms of the surface structure in many cases do not all lie in the same plane. But would an atom above some plane (parallel to the surface) be symmetrically related to an atom on the other side of the plane? Strictly speaking the atoms could not be symmetrically equivalent since one is closer to the substrate than the other and is therefore in a different force field. From this point of view one would say that only the seventeen strictly two-dimensional space groups would be useful. However, it is frequently so, in triperiodic crystallography, that the symmetry of a crystal structure closely approximates a symmetry that is higher than its true symmetry and that the use of the higher-symmetry space group is of great help in determining the structure. From this point of view one would say that the 80 diperiodic groups in three dimensions are likely to be useful in the solution of diperiodic surface structures. Their application to this field was suggested to the writer by A. L. Patterson.

There follow (*i*) a summary tab!e, Table I; (*ii*) a diagram of net types, Fig. 3; (*iii*) an explanation of terms and symbols used in the

* For a survey of some of this work, see Low-energy Electron Diffraction, by A. U. MacRae, Science, **139**, 1963, pp. 379–388.

Table I—Summary Table of the 80 Diperiodic Groups in Three Dimensions

Net	Diperiodic Group (DG) Number	Full Hermann-Mauguin Symbols	Triperiodic-Group Schoenflies Symbol, ITXRC Number and Orientation, if other than that given in ITXRC		Symbol Proposed by A. Niggli	Weber Number*
Oblique	1	$P1$	C_1^1 - 1		$1P1$	1
	2	$P\bar{1}$	C_i^1 - 2		$1P\bar{1}$	2
	3	$P211$	C_2^1 - 3	1st setting	$1P2$	8
	4	$Pm11$	C_s^1 - 6	1st setting	$mP1$	3
	5	$Pb11$	C_s^2 - 7	1st setting	$aP1$	4
	6	$P2/m\,11$	C_{2h}^1 - 10	1st setting	$mP2$	12
	7	$P2/b\,11$	C_{2h}^4 - 13	1st setting	$aP2$	13
Rectangular	8	$P112$	C_2^1 - 3	2nd setting	$1P12$	9
	9	$P112_1$	C_2^2 - 4	2nd setting	$1P12_1$	10
	10	$C112$	C_2^3 - 5	2nd setting	$1C12$	11
	11	$P11m$	C_s^1 - 6	2nd setting	$1P1m$	5
	12	$P11a$	C_s^2 - 7	2nd setting $\bar{c}ba$	$1P1g$	6
	13	$C11m$	C_s^3 - 8	2nd setting	$1C1m$	7
	14	$P11\,2/m$	C_{2h}^1 - 10	2nd setting	$1P12/m$	14
	15	$P11\,2_1/m$	C_{2h}^2 - 11	2nd setting	$1P12_1/m$	15
	16	$C11\,2/m$	C_{2h}^3 - 12	2nd setting	$1C12/m$	16
	17	$P11\,2/a$	C_{2h}^4 - 13	2nd setting $\bar{c}ba$	$1P12/g$	18
	18	$P11\,2_1/a$	C_{2h}^5 - 14	2nd setting $\bar{c}ba$	$1P12_1/g$	17
	19	$P222$	D_2^1 - 16		$1P222$	33
	20	$P222_1$	D_2^2 - 17	bca	$1P222_1$	34
	21	$P22_12_1$	D_2^3 - 18		$1P22_12_1$	35
	22	$C222$	D_2^6 - 21		$1C222$	36
	23	$P2mm$	C_{2v}^1 - 25		$1P2mm$	19
	24	$Pmm2$	C_{2v}^1 - 25	bca	$mP12m$	23
	25	$Pm2_1a$	C_{2v}^2 - 26	$\bar{c}ba$	$mP12_1g$	24
	26	$Pbm2_1$	C_{2v}^2 - 26	$a\bar{c}b$	$aP12_1m$	25
	27	$Pbb2$	C_{2v}^3 - 27	$a\bar{c}b$	$aP12g$	26
	28	$P2ma$	C_{2v}^4 - 28		$1P2mg$	20
	29	$Pam2$	C_{2v}^4 - 28	$a\bar{c}b$	$bP12m$	27
	30	$Pab2_1$	C_{2v}^5 - 29	$a\bar{c}b$	$bP12_1g$	28
	31	$Pnb2$	C_{2v}^6 - 30	bca	$nP12g$	29
	32	$Pnm2_1$	C_{2v}^7 - 31	$a\bar{c}b$	$nP12_1m$	30
	33	$P2ba$	C_{2v}^8 - 32		$1P2gg$	21
	34	$C2mm$	C_{2v}^{11} - 35		$1C2mm$	22
	35	$Cmm2$	C_{2v}^{14} - 38	bca	$mC12m$	31
	36	$Cam2$	C_{2v}^{15} - 39	bca	$aC12m$	32
	37	$P2/m\,2/m\,2/m$	D_{2h}^1 - 47		$mP2mm$	37
	38	$P2/a\,2/m\,2/a$	D_{2h}^3 - 49	cab	$aP2mg$	38
	39	$P2/n\,2/b\,2/a$	D_{2h}^4 - 50		$nP2gg$	39
	40	$P2/m\,2_1/m\,2/a$	D_{2h}^5 - 51	$a\bar{c}b$	$mP2mg$	40
	41	$P2/a\,2_1/m\,2/m$	D_{2h}^5 - 51		$aP2mm$	41
	42	$P2/n\,2/m\,2_1/a$	D_{2h}^7 - 53	$a\bar{c}b$	$nP2mg$	42
	43	$P2/a\,2/b\,2_1/a$	D_{2h}^8 - 54	cab	$aP2gg$	43
	44	$P2/m\,2_1/b\,2_1/a$	D_{2h}^9 - 55		$mP2gg$	44
	45	$P2/a\,2_1/b\,2_1/m$	D_{2h}^{11} - 57	bca	$aP2gm$	45
	46	$P2/n\,2_1/m\,2_1/m$	D_{2h}^{13} - 59		$nP2mm$	46
	47	$C2/m\,2/m\,2/m$	D_{2h}^{19} - 65		$mC2mm$	47
	48	$C2/a\,2/m\,2/m$	D_{2h}^{21} - 67		$aC2mm$	48
Square	49	$P4$	C_4^1 - 75		$1P4$	58
	50	$P\bar{4}$	S_4^1 - 81		$1P\bar{4}$	57
	51	$P4/m$	C_{4h}^1 - 83		$mP4$	61
	52	$P4/n$	C_{4h}^3 - 85		$nP4$	62
	53	$P422$	D_4^1 - 89		$1P422$	67
	54	$P42_12$	D_4^2 - 90		$1P42_12$	68
	55	$P4mm$	C_{4v}^1 - 99		$1P4mm$	59

