TRANSMISSION SYSTEMS FOR COMMUNICATIONS

BELL TELEPHONE LABORATORIES

Transmission Systems for Communications

BELL TELEPHONE LABORATORIES

Transmission Systems for Communications

Transmission Systems for Communications

Revised
Fourth Edition

Members of the Technical Staff
Bell Telephone Laboratories

Bell Telephone Laboratories, Incorporated

Prepared for Publication by
Western Electric Company, Inc.
Technical Publications
Winston-Salem, North Carolina

First Edition, 1954
Printed as Informal Notes

Second Edition, 1958

Third Edition, January 1964
Two Printings

Fourth Edition, February 1970
Revised Fourth Edition, December 1971

Printed in the United States of America

Preface

Objective

The objective of this text is to provide training material for communications engineers in the transmission field. Specifically, it is designed to serve as the text for a one-year basic course of 64 classroom hours in transmission systems design. An evolving version of this course has been taught for many years by Members of the Technical Staff as part of the in-house continuing education program for technical employees of Bell Telephone Laboratories.

The background assumed is an MS degree in electrical engineering or in a closely related field, but the course is intended to accommodate participants with a considerable range of prior training, and many students with backgrounds in physics, mechanical engineering, and other disciplines have completed the course without undue difficulty. A few pre-course introductory lectures in terminology and the general concepts of telecommunications have been found useful.

Plan of Text

In this edition, broadly useful basic information has been separated from specific system applications to a much greater extent than in earlier editions. Thus, the first eleven chapters, about one-third of the total, are devoted to this fundamental material. It is recognized that there is some risk of student frustration in the consequent delay in getting to applications, especially in what is advertised as a design course, but it is felt that the risk is more than offset by the advantages of a more orderly development of the subject.

The first three chapters describe the environment in which transmission systems operate, establish basic definitions and terminology, and describe the methods by which the transmission objectives for the message channels are established. Chapter 4 deals with voice-frequency transmission. Chapters 5 through 11 treat the methods of

processing signals into forms suitable for transmission over high-frequency lines, the characteristics of multiplexed speech signals, and the major impairment mechanisms of noise, nonlinearities, and crosstalk. At present, and for at least the next several years, voice signals will continue to make up the bulk of the traffic. Therefore, these chapters emphasize voice signal transmission even though other messages, such as data and PICTUREPHONE ®, are becoming increasingly important.

Following this essentially introductory material are three major sections on systems applications, divided according to the principal modes of transmission in use today, i.e. analog AM, analog FM, and digital. The text concludes with a discussion of some special techniques and more detailed treatments of the characteristics of video and wideband data transmission.

The presentation throughout is firmly based on the realities of current practice. The authors and editors are fully aware that the students will be designing transmission systems for the future, but experience suggests that these are most likely to be realized by the advance and evolution of present techniques rather than by a wholesale discarding of these methods. It is also necessary to recognize that, to be useful, any new system must be compatible with the enormous existing plant. The approach to system design therefore is based on what is in actual use, or what has a very high probability of use in the relatively near future.

State of the Art

Despite the seeming conservatism of the last statements, transmission design is moving rapidly in new and exciting directions. Over the last decade a complete generation of systems based on discrete solid state devices has been designed, manufactured, and placed in service. So complete has been the revolution from the electron tube technology of the 1940s and early 1950s that hardly a tube can be found in the new systems. Even the last holdouts, the higher power transmitting stages in microwave radio systems, may soon be replaced by bulk effect solid state devices.

We are now beginning to see the first widespread use of thin film tantalum and monolithic silicon integrated circuits. Circuits an order of magnitude more complex than those in use a few years ago and components with precision previously unattainable are becoming available in inexpensive and reliable devices. There is no doubt that

we are at the start of a new generation of systems as different from those based on discrete solid state devices as they in turn differed from the old electron tube systems. As always, when such basic changes occur, the great challenge lies not in designing just somewhat better or less costly versions of the old systems, but in exploiting the opportunities presented by the new devices to realize systems that would be completely impractical without them.

Acknowledgments

The preparation of this edition was started in October 1967. As in previous editions, the material was written, revised, checked, and edited by teams of specialists with expert knowledge in the subject matter of each chapter. Valuable contributions were made by several instructors based on many years of experience in teaching the course. All through this process, materials and ideas were taken from numerous internal and published sources. Thus, recognition of each individual contribution would be practically impossible. In the fullest sense this book is the product of the Members of the Technical Staff of the whole Bell Telephone Laboratories.

E. F. O'Neill
Executive Director
Transmission Division

Merrimack Valley Laboratory
September 1969

Contents

Chapter 1

Transmission System Environment

A transmission system in its simplest form is a pair of wires connecting two telephones. More commonly, a transmission system is a complex aggregate of electronic gear and the associated medium, which together provide a multiplicity of channels over which many customers' messages and associated control signals can be transmitted.

In general, a call between two points will be handled by connecting a number of different transmission systems in tandem to form an overall transmission path or connection between the two points. The way in which these systems are chosen and interconnected has a strong bearing on the characteristics required of each system. This is true because each element in the connection will degrade the message to some extent. It follows that the relationship between performance and cost of a transmission system cannot be considered in terms of that system alone but must also be viewed with respect to the relation of the system to the building up of a complete connection.

To provide the service which permits people or machines to talk together at a distance, the telephone system must supply the means and facilities for connecting the particular customer stations at the beginning of the call and disconnecting them when the call is completed. Switching, signaling, and transmission functions are involved. The switching function includes identifying and connecting the customers to a suitable transmission path. The signaling function involves supplying and interpreting the control and supervisory signals needed to perform this operation. The transmission aspect, which is the concern of this text, deals with the transmission of the customer's message and these control signals.

The design of new transmission systems is constrained by the fact that they must be compatible with an existing multibillion dollar

plant and by the fact that they must perform a number of functions, such as transmission of signaling information and transmission of various messages, e.g., telephone, narrowband and wideband data, telephoto, or television. Thus, the solution of design, manufacturing, and operations problems for a specific system will generally require knowledge of other systems in the telephone plant, both existing and planned. Moreover, it is necessary to design systems having good performance with very high reliability and simple maintenance procedures, and to provide equipment which will verify the performance by tests made on a routine basis.

1.1 TELEPHONE SERVICE

Connection Description

A connection may involve merely voice-frequency transmission between telephones through a single end (central) office, or it may involve a multiplicity of links including several offices, voice-frequency paths, and carrier systems.

The telephone set converts acoustic energy into an electrical analog signal. The set also converts a received signal to its acoustic form. In addition, it generates supervisory signals (on-hook and off-hook) and the address information used by the switching system to establish connections.

The customer loop provides a path for the two-way speech signals and the ringing, switching, and supervisory signals. Since the telephone set and customer loop are permanently associated, their combined transmission properties can be adjusted to meet their share of the message channel objectives. For example, the greater efficiency of an improved telephone set compensates for increased loop loss and thus permits longer loop lengths or use of finer gauge wire.

The small percentage of the time (of the order of 10 per cent during busy hours) that a customer loop is used has led to the consideration of line concentrators for introduction between the customer and the central office. The concentrator allows many customers to share a limited number of lines to the central office. The line from the concentrator to the central office is, in effect, a trunk. The essential difference between a loop and a trunk is that a loop is permanently associated with a particular customer, whereas a trunk is a common usage connection.

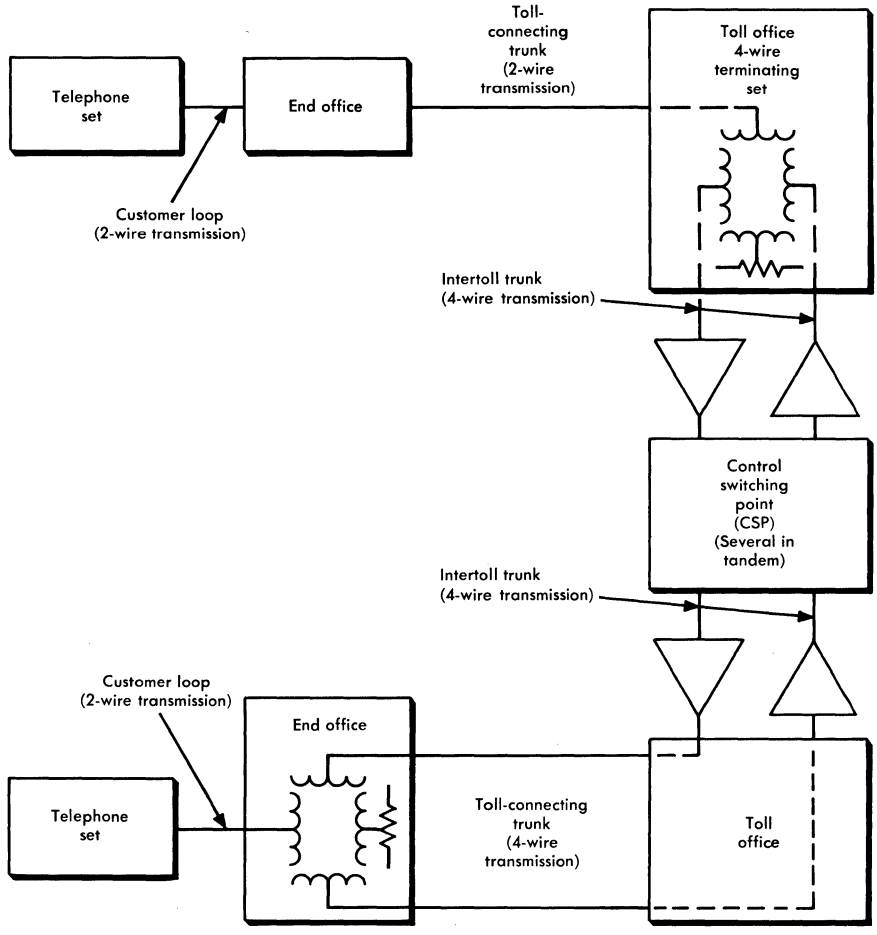


FIG. 1-1. Intercity customer-to-customer telephone connection.

Trunks of various types are used to interconnect end offices and toll centers. A direct interoffice trunk connects one end office to another end office; a tandem trunk connects an end office to an intermediate or tandem office; and a toll-connecting trunk connects an end office to a toll office. In toll transmission language, toll-connecting trunks are also described as terminating trunks.

Up to the point where the signals are connected to intertoll trunks in the toll office, the message and supervisory signals may be handled on a two-wire basis (the same pair of wires is used for both directions of transmission), or on a four-wire basis (separate transmission paths

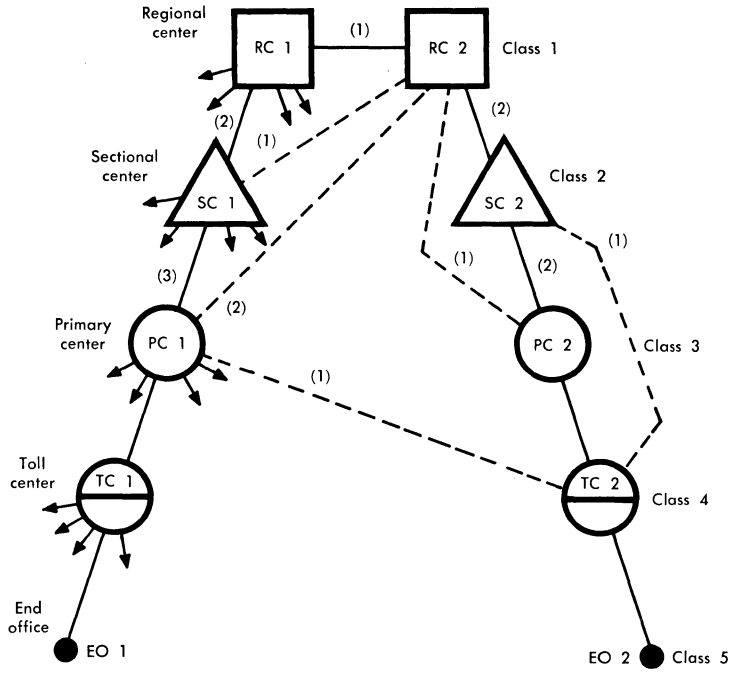
for each direction). At the toll office after appropriate switching and routing, the signals are generally connected to intertoll trunks by means of a four-wire terminating set, which splits apart the two directions of transmission so that the long-haul transmission may be accomplished on a four-wire basis. Through these intertoll trunks, the signals are transmitted to remote toll-switching centers (which in turn may be connected by intertoll trunks to other switching centers) and ultimately reach the recipient of the call through a toll-connecting trunk, an end office, another four-wire terminating set and local switching equipment, and a final customer loop, as indicated by Fig. 1-1.

In the present toll-switching plan there are five ranks or classes of switching centers. The highest rank is the regional center. The lowest rank, called the end office, is the telephone exchange in which the customer loops terminate. The chain of switching centers and an illustration of how a call might be routed is shown in Fig. 1-2. The order of choice at each control center is indicated in the figure by the numbers in parentheses. In the example, there are ten possible routes for the call, only one of which requires the maximum of seven intermediate links (toll trunks in tandem, excluding the two terminating links at the ends of the connection). Note that the first choice route involves two intermediate links. In many cases a single direct link, which would be the first choice, exists between the two toll centers.

The types of facilities that might be involved in various connections can be seen by reference to Fig. 1-3. The simplest connection would be a call between telephone sets 1 and 2, both working out of end office 1, in which no trunks would be involved. An interoffice call between sets 1 and 3 in city A would use two trunks, the connection being made via a tandem office. These trunks could be either voice-frequency circuits, possibly equipped with repeaters, or carrier circuits which combine a number of telephone channels into a single wideband channel.

Next, consider a call originating at telephone set 1 in city A and reaching telephone set 4 in city E. The path begins at a customer loop working into end office 1. From there it uses a toll-connecting trunk to the toll center. Between city A and city E there are a number of routes. If the two cities have a high community of interest, there would be direct trunks between them. Figure 1-3 shows that in this case the two cities are linked by N carrier.* An alternate route, which

*Different types of carrier systems are identified by letter designations. Many of these types are discussed in later chapters.

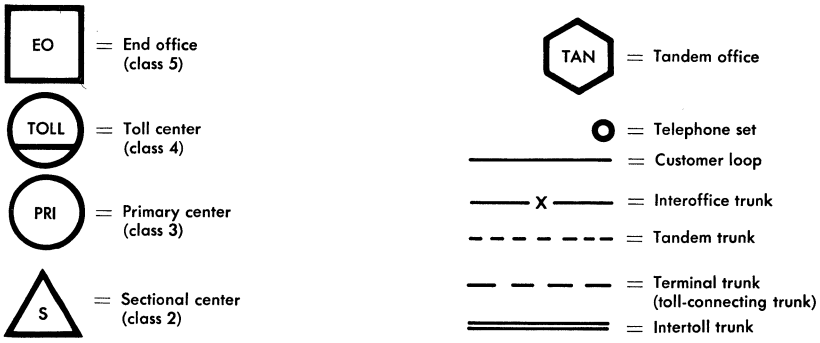


Notes:
1. Numbers in ( ) indicate order of choice of route at each center for calls originating at EO 1.
2. Arrows from a center indicate trunk groups to other lower rank centers that home on it. (Omitted in right chain.)
3. Dashed lines indicate high-usage groups.

FIG. 1-2. Choice of routes on assumed call.

happens to employ K carrier, is also shown, via a primary center. Out of this primary center (class 3) there might be direct, high-usage trunks on O carrier to city E. Alternatively, use would be made of *final trunks** to a sectional center (class 2) at city C, from which connection might be made to city E through another primary center. These latter trunks might be provided by a coaxial carrier system or a microwave radio system.

*In this instance, final has the connotation that these trunks are the last means chosen to get the message through, not being used unless all other circuits are busy. Very often, however, the final route is the only one provided; in that case it is, of course, the first choice.



For clarity in this diagram, nonstandard symbols are used for the end offices.
See Fig. 1-2 for standard usage.

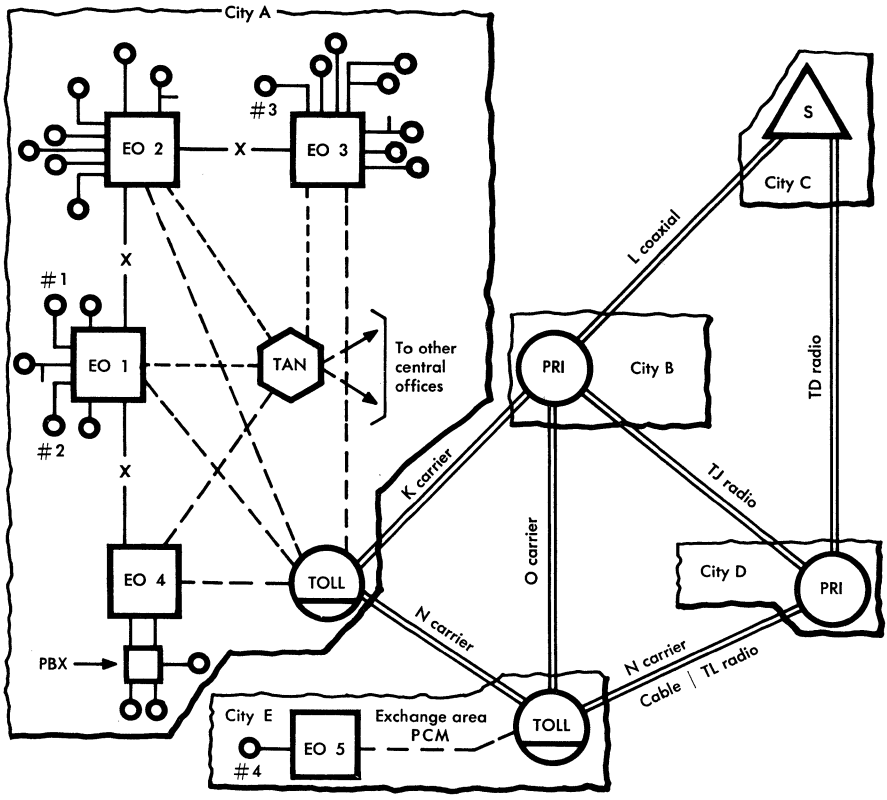


FIG. 1-3. A simplified telephone system.

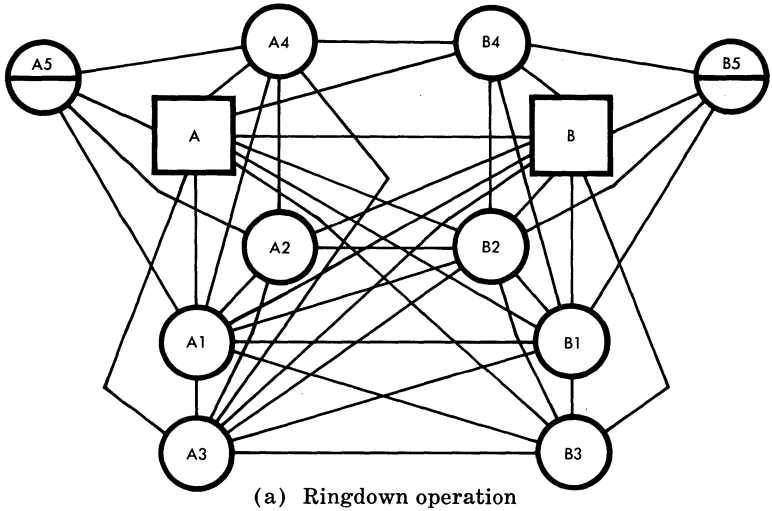
Switching Plan for Distance Dialing

The plan used to connect toll offices has a large bearing on the performance required of both local and toll transmission systems. In early practice, toll circuits were operated manually by operators on a so-called *ringdown* basis. With such arrangements the number of circuits that could be connected in tandem was severely limited, and relatively little use was made of alternate routing. Speed of service was comparatively slow, and trunks were inefficiently used in many cases.

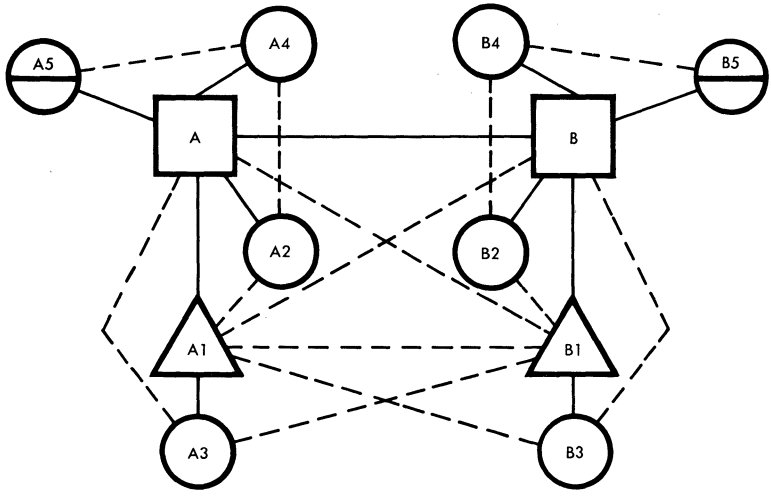
Automatic switching of toll circuits permits the use of alternate routes, so that small trunk groups can be operated at large trunk group efficiency with attendant economies. An example of the impact of toll dialing on the trunk layout is shown in Fig. 1-4. The upper diagram (a) shows the circuit groups that would be required to handle an assumed flow of traffic on a manual ringdown basis. The lower diagram (b) shows the circuit groups that would be required for the same traffic using toll dialing. Final trunk groups are provided between each lower ranking office and the higher ranking office on which it homes. All regional centers are interconnected with final trunk groups. High-usage groups are provided between any two offices that have sufficient community of interest. Final trunk groups carry traffic for which they are the only route, and also overflow traffic for which they are the "last choice" route. In (a) there are 42 different circuit groups. In (b) there are 24 circuit groups used on a more efficient basis.

The probability that a call will require more than n links in tandem to reach its destination decreases rapidly as n increases from 2 to 7. First, a large majority of toll calls are between end offices associated with the same regional center. The maximum number of toll trunks in these connections is therefore less than seven. Second, even a call between telephones associated with different regional centers is routed over the maximum of seven intermediate toll links only when all of the normally available high-usage trunk groups are busy. The probability of this happening in the case illustrated in Fig. 1-2 is only p^5, where p is the probability that all trunks in any one high-usage group are busy. Finally, many calls do not originate all the way down the line since each higher class of office will usually have class 5 offices homing on it and will act as a class 4 office for them.

Figure 1-5 makes these points more specific. The middle column of this table shows, for the fictitious system of Fig. 1-2, the probability



(a) Ringdown operation



(b) Nationwide toll dialing

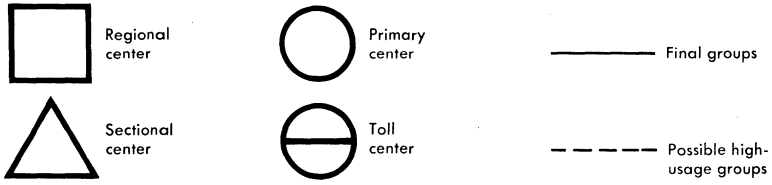


FIG. 1-4. Typical intertoll trunk networks.

Number of intermediate links, n	Probability	
	Fig. 1-2	1961 study
Exactly 1	0.0	0.50
2 or more	1.0	0.50
Exactly 2	0.9	0.30
3 or more	0.1	0.20
4 or more	0.1	0.06
5 or more	0.0109	0.01
6 or more	0.00109	0
Exactly 7	0.00001	0

FIG. 1-5. Probability that n or more links will be required to complete a toll call in example cited.

that the completion of a toll call will require n or more links between toll centers, for values of n from 1 to 7. In computing these probabilities, the assumptions are: (1) the chance that all trunks in any one high-usage group are simultaneously busy is 0.1; (2) the solid line routes are always available; and (3) of the available routes the one with the fewest links will always be selected. The figures in Fig. 1-5 illustrate that connections requiring more and more links become increasingly unlikely. These numbers are, of course, highly idealized and simplified. Actual figures from a Bell System study made in 1961 are shown in the last column of Fig. 1-5. These numbers represent the probability of encountering n links in a completed toll call between an office near White Plains, New York, and an office in the Sacramento, California region. The assumption was made that all traffic had alternate routing available and that blocking due to final groups was negligible. Note that at that time 50 per cent of the calls were completed over only one intermediate link. This is not possible in the system shown in Fig. 1-2, where it may be assumed that the traffic volume does not yet justify a direct trunk between toll centers. The maximum number of links involved in this particular system was five, and this number was required by only 1 per cent of the calls.

The switching pattern that has been described imposes strict transmission requirements on the toll trunks. Up to seven toll trunks may be connected in tandem, and successive calls between the same two telephones may take different routes and encounter different

numbers and kinds of circuits. When calls are routed over the maximum number of links, the loss must not be excessive. Also, the transmission quality should not vary greatly over the different possible routes that a call might take. If unsatisfactory transmission should occur, it will not be observed by an operator as in the past, and the customer's attempt to report unsatisfactory transmission will disconnect the impaired circuit, making identification of the source of trouble very difficult. It is therefore necessary to provide equipment which will test trunks on a routine basis.

1.2 SPECIAL SERVICES

Services other than residence, coin, or non-PBX business telephones are, broadly speaking, special services. Most of these are used for voice communication and differ from standard service principally in the arrangements for connecting into the direct distance dialing (DDD) network (some do not connect at all). In order to meet customer-to-customer transmission objectives, it is usually necessary to furnish specially engineered circuits for such service. Some special services may require more or less bandwidth than is commonly used for voice transmission.

Data Service

In addition to voice, program, and television channels, the Bell System provides facilities for the transmission of telephotograph, facsimile, teletypewriter, and digital data. These facilities may include special transmitting equipment which encodes the machine information into electrical analog or digital signals for transmission, and decoding equipment which translates the signal back into the original machine language.

The speed at which data can be transmitted is a direct function of the bandwidth available. The usual voice channel will transmit data at speeds of about 2000 bits per second without special treatment. Speeds of 4800 bits per second and higher can be achieved in a single voice channel by adding gain and delay equalization. Since data is usually transmitted in the form of pulses, impulse noise and delay distortion may cause errors. Thus, channels which meet voice transmission requirements may have limited application for data.

Voiceband data may be transmitted either over regular switched telephone channels or over specially equalized private line facilities.

Wideband data channels are obtained by using the frequency spectrum normally assigned to 6, 12, or 60 telephone channels. Special terminals and equalization are required and generally are supplied on a private line basis. Switched service is available between some large cities, and rapid expansion is anticipated.

Television Service

Television transmission in the Bell System involves connecting studios, the broadcaster's master control center, transmitters, and telephone company television operating centers (TOC) within cities, and then interconnecting cities by means of nationwide television facilities. Figure 1-6 shows a typical intracity layout for a large broadcaster. The local television links are usually video frequency systems rather than carrier or radio. Two-way connections between the master control location and the studio are often required for programming purposes. For network operation, connecting circuits are required between the master control room and the TOC where connection to the intercity facility is made.

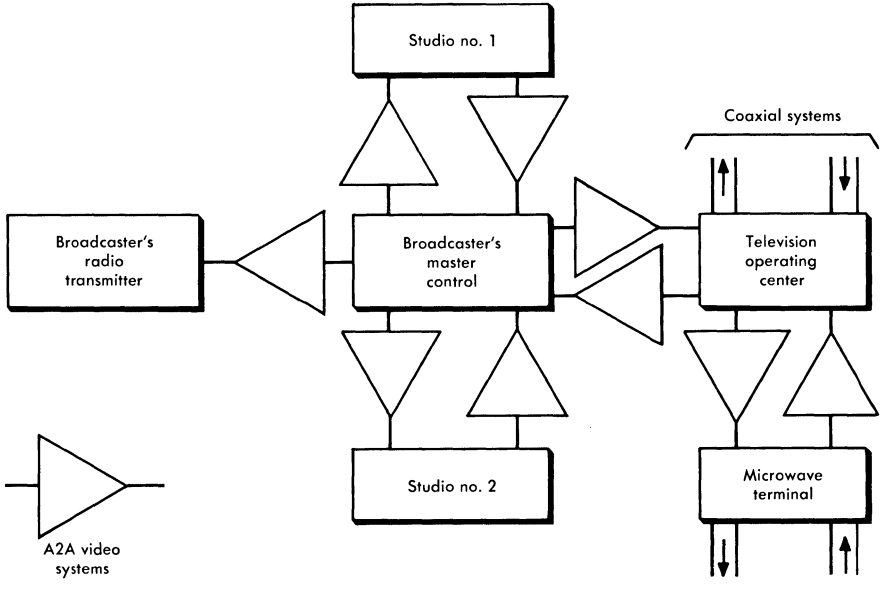


FIG. 1-6. Intracity television circuits.

The intercity channels may be either direct connections between cities or round robin channels with cities connected in a closed loop. These connecting facilities are formed and reformed each day, depending on the broadcaster's requirements. Thus, links must be connected in tandem in different ways on a day-to-day or hour-to-hour basis. It is therefore not usually practicable to line up or equalize on an overall basis. Instead, each link must be capable of a transmission quality such that when all the necessary links are connected in tandem, the signal will have a very small probability of being unduly degraded.

1.3 IMPACT OF SYSTEM MULTIPLICITY

In the preceding discussion it was seen that customer-to-customer communications channels can involve a multiplicity of different systems connected in many ways. It also was seen that local plant (i.e., telephone sets, customer loops, and end offices) is basic to every connection; its efficiency and uniformity with respect to loss, noise, and impedance (to mention but a few of the factors that must be considered) affect the entire system. Transmission systems used to interconnect central and toll offices often include terminals and sections of line which in turn are composed of numerous more-or-less identical repeater sections.

This composition of the overall connection gives rise to two problems that have a major bearing on everything that is done in the design of transmission systems. First, the accumulation of performance imperfections from a large number of systems leads to severe requirements on individual units and to great concern with the mechanisms causing imperfections and the ways in which imperfections accumulate. Second, the variable complement of systems forming overall connections makes the problem of economically allocating tolerable imperfections among these systems quite complex. Deriving objectives for a connection of fixed length and composition is a problem involving customer reactions and economics. However, when the transmission objectives must be met for connections of widely varying length and composition, the problem of deriving objectives for a particular system becomes an even more complex statistical study involving considerable knowledge of plant layout, operating procedures, and the performance of other systems.

Chapter 2

Transmission Fundamentals

The primary function of a transmission system is to provide circuits having the capability of accepting information-bearing electrical signals at one point and delivering related signals bearing the same information to a distant point. Some present-day transmission systems, such as N carrier, TD2 radio, etc., have been mentioned in Chap. 1. It has been emphasized that long distance conversations often require the tandem connection of several such systems.

System design is concerned primarily with the *terminals* which process the signals at each end of the transmission medium and with the *repeaters* which perform related functions at intermediate points. The properties of the transmission medium are basic considerations in every system design. More often than not, however, a new system design will be constrained to a choice among a limited number of standardized transmission media. The areas in which the system designer will find the greatest challenge to his skills, imagination, and ingenuity are the terminals and repeaters. This chapter introduces some of the language and concepts used in the evaluation of transmission performance, illustrates their application in defining the properties of typical transmission media and circuits, and concludes with a discussion of signal magnitudes, their measurement and manipulation.

2.1 POWER AND VOLTAGE RELATIONS IN LINEAR CIRCUITS

Since most of a transmission system is comprised of a tandem connection of several two-port networks, it is necessary to briefly

review many of the relationships commonly used with such circuits. The input-output relations or transfer characteristics of the individual two-port networks are of primary interest for system analysis. The resulting transfer characteristic of the tandem connection of several two-ports can then be found by a simple product of the appropriate transfer characteristics of the networks.

Some of the mathematical relations necessary for the evaluation of system performance can be explained in terms of the simple circuit diagram of Fig. 2-1. A source (generator) is characterized by its open circuit voltage, V_s, and its internal impedance, Z_s. A load is characterized by its impedance, Z_L.

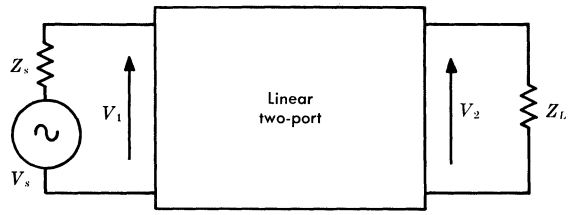


FIG. 2-1. Terminated two-port circuit.

Energy is transferred from source to load via a linear transducer. The transducer may take on a wide variety of forms, ranging from a simple pair of wires to a complex assortment of cables, amplifiers, modulators, filters, and similar circuits. The four terminals are associated in pairs; the pair connected to the source is commonly called the input port, and the pair connected to the load is referred to as the output port.

The circuit is linear if the relation between the output signal (response) and the input signal (stimulus) is determined by a set of linear differential equations with constant coefficients. In a linear circuit, signals may be represented by the Fourier series as a summation of terms of the form

$$V_k = E_k \cos(\omega t + \phi_k) \tag{2-1}$$

or, more conveniently, as a summation of terms of the form

$$V_k = E_k e^{j(\omega t + \phi_k)} \tag{2-2}$$

The input-output relation for each term in such a summation is not dependent on the presence or magnitude of other similar terms.

For example, if the generator voltage of Fig. 2-1 is represented by

$$V_s = E_s\, e^{j\omega t} \tag{2-3}$$

the ratio of V_1 to V_2 is given by

$$V_1/V_2 = (E_1\, e^{j(\omega t+\phi_1)})/(E_2\, e^{j(\omega t+\phi_2)})$$

$$= (E_1/E_2)\, e^{j(\phi_1-\phi_2)} \tag{2-4}$$

The ratios encountered in telephone transmission are often very large, and the numerical values involved are awkward. Moreover, it is frequently necessary to form the products of several ratios. The expression and manipulation of voltage or power ratios is simplified by the use of logarithmic units. The natural logarithm of the ratio of Eq. (2-4) is a complex number.

$$\theta = \alpha + j\beta = \ln(E_1/E_2) + j(\phi_1 - \phi_2) \tag{2-5}$$

The real and imaginary parts of Eq. (2-5) are uniquely identifiable, which is to say,

$$\alpha = \ln(E_1/E_2)$$

$$\beta = \phi_1 - \phi_2 \tag{2-6}$$

When this measure of a voltage (or current) ratio is used, α is said to be expressed in nepers and β in radians.

The Decibel

The logarithmic unit of signal ratio which finds wide acceptance is the *decibel.* Strictly speaking, the decibel (dB) is defined only for power ratios; however, as a matter of common usage, voltage or current ratios also are expressed in decibels. The precautions required to avoid misunderstanding of such usage are developed.

If two powers, p_1 and p_2, are expressed in the same units (watts, microwatts, etc.), then their ratio is a dimensionless quantity, and as a matter of definition,

$$D = 10 \log\,(p_1/p_2) \qquad \text{dB} \tag{2-7}$$

where log denotes logarithm to the base 10, and D expresses the relative magnitudes of the two powers in decibels.* If an arbitrary power is represented by p_0, then

$$D = 10 \log (p_1/p_0) - 10 \log (p_2/p_0) \quad \text{dB} \tag{2-8}$$

Each of the terms on the right of Eq. (2-8) represents a power ratio expressed in dB, and their difference is a measure of the relative magnitudes of p_1 and p_2. Clearly, the value of this difference is independent of the value assigned to p_0. In short, Eq. (2-7) is a measure of the *difference in dB* between p_1 and p_2.

When a voltage, expressed as in Eq. (2-1), appears across an impedance, $Z_k = R_k + jX_k$, the power dissipated in the impedance is equal to

$$p_k = \frac{E_k^2}{2R_k(1 + X_k^2/R_k^2)} = \frac{1}{2} R_k I_k^2 \tag{2-9}$$

where $I_k = E_k / \mid Z_k \mid$

Substitution in Eq. (2-7) gives

$$D = 20 \log (E_1/E_2) - 10 \log (R_1/R_2) - 10 \log \frac{1 + X_1^2/R_1^2}{1 + X_2^2/R_2^2} \quad \text{dB} \tag{2-10}$$

$$D = 20 \log (I_1/I_2) + 10 \log (R_1/R_2) \quad \text{dB} \tag{2-11}$$

Let

$$\begin{aligned} D_E &= 20 \log (E_1/E_2) \\ D_R &= 10 \log (R_1/R_2) \\ D_X &= 10 \log (1 + X_1^2/R_1^2)/(1 + X_2^2/R_2^2) \end{aligned} \tag{2-12}$$

Consider the statement: The difference between E_1 and E_2 is 20 dB. What is meant? There are three possibilities:

1. $Z_1 = Z_2$

In this case $D_R = D_X = 0$, $D = D_E = 20$ dB and the meaning is clear.

*It follows that one neper = 8.686 dB, or very roughly 1 bel.

2. $Z_1 \neq Z_2$, but $(X_1/R_1) = (X_2/R_2)$
   This is a common case occurring most often with $X_1 = X_2 = 0$. Voltmeters calibrated in dB can give D_E, but the value should be corrected by subtracting D_R.
3. $(X_1/R_1) \neq (X_2/R_2)$
   Clearly $Z_1 \neq Z_2$ even if $R_1 = R_2$. In this case both D_R and D_X need to be evaluated and subtracted from D_E in order to obtain D.

Situations represented by cases 2 and 3 can (and usually do) result in misunderstanding when voltage or current ratios are expressed in dB.* Statements of the type just quoted must be qualified if the recipient of this information is to be sure that the value quoted is in fact D and not the uncorrected value, D_E.

Loss, Delay, and Gain

There are several different methods of describing the transfer characteristic of a two-port network. In general, this will require specification of four complex quantities such as y or h parameters. However, in many cases where the network environment (such as source and load impedances) is controlled, the transfer can often be characterized by a frequency-dependent complex number describing the loss (or gain) and phase shift through the network. Since such description is of limited flexibility, several different means of describing such a characteristic have come into use, each having merit for a particular set of circumstances.

Insertion Loss and Phase Shift. Referring again to Fig. 2-1, suppose that for a particular value of the voltage, V_s, it has been determined that power, p_2, is delivered to the load, Z_L. Suppose then that the transducer has been removed and the source connected directly to the load, and the power delivered to Z_L has been determined to be p_0. The difference in dB between p_0 and p_2 is called the *insertion loss* of the transducer, i.e.,

$$\text{Insertion loss in dB} = 10 \log (p_0/p_2) \tag{2-13}$$

Since the first condition is satisfied, there is no ambiguity in expressing insertion loss as a voltage ratio. If V_s is expressed by Eq. (2-3), then V_0 and V_2, corresponding respectively to p_0 and p_2, are

*This remark applies equally to values expressed in nepers, but these units usually are encountered in situations which satisfy condition 1.

expressed by Eq. (2-2). Proceeding as in Eqs. (2-4) and (2-6) yields a restatement of the insertion loss and a definition of the *insertion phase shift:*

$$\text{Insertion loss} = 20 \log (E_0/E_2) \quad \text{dB} \tag{2-14}$$

$$\text{Insertion phase} = 57.3\ (\phi_0 - \phi_2) \quad \text{degrees} \tag{2-15}$$

If the transducer of Fig. 2-1 furnishes gain, then $E_2 > E_0$, and the insertion loss values are negative. In order to avoid talking about negative loss, it is customary to write

$$\text{Insertion gain} = 20 \log (E_2/E_0) \quad \text{dB} \tag{2-16}$$

If complex gain is expressed in the form of Eq. (2-5), the phase will be the negative of the value found in Eq. (2-15). Unfortunately, there is no standard name which clearly distinguishes between the phase calculated from a loss ratio and that calculated from a gain ratio. The ambiguity is entirely a matter of algebraic sign and can always be resolved by observing the effect of substituting a shunt capacitor for the transducer. This gives a negative sign to the value of ϕ_2 and a positive change in the phase of Eq. (2-15).

Phase and Envelope Delay. The *phase delay* and *envelope delay* of a circuit are defined as

$$\text{Phase delay} = \beta/\omega$$

$$\text{Envelope delay} = d\beta/d\omega$$

where β is in radians, ω is in radians per second, and delay is therefore expressed in seconds. In accordance with the sign convention adopted previously, both the phase and the envelope delay of an "all-pass" network are positive at all finite frequencies.

Phase and Group Velocity. For cables or similar transmission media, the phase shift is usually quoted in radians per mile. In this case, phase and envelope delays are expressed in seconds per mile. Their reciprocals are called *phase velocity* and *group velocity,* respectively, and the units are miles per second.

Available Gain. The maximum power available from a source of internal impedance, Z_s, is obtained when the load connected to its

terminals is equal to Z_s^*; i.e., if

$$\begin{aligned} Z_s &= R_s + jX_s \\ Z_s^* &= R_s - jX_s \end{aligned} \tag{2-17}$$

For an open circuit generator voltage having an rms value, E, the maximum available power is

$$p_{as} = E^2/4R_s \tag{2-18}$$

The power actually delivered to Z_L in Fig. 2-1 also will be maximized if the output impedance of the transducer (the equivalent Thevenin generator impedance) is conjugate to Z_L. Designating this power as p_{a2} leads to a definition of *available gain*, g_a, as:

$$g_a = 10 \log (p_{a2}/p_{as}) \tag{2-19}$$

Transducer Gain. Ordinarily the impedances do not meet the conjugacy requirements, and it is necessary to define the *transducer gain*, g_t, of the two-port circuit as:

$$g_t = 10 \log (p_L/p_{as}) \tag{2-20}$$

where p_L is the power actually delivered to the load. Transducer gain is dependent on load impedance and can never exceed available gain. Transducer gain is equal to available gain only when the load impedance is equal to the conjugate of the network output impedance.

Power Gain. Finally, *power gain*, g_p, is defined as:

$$g_p = 10 \log (p_L/p_1) \tag{2-21}$$

where p_1 is the power actually delivered to the input port of the transducer. The power gain is equal to the transducer gain of a network when the input impedance of the network is equal to the conjugate of the source impedance. The power gain is equal to the insertion gain of the network when the input impedance of the network is equal to the load impedance connected to the output of the network.

2.2 TRANSMISSION LINES

Today, the functions of the transmission line usually are realized by using cable pairs or coaxial conductors. Since their characteristics are described properly by parameters and equations originally developed for open wire lines, it is common practice to include such media under the broad classification, transmission lines.

Uniformly Distributed Lines

The transmission characteristics of lines are determined by such properties as conductivity, diameter and spacing of conductors, and the (lossy) dielectric constant of the insulation. These properties in turn determine the electrical *primary constants*, R, L, G, and C, representing the uniformly distributed series resistance, series inductance, shunt conductance, and shunt capacitance. It is common practice to express these constants in ohms, henries, etc. per mile of cable. The transmission characteristics, or secondary constants, are calculated from the primary constants by use of the following equations:

$$\text{Characteristic impedance, } Z_I, = \sqrt{\frac{R + j\omega L}{G + j\omega C}} \qquad (2\text{-}22)$$

$$\text{Propagation constant, } \gamma = \alpha + j\beta = \sqrt{(R + j\omega L)\ (G + j\omega C)} \qquad (2\text{-}23)$$

The characteristic impedance is a complex quantity, expressed in ohms and independent of length. Its value approaches a constant, $Z_0 = \sqrt{L/C}$, as the frequency is raised. The propagation constant expresses the attenuation, α, in nepers per mile and the phase shift, β, in radians per mile.

If the terminating impedances, Z_s and Z_L, of Fig. 2-1 are both equal to Z_1, then the insertion loss (and phase) of a length l of cable is equal to γl. If these impedances have some other value, say Z_T, then the insertion loss [1] is given by the voltage (current) ratio:

$$\frac{V_0}{V_2} = e^{\gamma l}\ (1 - \rho^2)/(1 - \rho^2 e^{-2\gamma l}) \qquad (2\text{-}24)$$

where

$$\rho = (Z_T - Z_1)/(Z_T + Z_1) \qquad (2\text{-}25)$$

The variation of Z_I and γ with frequency will usually result in a significant difference between the attenuation and the insertion loss. Moreover, the insertion loss will not be proportional to cable length.

The term ρ in Eq. (2-24) is called the *reflection coefficient.* Its magnitude is a measure of the fractional energy loss at an impedance mismatch. In the present example, there are equal reflections at the input and output ports. Reflection effects also may be stated in dB by defining a *return loss:*

$$\text{Return loss} = 20 \log (1/|\rho|) \tag{2-26}$$

Only the magnitude of the reflection coefficient, $|\rho|$, is of interest in ordinary situations. However, the numerator of Eq. (2-25) has been written in a form which gives the correct sign to *voltage* reflections.

The actual voltage across the load is equal to the voltage which would be delivered to a matched impedance plus the voltage reflected back to the source. As Z_T approaches zero, the reflection coefficient approaches -1, and the (measured) voltage, V_2, approaches zero. Conversely, as Z_T approaches infinity, ρ approaches $+1$, and the load voltage approaches its open circuit value equal to twice the load voltage under matched conditions.

Finally, the reflected energy is subject to multiple reflections at both ports. These reflections are accounted for by the *interaction factor,* the denominator of Eq. (2-24).

Twisted Pair Cable

A cable pair is made by twisting together two insulated conductors, usually of high purity copper. The insulation may be wood pulp formed on the conductors in a process similar to paper making, or it may be plastic formed by an extrusion process. (Polyethylene is a widely used plastic insulation.) Neighboring pairs are twisted with different pitch (twist length) in order to limit electrical interference (crosstalk) between them. The pairs are stranded in units, and the units are then cabled into cores. The cores are covered with various types of sheaths depending on the intended use. Polyethylene-insulated cables (PIC) are made in sizes from 6 to 900 pairs while pulp-insulated cables come in sizes from 300 to 2700 pairs. Unit sizes range from 6 to 50 pairs for PIC, and 25 to 100 pairs for pulp cables. Common wire sizes used are 19-, 22-, 24-, and 26-gauge, and in rare instances 16-gauge is used.

The primary constants of twisted pair cable are subject to manufacturing deviations, and change with the physical environment such

as temperature, moisture, and mechanical stress. The inductance, L, is of the order of 1 millihenry per mile for low frequencies; and the capacitance, C, has two standard values of 0.066 and 0.083 microfarads per mile although lower capacitance cables are under development.

Of the primary constants, only C is relatively independent of frequency; L decreases to about 70 per cent of its initial value as frequency increases from 50 kHz to 1 MHz and is stable beyond; G is very small for PIC and roughly proportional to frequency for pulp insulation; and R, approximately constant over the voiceband, is proportional to the square root of frequency at higher frequencies where skin effect and proximity effect dominate.

Loading. A detailed study of the attenuation of cable pairs, based on Eqs. (2-22) and (2-23) and typical values for the primary constants, shows that a substantial reduction of attenuation can be obtained by increasing the value of L. Minimum attenuation requires a value of L nearly 100 times the value obtained in ordinary twisted pair. The realization of such a value on a uniformly distributed basis is impractical. Instead, the desired effect is obtained by "lumped" loading, that is, by inserting series inductance periodically along the pair.

Loading arrangements are specified by a code letter designating the distance between loading coils and by numbers which indicate the inductance value and wire gauge. For example, the designation of one loading arrangement is 19H88. The number 19 specifies 19-gauge wire; H indicates the spacing between coils; the number 88 refers to the inductance (in millihenries) of the coils. Loading arrangements that may be found in use include H44, H88, D88, and B135, where H is the designation for 6000-foot, D for 4500-foot, and B for 3000-foot spacings. Figure 2-2 illustrates the improvement obtainable in loss and loss-frequency characteristic with loading.

Similar effects can be obtained by negative impedance loading. In this approach, a two-terminal active circuit is connected in series with the conductors of the pair. The inserted impedance gives a very good approximation to a negative resistance having a value slightly less than the series resistance of the pair.

In terms of the equations given previously, loading may be thought of as a means of approximately realizing the available gain of the system [Eq. (2-19)] or of realizing reflection gain [Eq. (2-24)] by introducing discontinuities at which $|\rho| > 1$. Detailed calculations of the performance of loaded cables are more conveniently made by

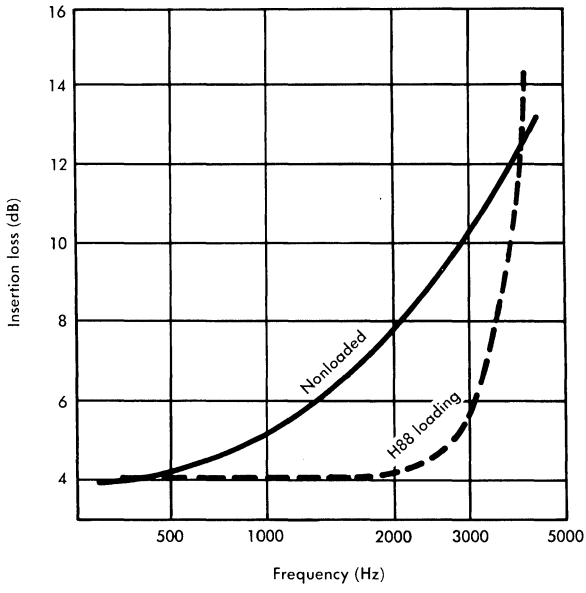


FIG. 2-2. Insertion-loss characteristics of 12,000 feet of 26-gauge BST cable measured between impedances of 900 ohms and 2 μF.

matrix methods, and these have been used to produce tables of the characteristics of common cables for a variety of loading conditions. Each type of loading has associated with it a cutoff frequency above which the attenuation increases very rapidly. The negative impedance devices require power at each loading point.

Bridged Taps. An irregularity frequently found in cables serving customer locations is the *bridged tap.* This consists of another pair of wires which are connected in shunt to the main cable pair at any point along its length. This pair may or may not be used at some future time, depending on the way in which service demands develop. In any case only one of the pairs going away from the bridging point is likely to be used at any given time. The other hangs open-circuited across the working pair and introduces bridging loss. In order to limit transmission impairment, there are rigid rules concerning the number, length, and location of bridged taps allowable on pairs assigned to various kinds of service.

Shielding. The possibility of electrical interference (crosstalk) between cable pairs has already been mentioned. When signal frequencies of the order of 1 MHz or higher are involved, crosstalk coupling may become a dominant design consideration. In such cases, the transmission medium may be isolated from interfering circuits by shielding.

Shielded, balanced pairs are frequently included in twisted pair cable in order to provide a satisfactory medium for the transmission of baseband television (video) signals. In video service the signal spectrum extends from near 0 to about 4.5 MHz. The 16 PEVL pair consists of two 16-gauge conductors insulated with expanded polyethylene and surrounded by a longitudinal-seam copper shield. Both the balance and the shielding are needed for interference reduction at low frequencies. The heavy gauge of the conductor and the low capacitance result in relatively low loss, about 18 dB per mile at the top of the band.

Coaxial Cable

At higher frequencies, the isolation between transmission paths can be achieved very efficiently by the use of coaxial conductors. The coaxial unit consists of a center conductor surrounded by a concentric outer conductor. For long-haul service, the standard unit has an inner conductor of 10-gauge copper wire and an outer conductor of solid copper with a diameter of 0.375 inch. The center conductor is supported by insulating discs located at approximately 1-inch intervals. A coaxial cable may contain 20 coaxial conductors and groups of twisted pairs which can be used to pass control and alarm signals to or from remote repeaters.

At normal operating frequencies, the coaxial outer conductor provides excellent shielding against extraneous signals. However, at low frequencies where skin depth is comparable to the thickness of the outer conductor, the shielding is ineffective. For this and economic reasons, the coaxial conductor, which is very desirable at radio frequencies, usually loses favor to twisted pair cable at lower frequencies.

The primary constants of a coaxial transmission line are under better control and are less frequency dependent than those for twisted pairs. This is because of the inherently more consistent mechanical structure and because of the shielding from outside influences provided by the outer conductor. The capacitance, C, is independent of frequency and is a function of the conductor diameter

ratio and the permittivity of the dielectric. The inductance, L, is also practically independent of frequency over the normal frequency ranges used with coaxial lines. This inductance increases at very low frequencies (audio and below), but the poor shielding qualities of the outer conductor at these frequencies make the use of coaxial lines at these low frequencies unattractive. The conductance, G, is a function of the dielectric used between the coaxial conductors. For air-insulated lines, this conductance is negligible at all frequencies usually used on such cables. The resistance, R, remains as the important frequency-dependent primary constant. Because of skin effect, R increases as the square root of frequency over the frequency range usually of interest. Rewriting Eq. (2-23) in its real and imaginary parts yields:

$$\gamma = \alpha + j\beta = \left\{\frac{1}{2}\left[\sqrt{(R^2 + \omega^2L^2)\ (G^2 + \omega^2C^2)} + RG - \omega^2LC\right]\right\}^{1/2}$$
$$+ j\left\{\frac{1}{2}\left[\sqrt{(R^2 + \omega^2L^2)\ (G^2 + \omega^2C^2)} - RG + \omega^2LC\right]\right\}^{1/2}$$

Evaluating the real part, α, for $G = 0$ and $R << \omega L$

$$\alpha \approx \frac{R}{2}\sqrt{\frac{C}{L}} \approx \frac{R}{2Z_0} \qquad \text{nepers/unit length}$$

The loss in nepers (or dB) per unit length is thus directly proportional to R, making the loss in dB per unit length of coaxial cable directly proportional to the square root of frequency for the frequencies of common interest.

Doubling the cross-sectional dimensions of the coaxial conductors halves the resistance (skin effect causes the resistance to be inversely proportional to conductor surface area rather than cross-sectional area). As a consequence, the loss in dB per unit length is inversely proportional to the conductor diameters providing their ratio remains the same. (It can be easily shown that Z_0 is a function of the conductor diameter ratio.) As a calibration point, the loss of the 0.375-inch coaxial conductor at 1 MHz is approximately 4 dB per mile.

2.3 TRANSFORMERS AND HYBRID CIRCUITS

A simple transformer consists of two closely coupled inductive windings wound around some magnetic material chosen for its high permeability and low loss at specified signal levels and frequencies.

Transformers provide efficient coupling between circuits of different impedance levels or between balanced circuits and unbalanced circuits. A common application is shown in Fig. 2-3, which illustrates two transformer-coupled balanced circuits. Signals transmitted in normal fashion over circuit No. 1 or No. 2 are not affected by currents entering and leaving via the indicated center taps of the transformer windings. In earlier days such circuits were actually used for speech transmission and were called phantom circuits. Today, they are used mainly for transmitting d-c power or signaling information and are called simplex circuits.

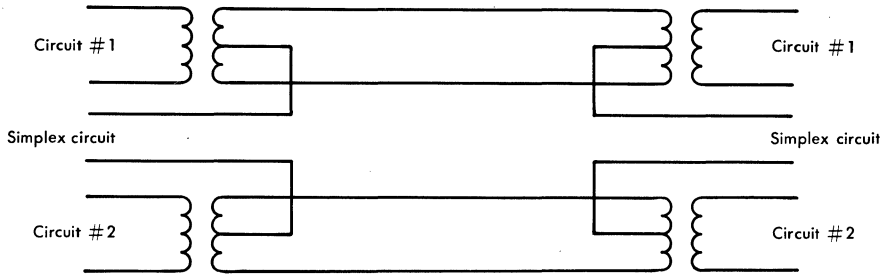


FIG. 2-3. Simplex circuit schematic.

The use of the precisely balanced transformer windings to obtain conjugacy between transmission paths results in the so-called *hybrid circuits*. These can be realized with a single transformer structure, but the impedance levels required are usually inconvenient. The more common realization uses two transformers connected as shown by the simplified diagram of Fig. 2-4. Transformers T_1 and T_2 each consist of at least three tightly coupled windings.

If $Z_1 = Z_2$ and $Z_3 = Z_4$, a proper choice of turns ratios will make port 1 conjugate to port 2, and port 3 conjugate to port 4. That is, if Z_1 is a source delivering power to port 1, a negligible part of this power will be received by impedance Z_2 and vice versa. Power flowing into the circuit at either port 1 or port 2 will be delivered to impedances Z_3 and Z_4 equally.

In one practical application, Z_3 is a bilateral two-wire line, and Z_4 is a fixed network whose only function is to match Z_3 and provide the necessary conjugacy. Impedances Z_1 and Z_2 represent a four-wire line using separate pairs for the two directions of transmission. The terms *trans-hybrid loss* and *through-balance* are used to describe the effec-

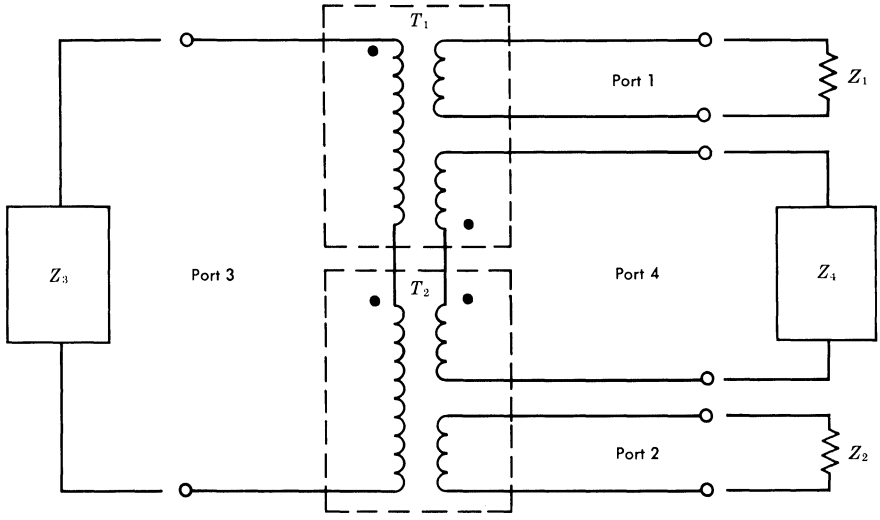


FIG. 2-4. Hybrid circuit using two transformers.

tiveness of this circuit. Losses of 50 dB between impedances Z_1 and Z_2 are realizable. In central offices where Z_3 is different for every call that is set up, much lower values are common.

2.4 TRANSMISSION LEVEL

Throughout this chapter, the discussion has been in terms of relative signal magnitudes, e.g., the magnitude of the output signal relative to that of the input signal. In a long system where a string of losses and gains produced by cable sections, repeaters, and other transducers is encountered, it is necessary to account for the signal magnitude at many points relative to its magnitude at other points. This is accomplished by choosing a datum, called the *transmission level point,* and determining the relative signal magnitude at any other point by the simple process of forming the algebraic sum of the gains, expressed in dB, encountered by any signal traversing the system. The resulting summation is a number which defines the transmission *level* (in dB) at that particular point. It must be emphasized that although absolute magnitudes are determined by the applied signal, the relative magnitudes within the system are uniquely determined by the transmission level.

In the interest of convenience and uniformity, the reference point is defined as the *0-dB transmission level point.* Although the reference point was at one time a point accessible to probes and measuring instruments, it is seldom so today. As a consequence of improving transmission, it is now customary to consider the outgoing side of the toll-transmitting switch as the —2 dB transmission level point. Signal magnitudes measured at this switch are 2 dB lower than would be measured at the reference level point if such a measurement were possible. The following abbreviations for 0-dB transmission level point are frequently encountered: zero level, zero level point, 0-dB TL, and 0 TLP. To put the concept of transmission level in the form of a definition:

> The transmission level of any point in a transmission system is the ratio (in dB) of the power of a signal at that point to the power of the same signal at the reference point.

Typical levels encountered in class 4 or higher offices are shown in Fig. 2-5. Here it is assumed that switching is on a two-wire basis and that transmission between the offices is on a four-wire basis. The via net loss (VNL), to be defined in a later chapter, can be assumed to be of the order of 1 dB. Each direction of transmission has its own reference level point, and these are at different places in the two-wire portion of the trunk. In addition, four-wire transmission systems have standardized voice-frequency input and output level points. These levels are —16 dB and +7 dB, respectively. This standardization is necessary for proper administration and operation of the telephone network. For instance, it permits interchanging (patching) different carrier systems for restoration of service.

When two or more trunks are connected in tandem, the level is redefined as —2 dB at the outgoing side of each outgoing toll switch, as shown in Fig. 2-6 for one direction of transmission.

It should also be noted that, although the power at the outgoing toll-office switch will be at an audio frequency, the corresponding signal power at any given point in a broadband carrier system may be at some carrier frequency. This signal power, nevertheless, can be measured or computed, thus specifying the transmission level in accordance with the definition. Unless otherwise stated, the transmission level is determined at a frequency of 1000 Hz or at a corresponding frequency obtained by modulation of 1000 Hz in the system.

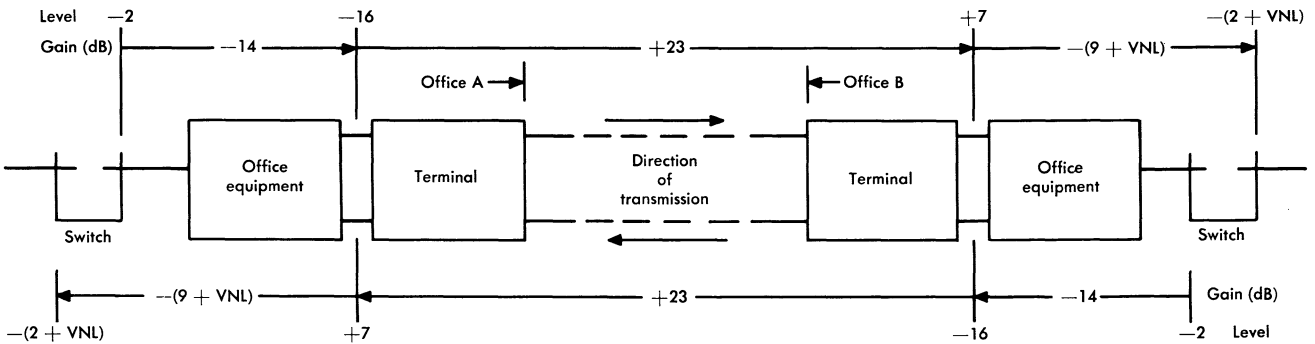


FIG. 2-5. Transmission levels: 4-wire trunk with 2-wire switching.

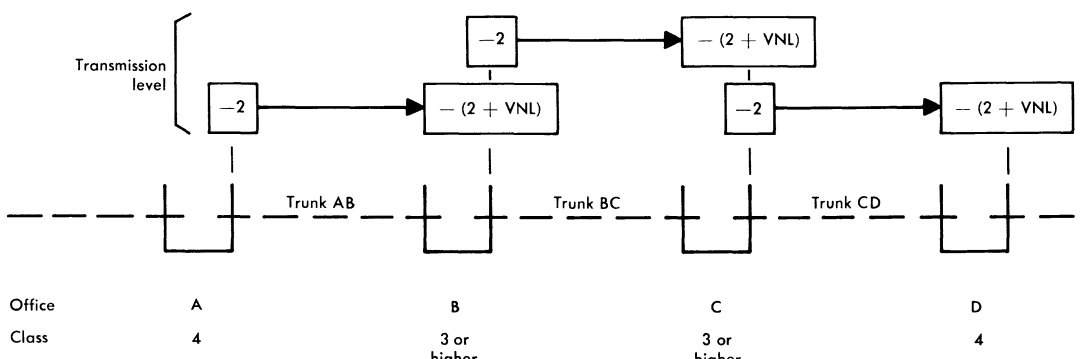


FIG. 2-6. Transmission levels in a tandem connection.

2.5 SIGNAL AND NOISE MEASUREMENT

A means of characterizing the signals to be handled by a transmission system is just as important as a knowledge of how these signals are affected by the circuits comprising the system. A detailed discussion of the nature of typical signals found in telephone message channels is given in Chap. 3; however, before such factors can be discussed, it is necessary first to define some of the various measures of signal magnitude and to combine and relate these measures to the transfer characteristics discussed so far.

Magnitudes

Since telephone circuits operate with signal powers which rarely are as large as 0.1 watt and which may be lower than 10^{-12} watts, the use of the watt as a unit of measurement is awkward. A more convenient unit is the milliwatt, or 10^{-3} watts. Many operations with signal magnitudes can be further simplified by expressing power in relative dB. This is accomplished by setting p_0 in Eq. (2-8) equal to 1 milliwatt. Then, the terms on the right side express the powers p_1 and p_2 in *dB relative to 1 milliwatt,* abbreviated dBm. Proceeding in analogous fashion from Eq. (2-4) yields expressions for E_1 and E_2 in *dB relative to 1 volt,* abbreviated dBV.

Expressing signal magnitude in dBm and system level in dB provides a simple method of determining signal magnitude at any point in a system. In particular, if the signal magnitude at 0 TLP is S_0 dBm, then the magnitude at a point whose level is L_x dB is

$$S_x = S_0 + L_x \qquad \text{dBm}$$

The abbreviation dBm0 is commonly used to indicate the signal magnitude in dBm at 0 TLP.

Volume

A periodic current or voltage can be characterized by any of three related values: the rms, the peak, or the average. The choice depends upon the particular problem for which the information is required. It is more difficult to deal with complex, nonperiodic functions like speech in simple numeric terms. The nature of the speech (or program) signal is such that the average, rms, and peak values, and the ratio of one to the other, are all irregular functions of time, so that one number cannot easily specify any of them. Regardless of the

difficulty of the problem, the magnitude of the telephone signal must be measured and characterized in some fashion that will be useful in designing and operating systems involving electronic equipment and transmission media of various kinds. Signal magnitudes must be adjusted to avoid overload and distortion, and gain and loss must be measured. If none of the simple characterizations is adequate, a new one must be invented. The characteristic unit used is called volume and is expressed in vu (volume units). It is an empirical kind of measure evolved to meet a practical need and is not definable by any precise mathematical formula. The volume is simply the reading of an audio signal on a volume indicator, called the vu meter, when the meter is read in a carefully specified fashion.

The development of the vu meter was a joint project of the Bell System and two large broadcasting networks. It was decided that the principal functions required of such a measuring device are:

1. Measuring signal magnitude in a manner which will enable the user to avoid overload and distortion.
2. Checking transmission gain and loss for the complex signal.
3. Indicating the relative loudness with which the signal will be heard when converted to sound.

In practice it is found that the vu meter can be used equally well for all speech, whether male or female. There is some difference between music and speech in this respect, and so a different reading technique is used for each.

For convenience, the meter scale is logarithmic, with a 10-log scale. That is, readings bear the same relationship to each other as do decibels; however, the scale units are in vu, not in dB. It is true that the meter will measure a continuous sinusoid imposed upon it. It is also true that a correlation between the volume of a talker and his long-term average power or his peak power can be established. Such correlations are valuable, but the fact that they exist should not be allowed to confuse the real definition of volume and vu. Putting it as simply as possible, a —10 vu talker is one whose signal is read on a volume indicator (by someone who knows how) as —10 vu. It should be noted that the vu meter has a flat frequency response over the audible range and is not frequency weighted in any fashion.