Table I—Continued

Net	Diperiodic Group (DG) Number	Full Hermann-Mauguin Symbols	Triperiodic-Group Schoenflies Symbol ITXRC Number and Orientation, if other than that given in ITXRC	Symbol Proposed by A. Niggli	Weber Number*
Square (cont.)	56	$P4bm$	C_{4v}^{2} - 100	$1P4gm$	60
	57	$P\bar{4}2m$	D_{2d}^{1} - 111	$1P\bar{4}2m$	63
	58	$P\bar{4}2_1m$	D_{2d}^{3} - 113	$1P\bar{4}2_1m$	64
	59	$P\bar{4}m2$	D_{2d}^{5} - 115	$1P\bar{4}m2$	65
	60	$P\bar{4}b2$	D_{2d}^{7} - 117	$1P\bar{4}g2$	66
	61	$P4/m\,2/m\,2/m$	D_{4h}^{1} - 123	$mP4mm$	69
	62	$P4/n\,2/b\,2/m$	D_{4h}^{3} - 125	$nP4gm$	70
	63	$P4/m\,2_1/b\,2/m$	D_{4h}^{5} - 127	$mP4gm$	71
	64	$P4/n\,2_1/m\,2/m$	D_{4h}^{7} - 129	$nP4mm$	72
Hexagonal	65	$P3$	$C_3{}^1$ - 143	$1P3$	49
	66	$P\bar{3}$	C_{3i}^{1} - 147	$1P\bar{3}$	50
	67	$P312$	$D_3{}^1$ - 149	$1P312$	54
	68	$P321$	$D_3{}^2$ - 150	$1P321$	53
	69	$P3m1$	C_{3v}^{1} - 156	$1P3m1$	51
	70	$P31m$	C_{3v}^{2} - 157	$1P31m$	52
	71	$P\bar{3}1\,2/m$	D_{3d}^{1} - 162	$1P\bar{3}1m$	55
	72	$P\bar{3}\,2/m\,1$	D_{3d}^{3} - 164	$1P\bar{3}m1$	56
	73	$P6$	$C_6{}^1$ - 168	$1P6$	76
	74	$P\bar{6}$	C_{3h}^{1} - 174	$mP3$	73
	75	$P6/m$	C_{6h}^{1} - 175	$mP6$	78
	76	$P622$	$D_6{}^1$ - 177	$1P622$	79
	77	$P6mm$	C_{6v}^{1} - 183	$1P6mm$	77
	78	$P\bar{6}m2$	D_{3h}^{1} - 187	$mP3m2$	74
	79	$P\bar{6}2m$	D_{3h}^{3} - 189	$mP32m$	75
	80	$P6/m\,2/m\,2/m$	D_{6h}^{1} - 191	$mP6mm$	80

* Useful for cross-comparison of this list with those of Weber (1929), C. Hermann (1928) and Alexander and Herrmann (1929) since the equivalence among these three is given in the last reference.

tables, Table II; (*iv*) samples of the systematic ITXRC "tables" for the 80 diperiodic groups in three dimensions, adapted from the three-dimensional space groups by making the appropriate modifications; and (*v*) an annotated list of references.