Noise

The measurement of noise, like the measurement of volume, is an effort to characterize a complex signal. The noise measurement is further complicated by an interest, not in the absolute magnitude of

the noise power, but rather in how much it annoys the telephone user. Consider the requirements of a meter which will measure the subjective effects of noise:

1. The readings should take into consideration the fact that the interfering effect of noise will be a function of frequency spectrum as well as of magnitude.
2. When dissimilar noises are present simultaneously, the meter should combine them to properly measure the overall interfering effect.
3. When different types of noise cause equal interference as determined in subjective tests, use of the meter should give equal readings.

The 3-type noise measuring set is essentially an electronic voltmeter with (1) frequency weighting, (2) a detector approximating an rms detector, and (3) a transient response similar to that of the human ear. These three characteristics cause the noise measurement to approximate the interfering effect that the noise would create for the average telephone user.

Interference is made up of two components: annoyance and the effect of noise on intelligibility. Both are functions of frequency, and therefore frequency weighting is included in the set. Annoyance is measured in the absence of speech by adjusting the level of a given tone until it is as annoying as a reference 1000-Hz tone. This is done for many tones and many observers, and the results are averaged and plotted. A similar experiment is done in the presence of speech at the average received volume to determine the effect of noise on articulation. The results of the two experiments are combined and smoothed, resulting in the *C-message weighting* curve shown in Fig. 2-7. The experiments are made with a 500-type telephone; therefore, the weighting curve includes the frequency characteristic of this telephone as well as the hearing of the average subscriber. The remainder of the telephone plant is assumed to provide transmission which is essentially flat across the band of a voice channel. Therefore, the C-message weighting is applicable to measurements made almost anywhere except across the telephone receiver.

The significance of the weighting curve of Fig. 2-7 is that, for example, a 200-Hz tone of given power is 25 dB less disturbing than a 1000-Hz tone of the same power. Hence, the weighting network incorporated in the noise meter will have 25 dB more loss at 200 Hz than at 1000 Hz.

Other weighting networks can be substituted on a plug-in basis.

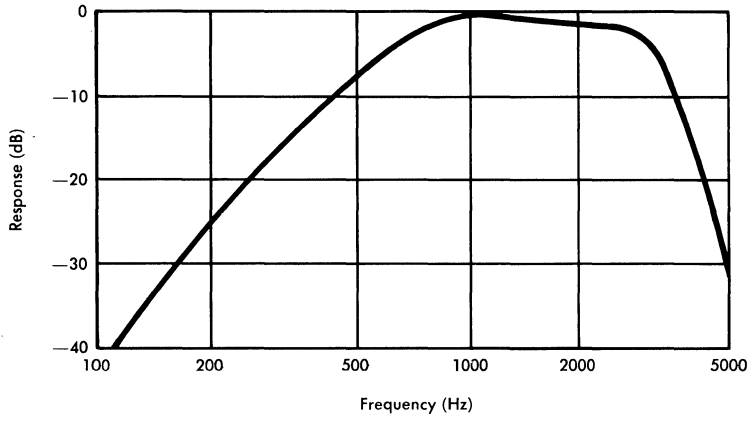


FIG. 2-7. C-message frequency weighting.

For example, the 3 KC FLAT network may be used to measure the power density of white noise. This network has a nominal low-pass response down 3 dB at 3 kHz and rolls off at 12 dB per octave. The effective response to white noise is almost identical to that of an ideal (sharp cutoff) 3 kHz low-pass filter.

Bands of noise are used in the determination of how different noises contribute to interference. Closest agreement between the judgment of the telephone user and the reading of the noise measuring set is obtained if the noises are added on a power basis. That is, if two tones have an equal interfering effect when applied individually, then the effect when both are present would be 3 dB worse than for each separately.

The third subjective factor which affects the manner in which noise must be measured is the transient response of the human ear. It has been found that, for sounds shorter than 200 milliseconds, the human ear does not fully appreciate the true power in the sound. For this reason the meter on the noise measuring set (as well as the vu meter) is designed to give a full indication on bursts of noise longer than 200 milliseconds. For shorter bursts, the meter indication decreases.

These three characteristics of the 3-type noise measuring set—frequency weighting, power addition, and transient response—essentially prescribe the way message circuit noise is to be measured. This is not yet enough; a noise reference and a scale of measurement must also be provided.

The chosen reference is 10^{-12} watts or -90 dBm. The scale marking is in decibels and measurements are expressed in decibels above reference noise (dBrn). A 1000-Hz tone at a level of -90 dBm will give a 0 dBrn reading regardless of which weighting network is used. For all other measurements the weighting must be specified. The notation dBrnc is commonly used when readings are made using the C-message weighting network.

As with dBm power readings, the vu and dBrn readings may be taken at any transmission level point and referred to 0 dB TLP by subtracting the level from the meter reading. Thus a typical noise reading might be 25 dBrn at 0 dB TLP, abbreviated 25 dBrn0. Similarly, values of dBrnc referred to 0 TLP are identified as dBrnc0.

2.6 POWER AND VOLTAGE SUMMATION

The preceding discussion has emphasized the merits of expressing signal magnitudes in relative dB or in dBm. However, when the need arises to determine the sum of two signals stated in dBm, there are disadvantages. Although the necessary steps are straightforward, they also are time consuming. Specifically, suppose it is known that powers p_1 and p_2 are flowing in a circuit but these powers are expressed as P_1 and P_2 dBm, respectively. It is desired to express the sum, p, of p_1 and p_2 as P dBm. The shorthand notation

$$P = P_1 \text{ "+" } P_2 \tag{2-27}$$

will be used to indicate the procedure:

$$P = 10 \log [\log^{-1}(P_1/10) + \log^{-1}(P_2/10)] \tag{2-28}$$

Assume that $p_1 \geq p_2$ and write

$$p = p_1 (1 + p_2/p_1) \tag{2-29}$$

which is equivalent to

$$P = P_1 + D \tag{2-30}$$

where

$$D = 10 \log (1 + p_2/p_1) \tag{2-31}$$

The maximum value of D is 10 log 2 or about 3 dB. Values of this function are shown by Fig. 2-8. The curve is plotted with D as the ordinate and $P_1 - P_2$ as the abcissa. In effect, it is a curve of relative dB. If only the difference, $P_1 - P_2$, is known, the graph still provides the information that P exceeds P_1 by D dB.

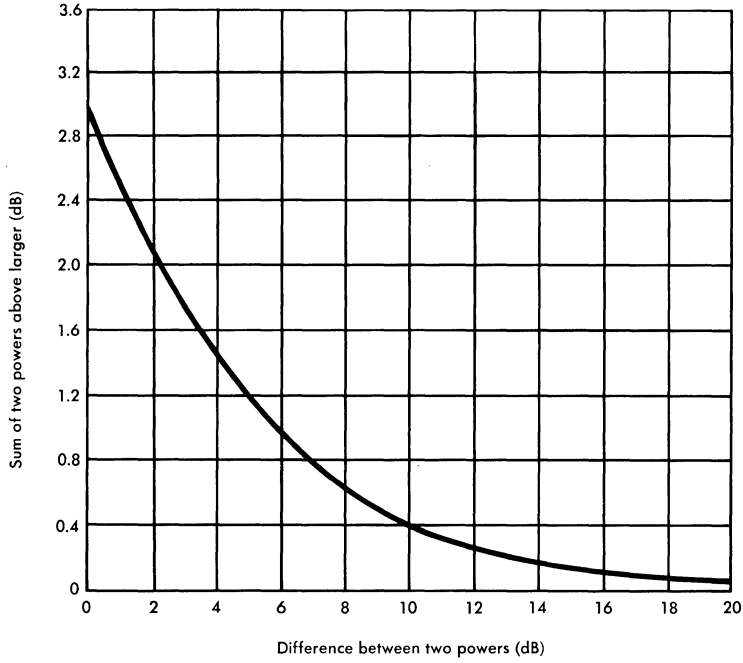


FIG. 2-8. Sum of two powers expressed in dB.

Occasionally, it is necessary to determine the power of the sum of two voltage or current waveforms. Obviously, if the two waveforms are uncorrelated, the total power is simply the sum of the powers of the components. However, if the waveforms are correlated, the sum can lie anywhere between zero and 3 dB more than the power addition result. For example, consider two waveforms of equal frequency and amplitude, but 180 degrees out of phase. The power of the sum of the waveforms is obviously zero. If instead, the two waveforms are exactly in phase, the resulting voltage is exactly doubled and the power quadrupled. Such addition is called in-phase or voltage addition.

To emphasize the voltage summation, Eq. (2-27) may be replaced by

$$P = P_1 \,{}^{\prime\prime\prime}{+}^{\prime\prime\prime}\, P_2 \tag{2-32}$$

to serve as a reminder that, in Eqs. (2-28) and (2-31), the value 10 is replaced by 20. Otherwise, the same procedure is followed and the curve of Fig. 2-9 is developed.

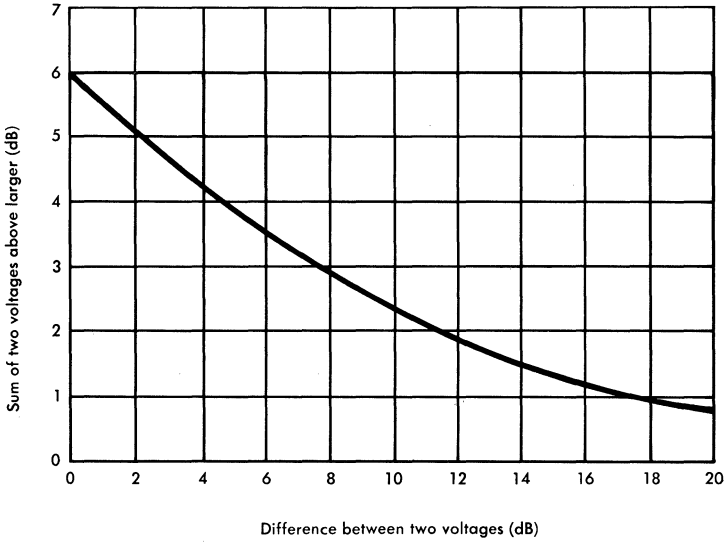


FIG. 2-9. Sum of two voltages expressed in dB.

There are many situations where an adequate estimate of D can be made by recalling that

$$\log (1 + x) = 0.4343 \left(x - \frac{x^2}{2} + \frac{x^3}{3} - \cdots\right) \tag{2-33}$$

and that for small x, terms beyond the first can be neglected. For example, in summing the in-phase voltages E_1 and E_2 with $E_1 > E_2$,

$$D = 20 \log (1 + E_2/E_1) \approx 8.686\, E_2/E_1 \tag{2-34}$$

If the difference between E_1 and E_2 is 20 dB, then the difference between their sum and E_1 is about 0.9 dB.

REFERENCES

1. Llewellyn, F. B. "Some Fundamental Properties of Transmission Systems," *Proc. IRE*, vol. 40 (March 1952), pp. 271-283.
2. Chinn, H. A., D. K. Gannett, and R. M. Morris. "A New Standard Volume Indicator and Reference Level," *Bell System Tech. J.*, vol. 19 (Jan. 1940), pp. 94-137.
3. Cochran, W. T. and D. A. Lewinski. "A New Measuring Set for Message Circuit Noise," *Bell System Tech. J.*, vol. 39 (July 1960), pp. 911-932.
4. Coolidge, O. H. and G. C. Reier. "An Appraisal of Received Telephone Speech Volume," *Bell System Tech. J.*, vol. 38 (May 1959), pp. 877-897.

Chapter 3

The Message Channel

The basic building block of the telephone transmission system is the message channel. In the broad, classical sense, a message is considered to be the entity to be transmitted from source to ultimate destination, whether this entity be speech, pictures, written word, or data [1]. In telephone terminology, the word *message* has often been used to denote speech and to distinguish speech channels from other types such as teletype, program, or video. In this book, a *message channel* is defined as a standard voice-frequency telephone channel whose performance requirements are primarily determined by the necessity of providing adequate telephone service, although other voice-frequency messages such as voiceband data and signaling must also be considered.

3.1 NATURE OF THE MESSAGE CHANNEL SIGNAL

Basically, the message channel is capable of satisfactorily carrying any of the following messages:

1. Telephone speech signal
2. Voice-frequency signaling
3. Voiceband data

The telephone speech signal is, of course, the most commonly encountered.

The Telephone Speech Signal

The telephone set, described in Chap. 4, serves as the transducer for converting the acoustic energy from the customer to an electrical signal which can be transmitted along wires. Generally, a direct

current is sent from the telephone office over the wires leading to a customer's set. Changes in the acoustic pressure cause changes in the resistance of the carbon transmitter which in turn modulates the direct current. The signal delivered to a telephone office consists of a direct current (which indicates off-hook conditions) modulated at an audio rate. It is this modulated current which is commonly referred to as the telephone speech signal.

The telephone speech signal at the central office has most of its energy concentrated in a band of frequencies from about 100 Hz to 5 kHz. This is a result of both the characteristic of the human voice and the bandlimiting introduced by the telephone set and loop. This bandwidth is much more than is needed for intelligibility, and it is advantageous to further limit the bandwidth to improve performance in the presence of interference and noise. The optimum trade-off between economics and quality of transmission generally occurs when the telephone speech signal is bandlimited to the range of 200 to 3300 Hz. Thus, it can usually be assumed that the telephone speech signal has all of its significant energy contained in this frequency band.

The time-varying waveform associated with the speech signal is not as easy to characterize. The audio frequencies making up the basic speech signal are amplitude modulated at a syllabic rate (several times per second). In addition, the speaker's pauses between phrases and sentences result in the speech energy being concentrated in "talkspurts" of about 1 second average duration separated by gaps of a second or so. Thus, the speech signal consists of randomly spaced bursts of energy of random duration. As a consequence, accurate measure of the speech signal is difficult at best.

Regardless of these problems, the magnitude of the telephone speech signal must be measured and characterized in some fashion which will be useful in designing and operating transmission systems. As discussed in Chap. 2, the vu meter was developed for this purpose. The vu meter is designed to measure the approximate rms voltage averaged over a syllabic interval. As a consequence, the meter gives no indication of the activity of the talker but only tells how loud he is when he talks. A talker who pauses a lot may have a higher volume in vu than one who talks endlessly, yet the endless talker may have a higher average power. Relating volumes in vu to average power is usually done by first converting a continuous talker from vu to average dBm by the relation:

$$\text{Average power} = \text{vu} - 1.4 \quad \text{dBm} \tag{3-1}$$

where the 1.4-dB conversion factor is the result of empirical tests with a variety of talkers reading text. The average power of a telephone talker who listens part time is related to that of a continuous talker by introducing a load activity factor, τ_L, to obtain:

$$\text{Average power} = \text{vu} - 1.4 + 10 \log \tau_L \quad \text{dBm} \qquad (3\text{-}2)$$

The peak factor for speech, defined as the ratio of peak to average power, is relatively high and is a function of talker activity. It has been empirically determined that the peak factor for a typical continuous talker is approximately 19 dB. For a talker of lower activity, peak magnitudes are not affected, but average power is lowered by the load activity factor, τ_L.

Talker volumes in the telephone plant have been found to be normally distributed in vu with a mean between −14 and −25 vu at 0 TLP and a standard deviation between 4 and 6.5 dB depending upon the geographical area and type of talker. The telephone load activity factor, τ_L, is also subject to some variations. A value of 0.25 has been traditionally used for a typical load activity factor.

From these considerations, it would not be unusual to expect a transmission system to accommodate a −40 vu talker having an average power of −41.4 dBm and not distort on a +10 dBm peak due to a louder talker. Signal-to-noise ratios, which are often used as a performance criterion for communication systems, are usually of very limited use in characterizing a telephone transmission channel. This is primarily due to two factors. As has already been discussed, signal power can fluctuate widely so that the signal-to-noise ratio is far from constant. Then too, subjective tests have shown that noise or any disturbance is most annoying on a telephone channel during the quiet intervals when no one is talking. It has become standard in the Bell System to specify the maximum noise power allowed in a message channel rather than to use signal-to-noise ratios. In the case of data signals, where the levels are much more closely controlled and noise during "quiet" intervals is not so important, signal-to-noise ratios could be used advantageously. However, the levels are such that the proper signal-to-noise ratio results if absolute noise power is controlled.

Voice-Frequency Signaling

In addition to the speech signal, a telephone transmission system must also pass special supervisory signals to the far end. In the

transmission loop between a central office and the customer, this signaling takes the form of direct currents for supervision (on- or off-hook) and for addressing (dial pulsing), and 20-Hz alternating currents for ringing. This information at the central office may be transferred to special signaling leads called E and M leads. The d-c state of the M lead at the originating station determines the state of the E lead at the called station and vice versa. Although a separate channel could be established for signaling (and common channel signaling may be very attractive in the future), it has been found most convenient to send this information through the standard voice message channel. Since the channel has a low-frequency cutoff in the vicinity of 200 Hz, the d-c state of the E and M leads must be converted. The Single Frequency Signaling System (SF) is the standard supervisory system for carrier transmission systems.

SF Supervision. The SF system uses the M lead status to key a 2600-Hz sinusoid on the line. The presence of this wave signifies an on-hook condition, whereas its absence indicates the off-hook condition. This means that an idle message circuit presents a 2600-Hz sinusoid on each direction of the transmission system. The level of each of these waves is nominally —20 dBm0 so that a 1000 channel system carrying only idle circuits must handle 10 dBm0 of idle power in each direction. It is essential that consideration of these supervisory signals be made part of the system design requirements.

Addressing. The address or number of the called party is generated by the customer at the telephone. This can be done in a variety of ways including dictating the number to an operator, dialing, and TOUCH-TONE®. With the centralized automatic message accounting (CAMA) system, the address may also include the calling number.

Dialing of a telephone interrupts the loop current between the central office and the customer. In some types of offices, this is used to pulse the M lead and in turn pulse the 2600-Hz tone. When this dial pulsing is used on toll facilities, the level of the tone when present is nominally —12 dBm0.

A more common means of sending address information over toll facilities is to convert the dial pulsing or TOUCH-TONE signals into the Multifrequency Keypulse System (MF or MFKP). This system uses six audio frequencies (700, 900, 1100, 1300, 1500, and 1700 Hz) to transmit the address information. The system operates at seven digits per second with each digit transmitted as a 68-millisecond burst of two of the six frequencies. These signals are transmitted at —6 dBm0 for each frequency or —3 dBm0 for the pair.

Such high levels are tolerable because of their short duration. In order to utilize expensive trunks efficiently, the MF address is usually transmitted in a 1 to 1-1/2 second burst. This means that the customer dialed number is stored at the office until completely dialed and then transmitted. Similarly, operator dialed calls are placed on MFKP buttons and transmitted when the operator signifies the address is complete.

The Voiceband Data Signal

The growth of data signals in the telephone plant has been significant in the past several years and is expected to become even more important in the future. There are two approaches to the problem of successful data communication. The first is to design a data transmission plan that is compatible with existing transmission facilities. This is a problem of optimum data set design rather than system design and is not pursued here. The other approach is the placement of additional restraints on a communication channel for data communications. If carried to extremes, this second approach would lead to the requirements of ideal transmission channels. As a practical matter, the degree of "tailoring" of voice message circuits to accommodate data must be severely limited by economics. However, the requirements on transmission systems have been upgraded continuously to reflect improvements in both telephone service and data service. Thus, new requirements have been devised in such a way as to optimize the channel performance for both data and voice telephone transmission.

An example is the case of impulse noise in a telephone system. Such impulses are commonly encountered near switching offices due to operation of relays and may be coupled into transmission systems. The effect of such impulse noise in voice telephone circuits is the introduction of "clicks" which are subjectively acceptable unless the power of the impulse becomes significant. In a data circuit, however, the presence of impulses results in errors which are rather serious even if the average power of the impulse train is low due to the low duty cycle of the impulses. Obviously, if voice circuits are to be compatible with data transmission, they must meet impulse noise requirements for data.

Another example is the restriction on waveform distortion for data transmission. As will be seen in Chap. 5, the demodulation of an SSB

signal requires reinsertion of the carrier at the exact modulating carrier frequency. Any phase or frequency error results in quadrature distortion of the signal waveform. Such waveform distortion has little effect on the speech signal but is serious for data transmission. In the interest of making voice circuits more compatible with data transmission, the reinserted carrier frequency stability requirement for new multiplex equipment has been tightened considerably. Although the net result is an improvement of the quality of the circuit, the subjective improvement on voice circuits is smaller than would justify the additional design effort required. However, existing technology in phase-locked oscillators makes the additional effort well worthwhile for the improvements in data transmission, even if data presently makes up a small percentage of the total system load. In general, improvements made in the telephone transmission system to accommodate data are limited by economic and practical considerations.

Types of Voiceband Data. It is necessary when designing systems to have knowledge of the waveforms of data signals that are commonly encountered. The basic digital data signal consists of a train of pulses which represent in coded form the data to be transmitted. This basic waveform consists of frequency components from direct current to some reasonably high frequency determined by the "sharpness" of the pulse edges. To transmit this information on a telephone channel requires that this bandwidth be confined to the range of a message channel, i.e., 200 to 3300 Hz, and that the data signal tolerate significant phase distortion near the band edges. The high-frequency components are controlled by suitable pulse shaping and by restricting the maximum pulse repetition frequency. The low-frequency components are transmitted by modulating a voice-frequency carrier with the data signal.

A simple, but not often used, form of data transmission is amplitude modulation of an audio carrier as embodied in an on-off system for transmitting binary data (n-ary data can be transmitted with n different transmission levels allowed for the carrier). The peak power of such a signal is 3 dB above the highest allowed carrier power. The long-term average power is dependent on the probability of the various pulse levels.

A more common means of converting a data signal for transmission on a message channel is to shift the frequency of an audio carrier by the digital signal. This approach, called frequency shift keying, is used by the 202-type data set to generate a binary signal with 1200 Hz corresponding to "mark" and 2200 Hz corresponding to

"space." The standard power of the carrier when present in such systems is —13.0 dBm0. The peak values of such signals are, of course, only 3 dB higher than the average value so that data peaks are of considerably lower amplitude than voice signal peaks.

Another means of preparing the data signal for transmission is to shift the phase of the audio frequency carrier. Such a plan is generally more efficient than amplitude or frequency modulation and thus is often used to provide higher capacity on message channels. The 201 family of data sets uses a four-phase modulation method. Basically, carrier amplitudes are the same as discussed for frequency-shift systems, and the previous comments apply.

Narrowband Data. The use of teletype (TWX) and telegraph requires the transmission of digital data at relatively low speeds (less than a few hundred bits per second). For these applications, multiplex systems have been developed whereby several narrowband data signals share a single message channel. Constraints must be placed on the multiplexing details to assure that the multiplexed narrowband data signal approximately resembles a typical voice signal regarding average power, single frequency energy, and the distribution of energy across the message channel bandwidth.

For example, the average power on each of N narrowband channels being multiplexed on a message channel must be 10 log N dB less than that allowed for a message channel. Similarly, wideband data occupying N message channels must also approximately satisfy message channel constraints with an allowable total average power of 10 log N dB greater than that allowed for a single message channel but without concentration of the energy at a single frequency.

3.2 MESSAGE CHANNEL OBJECTIVES

The telephone speech signal is usually delivered to the telephone company in the form of audible sounds impinging on the telephone transmitter. It is the telephone company's responsibility to deliver a replica of this sound to the ear of the called customer. How well a customer can talk and hear over a channel and his opinion of the grade of transmission, will depend on:

1. Received acoustic speech pressure, which is a function of the efficiency of the transmitter and receiver and of the electrical loss between them, as well as of the acoustic speech pressure of the talker.

2. The amount and character of the noise introduced.
3. The frequency response, bandwidth, amplitude distortion, and (to a small extent) phase or delay distortion.
4. The magnitude and delay of the echo.
5. The crosstalk heard, especially crosstalk which is either intelligible or seems nearly so.

There are imperfections other than those listed which should be considered in a study of message channel objectives. Tones of various frequencies and character, rapid gain and phase changes, and clicks, for example, impair message transmission. Furthermore, other special services such as telegraph, program, and telephotograph impose additional requirements beyond those set by voice telephone service.

The relationship between message channel performance objectives and design requirements for a specific transmission system is not as simple as it seems on first examination. One of the reasons for this is that the objectives are for the extensive Bell Telephone System while the largest subunit of this system to be designed at any one time is a particular transmission system or facility. Thus, the overall objectives must be allocated to particular parts of the system with the particular allocation plan determined by both economic and feasibility factors. This allocation is further complicated by the fact that the Bell System objectives are constantly being upgraded. It is important that any system design be compatible with these objectives for the life of the system. Thus, performance requirements for a specific system may include allowances for anticipated changes in future objectives. Because of the wide scope of Bell System objectives, a significant segment of effort is devoted to setting and evaluating such objectives as well as relating them to specific system design requirements.

Grade of Service Concept

A concept known as *grade of service* is commonly used to determine acceptable message channel objectives. It combines the distribution of customer opinion with the distribution of plant performance to obtain the expected percentage of customer opinion in a given category. Grade of service is defined by the integral

$$\int_{-\infty}^{\infty} P(R \mid x)\, f(x)\, dx \tag{3-3}$$

where $P(R \mid x)$ is the probability distribution that a customer would place a given stimulus in a given opinion category, R, and $f(x)$ is the

probability density function of obtaining that stimulus. The distribution function, $P(R \mid x)$, is obtained through subjective tests which reveal typical customer reactions to various controlled levels of the given stimulus (which may be noise, volume, bandwidth, or any other stimulus for which objectives are to be obtained). The remaining density function, $f(x)$, characterizes overall system performance for a given stimulus. In setting equipment performance criteria, the problem must be worked backwards to determine an adequate $f(x)$ for a given grade of service. The grade of service for an existing system is generally obtained through the process shown in Fig. 3-1.

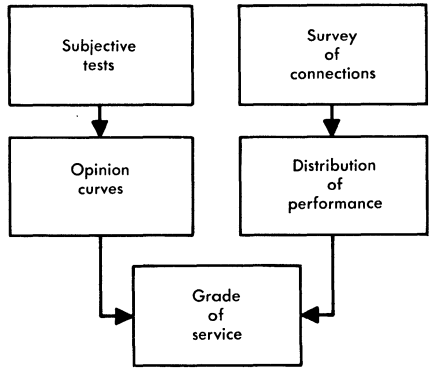


FIG. 3-1. Approach used in deriving the grade of service.

The evaluation of Eq. (3-3) is simplified if, as is often the case, the distribution $P(R \mid x)$ and density $f(x)$ are both normal. For example, on the basis of subjective testing, assume that for good or better service $P(R \mid x)$ is 0.5 when x equals P_0 (the mean) and has a standard deviation (due to differences in subjects) of σ_P. Assume further that the facility provides a stimulus of mean F_0 with a standard deviation of σ_F. What is the grade of service for good or better performance? In most practical cases, the stimulus will be expressed in logarithmic (dB) quantities and will be normally distributed in dB. For this example, it is assumed that the stimulus is such that the higher it is, the more undesirable (this would be true for noise or loss but not necessarily for talker volumes). Under these conditions, a simple technique may be used to calculate grade of service.

Assume that both $f(x)$ and $P(R \mid x)$ are normal in dB with means F_0 and P_0 and standard deviations σ_F and σ_P, respectively. Sub-

stitution of these distributions into Eq. (3-3) results in a double integral which must be evaluated to obtain grade of service. It can be shown that the resulting distribution of satisfaction, $Z(x)$, is also normally distributed in dB with a mean given by

$$Z_0 = F_0 - P_0 \tag{3-4}$$

and a standard deviation given by

$$\sigma_Z = \sqrt{\sigma_F^2 + \sigma_P^2} \tag{3-5}$$

An illustration of the probability of $Z(x)$ is given in Fig. 3-2.

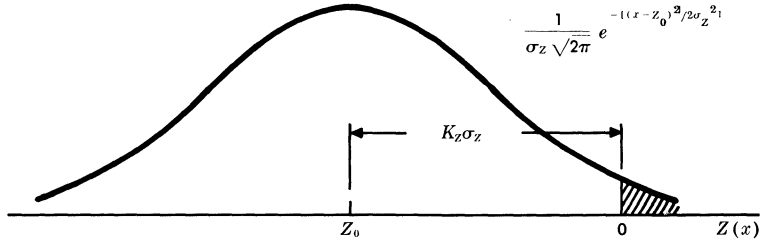


FIG. 3-2. Normal probability density function.

Clearly, when $Z(x)$ is less than zero, the stimulus is less than that which still gives good or better service. Thus, it is only for $Z(x)$ greater than zero that the stimulus is too great to give good or better service. Then, as indicated in Fig. 3-2

$$\text{Grade of service good or better} = 1 - \frac{1}{\sigma_Z\sqrt{2\pi}} \int_0^\infty e^{-\left[(x - Z_0)^2/2\sigma_Z^2\right]}\, dx \tag{3-6}$$

This integral is tabulated in normal tables for a given Z_0 and σ_Z. Generally, the telephone system goal is to provide a grade of service of about 95 per cent good or better, i.e., less than 5 per cent in the fair category and a negligible number in the poor category.

The implication so far is that grade of service can be related to the transmission performance for any *single* stimulus while all other possible degradations to transmission are ignored. The assumption

that these stimuli are independent of each other is, of course, not usually valid. Such interaction effects are usually minimized by holding all stimuli (other than the one under investigation) at typical or nominal values and changing the one stimulus under investigation. This is the basic approach used in subsequent discussions of the most important transmission stimuli, although often other stimuli must also be considered.

Received Volume

The basic transmission problem is to provide the proper signal magnitude at the receiver. This is achieved by controlling the loss of the telephone connection over reasonable limits and, of course, is a strong function of the design of the telephone station sets themselves. Expected ranges of such losses can be determined by a series of subjective tests. The results of one series of tests are shown in Fig. 3-3. The mean value for each transition is determined by reading the volume at the 50 per cent point. The standard deviation of the nearly normal distribution can be found by noting the difference (in vu) between the 50 per cent point and the 84 (or 16)

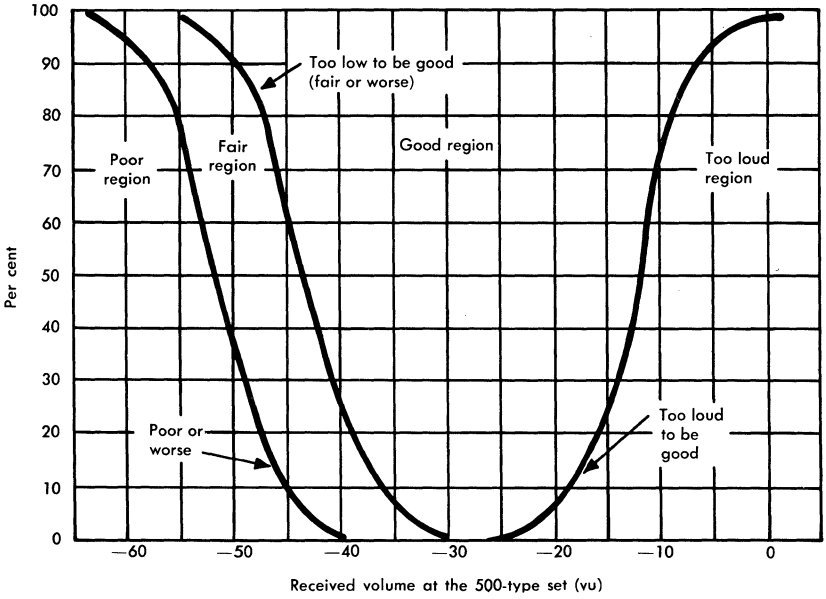


FIG. 3-3. Judgment of received volume.

per cent point. From Fig. 3-3, this amounts to about 5 vu (or 5 dB). From this data, an estimate of customer satisfaction for a particular distribution of received volumes over a number of calls can be made to determine the quality.

Noise

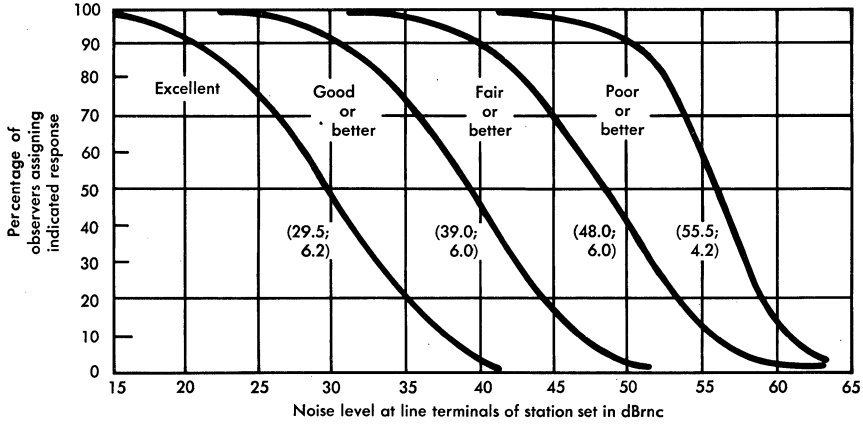
Noise, in the most general sense, is any signal present in a communication channel other than the wanted signal. Its effect will depend upon the receptor of the wanted signal, be it a human ear or a machine. In order to provide good service to all classes of service, it is necessary to measure the noise on a scale which reflects the amount of impairment introduced. In general, the impairment introduced in speech transmission is reflected in the short-term rms average value of the noise. The 3-type noise measuring set with C-message weighting was designed to measure noise in this manner. In general, the impairment introduced in data transmission is reflected in the peak values of noise above a given level. The 6-type noise measuring set was designed to count these peak values. Each of these types is discussed in turn.