The full set of 80 diperiodic groups in three dimensions has been bound separately and is available from the Circulation Manager, Bell System Technical Journal, Bell Telephone Laboratories, Incorporated, 463 West Street, New York 14, N. Y. It is anticipated that these groups will be included in a later volume of the International Tables.

III. TABLES OF THE 80 DIPERIODIC GROUPS IN THREE DIMENSIONS

In the sample tables, the usage and notation of the International Tables for X-ray Crystallography for the three-dimensional space groups have been followed as closely as possible. Chosen directly from the

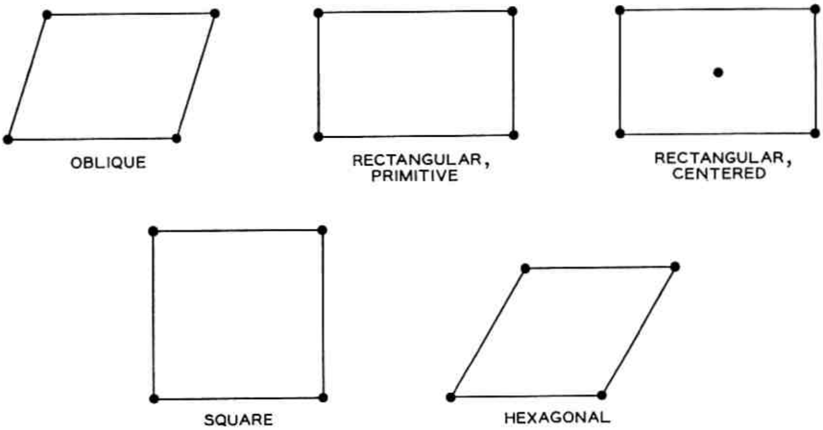


Fig. 3 — The five nets.

ITXRC Tables for the 230 space groups, these tables carry the same atom-position lettering. Letters of forbidden positions will therefore be missing.

In the oblique and rectangular system, the ITXRC convention of listing the symmetry symbols in the order *a*, *b*, *c* has not been retained. Holser (1958,b) chose to permute these so that the first symbol referred to the *c* axis. The justification for this is that in the plane groups the *c* axis is unique and therefore should be put first as in, for example, the tetragonal system (e.g., 4*mm*).

The possibility of confusion with the 230 three-dimensional groups will probably be avoided in all cases by the context. However, to aid in the distinction, the plane groups have been numbered, DG1, DG2, etc. The same letters could be used to distinguish DG*Pmm*2, for example, from the three-dimensional C_{2v}^1 - *Pmm*2, but since, in all cases, the two groups do in fact comprise the same symmetry operations, such a distinction may be undesirable.

The order of the DG list is that of the ITXRC which, in turn, is the Schoenflies order.

After this paper was in galley form a communication was received from A. Niggli to whom a manuscript copy had been sent. Niggli favors placing before the lattice symbol (*P* or *C*) that symbol referring to the glide plane or mirror plane which lies in the plane of diperiodicity and therefore occurs only once. This occurs in 37 of the 80 groups. This would be another way of distinguishing these groups from the triperiodic groups. The symbol proposed by Niggli is also listed in Table I.

TABLE II—SYMBOLS USED IN THE 80 DG TABLES

Symmetry Elements	Diperiodic Group Symbol	Symbol in the Symmetry Diagram	
		Normal to the Paper	Parallel to the Paper
nirror	m	— | /	┐
lide plane*	a	¦	↴
	b	---	↰
	n	not allowed	⬈
-fold rotation axis	2	⬮	← ↑
-, 4- and 6-fold rotation axes	3, 4, 6	▲ ◆ ⬢	not allowed
-fold screw axis	2_1	not allowed	↼ ↾
enter of symmetry	$\bar{1}$	○	○
-, 4- and 6-fold inversion axes†	$\bar{3}, \bar{4}, \bar{6}$	△ ◈ ⬣	not allowed
Center, on 2-fold axis		⬮	←o– ↕

* This operation combines reflection with translation of ½ the length of the cell in the lirection indicated by the letter. The diagonal glide, n, combines reflection with translaion of ½ of the length of the cell in both the a and b directions.