Message Circuit Noise. The value of noise read on the 3-type noise measuring set is a normalized measure of the annoying effects of noise on speech transmission. Message circuit noise is defined as the short-term average noise level as measured with a 3-type noise measuring set, or its equivalent, using C-message frequency weighting. The relationship between the meter reading and customer opinion of the impairment was determined by subjective tests.

The customer-to-customer noise objectives are based on the approach discussed previously whereby subjective tests are performed to provide estimates of customer opinion of various levels of message circuit noise, and noise surveys are independently made to determine the noise performance of the plant. These results are combined to give insight to the present customer satisfaction in terms of grade of service and to indicate areas where an increase in customer satisfaction is necessary.

The subjective tests were conducted using 500-type telephone sets with constant received volume and varying noise which was a composite of power hum, switching office noise, and thermal noise [2]. The quality of the circuit with noise was judged as excellent, good, fair, poor, or unsatisfactory. The results of these tests plotted in cumulated categories are shown in Fig. 3-4. Presented in this way, the curves show the proportion of excellent, good or better, fair or

better, and poor or better judgments at the particular noise levels. A good model of opinion is obtained by fitting normal distribution functions to the data points. As such, each curve is defined by the 50 per cent point (the mean) and the standard deviation of a normally distributed random variable.



Notes:
1. Values in parentheses indicate average and standard deviation.
2. Received volume constant, —28 vu.

FIG. 3-4. Noise judgment curves.

A survey of noise performance of toll connections was also taken and is shown in Fig. 3-5 [3]. The result of this survey combined with subjective tests gives the grade of service, also shown in Fig. 3-5. This indicates that although the grade of service is an impressive 97 per cent good or better overall, it is only 88 per cent good or better in the long calls. The reason for the poorer performance in the long calls is that mean noise doubles with a doubling of airline distance. Since the number of calls made decreases rapidly with distance, these long distance calls, although individually important, have a small effect on the total grade of service calculation.

Carrier Objectives. The long-haul and short-haul carrier noise objectives were derived to satisfy the customer-to-customer requirements when these systems are switched together to form connections of 1000 to 4000 circuit miles. They specify not only the noise due to the facility portion of a trunk but also that due to a representative set of multiplex terminals.

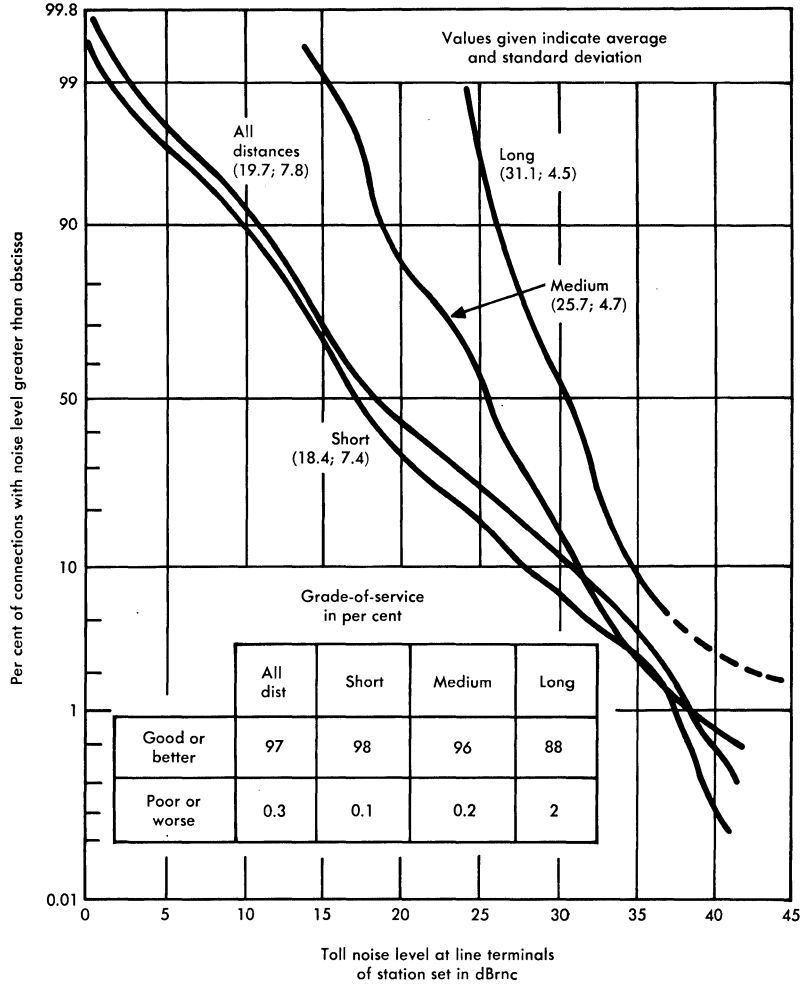


FIG. 3-5. Estimated toll noise distribution from 1962 connection survey data.

A short-haul carrier system is defined as one designed for use over distances less than 250 miles. A long-haul carrier system is defined as one designed to be used for distances greater than 250 miles. The optimum allocation of the total noise between short-haul and long-haul facilities is not directly proportional to length. Economic considerations for short-haul facilities make it desirable to allow more noise per mile on such facilities. Since the overall connection is very unlikely to include more than a few short-haul

facilities in tandem, the higher allowed noise per mile for the short distances does not seriously tighten the noise requirements on the long-haul facilities.

The long-haul and short-haul carrier facilities are assumed to have noise distributions which are normally distributed at any given route mileage. The mean values of these distributions are assumed to vary as a function of distance as shown in Fig. 3-6; the standard deviations are assumed to be a constant 4 dB at all distances.

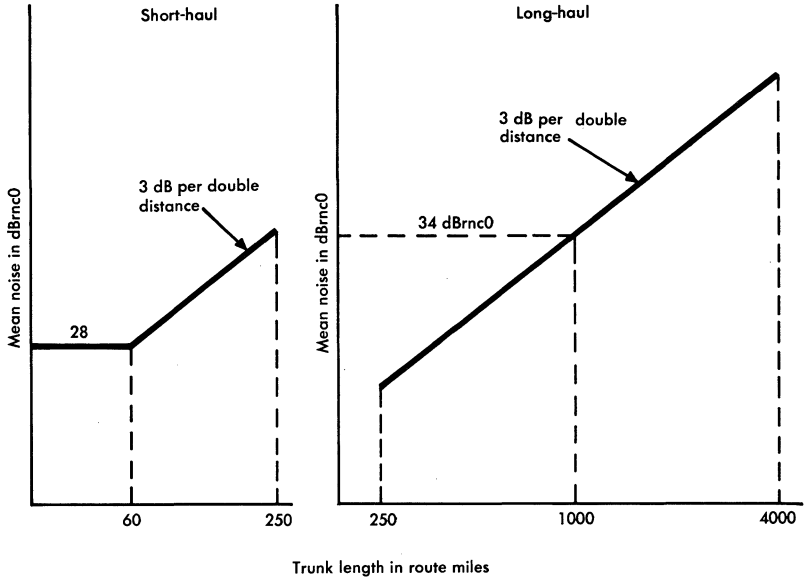


FIG. 3-6. Length dependence of the mean values of the carrier noise objectives (standard deviation = 4 dB for all lengths).

Mean values of the noise objective are chosen at 60 route miles and 1000 route miles for short-haul and long-haul carrier, respectively. This mean value was chosen for a grade of service of 95 per cent good or better and is 28 dBrncO for a 60 route-mile short-haul carrier system and 34 dBrncO for a 1000 route-mile long-haul carrier system. The mean noise for a system of other lengths is then obtainable from Fig. 3-6 and is directly proportional to length for all systems longer than 60 route miles.

The division of allowed noise between short-haul and long-haul systems was based upon the noise performance of the types of

systems presently under development and the economics of noise reduction in those systems. A different pairing of the objectives could also give the 95 per cent good or better grade of service but would not result in the most economical solution.

It should be noted that the objectives are stated at 60 and 1000 route miles rather than at the traditional maximum lengths of 250 and 4000 miles. The reason for this shift is to reduce the sensitivity of the objectives on the assumed slope of 3 dB per double distance. Some of the future systems, particularly PCM systems, do not exhibit a 3-dB noise increase with doubling distance. Thus, the new objectives based on more typical system lengths are much less sensitive to changes in noise slope with distance.

Impulse Noise. Impulse noise is defined as any burst of noise which exceeds the rms noise level by a given magnitude. This magnitude is nominally 12 dB for a 3-kHz bandwidth.

Impulse noise objectives are based primarily on the error susceptibility of data signals. This susceptibility depends upon the type of data set used and the characteristics of the transmission media. Data sets that employ different types of modulation, that operate at different bit rates, etc., will not all perform in the same manner when subjected to impulse noise. The important characteristics of the transmission media are the amount of delay and attenuation distortion, and the rms signal-to-peak impulse noise amplitude ratio. The impulse noise objectives are therefore determined by the Bell System data sets which are most susceptible to impulse noise and by the knowledge of other transmission impairments of the plant.

It is impractical to measure the exact peak amplitude of each noise pulse or to count the number that occur. Large numbers of detailed measurements have shown that certain bounds exist for the distributions of peak noise amplitudes in the ranges of interest. Studies have shown that expected digital error rates in the absence of other impairments are approximately proportional to the number of impulses which exceed the rms data signal by about 2 dB. The objectives are therefore stated in terms of the number of counts above a given threshold. Thus, if system objectives on loss, background noise, etc., are met, and if the impulse counts at the specified thresholds are within the limits given, the data transmission error rate should be small.

Voiceband Data. For the higher speed voiceband data sets, an error rate of 10^{-5} results from an average impulse rate of 1.5 per

minute. Allowances must be made for finite counting rates of practical impulse counters, and a sufficient measurement interval must be provided. Impulse noise threshold is therefore defined as that threshold at which the observed number of counts on a 6-type counter is 5 in 5 minutes (for trunk groups, etc.) or 15 in 15 minutes (for loops or single channels).

The sporadic nature of impulse noise requires that one of two conditions be met in measurements made to estimate it. Either a long measurement interval is needed or several similar channels must each be measured for a relatively shorter interval. This is the reason for requiring 15-minute measurements on loops but only 5-minute measurements on trunks.

The overall customer-to-customer objective is no more than 15 counts in 15 minutes on at least 80 per cent of all calls at a threshold 6 dB below the received signal. To meet this overall objective, it is necessary to establish the threshold for a loop measurement at 59 dBrnc0 for no more than 15 counts in 15 minutes. The trunk allocation is somewhat more involved since the threshold is dependent on trunk length, hence trunk loss. The basic objective to be applied to trunk groups is 5 counts or less in 5 minutes on at least 50 per cent of the trunks in each group at the thresholds tabulated in Fig. 3-7. In determining acceptable impulse noise, the relative performance of the various carrier systems is recognized.

Trunk length (miles)	Threshold level (dBrnc0)
0-125	58
125-1000	59
1000-2000	61
over 2000	64

FIG. 3-7. Impulse noise threshold for no more than 5 counts in 5 minutes on carrier trunk facilities.

Other Data Bandwidths. Impulse noise objectives for narrowband and wideband data are set in a manner similar to that used for voiceband data. The basic difference is in the bandwidth of the impulse counter. In all cases, the counter bandwidth corresponds to

that of the desired data channel. Because the impulse noise encountered in the telephone plant does not have a flat frequency spectrum, frequency corrections cannot be easily made for different bandwidths.

Frequency Response

Frequency response objectives are specified in the frequency domain and are usually broken into two parts: (1) limiting frequencies and (2) inband amplitude distortion. Limiting frequencies define the points at the edges of the transmission band where the loss relative to 1000 Hz is 10 dB or less. Inband amplitude distortion defines permissible deviations in the amplitude response from the 1000-Hz value.

Limiting Frequencies. From the standpoint of transmission quality, the bandwidth should not change appreciably for different connections. The major constraints on bandwidth are: (1) the 4-kHz channel spacing used in most carrier facilities, (2) the state of the art in filter design, and (3) cost. The limiting frequency objectives are a compromise among these factors.

The frequency response performance of trunks depends on the composite effect of carrier facilities in tandem, signaling equipment, terminating sets, office wiring, and other equipment which may be in a trunk.* As a result of the variability among trunks, there has been limited application of limiting frequency objectives. Where objectives have been expressed, they have typically indicated that the loss at 200 and 3300 Hz shall not exceed the 1000-Hz loss by more than 10 dB on some percentage of the trunks. Statistical control is usually maintained indirectly by the design objectives for the individual pieces of equipment. These objectives must be better than those for a complete trunk. For example, the bandwidth objective for the N3 system is 3 dB at 200 Hz and 3450 Hz [5].

Inband Distortion. Variations in the loss across the message channel must be controlled to give good quality voice transmission and reasonable error performance for data transmission. When a trunk is placed into service, circuit tests are frequently made to check the transmission at selected frequencies. The frequencies and acceptable limits depend on the type of facility.

*Frequency response distortions for Bell System intertoll trunks have been measured [4].

Echoes and Loss

To this point, delay has been ignored as a source of transmission impairment. Subjective tests have indicated that delay, as such, does not unduly impair conversation if the round-trip delay is less than 600 milliseconds. Delays of this magnitude do not occur in land-based plant but are significant in satellite systems. As discussed in the following, it is the effect of echoes in the presence of delay that introduces transmission impairment.

Echoes and Singing. An echo may be produced in a transmission system wherever there is an impedance discontinuity. In general, a telephone transmission system is composed of a number of transmission facilities in tandem with flexible switching possible between facilities. A typical end-to-end connection is shown in simplified form in Fig. 3-8. Note that the signal can encounter a significant impedance discontinuity at every switching office. However, it is economically feasible and highly desirable to design the trunk plant to avoid serious impedance discontinuities at these switches. This is accomplished by building out the office wiring and using suitable padding with the trunks. By these means, return losses of about 20 dB are obtained in class 4 and higher offices.

The final end office (class 5) presents a different problem. Here the customer loops present a large variation from one loop to the next. This is due to differences in loop lengths, wire gauges, environment, etc. As a consequence, it is economically unattractive to maintain an average return loss at the end office of more than 11 dB with a standard deviation of 3 dB over that portion of the voice-frequency spectrum which is most important from the echo standpoint, i.e., 500 to 2500 Hz. This impedance mismatch causes a reflection, or echo, of a talker's signal to be returned to him on the channel to which he is listening. Figure 3-8 illustrates such an echo path. If the echo is not delayed (i.e., if the connection is short enough), it is indistinguishable from sidetone and is not annoying. If, however, it is sufficiently delayed in making the round trip, it can be annoying and can interfere with the talker's normal process of speech.

An echo traveling in the opposite direction to the signal is often called a talker echo when it is heard by the talker. If the talker echo is again reflected so that it travels in the same direction as the desired signal, it is called a listener echo when it is heard by the listener. It has been observed that telephone circuits designed to limit talker echo usually provide adequate listener echo performance

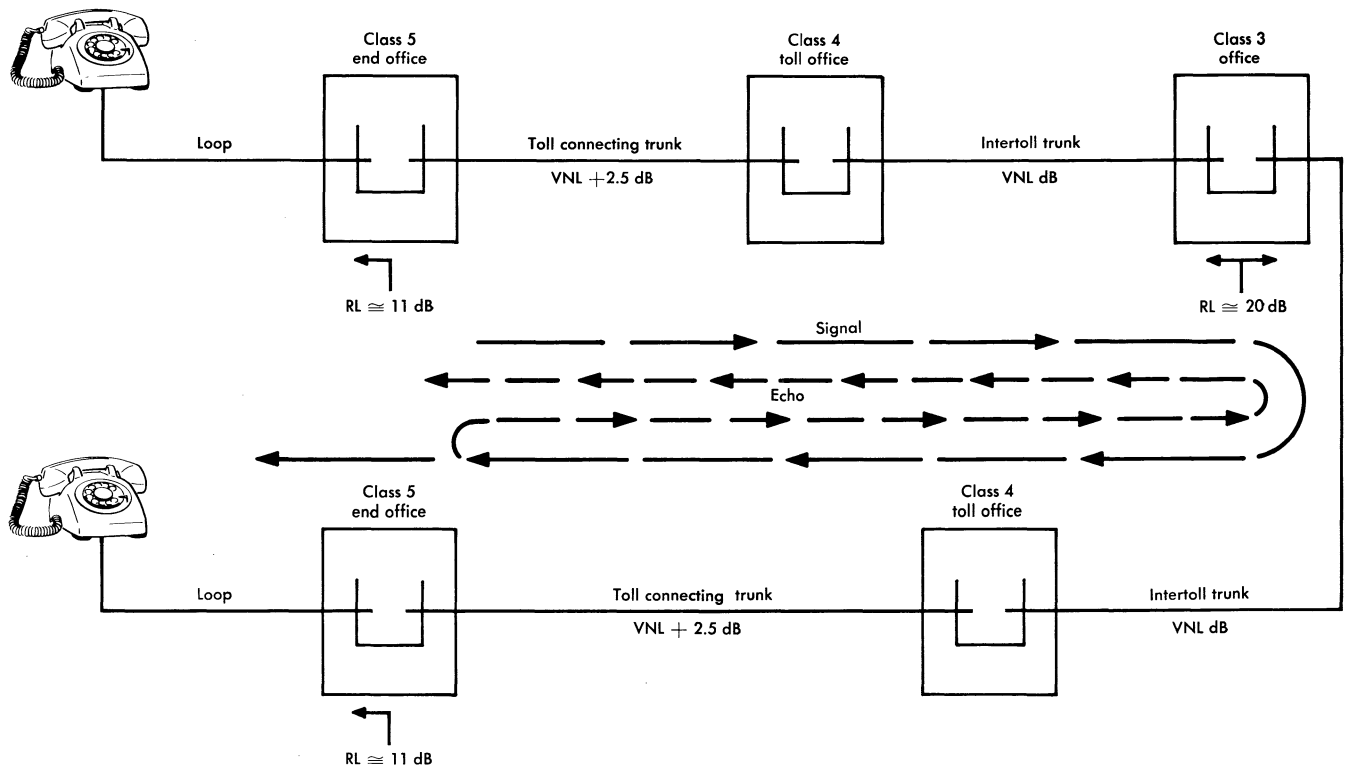


FIG. 3-8. Return losses and echo path in a toll connection.

for voice transmission. However, this is not necessarily true for data transmission, and return loss on some data loops must be improved by the use of impedance correcting networks.

At frequencies outside the 500- to 2500-Hz band, the return losses may be poorer than the 11 dB figure quoted previously. Experience indicates that the important frequency bands are from 250 to 500 Hz and from 2500 to 3200 Hz. At these frequencies, the return loss is usually degraded although circuit transmission is still appreciable. Multiple echoes at these frequencies, although not subjectively annoying, may lead to a sing or near-sing condition, especially under special termination conditions, e.g., when an operator's set is connected to the line. It is not sufficient to prevent singing—it is further necessary in a customer connection to provide enough margin to avoid the hollow or "rain-barrel" effect which is characteristic of circuits having the poor transient response associated with a near-sing condition. It is, therefore, necessary to provide some loss (of the order of 4-dB one-way loss) even with short toll lines where delay is negligible and echo is unimportant. In general, on longer circuits, echo objectives rather than singing protection determine the required loss.

There is no distinction between two-wire and four-wire portions of the system shown in Fig. 3-8. However, an echo is only annoying if it is heard. Thus, in a two-wire system, any reflections will result in an echo. On the other hand, a four-wire system will not transmit an echo unless it can appear in the opposite transmission path. Such appearances can only come about at terminating sets which convert two-wire to four-wire or vice versa. Thus, the terminating sets are often incorrectly blamed for the echo which is basically a result of poor return loss.

Echo Objectives. The magnitude of the echo that a talker hears will depend on the echo path loss, which is the sum of the return loss at the distant hybrid and the round-trip loss in the circuit. For the present, the losses of the loop at the talker's end of the connection will be ignored. As previously mentioned, however, the customer's tolerance of the echo depends not only on the echo magnitude, but also on the round-trip delay between echo and original signal. If the delay cannot be reduced and if return losses have been improved as much as economically possible by impedance balancing, the echo magnitude can be decreased by increasing the electrical loss between the talker and the point where the mismatch occurs, at the unavoidable cost of reducing received volumes.

The objective which has been selected is that the probability of customer dissatisfaction with the echo performance of a circuit shall be held to 1 per cent or less when echo is present. The value of loss which must be inserted to meet this objective for an echo of given delay can be determined by taking into account the following:

1. The average value and standard deviation of the customer's tolerance to echoes as a function of delay.
2. The average return loss producing the echo and the standard deviation of this distribution.
3. The deviations in toll trunk losses from an assigned nominal value.

A distribution resulting from the combination of these component distributions (which are nearly normal) makes it possible to select the average loss required to reduce to 1 per cent the chance of customer annoyance by echo (customer and circuit being selected at random).

The method of computing this average loss must begin with some knowledge of the results of subjective tests. Figure 3-9 gives the round-trip loss required by the average person as a function of delay [6].

Round-trip delay (ms)	Mean required loss (dB)
0	1.4
20	11.1
40	17.7
60	22.7
80	27.2
100	30.9

FIG. 3-9. Subjective reaction to echo delay.

It was also found that reaction to echo was normally distributed with a standard deviation of 2.5 dB for all delays. By combining this statistical distribution with those of return-loss and trunk loss deviations, it can be shown that the value of the overall loss which just meets the talker echo requirements is given by the following equation:

$$\text{Loss} = \frac{M_E(d) - M_{RL} + 2.33\sqrt{\sigma_E^2 + \sigma_{RL}^2 + N\sigma_L^2}}{2} \qquad (3\text{-}7)$$

where

$M_E(d)$ = echo tolerance in dB as a function of delay (Fig. 3-9)
M_{RL} = average return loss (11 dB)
σ_E = standard deviation of echo tolerance distribution (2.5 dB)
σ_{RL} = standard deviation of the return loss distribution (3.0 dB)
σ_L = standard deviation of the two-way loss per trunk (2.0 dB)
N = number of trunks

The values given after each parameter are those used in the original derivation of the overall loss.

From these values, curves of the required loss as a function of delay and number of trunks are obtained as shown in Fig. 3-10. The increase in required loss with number of trunks is compensation for the increased loss variability with increased number of trunks. In practice, a series of linear approximations to the curves of Fig. 3-10 is used. The approximate curve for a single trunk was derived by considering:

1. The need for increased loss at low delays to prevent near-singing.
2. The need for control of noise, crosstalk, and system loading.

The approximate curves for more than one trunk are derived by adding 0.4 dB for each additional trunk to the loss required for a single trunk. This loss is approximately the difference between the exact loss curves. The approximate curves are given by:

$$\text{Net loss} = 0.102 \times \text{round-trip delay} + 0.4 \times \text{number of trunks} + 4.0 \qquad (3\text{-}8)$$

This equation is used for connection round-trip delays of up to 45 milliseconds. Above that delay, the required loss may preclude adequate received volume, and the trunk is equipped with an echo suppressor.

Via Net Loss. It is desirable to assign the overall net loss so that each trunk of a connection operates at the lowest loss practicable considering its length and the type of facilities. The procedure is to assign half of the constant in Eq. (3-8) to each toll-connecting trunk (4 dB total on each connection)* and to assign the remainder of the

*The recent change in the method of defining the losses of toll-connecting trunks has necessitated assignment to the toll-connecting trunk of equipment having 0.5-dB loss, resulting in a design value equal to VNL + 2.5 dB. Since this equipment was previously assigned to the loop, the customer-to-customer loss has not changed.

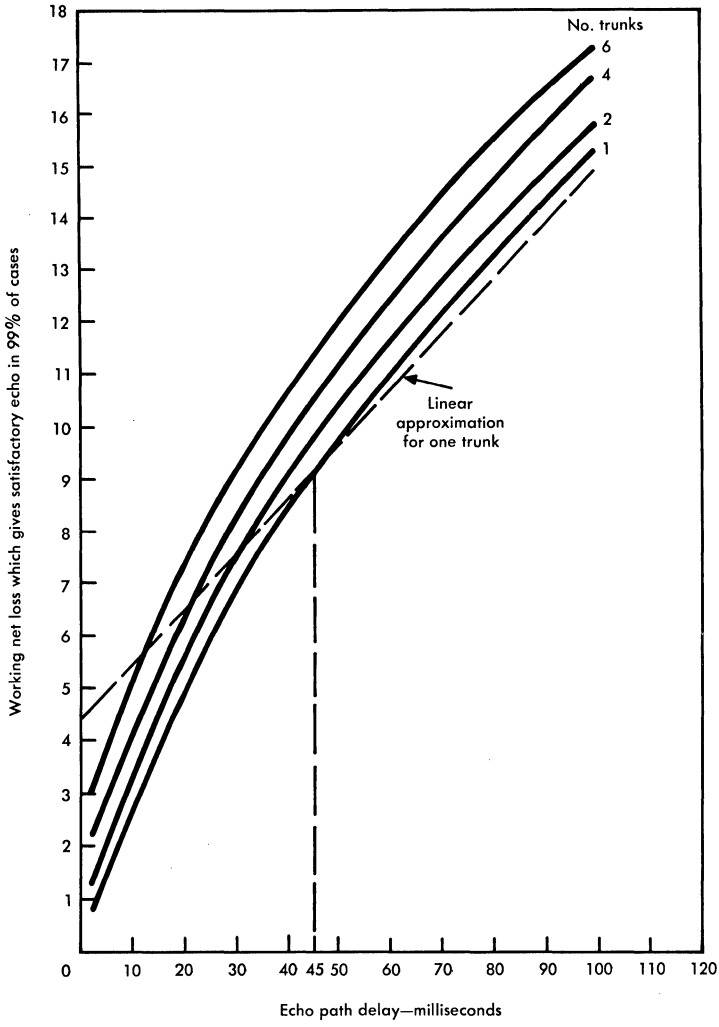


FIG. 3-10. Relationship between echo path delay and permissible working one-way loss for a trunk.

net loss to all trunks, including toll-connecting trunks, in proportion to the echo path delay of each trunk. This remainder is defined as via net loss and is commonly abbreviated VNL.

$$\text{VNL} = 0.102 \times \text{echo path delay of trunk} + 0.4 \text{ dB} \qquad (3\text{-}9)$$

Since the echo path delay of a circuit is related to the length of the circuit, this equation is usually given in terms of length and via net loss factor (VNLF) as:

$$\text{VNL} = \text{VNLF} \times \text{one-way distance} + 0.4 \text{ dB} \qquad (3\text{-}10)$$

where

$$\text{VNLF} = \frac{2 \times 0.102}{\text{velocity of propagation}}$$

The velocity of propagation must allow for the delay in an average number of terminals as well as that of the facilities. The accepted value for the VNLF for most carrier facilities is 0.0015 dB per mile.

Echo Suppressors. As discussed in Chap. 4, echo suppressors are used to reduce the amount of loss needed in a trunk. However, they are economically justifiable only in long-delay connections.

In the DDD network, the maximum round-trip delay within any regional center area is sufficiently short so that echo suppressors are not needed. However, there is a possibility of greater than 45-millisecond delay in calls routed between regional center areas. Thus, all regional center-to-regional center trunks are equipped with echo suppressors as are high-usage trunks which are over 1565 miles long or have a VNL calculation of 2.6 dB or greater.

Crosstalk

When a customer places a call, he expects that his conversation will not be overheard by others. The reception of intelligible crosstalk creates annoyance and violates privacy. Although nonintelligible crosstalk does not violate privacy, it is annoying; because of its syllabic nature, the customer thinks he could understand it if he were to listen intently. Thus, the amount of crosstalk has to be kept to a minimum.

Whether crosstalk is actually heard or not will depend upon many factors. These factors can be divided into two classes: (1) those which affect the volume and frequency of occurrence of the received crosstalk, and (2) those which affect the listener's ability to hear it.

The factors which affect the volume and frequency of occurrence of crosstalk can be best shown by considering a simplified pair of telephone circuits having crosstalk as shown in Fig. 3-11. A more realistic picture would contain more disturbing circuits; however, the factors would be the same. For this picture, the disturbed circuit

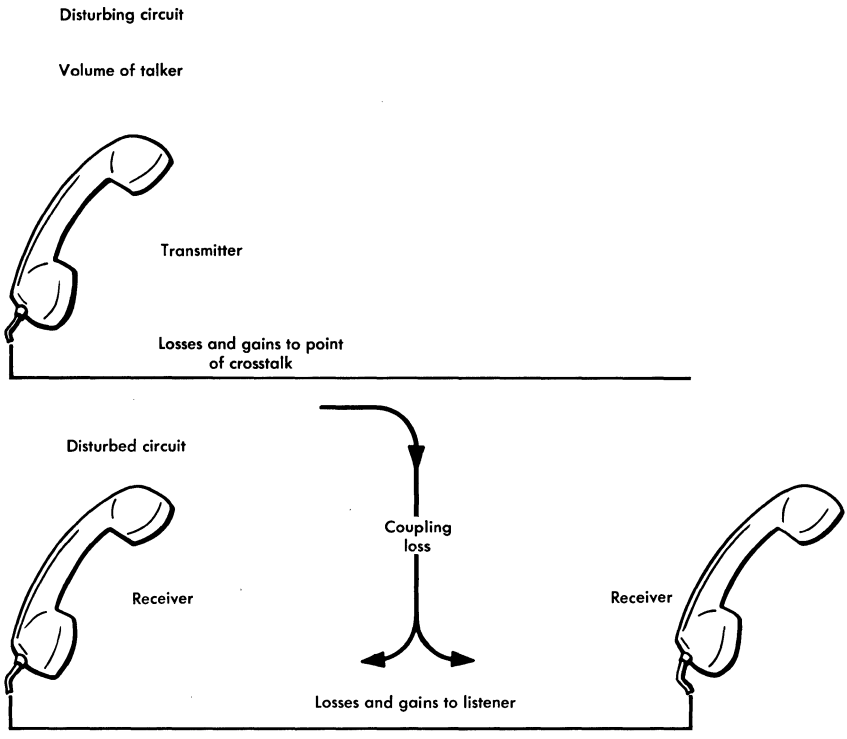


FIG. 3-11. Factors controlling the received crosstalk volume.

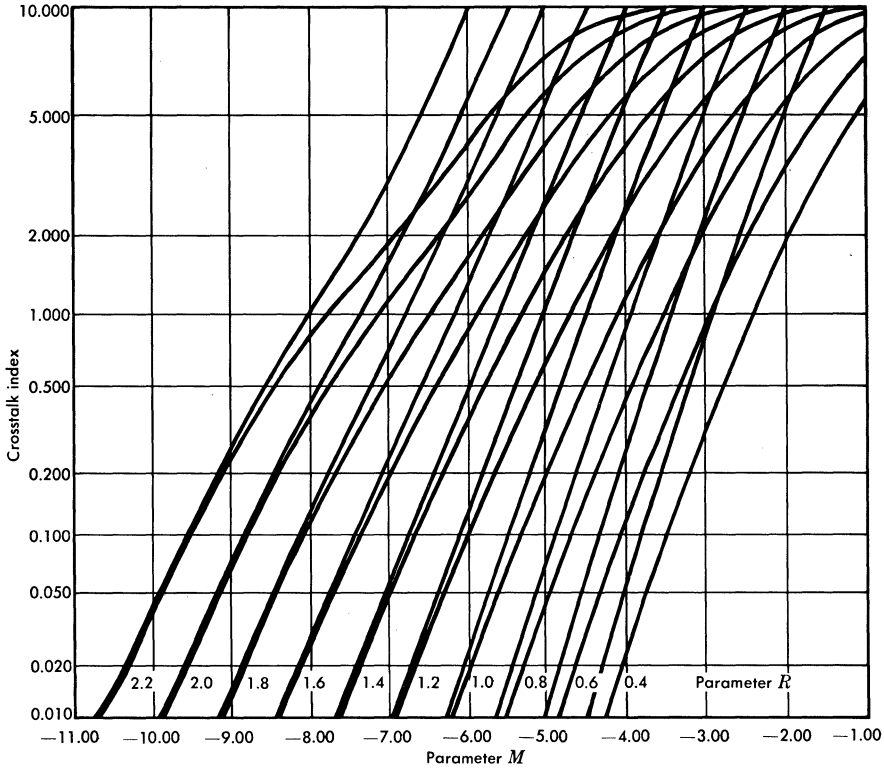
will receive crosstalk only if the disturbing circuit is active. The crosstalk received during this time will depend upon the volume of the disturbing talker, the loss to the point of crosstalk, the coupling loss between the two circuits, and the loss from the point of crosstalk to the listener. The ability of a customer to hear a given magnitude of crosstalk will depend upon his listening acuity in the presence of noise.*

Only one of these factors is uniquely associated with crosstalk and is controllable by the designer. This factor is the coupling loss. Thus, the crosstalk objectives are designed to restrict the coupling losses in a system to an amount that would limit the received crosstalk to a tolerably low amount.

*Implicit in the measurement of the crosstalk volume and listener acuity is the efficiency of the telephone set transmitter and receiver.

Trunk Objectives. The trunk crosstalk objectives are based on considerations of the effect of intelligible crosstalk. Very little information is known concerning the annoyance effects of nonintelligible crosstalk. It is suggested that the intelligible crosstalk objective be used for nonintelligible crosstalk except in two special cases, "babble" and "staggering advantage," where several dB more crosstalk can be tolerated.

In placing a call, a customer is assigned a trunk randomly. Thus, the event of a listener hearing intelligible crosstalk is a random event with a certain probability of occurrence. This probability or crosstalk index can be calculated by assuming values for the distributions of factors given in the previous paragraphs. The objective is a crosstalk index of 1 (1 per cent) for intertoll trunks and an index of 0.5 for all other trunks.