† Combined rotation through 360°/n (for $\bar{n}$) and inversion. Not equivalent to the two perations performed separately.

IV. EXPLANATION OF TERMS AND SYMBOLS USED ON THE 80 DG SHEETS (These are the same as those used in the ITXRC)

1. *Top of sheet, left to right:* Net-type, full Hermann-Mauguin diperiodic group symbol, diperiodic-group (DG) number. The Hermann-Mauguin symbol begins with a letter which indicates whether the net is primitive or centered and is followed by symbols for symmetry elements that relate to the c, a and b axis, in turn. The c axis is normal to the paper in the diagrams, the a axis is directed toward the bottom of the page, and the b axis is directed toward the right. In DG 46 ($P\ 2/n\ 2_1/m\ 2_1/m$), for example, the lattice is primitive, there is a two-fold axis parallel to c with a diagonal-glide plane normal to c, a two-fold screw axis parallel to a with a mirror plane normal to a, and a two-fold screw axis parallel to b with a mirror plane normal to b. In DG 16 ($C\ 11\ 2/m$) we have a centered net with a two-fold axis parallel to b and a mirror plane normal to b.

2. *Diagrams:* On the right, the distribution of the symmetry elements in the unit mesh. On the left, the distribution in the unit mesh of the points in the "general position" (x, y, z and points symmetrically

equivalent to it). Here, the value of x is taken, arbitrarily, to be a very small distance and y, a slightly larger distance, except in the oblique groups where the reverse choice has been made. The sign of z is indicated beside the "point" (small circle). In both diagrams, the $+x$ direction (a) is down the page, $+y$ (b) toward the right. A comma within the circle indicates that that point is of opposite handedness to the points without commas, as when derived from these by mirror plane or inversion operation. Where two points are related by a mirror lying in the plane of the paper, half of the circular symbol is marked with a comma, half left blank.

Below the diagrams are the lists of all possible points in this diperiodic group and equivalent point positions.

First column: Number of positions that are symmetrically equivalent, given the first position in the series.

Second column: Arbitrary identifying letter, conventionally the same as that first used by Wyckoff for this position.

Third column: The symmetry of each point in the group (if each point lies on a two-fold axis, this will be "2"; if each point lies in a mirror plane this will be "m"; etc.). This will always be "1" for the "general position" which, by definition, is the position of a point not lying on any symmetry element.

Fourth column: Coordinates of equivalent positions. Note that not every group has "special positions." Special positions occur when a particular value of x, y, or z results in a reduction of the number of equivalent positions due to symmetry.

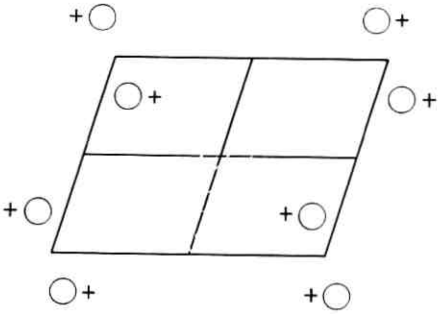
Fifth column: Conditions on hk which must be satisfied, for x-ray reflection to occur when the point positions in column 4 are occupied.

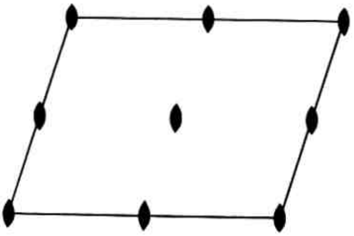
(References on page 559)

Oblique

P 2 1 1

DG3





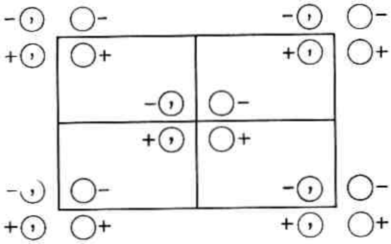
Number of positions, Wyckoff notation, and point symmetry			Co-ordinates of equivalent positions
2	*e*	1	$x,y,z;\quad \bar{x},\bar{y},z.$
1	*d*	2	$\frac{1}{2},\frac{1}{2},z.$
1	*c*	2	$\frac{1}{2},0,z.$
1	*b*	2	$0,\frac{1}{2},z.$
1	*a*	2	$0,0,z.$

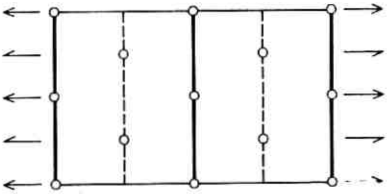
Conditions limiting possible reflections

General:

$\left.\begin{array}{l} hk: \\ h0: \\ 0k: \end{array}\right\}$ No conditions

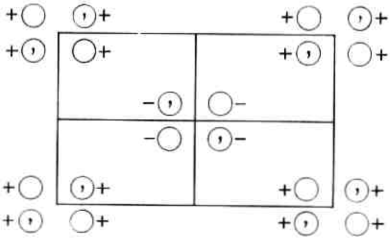
Special:
No conditions

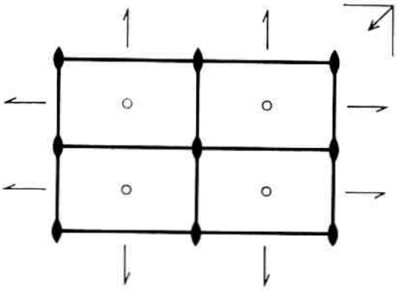




Number of positions, Wyckoff notation, and point symmetry			Co-ordinates of equivalent positions				Conditions limiting possible reflections
			$(0,0,0;\ \frac{1}{2},\frac{1}{2},0)+$				General:
8	j	1	$x,y,z;$	$x,\bar{y},z;$	$\bar{x},y,\bar{z};$	$\bar{x},\bar{y},\bar{z}.$	hk: $h+k=2n$ $h0$: $(h=2n)$ $0k$: $(k=2n)$
							Special: as above, plus
4	i	m	$x,0,z;$	$\bar{x},0,\bar{z}.$			No extra conditions
4	g	2	$0,y,0;$	$0,\bar{y},0.$			
4	e	$\bar{1}$	$\frac{1}{4},\frac{1}{4},0;$	$\frac{1}{4},\frac{3}{4},0.$			hk: $h=2n$; $(k=2n)$
2	b	$2/m$	$0,\frac{1}{2},0.$				No extra conditions
2	a	$2/m$	$0,0,0.$				

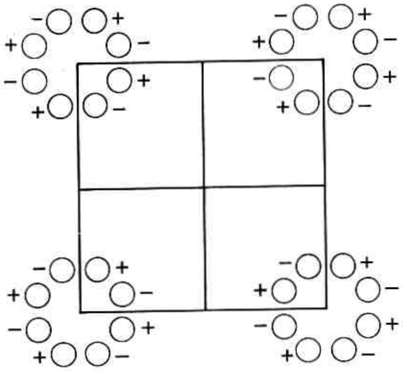
Rectangular $P\,2/n\,2_1/m\,2_1/m$ DG46

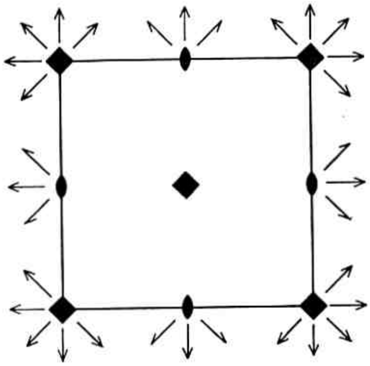




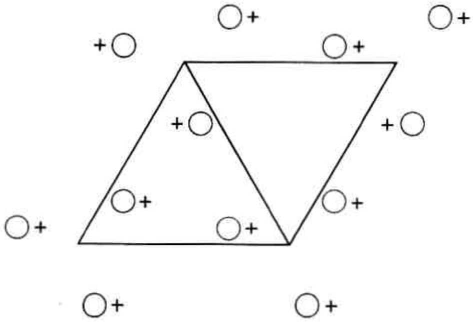
Number of positions, Wyckoff notation, and point symmetry			Co-ordinates of equivalent positions	Conditions limiting possible reflections
8	g	1	$x,y,z;\ \bar{x},\bar{y},z;\ \tfrac{1}{2}-x,\tfrac{1}{2}-y,\bar{z};\ \tfrac{1}{2}-x,\tfrac{1}{2}+y,\bar{z};$ $\bar{x},y,z;\ x,\bar{y},z;\ \tfrac{1}{2}+x,\tfrac{1}{2}+y,\bar{z};\ \tfrac{1}{2}+x,\tfrac{1}{2}-y,\bar{z}.$	General: hk: $h+k=2n$ $h0$: $(h=2n)$ $0k$: $(k=2n)$ Special: as above, plus
4	f	m	$x,0,z;\ \bar{x},0,z;\ \tfrac{1}{2}-x,\tfrac{1}{2},\bar{z};\ \tfrac{1}{2}+x,\tfrac{1}{2},\bar{z}.$	no extra conditions
4	e	m	$0,y,z;\ 0,\bar{y},z;\ \tfrac{1}{2},\tfrac{1}{2}-y,\bar{z};\ \tfrac{1}{2},\tfrac{1}{2}+y,\bar{z}.$	
4	c		$\tfrac{1}{4},\tfrac{1}{4},0;\ \tfrac{3}{4},\tfrac{3}{4},0;\ \tfrac{1}{4},\tfrac{3}{4},0;\ \tfrac{3}{4},\tfrac{1}{4},0.$	hkl: $h=2n$; $k=2n$
2	b	mm	$0,\tfrac{1}{2},z;\ \tfrac{1}{2},0,\bar{z}.$	no extra conditions
2	a	mm	$0,0,z;\ \tfrac{1}{2},\tfrac{1}{2},\bar{z}.$	