Note: $\tau_l = 0.25$, and for each value of R, the parameter B assumes from left to right the value 0.5 and 1.0.

FIG. 3-12. Generalized crosstalk index chart (100 disturbers).

Generalized crosstalk index charts provide a graphical method for determining the crosstalk index from the coupling loss for any set of values for the distributions of the factors given previously. They were obtained by assuming that each of the parameters was normally distributed, and then showing that the defining integral for the crosstalk index was a function of five parameters. Two of these parameters are the activity factor and the number of disturbers. It was computationally convenient in making these charts to assume that the activity factor was a constant 0.25 and to make a separate chart for each value of the number of disturbers. For example, the crosstalk index chart for 100 disturbers is shown in Fig. 3-12. The remaining three parameters of the chart have no physical significance and have been given the symbols M, R, and B. They are related to the mean and variance of the various parameters. Charts are available for different values of the number of disturbers and activity factors.

Use of the Generalized Crosstalk Charts. As an example of the use of these charts, the mean near-end coupling loss is calculated using the following set of values* to yield a crosstalk index of 1 for 100 disturbers:

	Mean	Sigma
Talker volume at class 5 office	$M_{TV} = -16.8$ vu	$\sigma_{TV} = 6.4$ dB
Loss of circuit, class 5 to class 5	$M_l = 8.0$ dB	$\sigma_l = 3.0$ dB
Coupling loss	$M_{Cl} =$ (to be determined)	$\sigma_{Cl} = 4.5$ dB
Equivalent loop loss	$M_{l0} = 2.1$ dB	$\sigma_{l0} = 2.2$ dB
Noise at listener's telephone set	20.0 dBrnc	3.0 dB
Equivalent loop noise	$M_N = 20.5$ dBrnc	$\sigma_N = 3.0$ dB
Listener acuity without noise	$M_{INT} = -80.3$ vu	$\sigma_{INT} = 4.5$ dB
Disturbing circuit activity factor	$\tau_L = 0.25$	

*These values have been chosen as an example and should not be thought of as a representative set of values for an actual problem.

In order to determine the parameters M, R, and B, the following formulas are needed:

$$M = \frac{M_v - M_I}{\sigma_I}$$

$$R = \frac{\sigma_v}{\sigma_I}$$

$$B = \frac{5}{\sigma_v}$$

where

$$M_v = M_{TV} - M_{l1} - M_{Cl} - M_{l2} - M_{l0}$$

$$M_I = M_{INT} + M_N - 6.0$$

$$\sigma_I^2 = \sigma_{INT}^2 + \sigma_N^2$$

$$\sigma_v^2 = \sigma_{TV}^2 + \sigma_{l1}^2 + \sigma_{Cl}^2 + \sigma_{l2}^2 + \sigma_{l0}^2$$

By noting that for near-end crosstalk, $\sigma_{l1}^2 + \sigma_{l2}^2$ equals the variance of the total loss (3 dB), the values for R and B are:

$$R = \frac{\sqrt{(6.4)^2 + (3.0)^2 + (4.5)^2 + (2.2)^2}}{\sqrt{(4.5)^2 + (3.0)^2}} = 1.60$$

$$B = \frac{5}{\sqrt{(6.4)^2 + (3.0)^2 + (4.5)^2 + (2.2)^2}} = 0.576$$

With a crosstalk index of 1.0 and these values, Fig. 3-12 gives a value of -6.13 for M. Since for near-end crosstalk $M_{l1} + M_{l2}$ equals M_l, the total loss of the circuit, the mean coupling loss is

$$M_{Cl} = -\sigma_1 M - M_{INT} - M_N + M_{TV} - M_l - M_{l0} + 6.0$$

$$= -(5.40)(-6.13) - (-80.3) - 20.5 + (-16.8)$$

$$- 8.0 - 2.1 + 6.0$$

$$= 72.0 \text{ dB}$$

Means of relating specific facilities to a coupling loss are covered in greater detail in Chap. 11.

REFERENCES

1. Shannon, C. E. "Communication in the Presence of Noise," *Proc. IRE*, vol. 37 (Jan. 1949), pp. 10-21.
2. Lewinski, D. A. "A New Objective for Message Circuit Noise," *Bell System Tech. J.*, vol. 43 (Mar. 1964), pp. 719-740.
3. Nasell, I. "The 1962 Survey of Noise and Loss on Toll Connections," *Bell System Tech. J.*, vol. 43 (Mar. 1964), pp. 697-718.
4. Nasell, I., C. R. Ellison, Jr., and R. Holmstrom. "The Transmission Performance of Bell System Intertoll Trunks," *Bell System Tech. J.*, vol. 43 (Oct. 1968), pp. 1561-1613.
5. Bleisch, G. W. and C. W. Irby. "N3 Carrier System: Objectives and Transmission Features," *Bell System Tech. J.*, vol. 45 (July-Aug. 1966), pp. 767-799.
6. Huntly, H. R. "Transmission Design of Intertoll Telephone Trunks," *Bell System Tech. J.*, vol. 32 (Sept. 1953), pp. 1019-1036.

Chapter 4

Voice-Frequency Transmission

The generation, transmission, and reception of voice-frequency telephone signals are the oldest and most basic processes in the practice of telephone transmission. The evolution of methods and theory has provided a vast body of knowledge covering in great detail the solutions to many of the problems encountered. This chapter describes the essential parts of the voice-frequency plant, the important problems encountered in voice-frequency system design, and the solutions that are being applied to some of these problems.

The performance and limitations of the elements of the voice-frequency telephone plant are gaining a new importance that goes beyond the simple fact that they form a part of every telephone connection. New services, such as the transmission of data signals and the application of wideband transmission to existing plant, along with the need for improved transmission performance, require increased understanding of the voice-frequency plant.

4.1 TELEPHONE SET

The telephone set is composed of a transmitter, a receiver, an electrical network for equalization, and associated circuitry to control sidetone and to connect power and signaling. In the present day telephone transmitter, granules of carbon are held between two electrodes—one, a cup holding the granules and the other, a diaphragm. The contact resistance between the granules is changed by varying the sound pressure on the diaphragm. The resulting resistance variation modulates a battery current flowing between the electrodes, thereby translating the acoustic message into an electrical signal.

In the telephone receiver, the varying component of this current passes through a winding on a permanent magnet. The alternate strengthening and weakening of the magnetic field causes the diaphragm to vibrate; this generates sound waves corresponding to those delivered to the transmitter by the talker.

The transmission circuit of the telephone set [1] must separate the transmitter and receiver circuits to limit the amount of the talker's signal appearing in his own receiver (sidetone) and to block the direct current in the transmitter from the receiver. Subjective tests have shown that some coupling must be allowed between the transmitter and receiver to provide a controlled amount of sidetone. Too much sidetone causes the talker to lower his voice, thereby reducing the volume which the listener receives; too little sidetone makes telephone conversation seem unnatural and tends to cause people to talk too loudly. A common-battery anti-sidetone circuit for accomplishing this purpose is shown in Fig. 4-1. The three-winding transformer and the sidetone balancing network form a hybrid which places the transmitter in conjugate with the receiver. Capacitors in the balancing network prevent the direct current flowing in the transmitter from appearing in the receiver.

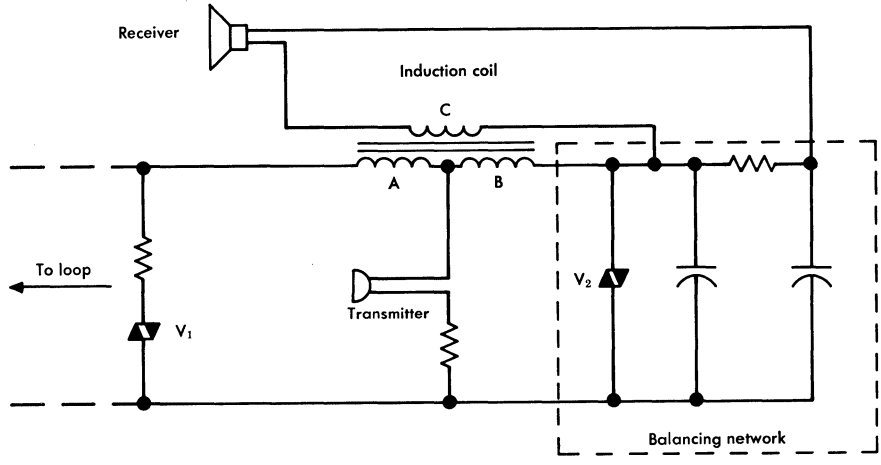


FIG. 4-1. Diagram of transmission circuit of 500D-type telephone set illustrating sidetone balance.

Figure 4-1 is a functional schematic drawn to illustrate the principle of the anti-sidetone circuit. The schematic circuit of the 500D-type telephone set is shown in Fig. 4-2. When the handset is on its

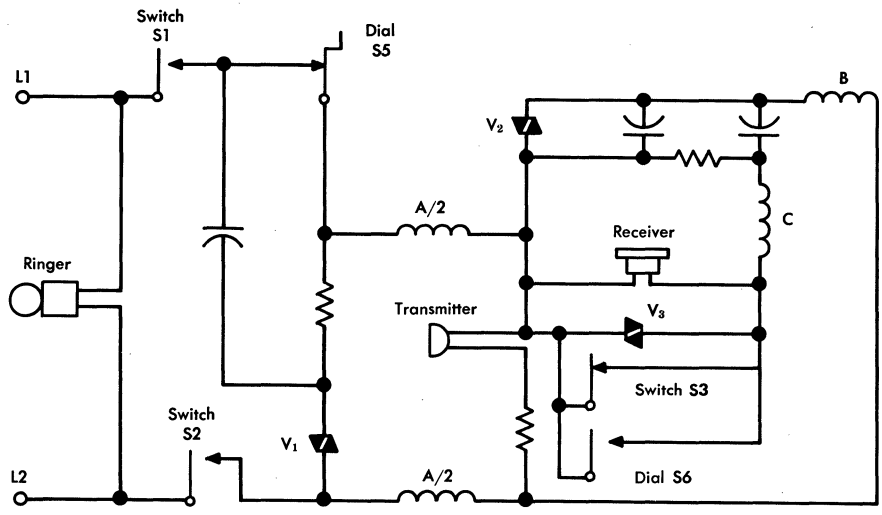


FIG. 4-2. Schematic diagram of 500D-type telephone set.

mounting, switchhook contacts S1 and S2 are open, and S3 is closed to protect the transmitter and receiver from ringing currents. Removal of the handset allows direct current from the central office (or a local battery) to pass through the transmitter and removes the short circuit from across the receiver. Dial contact S5 interrupts the battery current to form the dial pulses required to control the central office equipment. During dialing, contact S6 across the receiver is closed.

The 500-type telephone set incorporates a number of characteristics and features that represent improvements over earlier sets. Transmitting efficiency and receiving efficiency have been increased, and the frequency responses of both the transmitter and receiver have been extended. Components added to the basic common-battery antisidetone circuit are (1) a dial pulse filter to suppress high-frequency interference into radio sets, (2) a varistor, V_3, to suppress clicks in the telephone receiver, (3) an improved sidetone balancing network (necessitated by the improved transmission characteristics of the instruments), and (4) an equalizer employing two additional varistors, V_1 and V_2, to reduce transmitting and receiving efficiency on short loops.

This equalizer helps to solve an important transmission problem in telephone set design, namely, the interdependence of the transmitting and receiving efficiencies and the wide range of transmitter

currents caused by the large allowable variation in the resistances of customer loops. On long loops the direct current from the central office battery is low; the varistor impedances are therefore high, and the maximum telephone set efficiency is obtained. On short loops the high direct current results in low varistor impedances which shunt the speech currents and reduce the set efficiency. As Fig. 4-3 shows, the combined receiving response of a 26-gauge loop, station set, and equalizer is nearly constant for any loop length between 0 and 12,000 feet. The transmitting response variations are substantially less than would occur without the varistors. The overall effect is to make speech volumes at the central office and at the customer receivers less dependent on loop length. These volumes are highly variable because of the differences in customer talking habits and in the manner in which people hold the telephone transmitter to the mouth.

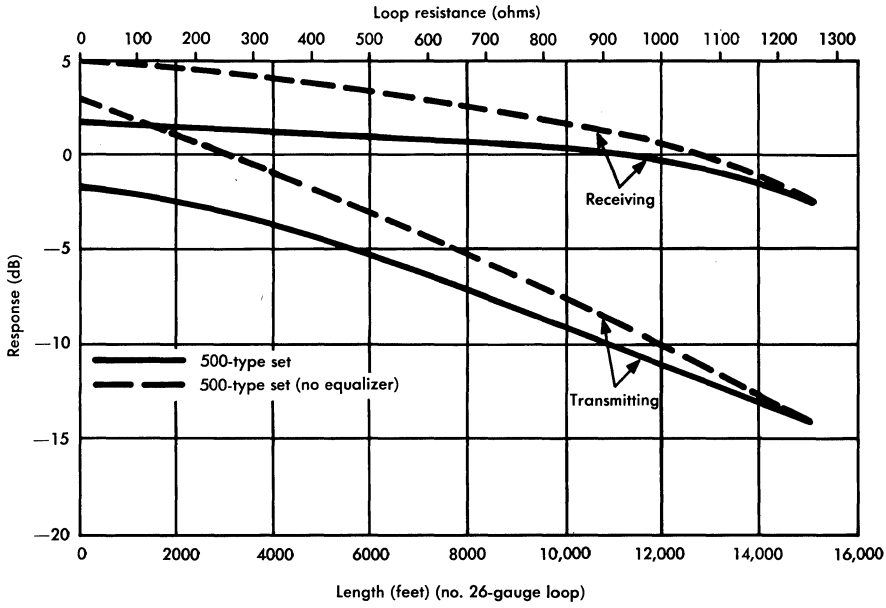


FIG. 4-3. Relative response of 500D-type telephone set.

Varistors V_1 and V_2 serve an additional purpose. By a mechanism similar to the one described for the equalizing function, they compensate for differences in customer loop impedances which would

otherwise tend to produce unbalance in the sidetone circuit. Prevention of significant unbalance is necessary since the greater efficiency of both the receiver and transmitter makes excessive sidetone more objectionable.

Connection and Performance

The quality of transmission over a telephone connection will depend on the received volume, the relative response at different frequencies, and the interferences. In a typical connection, the ratio of the acoustic pressure at the transmitter input to the corresponding pressure at the receiver output will depend upon:

1. The translation of acoustic pressure into an electrical signal across the customer loop.
2. The losses of the two customer loops, the central office equipment, and the trunks.
3. The translation of the electrical signal at the receiving telephone set to acoustic pressure at the receiver output.

Figure 4-4 shows the transmission-versus-frequency characteristics (normalized at 1000 Hz) at four separate points in a connection. Since, in this illustration, the trunk is assumed flat with frequency, the characteristic shown at point D shows somewhat less deviation from flat than would normally be encountered in practice.

The importance of speech volume and the losses in various parts of a telephone connection can be seen by examining the power level diagram for a typical telephone connection. Figure 4-5 shows the speech power at various points in a local connection involving a trunk between two central offices. Additional losses in the trunks or loops could further reduce the received volume and, depending on the magnitude of room noise at the receiving end, might require the talker to speak louder.

4.2 EXCHANGE AREA PLANT

An exchange area is sometimes defined as the area served by a single central office. More generally, the exchange area includes the metropolitan area served by one or more central offices. In the exchange area, customer loops connect each telephone with its central office, and trunks interconnect the offices. A typical pattern of loops and interoffice trunks is shown in Fig. 4-6. Loop and trunk transmission objectives are allocated separately.

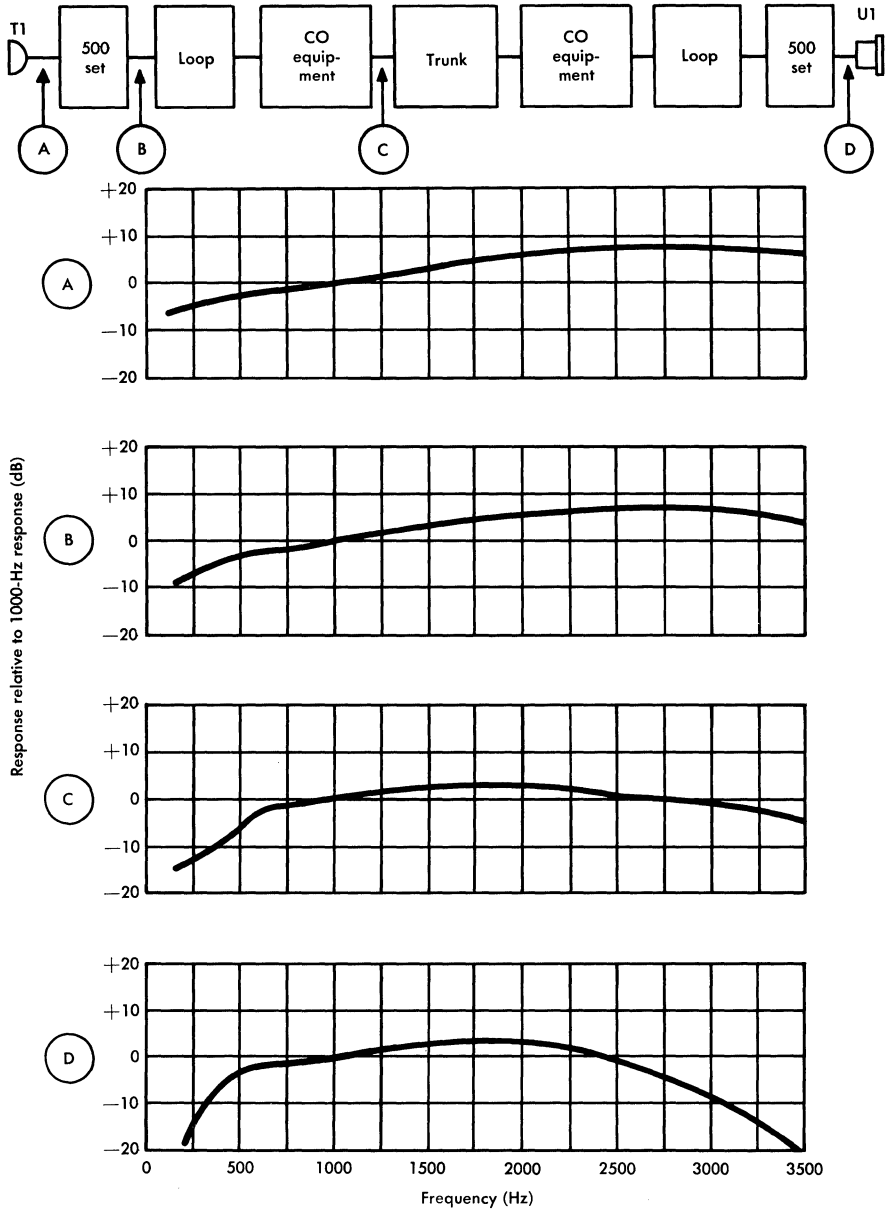


FIG. 4-4. Transmission-versus-frequency characteristics, normalized at 1000 Hz, for various points along a telephone connection (trunk is assumed distortionless).

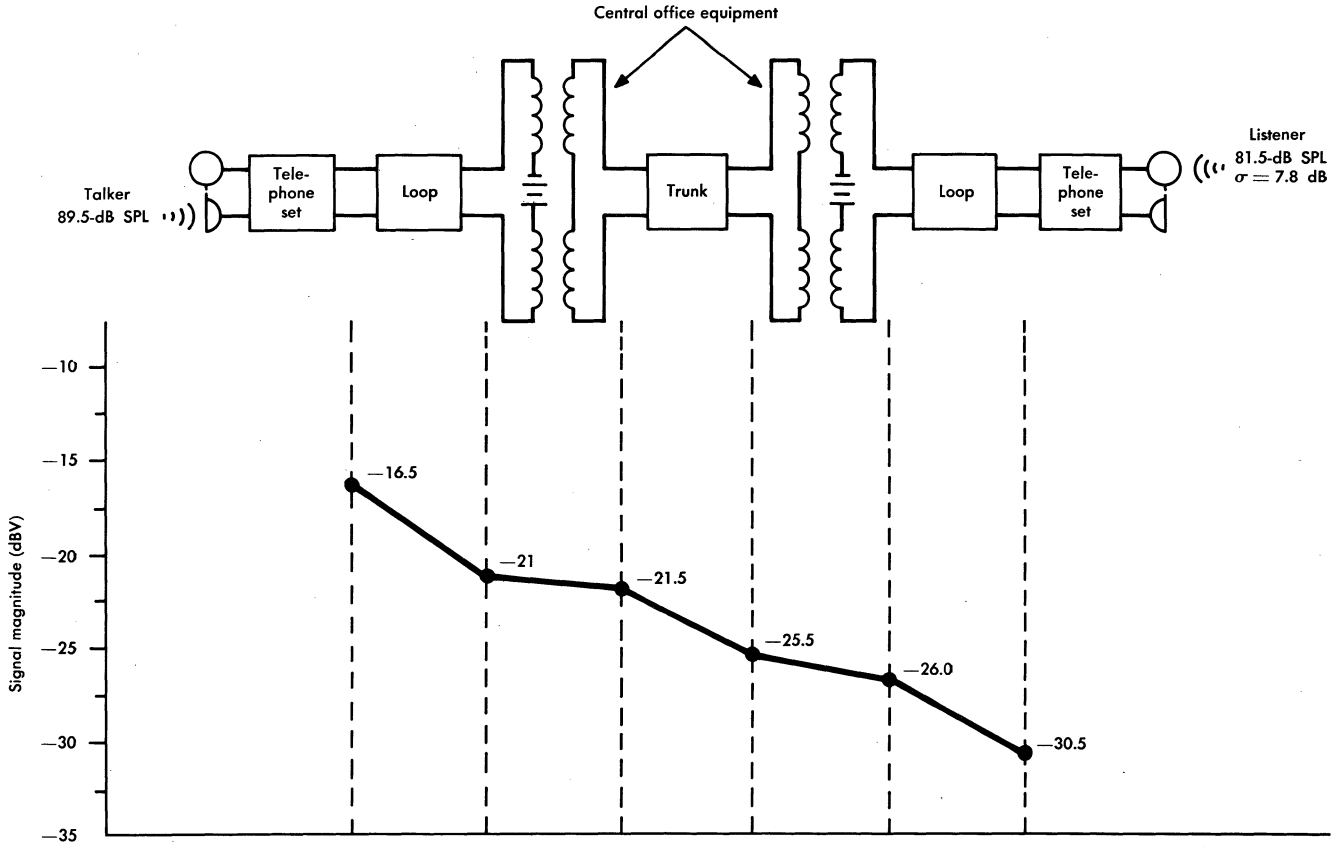


FIG. 4-5. Sound pressure levels and voltages in typical telephone connection.

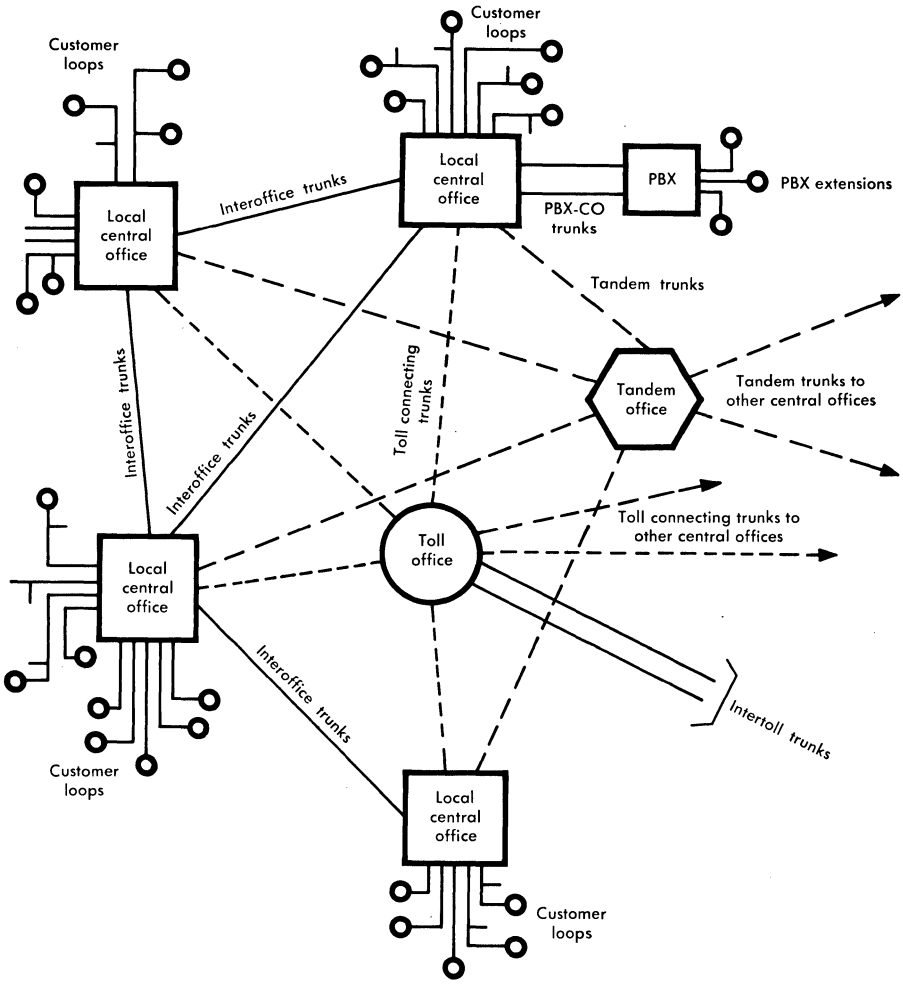


FIG. 4-6. Possible layout of loops and trunks in part of an exchange area.

Loop Design

The selection of wire sizes and loop lengths is based on loop resistance and a set of rules which take into account the correlation between resistance and attenuation, the type of office being engineered, the limitations of supervisory signaling and transmission, the use of loading coils, and the effects of resistance variations due to temperature changes. In panel and step-by-step offices, supervisory

signaling limits loops to maximum values of approximately 1200 ohms. In No. 5 crossbar and ESS offices, signaling would permit a loop resistance of about 1600 ohms. However, these offices ordinarily are engineered to a 1300-ohm limit for transmission reasons, and loading is required on loops 18,000 feet and longer. An allowance must be made for the resistance of loading coils and for cable resistance variations due to temperature changes.

For greatest economy in new plant layout, 26-gauge pairs are used wherever possible, but on a resistance design basis the maximum length of such loops is 15 kilofeet. A design concept known as Unigauge allows the exclusive use of 26-gauge pairs for loops of 30 kilofeet or less [2]. This is made possible by the use of a 72-volt office battery and a range extender. Loops longer than 24 kilofeet

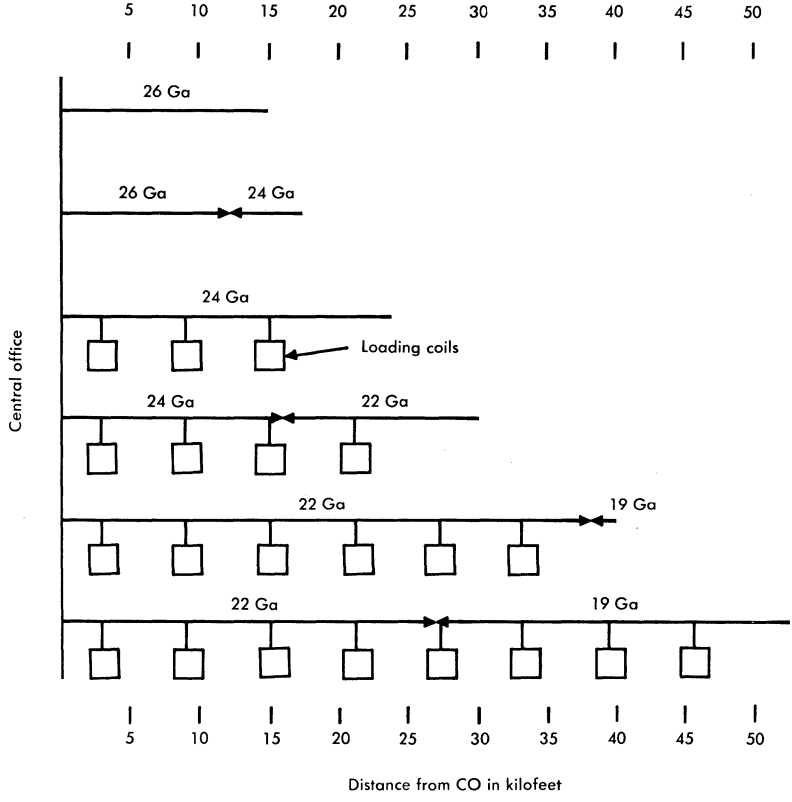


FIG. 4-7. Resistance design.

require some loading, but the first loading point is 15 kilofeet from the office. The range extender gives satisfactory transmission by providing gain and equalization in the talking path. Supervision and ringing signals bypass the voice repeater.

Range extenders are used on a "concentration" basis. That is, the number of range extenders available for a group of lines is substantially less than the number of lines served. The common control portion of the switching machine (No. 5 crossbar, for example) determines which calls require range extenders, and connects and enables them when required.

A comparison of the cable gauges and loading coil requirements for resistance design and Unigauge layouts is shown in Figs. 4-7 and 4-8.

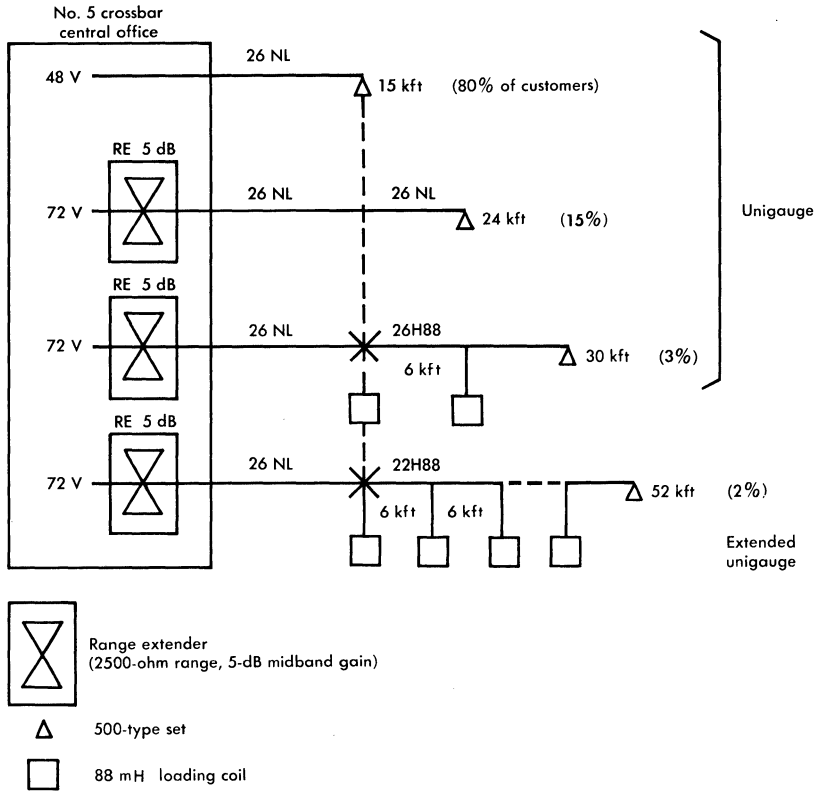


FIG. 4-8. Uniform gauge loop plant layout.

These figures show also the cable gauges and loading required to extend the loop length to 52 kilofeet.

Trunks in the Exchange Plant

The layout of the trunk plant is governed by transmission objectives because transmission loss compensation is usually required on exchange facilities before the d-c resistance limits for supervisory signaling are exceeded. A trunk extends from the outgoing side of the switch, or its equivalent in the originating office, to the outgoing side of the switch in the terminating switching office to which the trunk is connected. It therefore includes the switching path at the terminating end, the office equipment at each end, and the transmission media between the two offices. Transmission objectives for trunks between central offices (i.e., interoffice trunks) call for an average loss of 5 dB, with a maximum value of 7 dB. Variation of attenuation with frequency must be minimized. Meeting present objectives calls for loading many trunks between central offices and often necessitates the installation of repeaters.

The objectives for trunks between local and tandem offices are for a 3-dB nominal loss with maximum value of 4 dB. Trunks between tandem offices have a via net loss objective which provides sufficient loss (0.5 to 1.5 dB) to assure adequate stability and echo performance. Four-wire repeatered lines generally will be required to achieve this low loss.

Figure 4-9 gives the maximum length in miles and the corresponding resistance for various cable facilities having an insertion loss of 6 dB. Loss or resistance may govern, depending on the type of office involved; usually, unless repeaters are used, the transmission performance is the limiting factor. It should be noted that the data in Fig. 4-9 is illustrative only. Limiting values of cable length and resistance (last two columns) should include the effects of central office equipment. These are not included in the table because their values vary somewhat among systems.

Return Loss. The interconnections among loops and trunks set up by switching systems, and the infinitely variable nature of the interfaces thus created lead to one of the most difficult problems in voice-frequency system design, namely, the control of return losses at these interfaces.