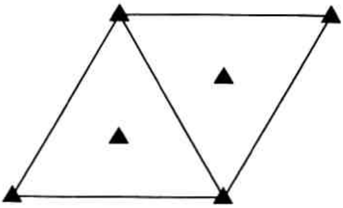
$P\,422$





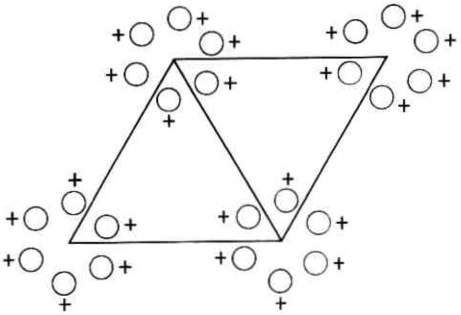
Number of positions, Wyckoff notation, and point symmetry			Co-ordinates of equivalent positions				Conditions limiting possible reflections
							General:
8	p	1	$x,y,z;$ $\bar{y},\bar{x},\bar{z};$	$\bar{x},\bar{y},z;$ $y,x,\bar{z};$	$\bar{x},y,\bar{z};$ $y,\bar{x},z;$	$x,\bar{y},\bar{z};$ $\bar{y},x,z.$	No conditions
							Special:
4	o	2	$x,\frac{1}{2},0;$	$\bar{x},\frac{1}{2},0;$	$\frac{1}{2},x,0;$	$\frac{1}{2},\bar{x},0.$	No conditions
4	l	2	$x,0,0;$	$\bar{x},0,0;$	$0,x,0;$	$0,\bar{x},0.$	
4	j	2	$x,x,0;$	$\bar{x},\bar{x},0;$	$\bar{x},x,0;$	$x,\bar{x},0.$	
4	i	2	$0,\frac{1}{2},z;$	$0,\frac{1}{2},\bar{z};$	$\frac{1}{2},0,z;$	$\frac{1}{2},0,\bar{z}.$	hk: $h+k=2n$
2	h	4	$\frac{1}{2},\frac{1}{2},z;$	$\frac{1}{2},\frac{1}{2},\bar{z}.$			No conditions
2	g	4	$0,0,z;$	$0,0,\bar{z}.$			
2	e	222	$\frac{1}{2},0,0;$	$0,\frac{1}{2},0.$			hk: $h+k=2n$
1	c	42	$\frac{1}{2},\frac{1}{2},0.$				No conditions
1	a	42	$0,0,0.$				

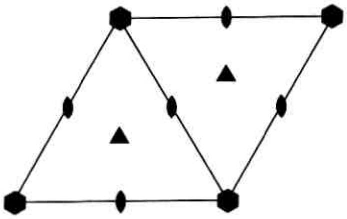




Number of positions, Wyckoff notation, and point symmetry			Co-ordinates of equivalent positions	Conditions limiting possible reflections
				General:
3	*d*	1	x,y,z; $\bar{y},x-y,z$; $y-x,\bar{x},z$.	No conditions
1	*c*	3	$\frac{2}{3},\frac{1}{3},z$.	
1	*b*	3	$\frac{1}{3},\frac{2}{3},z$.	Special:
1	*a*	3	$0,0,z$.	No conditions

$P\,6$





Number of positions, Wyckoff notation, and point symmetry			Co-ordinates of equivalent positions	Conditions limiting possible reflections
6	d	1	$x,y,z;\quad \bar{y},x-y,z;\quad y-x,\bar{x},z;$ $\bar{x},\bar{y},z;\quad y,y-x,z;\quad x-y,x,z.$	General: No conditions
3	c	2	$\frac{1}{2},0,z;\quad 0,\frac{1}{2},z;\quad \frac{1}{2},\frac{1}{2},z.$	
2	b	3	$\frac{1}{3},\frac{2}{3},z;\quad \frac{2}{3},\frac{1}{3},z.$	Special: No conditions
1	a	6	$0,0,z.$	

REFERENCES

The 80 diperiodic groups in three dimensions (in chronological order)

A. Speiser (1927), *Die Theorie der Gruppen von endlicher Ordnung*, 2nd edition, Berlin, Springer.

C. Hermann (1928), Zur systematischen Struktur theorie, III Ketten- und Netzgruppen, Z. f. Krist., **69**, p. 250. Lists point positions.

Ernst Alexander and Karl Herrmann (1928), Zur Theorie der flüssigen Kristalle, Z. f. Krist., **69**, p. 285.

C. Hermann (1928), Zur systematischen Struktur theorie, IV Untergruppen, Z. f. Krist., **69**, p. 533. Subgroups.

L. Weber (1929), Die Symmetrie homogener ebener Punktsysteme, Z. f. Krist., **70**, p. 309. Describes 80 plane groups.

Ernst Alexander and Karl Herrmann (1929), Die 80 zweidimensionalen Raumgruppen, Z. f. Krist., **70**, p. 328. Ibid, p. 460: Berichtung zu unserer Arbeit "Die 80 zweidimensionalen Raumgruppen." Comparative list of Weber, C. Hermann and A. and H. designations. Refers to H. Mark, Die Verwendung der Roentgenstrahlen for most point positions. Caution: Schoenflies symbols in this paper do not correspond with the Schoenflies symbols in ITXRC for many groups.

W. Cochran (1952, a), The Symmetry of Real Periodic Two-Dimensional Functions, Acta Cryst., **5**, p. 630. (Lists "46 reversal groups in two dimensions" using a *pgm* notation. Notes that Prof. Lonsdale has pointed out to him that these groups could be described by means of the symbols for the usual 230 space groups.)

W. Cochran and H. B. Dyer (1952, b), Some Practical Applications of Generalized Crystal-Structure Projections, Acta Cryst., **5**, p. 634.

N. F. M. Henry and Kathleen Lonsdale (1952), . . . Plane Groups in Three Dimensions, p. 56. International Tables for X-ray Crystallography, Vol. I, Kynoch Press. Calls attention to the existence of the 80 diperiodic groups in three dimensions.

K. Dornberger-Schiff (1956), On Order-Disorder Structures (OD-Structures), with an Appendix, "Proposal for International Symbols for the 80 Plane Groups in Three Dimensions," Acta Cryst., **9**, p. 593. (Compare Dornberger-Schiff, 1957, below.)

N. V. Belov and T. N. Tarkhova (1956), Groups with Color Symmetry, Kristallografiya, **1**, p. 4. Lists 46 groups, using color-group nomenclature, colored illustrations.

K. Dornberger-Schiff (1957), Zur OD-Strukture (Order-Disorder Structure) des Purpurogallin, Acta Cryst., **10**, p. 271. Gives table like the Tables 4.3 in the International Tables for the "Space Groupoid $P\{2_{\pm\frac{1}{2}}/b\}(\{2_1/n_{1,\pm\frac{1}{2}}\})\langle 2_1\rangle/a$."

W. T. Holser (1958, a), The Relation of Structure to Symmetry in Twinning, Z. Krist., **110**, p. 249.

W. T. Holser (1958, b), Point Groups and Plane Groups in a Two-Sided Plane and Their Subgroups, Z. f. Krist., **110**, p. 266. Lists ITXRC-type symbols, subgroups. Caution: Holser's first symbol always refers to the *c* axis.

N. V. Belov (1959), On the Nomenclature of the 80 Plane Groups in Three Dimensions, Translation in Soviet Physics: Crystallography **4**, p. 730. Original: Kristallografiya, **4**, p. 775. List of "Rational," Cochran, and Fedorov symbols. Groups arranged according to whether they are "two-color," "one-color" or "grey."

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 recently conducted theoretical studies of data transmission systems by computer simulation. Member, A.A.A.S., Sigma Xi, and Tau Beta Pi.

SIDNEY DARLINGTON, B.S., 1928, Harvard College; B.S. in E.E., 1929, Massachusetts Institute of Technology; Ph.D., 1940, Columbia University; Bell Telephone Laboratories, 1929—. He has been engaged in research in applied mathematics with emphasis on network theory and military and space electronics. He holds more than 20 patents in these fields. Fellow, IEEE; member, AIAA.