As discussed in Chap. 3, there are two phenomena associated with return loss. In one case, a poor return loss may cause instability (singing) or near-instability of amplifiers in the circuit. In the

Gauge and loading	Attenuation at 1000 Hz (dB per mile)	d-c resistance (ohms per mile)	Length for 5-dB insertion loss at 1000 Hz between 900 ohms (miles)	Total resistance of cable (ohms)
26NL	2.8	440	2.2	960
26H88	1.8	448	2.9	1280
24NL	2.3	274	2.8	760
24H88	1.2	282	4.2	1190
22NL	1.8	172	3.4	590
22H88	0.8	180	6.3	1140
19NL	1.3	85	4.3	370
19H88	0.42	93	11.8	1100

FIG. 4-9. Cable loss and resistance characteristics.

second case, echoes are produced which are subjectively objectionable in varying degrees. Echo return loss is defined as a weighted return loss over the band of frequencies between 500 and 2500 Hz, while singing return loss is important at all frequencies. The concern here is primarily with echo return loss and its control through voice-frequency system designs.

The impedance of a local loop terminated by a telephone set varies as a function of frequency, cable design, and loop length. Figure 4-10 illustrates these variations, showing how the impedance varies for three typical loops. Theoretically, these impedances could be adjusted by the inclusion of suitably designed networks, but this is economically unattractive because each loop would have to be so treated.

Individual loops, which may have different impedances, are connected by end-office (class 5) switching systems to trunk facilities, which may also have different impedances. However, there are fewer trunks and their costs are shared by many customers. Control of trunk impedance is therefore somewhat more economical; precision networks and impedance compensating networks of various designs are provided for this purpose.

Figure 4-11 illustrates two of the interfaces at which return losses are important. In Fig. 4-11(a), the compromise balancing network must of economic necessity be designed to an impedance value that represents a reasonable compromise in matching a large number of the highly variable impedances of connected local loops. In Fig. 4-11(b), the two-wire trunks can be treated economically to approach

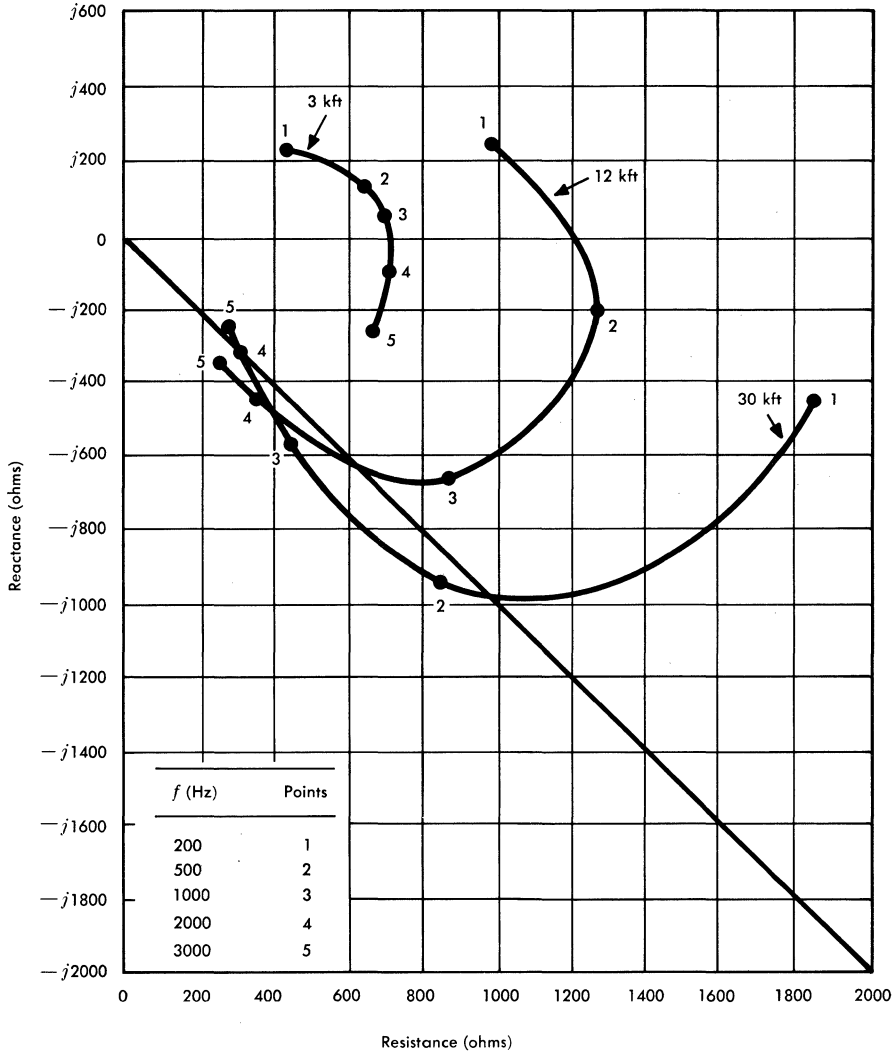
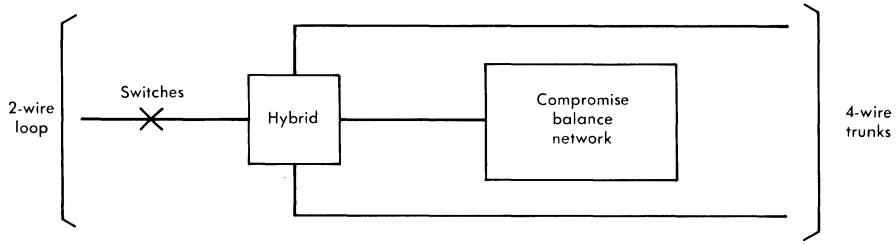
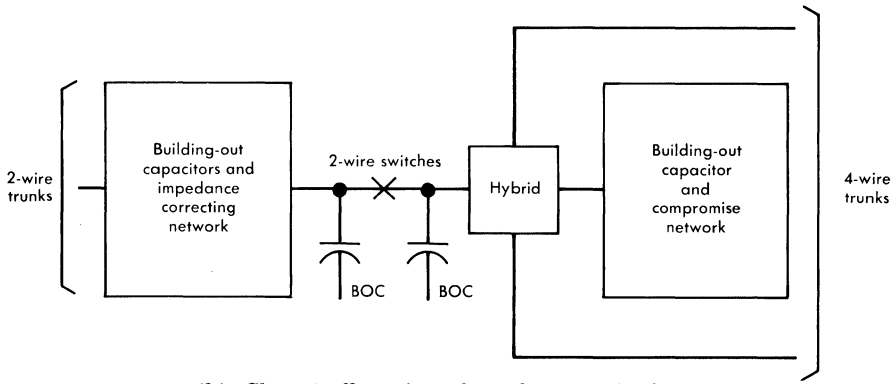


FIG. 4-10. Cable impedance with 500-set termination (24-gauge, nonloaded).

a more nearly constant impedance by the use of suitable building-out capacitors (BOC) and impedance-correcting networks. Typical echo return-loss values with large variations about the average are 11 dB at class 5 offices where loops are terminated, as shown in Fig. 4-11(a), and 15 dB at class 4 offices where intertoll and toll-connecting trunks are interconnected, as shown in Fig. 4-11(b).



(a) Class 5 office—interfaces between loops and trunks



(b) Class 4 office—interfaces between 2-wire and 4-wire trunks

FIG. 4-11. Some interfaces at which return losses are important.

The present design for toll-connecting trunks calls for a minimum loss of 2.5 dB in each trunk to pad out the generally poor impedances of the local plant. When necessary, the return loss of the trunk is increased by adding impedance-correcting networks at the toll office. The maximum loss of the toll-connecting trunks in the present plant is 4 dB. These transmission objectives can be achieved by any one of the following means:

1. If the normal loss of the trunk is less than 2 dB, a 2-dB loss is added at the toll office.
2. If the trunk has a normal loss of 2 to 4 dB, it will adequately mask the impedance mismatches of the local plant. Its own impedance, however, may not be very good, and in such cases it is necessary to add an impedance-correcting network at the toll office.

3. If the trunk loss exceeds 4 dB, one or more repeaters must be added and the trunk loss adjusted to equal the quantity (VNL + 2.5 dB). Here, VNL represents the via net loss computed for the toll-connecting trunk facility as if it were an intertoll trunk. The toll-connecting trunks entering four-wire switching centers frequently fall in this category since the hybrid (approximately 3.5-dB loss) is considered part of such a trunk. For typical lengths of toll-connecting trunks, the VNL will be 1 dB or less, so that the addition of 2 dB puts the trunk loss into the required 2- to 4-dB range. An impedance-correcting network is added to the trunk at the toll office if necessary.

In addition to these procedures, most voice-frequency toll-connecting trunks are loaded to minimize the attenuation variation across the voice channel. Where round-trip delays of more than 45 milliseconds are encountered, echo suppressors are used.

Echo Suppressors. An echo suppressor [3] is basically a pair of voice-operated switches which, while one subscriber is talking, insert a loss of 35 dB or more in the echo return path. In case both parties talk simultaneously, the talker whose signal is stronger at the echo-suppressor will control the switch and will be heard. Although they effectively suppress echoes, echo suppressors can introduce transmission impairments by sometimes clipping the beginning of words. There are two basic types, the full and the split echo suppressor. The full echo suppressor controls both directions of transmission. It seldom is located at the mid-point of the circuit. Typically, the delay between the near talker and the echo suppressor is 5 to 10 milliseconds while the distant talker may have as much as 35 milliseconds delay between him and the echo suppressor. The imbalance of switching control resulting from this arrangement is avoided in the split echo suppressor. This provides a sensing circuit and switch near each talker. Normally switch operation is controlled by the distant talker and introduces high loss in the return (echo) path. This blocks transmission from the talker nearer the switch. However, the sensing circuit continually compares signal levels in the transmitting and receiving paths, and in case of simultaneous talking it always gives the low-loss path to the louder talker. In the latest designs, this type of break-in occurs in 2 to 5 milliseconds. In older designs the louder talker had to persist for 50 milliseconds or longer in order to break in.

The normal operation of the echo suppressor must be inhibited for certain types of data transmission. This is accomplished by the "tone disabler" section of the echo suppressor circuit. A short burst of single-frequency energy, sent over the circuit immediately after the connection is set up, activates the disabler and prevents subsequent operation of the echo suppressor. Another type of disabling feature found in some offices permits the switching machine to control this function and prevent the buildup of circuits having two echo suppressors in tandem.

Under the present toll-switching plan, echo suppressors are used only between regional centers or on trunks between regions which otherwise would have a VNL design of more than 2.5 dB. When echo suppressors are used, the VNL is set equal to 0.5 dB or less.

Insertion Loss in an Exchange Area Telephone Connection

Intertoll trunks usually are designed to provide good impedance matches at their terminals and at intermediate points. Furthermore, intertoll trunks generally introduce no appreciable frequency distortion. Thus, the attenuation of an intertoll trunk usually determines its contribution to the effective loss of a circuit.

In the exchange plant, loops and trunks are normally electrically short and, as mentioned earlier, do not have particularly good impedance matches at junctions. Thus, the impedance at a particular point of a built-up connection will depend on the impedance terminations at remote points. Likewise, the power loss will be a function of the attenuation of the components and the reflection gains and losses at the numerous junctions [4].

Figure 4-12 shows a particular exchange area connection with the 1000-Hz impedance seen at each junction as computed from the equivalent tee network for each component. The 1000-Hz attenuation of each component is also shown. In this instance, the impedance requirement for maximum power transfer (namely, conjugate impedances) is approximately met at the junction between the trunk and the central office. At the other junctions, the situation becomes less ideal in proceeding toward the customer. A detailed computation shows the 1000-Hz insertion loss between the telephone sets in this connection to be 10.9 dB. The sum of the attenuations is 13.4 dB. This illustrates the fact that multiple reflections between component parts of an exchange area circuit combine in such a way as to make the summation of individual attenuations an unreliable measure of total loss.

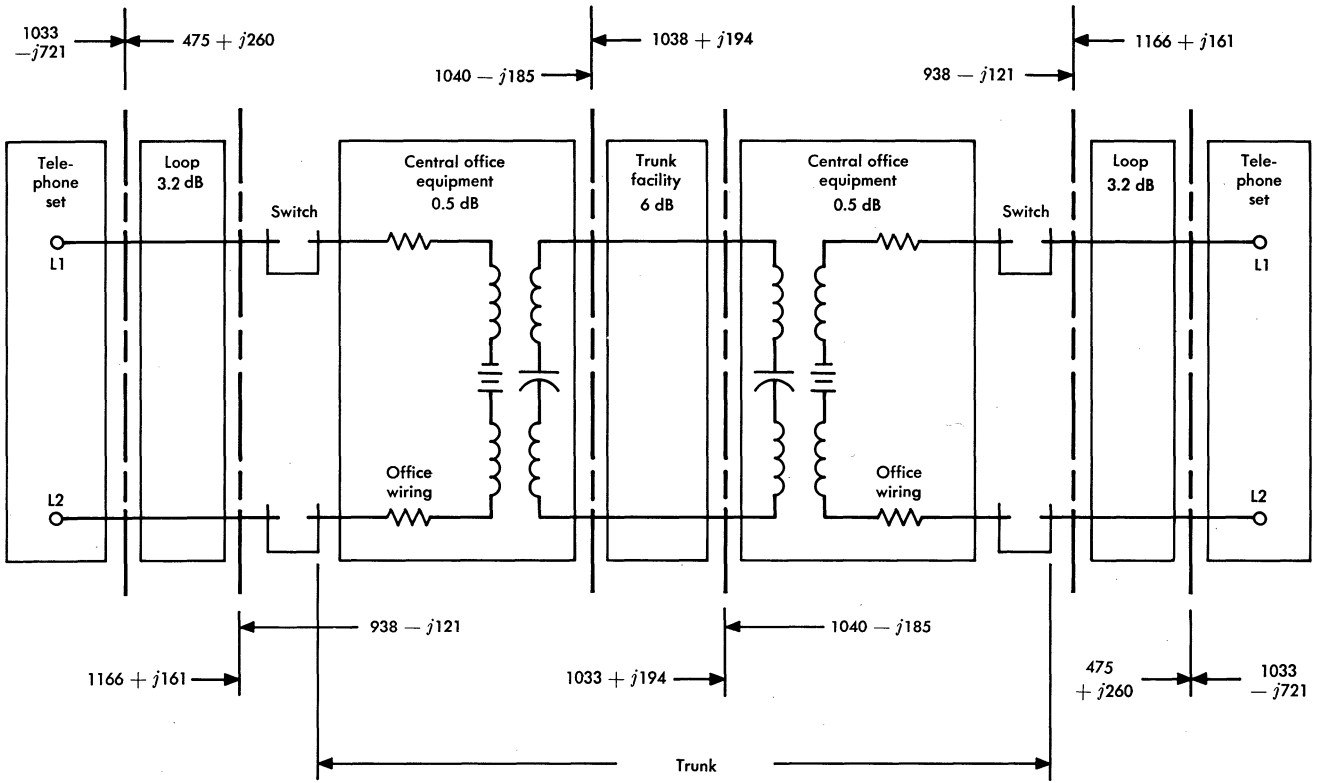


FIG. 4-12. Typical impedances at various points in a possible exchange area telephone connection.

Realization of the correct 1000-Hz insertion loss is, of course, only a first step toward the realization of a satisfactory transmission system. Insertion loss across the entire voice-frequency band and, for some types of signals, delay distortion must be controlled. In addition, circuit noise, crosstalk, and nonlinear distortion are a few of the factors which must be considered in the determination of overall circuit quality.

4.3 VOICE-FREQUENCY TRANSMISSION CIRCUITS

Telephone instruments and the voice-frequency facilities to which they connect may be either two-wire or four-wire circuits. The economic advantage of using only half as much copper dictates the use of the two-wire circuits in many voice-frequency trunks and in most loops, although some special applications (certain military installations, for example) have required the use of four-wire telephone sets and loops. The use of four-wire trunks is becoming of greater importance, but many trunks are still two-wire because of the copper savings.

Two-Wire Voice-Frequency Circuits

In order to meet the tandem and toll trunk loss requirements, it is often necessary to add gain to the transmission path. A number of special problems are encountered when this is required in a two-wire circuit. The gain introduced must be independent of the direction of transmission; i.e., the circuit must retain its bilateral properties for speech signals. Dial pulses must not be impaired and the equipment must be able to withstand the high voltage used to ring the customer's bell.

E-Type Repeaters. The most commonly used repeaters for this type of service are known as the E-type negative impedance repeaters. A composite schematic of E-type repeaters is shown in Fig. 4-13 [5]. Some of the designs omit some of the elements shown. The gain-producing circuits are called negative impedance converters (NIC). Each of these is a two-port amplifier with regenerative (positive) feedback. The impedance seen at one of the ports is approximately the negative of the impedance of the network connected to the second port. The relative magnitudes and phase angles of the impedances connected to the two ports interact to determine the amount of feedback, and improper combinations will cause oscillation. The manner

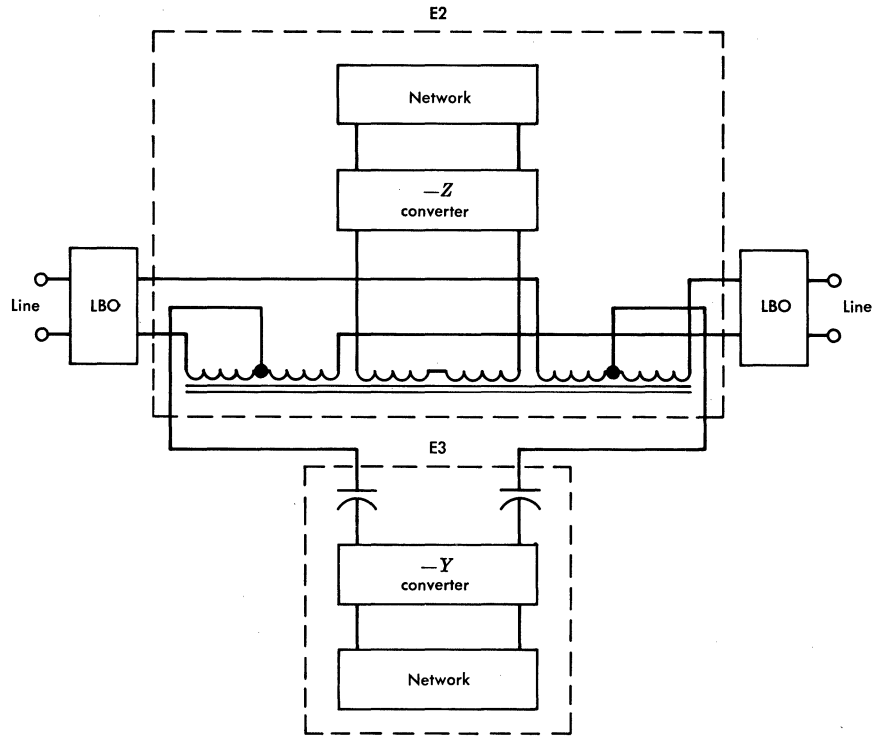


FIG. 4-13. Block schematic of E-type repeater.

in which the feedback is derived permits a loose classification into series ($-Z$) type and shunt ($-Y$) type converters. The distinction is between the extremes of impedance which can be connected to the input port without causing oscillation. The series type is open-circuit ($Z = \infty$) stable while the shunt type is short-circuit ($Y = \infty$) stable. Clearly, this is a matter of which port is defined as input, and, in general, interchanging the ports will invert the type of the converter.

The general E-type repeater is shown in Fig. 4-13. The electrically identical E1 and E2 repeaters insert series negative impedance in the line. They contain only those elements shown within the dotted lines labeled E2. Similarly, the E3 repeaters insert shunt negative impedance across the line and contain only the elements shown within dotted lines at the bottom of the figure. Either of these repeaters can be used by itself to provide bilateral gain; however,

a substantial impedance irregularity will be introduced in the line. Even with careful attention to spacing and gain adjustment, changes in cable characteristics with temperature can cause large variations in the loss of the circuit.

Improved performance can be obtained by using both types in a combination called the E23 repeater. The boxes labeled network in Fig. 4-13 contain a variety of resistors, inductors, and capacitors, with provision for connecting various combinations of these to the indicated terminals of the converters. When the proper combination is used, the repeater will provide the desired gain and a reasonably good match to the cable impedance.

In the E6 repeater [6] only resistors are used in the adjustable networks, and the impedance matching function is performed by a passive line-building-out (LBO) network separate from the gain unit. The LBO builds out the cable impedance, for any end section from zero to nearly a full loading section of cable, to appear as a fixed resistance of 900 ohms in series with 2 microfarads of capacitance, the desired impedance. The gain unit can then be designed as a simple 900-ohm, 2-μF converter with negative series resistances and negative shunt conductances to vary the amount of gain.

22-Type Repeaters. The 22-type repeater takes the general form shown in Fig. 4-14(a). Hybrid coils are used to split the two directions of transmission and separate amplifiers furnish gain for each direction. Equalization and gain can be adjusted independently for each direction. Good hybrid balance is required to prevent oscillation in the feedback loop containing two hybrids and two amplifiers in tandem. This balance is obtained by using networks which match closely the impedances connected to the conjugate hybrid ports. A repeater of this type is included in the range extender for the Unigauge system.

So far as external performance is concerned, the E6 and 22-type repeaters are equivalent. Repeater and cable impedances can be matched to achieve 35-dB return losses under fixed conditions. Maximum gains are limited to about 12 dB by cable characteristic changes due to temperature variations, changes in end terminations during switching, and crosstalk considerations.

The need for lower losses on exchange trunks and on special service lines such as carrier circuit extensions over voice-frequency lines to PBX's has required continued attention to the problems

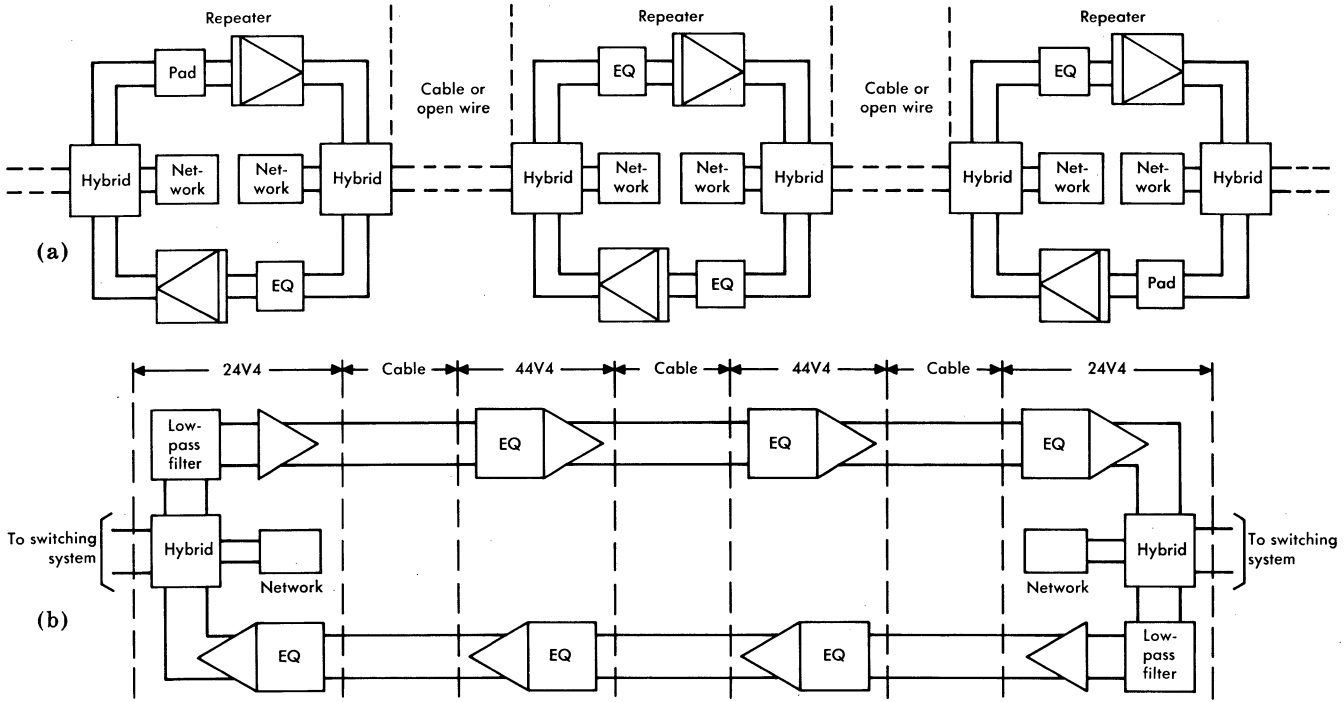


FIG. 4-14. Two-wire and four-wire methods of operation.

of designing low-loss voice-frequency facilities. Both loading coils and voice-frequency repeaters are used on these circuits to reduce their losses.

Figure 4-14 shows that in two-wire operation there is a multiplicity of singing or echo paths for each repeater. A given repeater is subject to instability due to its own loop transmission, its own plus that of the adjacent repeater, and so on until the end of the system. Thus, the tendency towards instability in two-wire systems increases as the length of the circuit and the number of repeaters are increased. As a result, singing and echo performance are the controlling factors in setting the repeater gains in most long two-wire voice-frequency circuits. To achieve reasonably high gains in the repeaters, careful attention must be paid to the installation and maintenance of the circuit, particularly with regard to the uniformity of loading and to the balancing or LBO networks at each repeater. Such practical considerations frequently dictate the use of four-wire systems equipped with four-wire repeaters.

Four-Wire Circuits

In four-wire operation the major singing path is that which extends around the circuit, from one end to the other. Thus, four-wire voice-frequency systems are inherently more stable than two-wire. As a result, repeater gains in four-wire systems are limited more by modulation distortion (due to nonlinear amplifiers), echo, and crosstalk into other systems than by singing considerations.

The minimum loss to which an intertoll trunk can be adjusted is neatly summarized by the via net loss. The minimum VNL for a facility is found by multiplying the VNLF by the facility length and adding 0.4 dB to allow for maintenance variation. Via net loss factors for the various types of voice-frequency circuits that have been described are tabulated in Fig. 4-15. The factor for carrier systems is also shown for comparison purposes.

Facility	VNLF (dB per mile)
Two-wire open wire (all wire sizes)	0.01
Two-wire 19H88	0.03
Four-wire 19H44	0.01
Carrier systems (all types)	0.0015

FIG. 4-15. Via net loss factors.

The use of carrier permits lower loss operation of trunks than does the use of voice frequency. Since carrier circuits also tend to be more economical, they are most commonly used now for intertoll trunks except where distances are short and cross sections are small.

V4 Repeaters. A number of types of voice-frequency repeaters have been developed, but only the 24V4 and 44V4 are now in large scale production. They are used as the basis for practically all new circuit designs for which this general class of repeaters is applicable. The remainder of the discussion of repeatered voice-frequency systems is confined to a description of the V4 design and its use.

Figure 4-14(b) shows the application of the 24V4 and the 44V4 repeaters to a trunk. The 44V4 repeaters are used in four-wire circuits to provide gain and equalization for each direction of transmission. The 24V4 repeaters provide equalization for one direction of transmission, gain for both directions, and terminating arrangements to connect the four-wire circuits to the two-wire telephone plant.

In both the 24V4 and the 44V4, the equalizers are plug-in units that may be selected to compensate for the loss-frequency characteristics of any of a variety of cable gauges and loading. Circuits used for data transmission may also require delay distortion equalization. This is usually made a part of the equalizer at one of the terminal repeaters. The flat loss of the combined cable section plus equalizers is then overcome by a plug-in two-stage transistor amplifier having a flat gain characteristic and an adjustable gain of 0 to 36 dB.

The 24V4 repeater is shown schematically in Fig. 4-16. A plug-in hybrid-type terminating set forms the interface between the two-wire and four-wire portions of the circuit. The balance achieved across the hybrid is dependent on the match between the impedance of the balancing network and the impedance of the connections on the two-wire side of the circuit. When the two-wire circuit is fixed, as it might well be in Fig. 4-14(a), a precision network may be used; then, a high degree of balance and excellent return-loss performance are achievable. However, widely varying balance and return losses occur when the two-wire side of the circuit is connected to a switching system, as indicated in Fig. 4-14(b). In these applications, a compromise network and a capacitance building-out network must be used to obtain the required average office balance. Figure 4-14(b) also shows the connection of a low-pass filter used to suppress out-of-band singing around the four-wire loop.

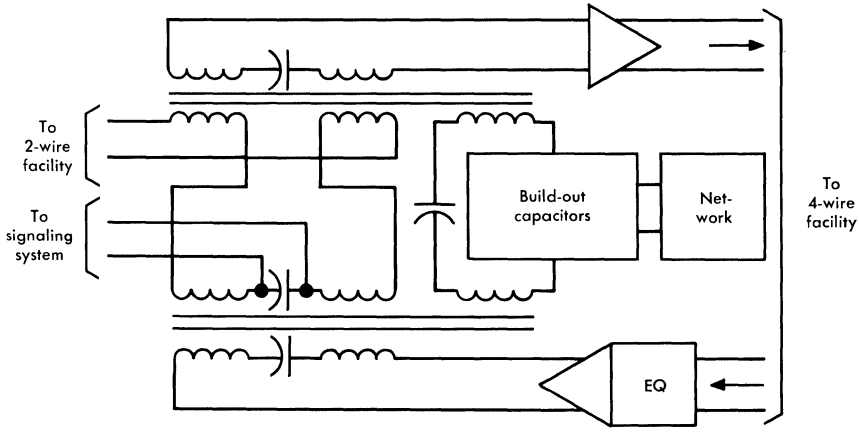


FIG. 4-16. Block schematic of 24V4 repeater showing hybrid-type terminating set.

Finally, connections to signaling equipment are shown in Fig. 4-16. A short discussion of the relation between signaling and voice-frequency system design follows.

Signaling

The d-c dial pulse and supervisory range of a signaling system are functions of the d-c loop resistance as well as of the shunt resistance and capacitance. Voice repeaters introduce added resistance and often shunt resistance and capacitance. The added series resistance must be minimized so that the supervisory range of the circuit is not appreciably reduced. Even when consideration has been given to minimum resistance design, shunt capacitance may often limit the pulsing range. Capacitors shunted across midpoints of repeat coils must charge and discharge during each pulse, introducing time constants which may distort pulses and thus decrease the overall range of the system. There are many different types of pulsing, and unless the requirements of all types are considered in the design of voice repeaters, the increase in the transmission range of the system may lead to a reduction of signaling range.

4.4 INDUCED INTERFERENCES IN EXCHANGE AREA PLANT

The importance of maintaining low losses in the telephone plant has been pointed out. To achieve satisfactory service, it is also important to control and minimize interferences. These include message

circuit noise, crosstalk, and power line induction. Some interferences (for example, certain types of noise) arise within a message channel; others, such as crosstalk and power line induction, are due to external influences. These topics appear frequently in later chapters. For the present, the discussion is limited to a brief description of the mechanisms by which such noise is induced and controlled in voice-frequency systems.

Mechanisms

As the preceding discussion has indicated, a transmission path can never be entirely isolated from the external world. It may parallel a power line; it usually lies in proximity to similar paths (loops and trunks in cables or on pole lines); and it passes through central offices in which switching apparatus creates sizeable transients. While not intended, coupling to these sources of interference always exists [7].

Although ground-return circuits were used originally in the telephone plant, these were soon given up in favor of the balanced pair. The balanced pair has the advantage that interferences induced equally in both wires of the pair are balanced out. This is illustrated in Fig. 4-17 where repeat coils are shown at both ends of the circuit. The signal voltage, e_{sig}, causes signal current, i_{sig}, to flow. The current

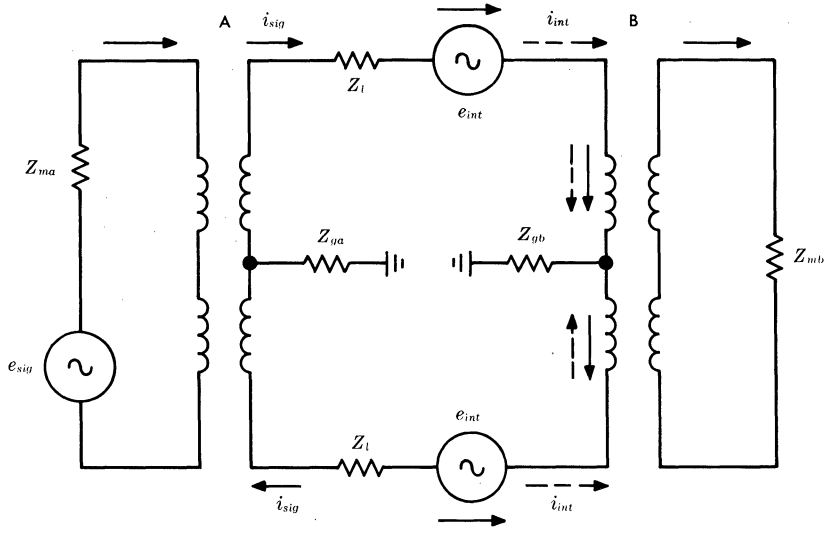


FIG. 4-17. Signals and interferences in balanced pair.

direction in the two wires is opposite, and thus the signal passes through the B repeat coil to Z_{mb}. An interference which is induced equally in each wire is indicated by the e_{int} generators. In this case, the currents flow in the same direction along the pair and cancel in the repeat coil. The currents which flow in opposite directions in the wires of the pair are known as *metallic circuit currents*. The currents which flow in the same direction along the pair are called *longitudinal currents*.

Induced interferences may reach the receiver and disturb transmission in two ways. In one case, the coupling between the source and one of the wires of the pair differs from the coupling between the source and the other wire; in the second case, there is an unbalance with respect to ground within the pair.