ROBERT W. DEGRASSE, B.S., 1951, California Institute of Technology; M.S., 1954, and Ph.D., 1958, Stanford University; Bell Telephone Laboratories, 1957–1960; Microwave Electronics Corp., 1960—. Mr. DeGrasse's work at Bell Laboratories was in research and development of solid state masers. He took part in the development of the ruby maser used in the Bell Laboratories receiving system for the Project Echo satellite communication experiments. Member, IEEE and Sigma Xi.

F.E. FROEHLICH, B.S., 1950, M.S., 1952, Ph.D., 1955, Syracuse University; Bell Telephone Laboratories, 1954—. Upon joining Bell Laboratories, Mr. Froehlich worked with magnetic core memory and magnetic core switching devices and circuits. In 1956 he became engaged in exploratory development of DATA-PHONE systems and subsequently had charge of groups conducting research and development in in the field of digital communications. He is now head of the high-speed data terminals department and is concerned with data transmission over wideband channels, digital coding and error control systems, and maintenance equipment for data services. Senior member, IEEE; member of Data and Telegraph Communication Committee and Communication Theory Committee; Chairman of the Monmouth County, N. J. subsection of the PTGCS. Member, American Physical Society, Sigma Xi, Phi Beta Kappa, Sigma Pi Sigma and Pi Mu Epsilon.

EUGENE I. GORDON, B.S., 1952, City College of New York; Ph.D., 1957, Massachusetts Institute of Technology; Bell Telephone Laboratories, 1957—. A member of the electron device laboratory, he is engaged in research in optical masers and optical modulation techniques. Member, American Physical Society, Phi Beta Kappa, Sigma Xi and IEEE.

STEPHEN E. HARRIS, B.E.E., 1959, Rensselaer Polytechnic Institute; M.S.E.E., 1961, Stanford University; Bell Telephone Laboratories, 1959—. Mr. Harris was engaged in work on microwave noise generation and later in work on the development of traveling-wave masers. Since September, 1960, he has been on leave of absence from Bell Laboratories to pursue doctoral studies at Stanford University. He is also presently serving as an acting professor of electrical engineering at Stanford. Member, Tau Beta Pi, Sigma Xi, Eta Kappa Nu, American Physical Society, Optical Society of America and IEEE.

JESSIE MACWILLIAMS, B.A., 1939, M.A., 1941, Cambridge (England); Ph.D., 1962, Harvard; Bell Telephone Laboratories 1956—. Mrs. MacWilliams has been concerned with writing computer programs for the analysis and synthesis of transmission networks. She is now engaged in data systems studies, particularly the study of algorithms for decoding systematic error-correcting codes. Member, Mathematical Association of America and American Mathematical Society.

E. O. SCHULZ-DUBOIS, Dipl. phys., 1950, and Dr. Phil. nat., 1954, Johann Wolfgang Goethe University (Germany); Purdue University, 1954–1955; Raytheon Manufacturing Co., 1956–1957; Bell Telephone Laboratories, 1957—. At Purdue Mr. Schulz-DuBois was engaged in paramagnetic resonance studies of irradiated semiconductors. At Raytheon he was concerned with the development of ferrite materials and devices. After joining Bell Laboratories his work was with paramagnetic materials, slow-wave structures, and ferrimagnetic isolators for application to solid state maser devices. More recently he was responsible for a group engaged in advanced development of traveling-wave masers and in related exploratory studies. Since September 1963 he has been on sabbatical leave as visiting professor at Technische Hochschule, Karlsruhe (Germany).

Erling D. Sunde, Dipl. Ing., 1926, Technische Hochschule, Darmstadt, Germany; American Telephone and Telegraph Co., 1927–1934; Bell Telephone Laboratories, 1934—. He has made theoretical and experimental studies of inductive interference from railway and power systems, lightning protection of the telephone plant, and fundamental transmission studies in connection with the use of pulse modulation systems. He is the author of *Earth Conduction Effects in Transmission Systems*, a Bell Laboratories Series book. Fellow, IEEE; member, A.A.A.S. and American Mathematical Society.

Elizabeth A. Wood, B.A., 1933, Barnard College; M.A. 1934 and Ph.D., 1939, Bryn Mawr College; D.Sc., 1963, Wheaton College; Bell Telephone Laboratories, 1943—. Mrs. Wood's first work at Bell Laboratories had to do with the development of techniques for producing quartz oscillator plates. Since then she has been using X-ray diffraction and optical methods for studying a wide variety of crystals, most of them of interest because of being ferroelectric. She is the author of *Crystal Orientation Manual* (Columbia University Press, 1963) and *Crystals and Light* (D. Van Nostrand, 1963). Member and past President (1957), American Crystallographic Association; Fellow, American Physical Society and Mineralogical Society of America.