If the voltages induced in the two wires of the pair are not equal, then the interference will appear as both a longitudinal and a metallic circuit current. Consider the case shown in Fig. 4-18 where it is assumed that unequal voltages are induced as a result of differences in exposure to the disturbing source. These voltages may be analyzed into a pair of voltages which would cause balanced longitudinal currents, plus a residual which would cause a metallic current. The circuit presents quite different impedances to these two types of generators. In Fig. 4-18, the total resulting current in the upper wire is shown as ten units and that in the lower wire as six units. As shown by the solid and dashed arrows, these induced currents can be divided into metallic currents of two units and longitudinal currents of eight units. The two units of metallic circuit current will appear as an interference in Z_{mb}. A 4 to 1 ratio of longitudinal to metallic current represents an unsatisfactory circuit. Ratios of 500 to 1 or 1000 to 1 would be more representative of plant performance.

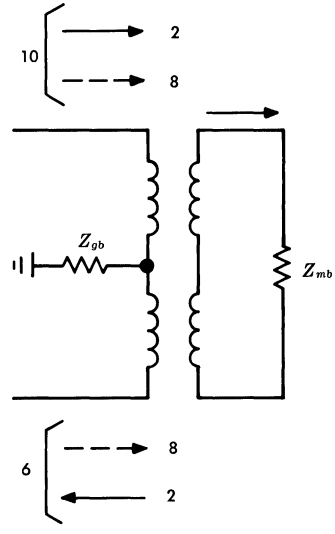


FIG. 4-18. Metallic and longitudinal currents.

A similar effect is obtained even when the voltages induced in the two wires are identical if there is an unbalance in the pair. This

unbalance may result from a difference in the resistances of the two halves of the transformer winding or from a poor splice in one of the wires. The different impedances result in different currents in the two wires. The difference can be minimized by making the common impedance, Z_{gb}, large relative to the unbalance in impedance of the pair.

The coupling between a source of interference and the wires of a pair may be capacitive, inductive, conductive, or various combinations of these. Coupling to power lines is usually predominantly inductive, and would be trivial if it were not for the great energy of the disturbing source. The coupling in cables and open-wire lines which produces crosstalk at voice frequencies is predominantly capacitive, although inductive coupling becomes important as carrier frequencies are approached.

Methods of Reduction

The preceding discussion of mechanisms is sufficient to suggest how induced interferences may be minimized. It is clearly advantageous to maintain balance within the pair itself. Similarly, the apparatus to which the pair connects, such as repeat coils, must be well balanced; a common criterion is that the balance of such apparatus should be 10 dB better than the best pair with which it will be used. Another obvious method, to be used when practicable, is to reduce the coupling between line and source by adequate spacing. It is one of the reasons for keeping a reasonable spacing between pairs of open-wire lines, thus reducing the crosstalk between them.

In addition to maintaining balance within the pair itself, it is important, as the preceding discussion has shown, to expose each wire of the pair equally to the interfering source. In open-wire lines this is done by intricate transposition systems. In cables the wires of a pair are twisted around each other and at various rates (i.e., various twist lengths), in order to achieve as much balance and randomization of coupling paths as possible. When telephone and power cables are laid in the same trench, a method called random separation is used for the same purpose.

In repeatered systems, a satisfactory overall signal-to-noise ratio can be obtained by judiciously spacing repeaters so that signal levels never become too low relative to the induced noise. (The same approach is used in systems where thermal rather than induced noise is controlling, as discussed in detail in subsequent chapters.) On the other hand, the signal levels must not be set so high as to make the repeatered circuit a source of excessive crosstalk.

For the sake of completeness, although they are not economically applicable to voice-frequency telephone circuits, two other methods of reduction might be mentioned. One is shielding, as exemplified by shielded video pairs. The other is crosstalk balancing—the deliberate introduction of coupling paths between pairs of a cable, with the magnitude and phase of the coupling chosen to balance out the unwanted coupling.

Conclusion

Two basic points should be apparent from this survey of typical methods and basic problems encountered in handling voice-frequency signals.

1. The wire plant used for exchange area and short-haul toll transmission is composed of a wide variety of wire sizes, cable and open-wire facilities, and repeater and loading arrangements.
2. The losses, impedance discontinuities, and interferences have been engineered and adjusted to be adequate for voice-frequency transmission.

During the coming years, increasing efforts will be made to adapt this extremely valuable plant so that it will provide improved transmission and new types of service. The performance, variability, and limitations of this existing plant will strongly affect the design and application of new systems.

REFERENCES

1. Bennett, A. F. "An Improved Circuit for the Telephone Set," *Bell System Tech. J.*, vol. 32 (May 1953), pp. 611-626.
2. Gresh, P. A., L. Howson, A. F. Lowe, and A. Zarouni. "A Unigauge Design Concept for Customer Loop Plant," *IEEE Trans. on Comm. Tech.* (Apr. 1968).
3. Holman, E. W. and V. P. Suhocki. "A New Echo Suppressor," *Bell Laboratories Record* (Apr. 1966).
4. Llewellyn, F. B. "Some Fundamental Properties of Transmission Systems," *Proc. IRE*, vol. 40 (Mar. 1952), pp. 271-283.
5. Merrill, J. L., Jr., A. F. Rose, and J. O. Smethurst. "Negative Impedance Telephone Repeaters," *Bell System Tech. J.*, vol. 33 (Sept. 1954), pp. 1055-1092.
6. Bonner, A. L., J. L. Garrison, and W. J. Kopp. "The E6 Negative Impedance Repeater," *Bell System Tech. J.*, vol. 39 (Nov. 1960), pp. 1455-1504.
7. Aikens, A. J. and D. A. Lewinski. "Evaluation of Message Circuit Noise," *Bell System Tech. J.*, vol. 39 (July 1960), pp. 879-909.

Chapter 5

Modulation

Communication signals must be transmitted over some medium separating the transmitter from the receiver. The information to be sent is rarely in the best form for direct transmission over this medium. Efficiency of transmission requires that this information be processed in some manner before being transmitted. Modulation may be defined as that process whereby a signal is transformed from its original form into a signal that is more suitable for transmission over the medium between the transmitter and receiver [1]. It may shift the signal frequencies for ease of transmission or to change the bandwidth occupancy, or it may materially alter the form of the signal to optimize noise or distortion performance. At the receiver this process is reversed by demodulation methods. In this chapter some of the modulation methods and modulation systems which are commonly used in telephone transmission systems are reviewed.

The process of modulation can be represented in two major forms (amplitude and angle) by

$$M(t) = a(t) \cos [\omega_c t + \phi(t)] \qquad (5\text{-}1)$$

Here $a(t)$ represents the amplitude of the sinusoidal carrier, and $\omega_c t + \phi(t)$ is the phase angle. Although both amplitude and angle modulation may be present simultaneously, an amplitude-modulated system is one in which $\phi(t)$ is a constant and $a(t)$ is made proportional to the modulating signal. Similarly, an angle-modulated system results when $a(t)$ is held constant and $\phi(t)$ is made proportional to the modulating signal. It is appropriate to discuss each of these two types separately in some detail.

5.1 PROPERTIES OF AMPLITUDE-MODULATED SIGNALS

Equation (5-1) can be rewritten for amplitude-modulated waves by ignoring the phase modulation term to obtain

$$M(t) = a(t) \cos \omega_c t \tag{5-2}$$

where the carrier is at the frequency f_c (equal to $\omega_c/2\pi$) and $a(t)$ is the modulating time function. Since the modulated wave is the product of $a(t)$ and a carrier wave, the process is often called product modulation. In the most general case, negative values are allowed for $a(t)$. When $a(t)$ changes sign, the phase of the carrier is reversed, i.e., changed by 180 degrees. Although it is shown later that restricting $a(t)$ to positive values in the time domain is equivalent to inserting a strong carrier in the frequency domain to produce a double-sideband transmitted carrier (DSBTC) wave, for the present, a more general analysis includes the double-sideband suppressed carrier (DSBSC) case.

If $a(t)$ is a single-frequency sinusoid of unit amplitude at the frequency f_m then

$$a(t) = \cos \omega_m t \tag{5-3}$$

The modulated wave is then

$$M(t) = \cos \omega_m t \cos \omega_c t \tag{5-4}$$

By a trigonometric identity this may be expanded to

$$M(t) = \frac{1}{2} \cos (\omega_c - \omega_m) t + \frac{1}{2} \cos (\omega_c + \omega_m) t \tag{5-5}$$

Equation (5-5) contains no component at the original carrier frequency, f_c, but only a side-frequency on either side of the carrier, spaced f_m hertz from the carrier, as shown in Fig. 5-1. However, if $a(t)$ has a d-c component, the resulting $M(t)$ will have a component at the carrier frequency.

The effect of product modulation, then, is to translate $a(t)$ in the frequency domain so that it is reflected symmetrically about f_c. That this is true for a complex $a(t)$ may be shown by considering a two-frequency $a(t)$.

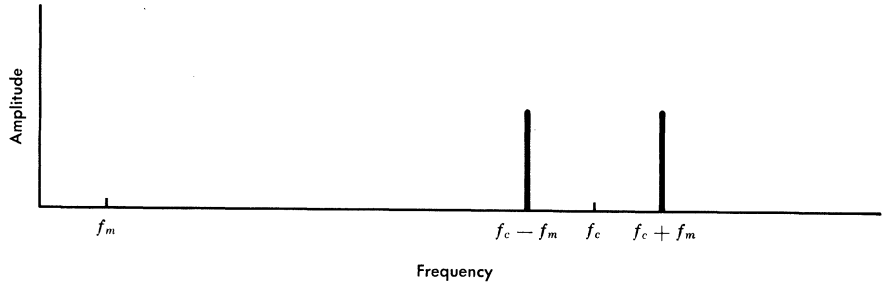


FIG. 5-1. Product modulator frequency spectrum—single-frequency modulating signal.

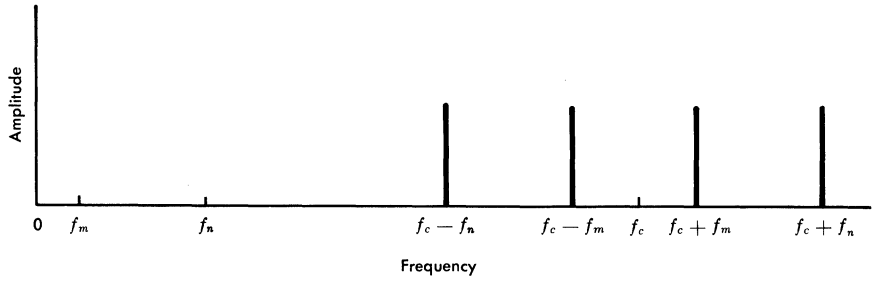
If $a(t)$ is a wave containing two frequency components (f_m and f_n), then

$$M(t) = \frac{1}{2}\cos(\omega_c - \omega_m)t + \frac{1}{2}\cos(\omega_c + \omega_m)t + \frac{1}{2}\cos(\omega_c - \omega_n)t + \frac{1}{2}\cos(\omega_c + \omega_n)t \qquad (5\text{-}6)$$

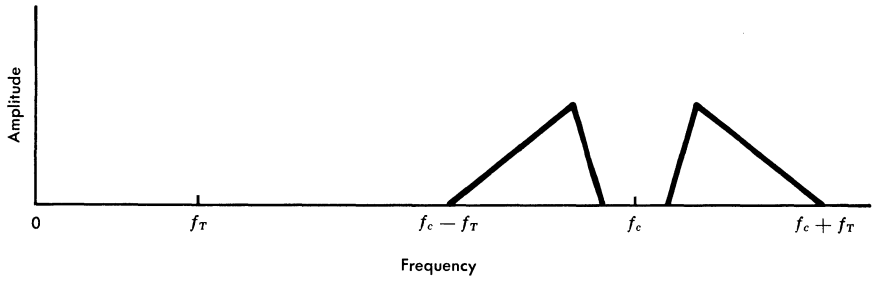
This modulated wave is depicted on the frequency scale in Fig. 5-2(a). The result is as if the two modulating frequency components were modulated independently and then added linearly. Thus, superposition holds, the product modulation process is quasi-linear, and it may be inferred that product modulation translates and reflects the baseband signal symmetrically about f_c without distortion. This is illustrated in Fig. 5-2, which shows the two-frequency case and also the more general case of a modulating wave with a continuous spectrum up to a frequency $f_T < f_c$. The resulting output signal is DSBSC.

A special case of DSBSC results if $a(t)$ is given a d-c component. To see this, consider the two-frequency case again, but let one of the frequencies equal zero (d-c). Since zero frequency at baseband corresponds to carrier frequency after translation, the modulated wave will have a component present at the carrier frequency, f_c. If it is further assumed that the amplitude of the zero frequency component is large enough, $a(t)$ will never go negative and the modulator will produce the familiar AM wave.

If either sideband in the DSBSC spectrum, Fig. 5-2(b), is rejected by a filter or other means, the result is a single-sideband (SSB) wave.



(a) Modulated spectrum of two modulating frequencies



(b) Modulated spectrum of modulating band of frequencies

FIG. 5-2. Product modulator frequency spectrum—complex modulating signal.

Basically, single-sideband modulation is pure frequency translation, with or without the inversion obtainable by selecting the lower rather than the upper sideband.

Up to this point, three types of amplitude-modulated signals have been mentioned: double-sideband with transmitted carrier (DSBTC), double-sideband suppressed carrier (DSBSC), and single-sideband (SSB). The properties of these three signals are further examined, and finally a fourth type known as vestigial sideband (VSB) is considered.

Double-Sideband with Transmitted Carrier

Although it is not mathematically the purest form of amplitude-modulated signal, DSBTC is perhaps the most familiar and will be taken up first. The waveform has upper and lower envelopes which do not overlap as long as the carrier is not overmodulated [2].

Consider a baseband signal (e.g., a complex wave with a continuous but bandlimited frequency spectrum) with a time function represented by $v(t)$, and, for simplicity, a maximum amplitude of unity.

The modulating function, $a(t)$, can be forced positive at all times by adding unity to $v(t)$ or, more generally, by substituting $1 + mv(t)$ for $a(t)$ in Eq. (5-2) to obtain

$$M(t) = [1 + mv(t)] \cos \omega_c t \tag{5-7}$$

where m is the *modulation index* and is equal to unity for 100 per cent modulation. For a single-frequency modulating wave, $v(t)$ becomes $\cos \omega_m t$, and the preceding equation can be expanded as follows:

$$\begin{aligned} M(t) &= \cos \omega_c t + m \cos \omega_m t \cos \omega_c t \\ &= \cos \omega_c t + \frac{m}{2} \cos (\omega_c - \omega_m) t + \frac{m}{2} \cos (\omega_c + \omega_m) t \end{aligned} \tag{5-8}$$

In many instances the use of exponential notation for periodic functions has advantages over the trigonometric notation which has been used thus far in this chapter. A particularly useful application is in the phasor representation of modulated waves as an aid in understanding the various modulation processes. A sinusoidal carrier, $\cos \omega_c t$, can be written as

$$\text{Re}\left[e^{j\omega_c t} \right]$$

where Re represents the real part of the complex quantity, and

$$e^{j\omega_c t} = \cos \omega_c t + j \sin \omega_c t$$

The exponential $e^{j\omega_c t}$ is a counterclockwise rotating phasor of unit length in the complex plane, and its real part is its projection on the real axis. This phasor is shown for several values of time in Fig. 5-3.

Now consider the amplitude-modulated wave of Eq. (5-8). This can be written in exponential notation as

$$\begin{aligned} M(t) &= \text{Re}\left[e^{j\omega_c t} + \frac{m}{2} e^{j(\omega_c - \omega_m)t} + \frac{m}{2} e^{j(\omega_c + \omega_m)t} \right] \\ &= \text{Re}\left[e^{j\omega_c t}\left(1 + \frac{m}{2} e^{j\omega_m t} + \frac{m}{2} e^{-j\omega_m t} \right) \right] \end{aligned}$$

In this form the carrier phasor is multiplied by the sum of a stationary vector and two rotating vectors of equal size which rotate

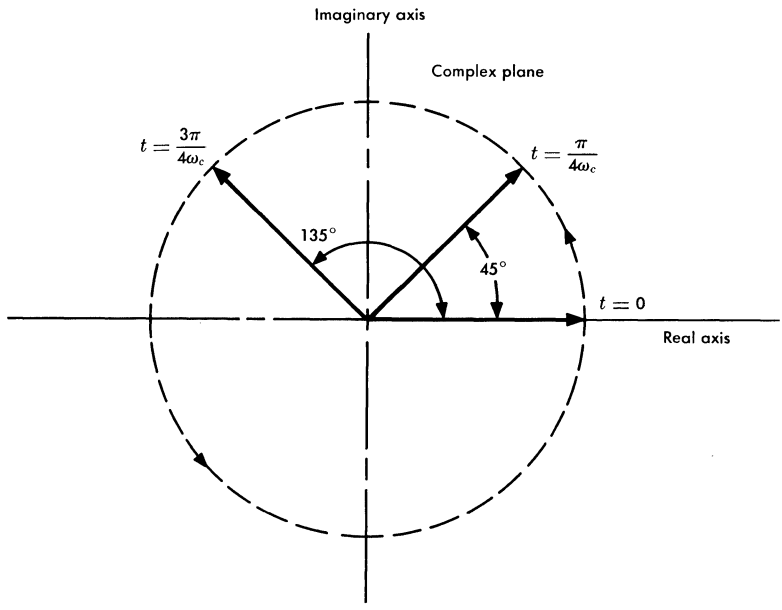


FIG. 5-3. Phasor diagram of $e^{j\omega_c t}$ for various times.

in opposite directions. As may be seen in Fig. 5-4, the sum of these three vectors is always real and, consequently, acts only to modify the length of the rotating carrier phasor. This produces amplitude modulation as expected.

At this point the average power in the carrier and in the side-frequencies should be considered. For a unit amplitude carrier and a circuit impedance such that average carrier power is 1 watt, the power in each side-frequency will be $m^2/4$ watts, or a total sideband power of $m^2/2$ watts. Thus, for 100 per cent modulation, only one-third of the total power is in the information-bearing sidebands. The sidebands get an even smaller share of the total power when the modulating function is a speech signal which has a higher peak factor than a sinusoid. The sideband power must then be reduced to a few per cent of the total power to prevent occasional peaks from over-modulating the carrier.

The DSBTC signal is sensitive to certain types of transmission phase distortion. It is not impaired (except for an absolute delay) by a transmission phase characteristic that is linear with the frequency.

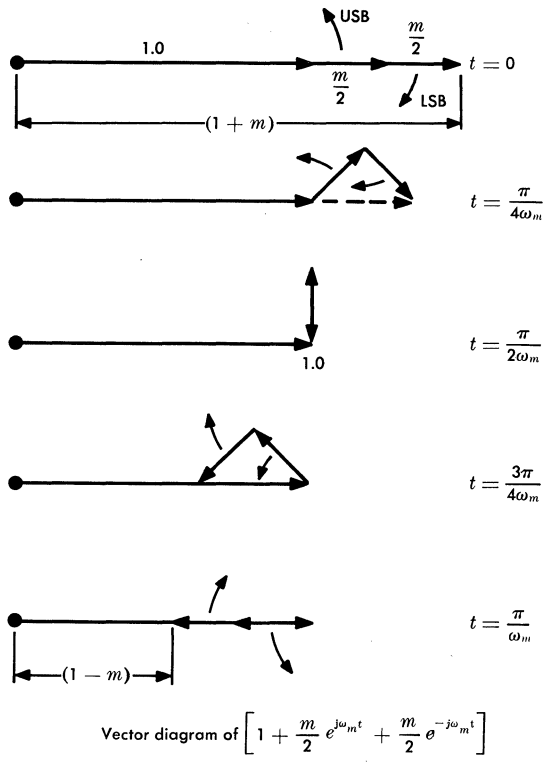


FIG. 5-4. Amplitude modulation—index of modulation = m.

The basic requirement is odd symmetry of phase about the carrier frequency, a condition that is met by a linear phase characteristic in any segment of its frequency range.

An interesting degradation occurs under certain extreme transmission-phase conditions. Suppose that the lower side-frequency vector in Fig. 5-4 is shifted clockwise by θ degrees, and the upper side frequency is shifted clockwise by $180 - \theta$ degrees. The resulting signal, Fig. 5-5, consists of a carrier phasor with the side-frequency vectors adding at right angles. The resultant vector represents a phase-modulated wave whose amplitude modulation has been largely cancelled, or washed out. A low-index DSBTC signal so distorted is indistinguishable from a low-index phase-modulated signal. This washout effect will be seen again in the case of product demodulation of DSBSC, as a result of a phase error in the local carrier used for

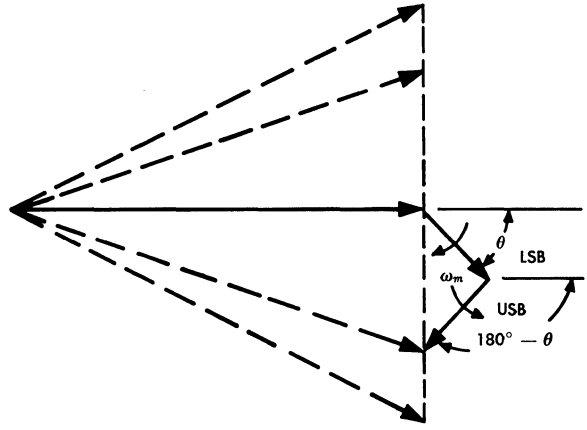


FIG. 5-5. Result of certain extreme phase distortion of DSBTC signal to produce phase modulation.

demodulation. This can be seen by shifting only the stationary unit phasor (the carrier) of Fig. 5-4 by 90 degrees to obtain the washout result of Fig. 5-5.

Double-Sideband Suppressed Carrier

The DSBSC signal requires the same transmission bandwidth as DSBTC, but the power efficiency is improved by the suppression of the carrier. This requires reintroduction of the carrier at the receiving terminal, which must be done with extreme phase accuracy to avoid the type of washout distortion just discussed. Examination of Fig. 5-4 shows that a phase error of the inserted carrier of θ degrees results in the effective amplitude modulation being reduced by the factor cos θ. If the phase error $= \Delta\omega_e t$ (the inserted carrier has a frequency error Δf_e) and the baseband signal is a single-frequency sinusoid, the demodulated signal will consist of two sinusoids separated by twice the error frequency Δf_e.

The difficulty of accurately reinserting the carrier is the greatest disadvantage of DSBSC and is probably the reason this form has not seen more use. However, the transmitted sidebands contain the information required to establish the exact frequency and, except for a 180 degree ambiguity, the phase of the required demodulating carrier. This is so by virtue of symmetry about the carrier frequency, even with a random modulating wave. One means of establishing the

carrier at f_c is to square the DSBSC wave, filter the component present at frequency $2f_c$, and electrically divide the frequency by two [3]. This can be shown by squaring Eq. (5-5) to obtain:

$$M^2(t) = \frac{1}{4}\cos^2(\omega_c + \omega_m)t + \frac{1}{4}\cos^2(\omega_c - \omega_m)t$$

$$+ \frac{1}{2}\cos(\omega_c + \omega_m)t\cos(\omega_c - \omega_m)t \qquad (5\text{-}9)$$

By a trigonometric identity the last term in Eq. (5-9) can be written:

$$\frac{1}{2}\cos(\omega_c + \omega_m)t\cos(\omega_c - \omega_m)t = \frac{1}{4}\cos 2\omega_c t + \frac{1}{4}\cos 2\omega_m t \qquad (5\text{-}10)$$

The $\cos 2\omega_c t$ term can be easily separated by filtering to yield a sinusoid at $2f_c$ whose frequency can be electronically halved. The same result would follow from a similar analysis of a complex modulating wave. It should be noted that a carrier thus derived will disappear in the absence of modulation.

Single-Sideband

The single-sideband signal is not subject to the modulation washout effect discussed in connection with the DSB signals. In fact, the local carrier at the receiving terminal is conventionally allowed to have a slight frequency error. This produces a frequency shift in each demodulated baseband component. If the error is kept within 1 or 2 Hz, the system is adequate for high-quality telephone circuits. However, the single-sideband method of transmission with a fixed or rotating phase error in demodulation does not preserve the baseband waveform at all. This may be seen by considering the phasor representing the upper sideband signal arising from a single baseband frequency component at f_m, Fig. 5-6. The dashed line represents the reference carrier phasor about which the sideband rotates with a relative angular velocity, ω_m.

If a strong carrier of reference phase is added to the received sideband (as could be done in the receiving terminal just ahead of an envelope detector), the envelope of the resultant wave will be nearly sinusoidal and will peak when the sideband phasor aligns itself with the carrier. An envelope detector would produce in the proper phase a nearly sinusoidal wave of frequency f_m.

If the carrier phase is advanced 90 degrees, the peaks in the demodulated wave will occur 90 degrees later, so that the baseband signal is retarded by 90 degrees. Although this does not distort the waveform of the single-frequency wave considered, in a complex baseband wave each frequency component will be retarded 90 degrees, which will cause gross waveform distortion as illustrated in Fig. 5-7. In this figure the baseband fundamental and the third harmonic are both shifted 90 degrees. Although an envelope detector is assumed here, similar results would follow from analyzing product detection of the SSB signal.

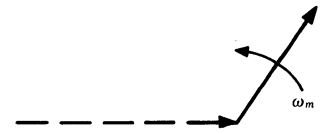
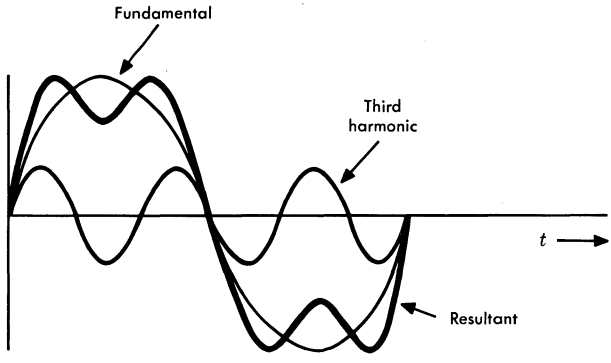


FIG. 5-6. Upper sideband and reference carrier phasors for SSB signal.

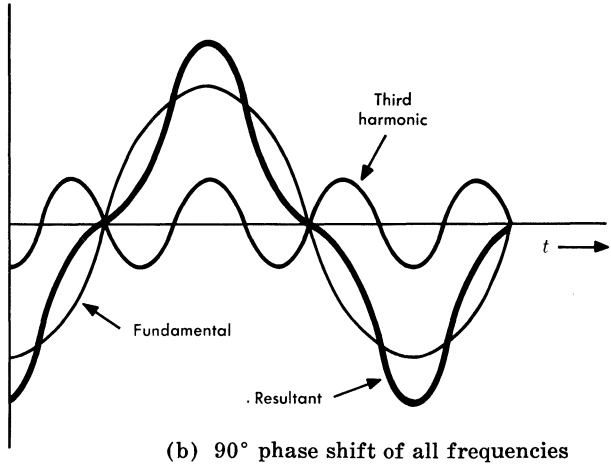
This type of distortion is called quadrature distortion. An SSB signal can be represented as two DSB signal pairs superimposed as in Fig. 5-8. One DSB pair is chosen so that its resultant aligns with the carrier, and the other pair has its resultant at right angles or in quadrature.

Thus, SSB signals inherently contain quadrature components. The first step in reducing or eliminating quadrature distortion is to introduce the local carrier in the proper phase to serve as a reference against which the in-phase or direct component may be identified. The second step is to design the demodulator to respond only to the direct component. Since the quadrature component results primarily in phase modulation of the carrier, it is important that the demodulator be insensitive to these phase perturbations. The use of a strong or *exalted* local carrier followed by an envelope detector is one way of approaching the desired condition. A more exact way to eliminate the quadrature component is to use a product detector.

Voice transmission is very tolerant of quadrature distortion. As a consequence, the design of early carrier systems allowed reintroduction of the carrier with a frequency error. The resulting severe quadrature distortion renders these systems unsuitable for transmission of accurate baseband waveforms and makes these systems theoretically unfit for data pulse transmission. Also, many data signals contain very low frequency or even d-c components. An SSB system will not transmit these components since practical filters cannot be built to suppress all of the unwanted sideband without



(a) Reference condition



(b) 90° phase shift of all frequencies

FIG. 5-7. Waveform distortion due to 90° reference carrier phase error causing 90° lag of all frequencies.

cutting into the carrier frequency (equivalent to d-c) and the equivalent low frequencies of the wanted sideband.

A common technique used in carrying data traffic on SSB channels is to modulate a subcarrier in the data terminal, using angle modulation or types of amplitude modulation which permit transmission of d-c components. This also solves the quadrature distortion problem, since the subcarrier is transmitted and used in the ultimate demodu-

lation in the receiving data terminal. Since the data subcarrier and the data sidebands travel the same path, the former provides the proper reference information for demodulating the latter, even in the presence of frequency shift. Of course, the baseband channel must be adequately equalized for delay and attenuation.

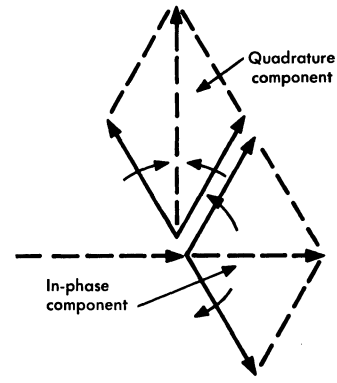


FIG. 5-8. Analysis of SSB signal into in-phase and quadrature components.

Single-sideband is the modulation technique usually used for frequency division multiplex of several message channels prior to transmission over broadband facilities. Actually, SSB techniques are often used for interim frequency translations in the multiplex terminal for purposes of convenient filtering [4]. The bandwidth of the signal, measured in octaves, may be increased or decreased by such translations.

Vestigial Sideband

Vestigial sideband (VSB) modulation is a modification of DSB in which part of the frequency spectrum is suppressed. It can be produced by passing a DSB wave through a filter to remove part of one sideband as shown in Fig. 5-9. The demodulation of such a wave results in addition of the lower and upper sideband components to form the baseband signal. To preserve the baseband frequency spectrum, it is necessary to keep the filter-cutoff characteristic symmetrical about the carrier frequency. This will result in the spectrum of the sideband vestige effectively complementing the attenuated portion of the desired sideband. For the same reason, and to avoid quadrature distortion, the phase must exhibit odd symmetry about the carrier frequency. As long as the cutoff is symmetrical about the carrier, it can be gradual (approaching DSB conditions), or sharp (approaching SSB conditions), or anywhere between these extremes.

The desired transmission characteristic is shared among the transmitting and receiving terminals and the transmission medium. The apportioning of the characteristic is determined by economics and signal-to-noise considerations.

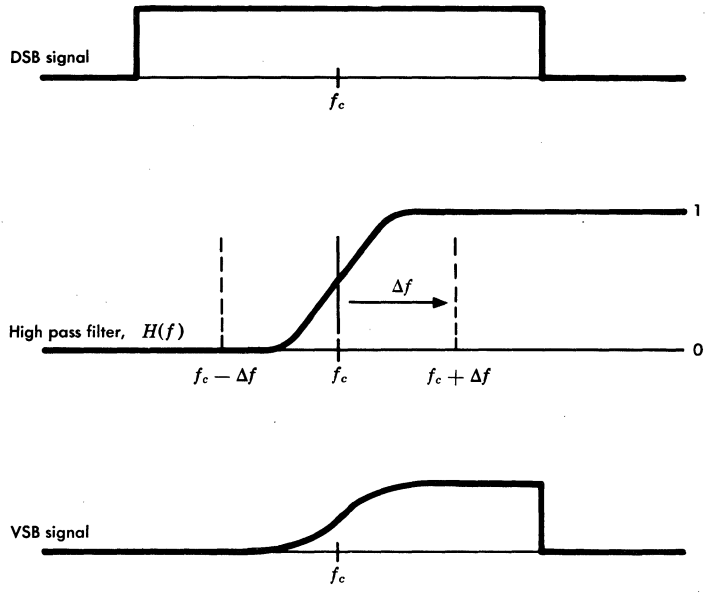


FIG. 5-9. Generation of VSB wave. For no distortion, $1 - H(f_c + \Delta f) = H(f_c - \Delta f)$.

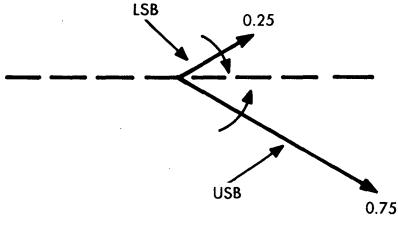


FIG. 5-10. VSB phasors for intermediate modulating frequency.

The VSB signal is similar to DSBTC for low baseband frequencies and to SSB for high baseband frequencies. In the cutoff region the behavior is as shown in Fig. 5-10. The upper and lower side-frequency vectors add to unity when they peak along the reference carrier line, and will, if properly demodulated, produce the same baseband signal as an SSB side-frequency of unit amplitude.

VSB has the virtue of conserving bandwidth almost as efficiently as SSB, while retaining the excellent low-frequency baseband characteristics of DSB. Although the ideal SSB signal should allow the sideband spectrum to extend all the way to the carrier frequency, practical limitations on filters and phase distortion make it impractical. Thus, VSB has become standard for television and similar

signals where good phase characteristics and transmission of low-frequency components are important, but the bandwidth required for DSB transmission is unavailable or uneconomical.

5.2 PROPERTIES OF ANGLE-MODULATED SIGNALS

Referring to Eq. (5-1) with $a(t)$ held constant,

$$M(t) = A_c \cos [\omega_c t + \phi(t)] \tag{5-11}$$

where $\phi(t)$ is the angle modulation in radians. If angle modulation is used to transmit information, it is necessary that $\phi(t)$ be a prescribed function of the modulating signal. For example, if $v(t)$ is the modulating signal, the angle modulation $\phi(t)$ can be expressed as some function of $v(t)$.

Many varieties of angle modulation are possible depending on the selection of the functional relationship between the angle and the modulating wave. Two of these are important enough to have the individual names of phase modulation (PM) and frequency modulation (FM).

Phase Modulation and Frequency Modulation

The difference between phase and frequency modulation can be understood by first defining four terms with reference to Eq. (5-11).

$$\text{Instantaneous phase} = \omega_c t + \phi(t) \quad \text{rad} \tag{5-12}$$

$$\text{Instantaneous phase deviation} = \phi(t) \quad \text{rad} \tag{5-13}$$

$$\text{Instantaneous frequency}^* = \frac{d}{dt} [\omega_c t + \phi(t)]$$

$$= \omega_c + \phi'(t) \quad \text{rad/sec} \tag{5-14}$$

$$\text{Instantaneous frequency deviation} = \phi'(t) \quad \text{rad/sec} \tag{5-15}$$

Phase modulation can then be defined as angle modulation in which the instantaneous phase deviation, $\phi(t)$, is proportional to the

*The instantaneous frequency of an angle-modulated carrier is defined as the first time derivative of the instantaneous phase.

modulating signal voltage, $v(t)$. Similarly, frequency modulation is angle modulation in which the instantaneous frequency deviation, $\phi'(t)$, is proportional to the modulating signal voltage, $v(t)$. Mathematically, these statements become, for phase modulation,

$$\phi(t) = kv(t) \qquad \text{rad} \tag{5-16}$$

and for frequency modulation,

$$\phi'(t) = k_1 v(t) \qquad \text{rad/sec} \tag{5-17}$$

from which

$$\phi(t) = k_1 \int v(t)\,dt \qquad \text{rad} \tag{5-18}$$

where k and k_1 are constants.

These results are summarized in the table of Fig. 5-11. This table also illustrates phase-modulated and frequency-modulated waves which occur when the modulating wave is a single sinusoid.

Type of modulation	Modulating signal	Angle-modulated carrier
(*a*) Phase	$v(t)$	$M(t) = A_c \cos [\omega_c t + kv(t)]$
(*b*) Frequency	$v(t)$	$M(t) = A_c \cos [\omega_c t + k_1 \int v(t)\,dt]$
(*c*) Phase	$A_m \cos \omega_m t$	$M(t) = A_c \cos (\omega_c t + kA_m \cos \omega_m t)$
(*d*) Frequency	$-A_m \sin \omega_m t$	$M(t) = A_c \cos \left(\omega_c t + \frac{k_1 A_m}{\omega_m} \cos \omega_m t \right)$
(*e*) Frequency	$A_m \cos \omega_m t$	$M(t) = A_c \cos \left(\omega_c t + \frac{k_1 A_m}{\omega_m} \sin \omega_m t \right)$

FIG. 5-11. Equations for phase- and frequency-modulated carriers

Figure 5-12 illustrates amplitude, phase, and frequency modulation of a carrier by a single sinusoid. The similarity of waveforms of the PM and FM waves shows that for angle-modulated waves it is necessary to know the modulation function; that is, the waveform alone

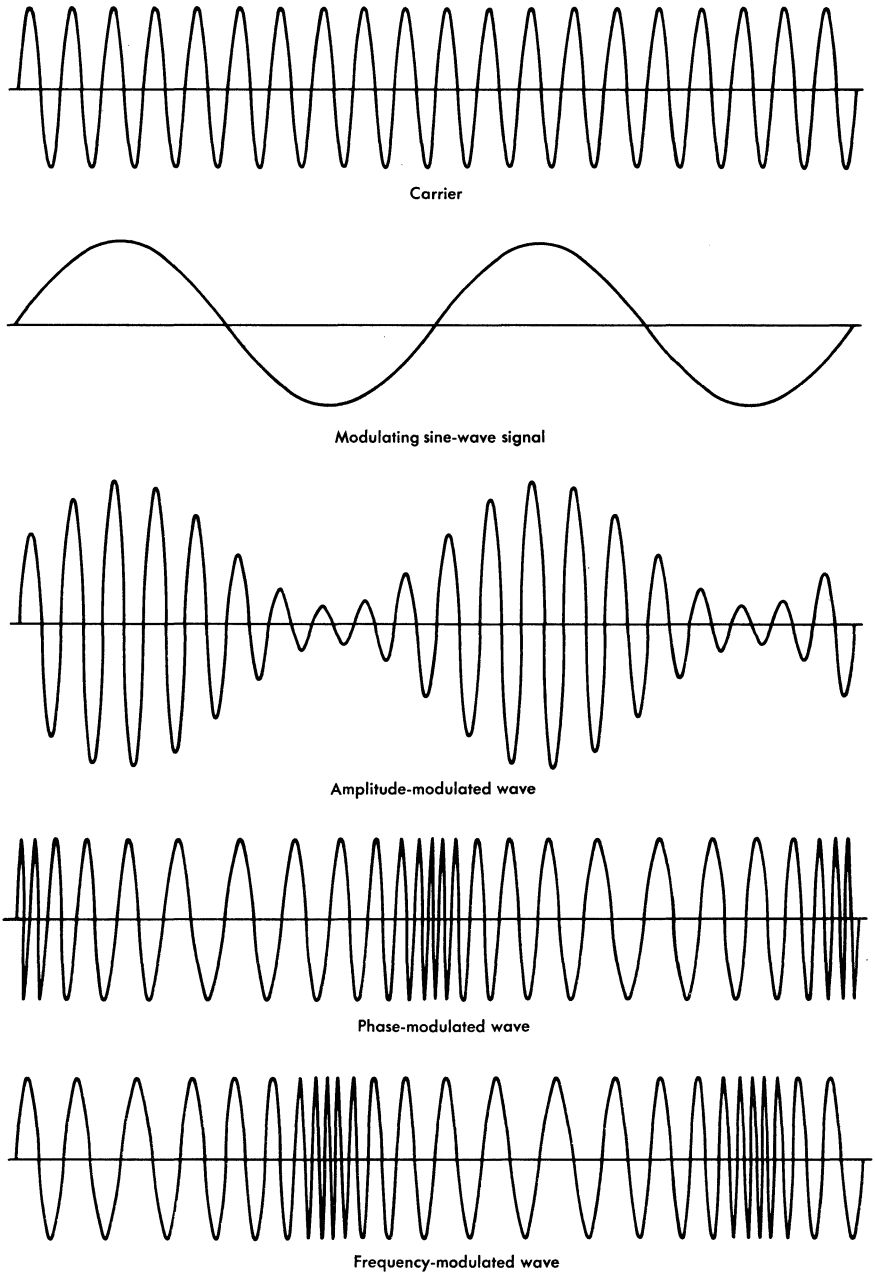


FIG. 5-12. Amplitude, phase, and frequency modulation of a sine-wave carrier by a sine-wave signal.

cannot be used to distinguish between PM and FM. Similarly, it is not apparent from Eq. (5-11) whether an FM or a PM wave is represented. It could be either. A knowledge of the modulation function, however, will permit correct identification. If $\phi(t) = kv(t)$, it is phase modulation, and if $\phi'(t) = k_1 v(t)$, it is frequency modulation.

Comparison of (c), (d), and (e) in Fig. 5-11 shows that the expression for a carrier which is phase or frequency modulated by a sinusoidal type signal can be written in the general form of

$$M(t) = A_c \cos (\omega_c t + X \cos \omega_m t) \tag{5-19}$$

where

$$X = kA_m \qquad \text{rad for PM} \tag{5-20}$$

and

$$X = \frac{k_1 A_m}{\omega_m} \qquad \text{rad for FM} \tag{5-21}$$

Here X is the peak phase deviation in radians and is called the index of modulation. For PM the index of modulation is a constant, independent of the frequency of the modulating wave; for FM it is inversely proportional to the frequency of the modulating wave. Note that in the FM case, the modulation index can also be expressed as the peak frequency deviation, $k_1 A_m$, divided by the modulating signal frequency, ω_m. The terms high-index and low-index of modulation are often used. It is difficult to define a sharp division; however, in general the term low-index is used when the peak phase deviation is less than 1 radian. It is shown later that the frequency spectrum of the modulated wave is dependent on the index of modulation.

When the modulation function consists of a single sinusoid, it is evident from Eq. (5-19) that the phase angle of the carrier varies from its unmodulated value in a simple sinusoidal fashion, with the peak phase deviation being equal to X. The phase deviation can also be expressed in terms of the mean square phase deviation, D_ϕ, which for this case is $X^2/2$. Similarly, the frequency deviation of a sinusoidally modulated carrier can be expressed either in terms of the peak frequency deviation, $k_1 A_m$ rad/sec or $k_1 A_m/2\pi$ Hz, or the mean square frequency deviation, D_f, which is $k_1^2 A_m^2/8\pi^2$ Hz^2. In the more general case where the modulation function is a complex signal (such as a multichannel telephone signal or noise), the peak value

of the modulation function, and hence the peak phase or frequency deviation, is not a precisely defined magnitude. For these cases, the phase or frequency deviation of the carrier is normally expressed in terms of the mean square or rms value.

Phasor Representation

A wave angle modulated by sinusoids can be represented by phasors as was done for the AM waves. Generally, the angle-modulated case is more complex as can be seen by expanding Eq. (5-19) into a Bessel series of sinusoids in the manner shown in Chap. 19. In the special case of very low index (X less than 1/2 radian), all terms after the first can be ignored, and the phasor diagram is very similar to that for an AM wave except for a phase reversal of one of the sidebands. In fact, it was pointed out in the AM discussion that if the inserted carrier of a DSBSC signal has a phase error of 90 degrees, severe washout occurs, and the previously amplitude-modulated wave has very little amplitude modulation but considerable phase (or angle) modulation. The approximate phasor diagram for a low-index angle-modulated system modulated by a single-frequency sinusoid at f_m is shown in Fig. 5-13 for several values of time. The resultant vector has an amplitude close to unity at all times and an index, or maximum phase deviation, of X radians. A true angle-modulated wave would include higher order terms and would have no amplitude variation. If X is small enough these terms are often ignored.

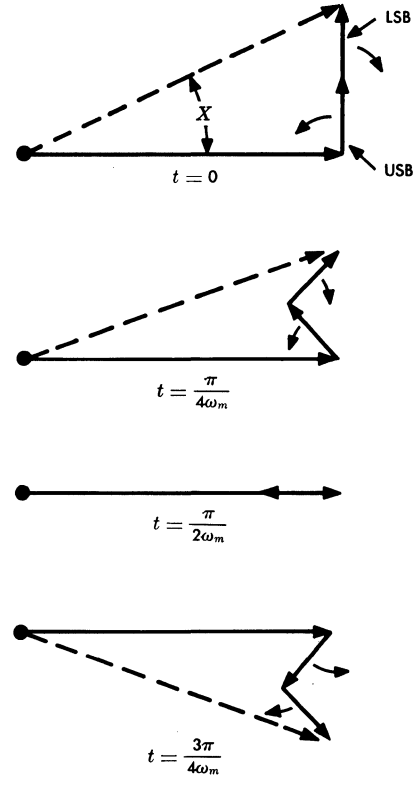


FIG. 5-13. Phase modulation — low index.

Several interesting conclusions may be observed by comparing the low-index angle-modulated wave with the AM signal shown in Fig. 5-4. Both types of modulation are

similar in the sense that they both contain the carrier and the same first order sideband frequency components. In fact, for the low-index case the amplitudes of the first order sidebands are approximately the same when the indices are equal ($X = m$). The important difference is the phase of the sideband components. It may be expected, therefore, that in the transmission of an FM or a PM wave the phase characteristic of the transmission path will be extremely important and that certain phase irregularities could easily convert phase-modulation components into amplitude-modulation components.

Average Power of an Angle-Modulated Wave

The average power of an FM or a PM wave is independent of the modulating signal and is equal to the average power of the carrier when the modulation is zero. Hence, the modulation process takes power from the carrier and distributes it among the many sidebands but does not alter the average power present. This may be demonstrated by assuming a voltage of the form of Eq. (5-11), squaring, and dividing by a resistance, R, to obtain

$$P(t) = \frac{M^2(t)}{R}$$

$$= \frac{A_c^2}{R} \cos^2 [\omega_c t + \phi(t)]$$

$$= \frac{A_c^2}{R} \left\{ \frac{1}{2} + \frac{1}{2} \cos [2\omega_c t + 2\phi(t)] \right\} \tag{5-22}$$

The second term can be assumed to consist of a large number of sinusoidal sideband components about a carrier frequency of $2f_c$ Hz; therefore, the average value of the second term of Eq. (5-22) is zero. Thus, the average power is given by the zero frequency term

$$P_{\text{avg}} = \frac{A_c^2}{2R} \tag{5-23}$$

This, of course, is the same as the average power in the absence of modulation.

Bandwidth Required for Angle-Modulated Waves

For the low-index case where the peak phase deviation is less than 1 radian, most of the signal information of an angle-modulated wave is carried by the first order sidebands. It follows that the bandwidth required is at least twice the frequency of the highest frequency component of interest in the modulating signal. This would permit the transmission of the entire first order sideband.

For the high-index signal a different method called the quasi-stationary approach must be used [5]. In this approach the assumption is made that the modulating waveform is changing very slowly so that static response can be used. For example, if 1 volt at baseband causes a 1-MHz frequency deviation (corresponding to $k_1 = 2\pi \times 10^6$ radians per volt-sec) and the modulating signal has a 1-volt peak, the peak frequency deviation will be 1 MHz. It is obvious that if the *rate of change of frequency is very small,* the bandwidth is determined by the peak-to-peak frequency deviation. It was mathematically proven by J. R. Carson in 1922 that frequency modulation could not be accommodated in a narrower band than amplitude modulation, but might actually require a wider band [6]. The quasi-stationary approach for large index indicates that the minimum bandwidth required is equal to the peak-to-peak (or twice the peak) frequency deviation.

Thus, for low-index systems ($X < 1$) the minimum bandwidth is given by $2f_T$, where f_T is the highest frequency in the modulating signal. For high-index systems ($X > 10$) the minimum bandwidth is given by $2\Delta F$, where ΔF is the peak frequency deviation. It would be desirable to have an estimate of the bandwidth for all angle-modulated systems regardless of index. A general rule (first stated by J. R. Carson in an unpublished memorandum dated August 28, 1939) is that the minimum bandwidth required for the transmission of an angle-modulated signal is equal to two times the sum of the peak frequency deviation and the highest modulating frequency to be transmitted. In the notation previously defined

$$B_w = 2(f_T + \Delta F) \qquad \text{Hz} \tag{5-24}$$

This rule (called Carson's rule) gives results which agree quite well with the bandwidths actually used in the Bell System. It should be realized, however, that this is only an approximate rule and that the actual bandwidth required is to some extent a function of the waveform of the modulating signal and the quality of transmission desired.

5.3 PROPERTIES OF PULSE MODULATION

In pulse modulation systems the unmodulated carrier is usually a series of regularly recurrent pulses. Modulation results from varying some parameter of the transmitted pulses, such as the amplitude, duration, shape, or timing. If the baseband signal is a continuous waveform, it will be broken up by the discrete nature of the pulses. In considering the feasibility of pulse modulation, it is important to recognize that the continuous transmission of information describing the modulating function is unnecessary, provided that the modulating function is bandlimited and the pulses occur often enough. The necessary conditions are expressed by the sampling principle, as subsequently discussed.

It is usually convenient to specify the signaling speed or pulse rate in *bauds*. By definition, the speed in bauds is equal to the number of signaling elements or symbols per second. Thus, the baud denotes pulses per second in a manner parallel to hertz denoting cycles per second. Note that all possible pulses are counted whether or not a pulse is sent since no pulse is also usually a valid symbol. Since there is no restriction on the allowed amplitudes of the pulses, a baud can contain any arbitrary information rate in bits per second. Unfortunately, the term bits per second is often used incorrectly to specify a digital transmission rate in bauds. For binary symbols, the information rate in bits per second is equal to the signaling speed in bauds. In general, the relation between information rate and signaling rate depends upon the coding scheme employed.

Sampling

In any physically realizable transmission system, the message or modulating function is limited to a finite frequency band. Such a bandlimited function is continuous with time and limited in its possible range of excursions in a small time interval. Thus, it is only necessary to specify the amplitude of the function at discrete time intervals in order to specify it exactly. The basic principle discussed here is called the sampling theorem, which in a restricted form states [7]:

> If a message that is a magnitude-time function is sampled instantaneously at regular intervals and at a rate at least twice the highest significant message frequency, then the samples contain all of the information of the original message.

The application of the sampling theorem reduces the problem of transmitting a continuously varying message to one of transmitting a discrete number of amplitude values. For example, a message bandlimited to f_T hertz is completely specified by the amplitudes at any set of points in time spaced T seconds apart, where $T = 1/2f_T$ [8]. The time interval, T, is often referred to as the Nyquist interval. Hence, to transmit a bandlimited message, it is only necessary to transmit $2f_T$ independent values per second.

The process of sampling can be thought of as the product modulation of a message function and a set of impulses, as shown in Fig. 5-14. The message function of time, $v(t)$, is multiplied by a train of impulses, $c(t)$, to produce a series of amplitude-modulated pulses, $s(t)$. If the spectrum (i.e., the Fourier transform) of $v(t)$ is given by $F(f)$ as shown in Fig. 5-14, the spectrum of the sampled wave, $s(t)$, will be as shown by $S(f)$ in the figure. The output spectrum, $S(f)$, is periodic on the frequency scale with period f_s, the sampling frequency. It is important to note that a pair of sidebands has been produced around d-c, f_s, $2f_s$, and so on through each harmonic of the sampling frequency. This figure also shows the need for $f_s > 2f_T$, so that the sidebands do not overlap. Note also that all sidebands around all harmonics of the sampling frequency have the same amplitude. This is a result of the fact that the frequency spectrum of an impulse is flat with frequency. In a practical case, of course, finite width pulses would have to be used for the sampling function, and the spectrum of the sampled signal would fall off with frequency as the spectrum of the sampling function does.

The amplitude-modulated pulse signal that results from sampling the input message may be transmitted to the receiver in any form which is convenient or desirable from a transmission standpoint. At the receiver the incoming signal, which may no longer resemble the impulse train, must be operated on to recreate the original pulse amplitude-modulated sample values in their original time sequence at a rate of $2f_T$ samples per second. To reconstruct the message, it is necessary to generate from each sample a proportional impulse, and to pass this regularly spaced series of impulses through an ideal low-pass filter with a cutoff frequency f_T. Examination of the spectrum of $S(f)$ in Fig. 5-14 makes the feasibility of this obvious. Except for an overall time delay and possibly a constant of proportionality, the output of this filter will then be identical to the original message. Ideally, then, it is possible to transmit information exactly,

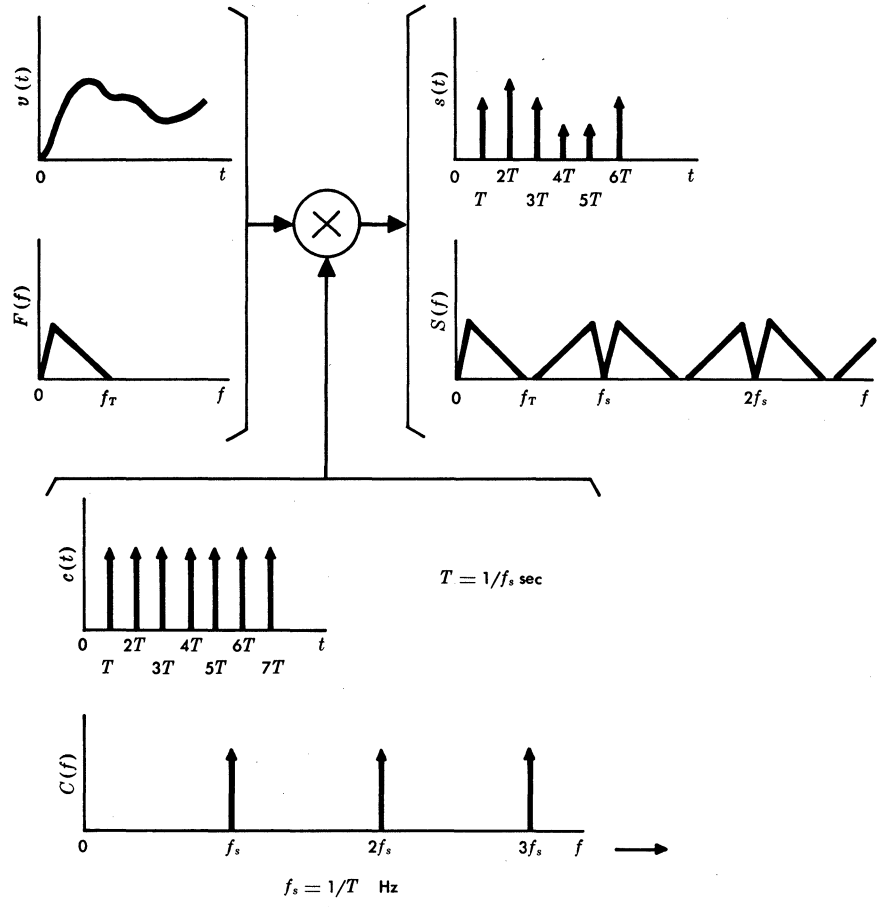


FIG. 5-14. Sampling with an impulse modulator.

given the instantaneous amplitude of the message at intervals spaced not further than $1/2f_T$ seconds apart.

Pulse Amplitude Modulation

In pulse amplitude modulation (PAM) the amplitude of a pulse carrier is varied in accordance with the value of the modulating wave as shown in Fig. 5-15(c). It is convenient to look upon PAM as modulation in which the value of each instantaneous sample of the modulating wave is caused to modulate the amplitude of a pulse. Signal

processing in time division multiplex terminals often begins with PAM although further processing usually takes place before the signal is launched onto a transmission system.

Pulse Duration Modulation

Pulse duration modulation (PDM), sometimes referred to as pulse length modulation or pulse width modulation, is a particular form of pulse time modulation. It is modulation of a pulse carrier in which the value of each instantaneous sample of a continuously varying modulating wave is caused to produce a pulse of proportional duration, as shown in Fig. 5-15(d). The modulating wave may vary

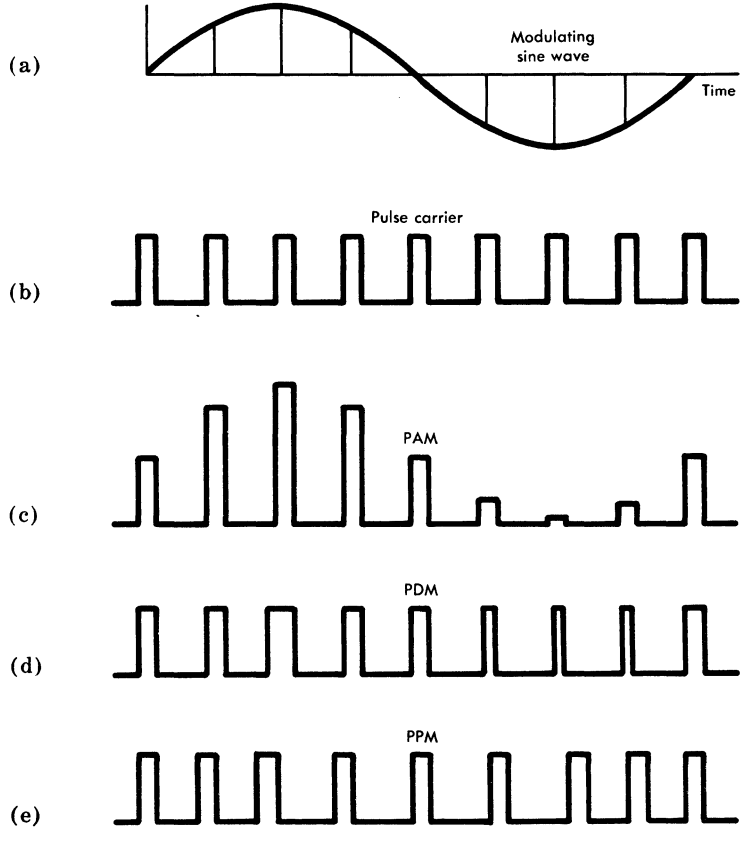


FIG. 5-15. Examples of pulse-modulation systems.

the time of occurrence of the leading edge, the trailing edge, or both edges of the pulse. In any case, the message to be transmitted is composed of sample values at discrete times, and each value must be uniquely defined by the duration of a modulated pulse.

In PDM, long pulses expend considerable power during the pulse while bearing no additional information. If this unused power is subtracted from PDM so that only transitions are preserved, another type of pulse modulation, called pulse position modulation, results. The power saved represents the fundamental advantage of pulse position modulation over PDM.

Pulse Position Modulation

A particular form of pulse time modulation in which the value of each instantaneous sample of a modulating wave varies the position of a pulse relative to its unmodulated time of occurrence is pulse position modulation (PPM). This is illustrated in Fig. 5-15(e). The variation in relative position may be related to the modulating wave in any predetermined unique manner. Practical applications of PPM systems have been on a modest scale, even though their instrumentation can be extremely simple.

If either PDM or PPM is used to time division multiplex several channels, the maximum modulating signal must not cause a pulse to enter adjacent allotted time intervals. In telephone systems with high peak factors, this requirement leads to a very wasteful use of time space. In fact, almost all of the time available for modulation is wasted because many of the busy channels may be expected to be inactive, and most of the rest will be carrying small signal power. Consequently, although PPM is more efficient than PDM, both fall short of the theoretical ideal when used for multiplexing ordinary telephone channels.

Pulse Code Modulation

Instead of attempting the impossible task of transmitting the exact amplitude of a sampled signal, suppose only certain discrete amplitudes of sample size are allowed. Then, when the message is sampled in a PAM system, the discrete amplitude nearest the true amplitude is sent. When this is received and amplified, it will have an amplitude slightly different from any of the specified discrete steps, because of the disturbances encountered in transmission. But if the noise and distortion are not too great, it will be possible to tell accurately

which discrete amplitude of the signal was transmitted. Then the signal can be reformed, or a new signal created which has the amplitude originally sent.

Representing the message by a discrete and therefore limited number of signal amplitudes is called *quantizing*. It inherently introduces an initial error in the amplitude of the samples, giving rise to quantization noise. But once the message information is in a quantized state, it can be relayed for any distance without further loss in quality, provided only that the added noise in the signal received at each repeater is not too great to prevent correct recognition of the particular amplitude each given signal is intended to represent. If the received signal lies between a and b and is closer to b, it is surmised that b was sent. If the noise is small enough, there will be no errors. Note therefore that in quantized signal transmission the maximum noise is selected by the number of bits in the code, while in analog signal transmission, it is determined by the repeater spacing and characteristics of the medium.

Coding and Decoding. A quantized sample can be sent as a single pulse having certain possible discrete amplitudes or certain discrete positions with respect to a reference position. However, if many discrete sample amplitudes are required (one hundred, for example), it is difficult to design circuits that can distinguish between amplitudes. It is much less difficult to design a circuit that can determine whether or not a pulse is present. If several pulses are used as a code group to describe the amplitude of a single sample, each pulse can be present (1) or absent (0). For instance, if three pulse positions are used, then a code can be devised to represent the eight different amplitudes shown in Fig. 5-16.

These codes are, in fact, just the numbers (amplitudes) at the left written in binary notation. In general, a code group of n on-off pulses can be used to represent 2^n amplitudes. For example, 7 pulses yield 128 sample levels.

It is possible, of course, to code the amplitude in terms of a number of pulses which have discrete amplitudes of 0, 1, and 2 (ternary or base 3), or 0, 1, 2, and 3 (quaternary or base 4), etc., instead of the pulses with amplitudes 0 and 1 (binary or base 2). If ten levels are allowed for each pulse, then each pulse in a code group is simply a digit or an ordinary decimal number expressing the amplitude of the sample. If n is the number of pulses and b is the base, the number of quantizing levels the code can express is b^n. To decode this code group, it is necessary to generate a pulse which is the linear sum of

Amplitude represented	Code
0	000
1	001
2	010
3	011
4	100
5	101
6	110
7	111

FIG. 5-16. Binary code representation of sample amplitudes

all the pulses in the group, each pulse of which is multiplied by its place value $(1,b,b^2,b^3 \ldots)$ in the code. Systems using codes to represent discrete signal amplitudes are called pulse code modulation or PCM systems.

REFERENCES

1. Panter, P. F. *Modulation, Noise, and Spectral Analysis* (New York: McGraw-Hill, Inc., 1965), p. 1.
2. Terman, F. E. *Electronic and Radio Engineering* (New York: McGraw-Hill, Inc., 1955), p. 523.
3. Rieke, J. W. and R. S. Graham. "The L3 Coaxial System—Television Terminals," *Bell System Tech. J.*, vol. 32 (July 1953), pp. 915-942.
4. Blecher, F. H. and F. J. Hallenbeck. "The Transistorized A5 Channel Bank for Broadband Systems," *Bell System Tech. J.*, vol. 41 (Jan. 1952), pp. 321-359.
5. Rowe, H. E. *Signals and Noise in Communication Systems* (Princeton: D. Van Nostrand Company, 1965), pp. 103, 119-124.
6. Carson, J. R. "Notes on the Theory of Modulation," *Proc. IRE* (Feb. 1922).
7. Black, H. S. *Modulation Theory* (Princeton: D. Van Nostrand Company, 1953), p. 37.
8. Oliver, B. M., J. R. Pierce, and C. E. Shannon. "The Philosophy of PCM," *Proc. IRE*, vol. 36 (Nov. 1948), pp. 1324-1331.

Chapter 6

Signal Multiplexing

Multiplexing in transmission systems is a means of ultilizing the same transmission medium for many different users. Before being placed on the transmission medium, each user's signal may have to be modified in some unique way so that it can be separated from all of the other signals at the opposite end of the transmission path. This separation involves basically the inverse of the original modification. There are several ways in which signals can be multiplexed, the most important being space division multiplex, frequency division multiplex, and time division multiplex.

6.1 SPACE DIVISION MULTIPLEX

Space division multiplex is simply the bundling of many physically separate transmission paths into a common cable. A telephone cable consisting of hundreds (or thousands) of twisted pairs constitutes a space division multiplex system since many conversations can be carried on the single cable although each is assigned a unique pair in the cable. Such a scheme is obviously economical when it is remembered that the transmission right-of-way absorbs a substantial part of the cost of any transmission system. The advantages of space division multiplex do not come free, however. First of all, the traffic must be combined into specific routes to achieve the desirably large channel cross sections. Secondly, achieving true isolation between transmission media separated by distances that can be 10^{-8} times their length is nearly impossible. As a consequence, such systems are subject to interference resulting from coupling between the channels. Chapter 11 discusses this undesired coupling in more detail.

It should be emphasized that space division multiplex is not confined to voice-frequency circuits. In fact, many of the high-capacity transmission systems (either frequency or time division multiplexed) can in turn be space division multiplexed on parallel facilities sharing the same right-of-way. A well known example of this is the L4-carrier system which puts thousands of message channels on a single coaxial line that is in turn part of a large cable containing many coaxials.

6.2 FREQUENCY DIVISION MULTIPLEX

In frequency division multiplex (FDM) each channel of the system is assigned a discrete portion of the transmitted frequency spectrum. Thus, many narrow bandwidth channels can be accommodated by a single wide bandwidth transmission system. To illustrate the concept of FDM, a telegraph channel (d-c to 100 Hz) and a telephone channel (200 to 3400 Hz) could easily be placed on a common transmission medium as long as the medium could handle all frequencies from d-c to 3400 Hz. To separate the signals at the receiving end would require only a low-pass filter for the telegraph channel and a high-pass filter for the telephone channel. A more common FDM problem, however, is to multiplex several channels, each of which occupies the same portion of the frequency spectrum. Then, the multiplexing operation involves a frequency shift of each channel before launching it on the broadband facility. At the receiving end, each channel has to be shifted back to the original frequency spectrum.

Although this discussion is primarily concerned with frequency division multiplex in the Bell System, there are many other applications. For example, frequencies are allocated among broadcast stations using a frequency division multiplex plan in which many stations can broadcast simultaneously over the same medium and yet be separated at a receiver for individual reception. Likewise, the frequency allocation of microwave stations is a form of FDM. These applications are not discussed here, although the microwave frequency allocation problem is discussed in a later chapter.

Before proceeding with the details of multiplexing, it is of interest to investigate the properties of a modulator and demodulator (*modem*), used in the Bell System for frequency translation in FDM. No attempt has been made to be inclusive, and it should be emphasized that other methods are available for performing the modulation

function. Although it is shown in Chap. 10 that with appropriate filters any nonlinear network can perform the basic modulation function, the product type of modem discussed here is undoubtedly the most widely used in telephone applications.

The Ring Modulator

As implied in the previous chapter, the modulating function can be efficiently accomplished by a circuit that takes the product of the carrier and the baseband signal. One of the most useful product-type circuits is the ring (lattice or double-balanced) modulator, which is shown in Fig. 6-1. If the diodes in the lattice are ideal, it is apparent that this modulator multiplies the baseband signal by $+1$ when the carrier supply is positive and by -1 when the carrier supply is negative. Thus, it is an ideal product modulator for a square wave carrier and the baseband signal. However, a square wave carrier, $c(t)$, can be expanded by a Fourier series into a summation of sinusoids at all odd harmonics of the basic square wave frequency

$$c(t) = a_1 \sin \omega_c t + a_3 \sin 3\omega_c t + a_5 \sin 5\omega_c t + \ldots \quad (6\text{-}1)$$

Multiplying by a baseband signal $v(t)$ gives:

$$M(t) = v(t)c(t)$$
$$= a_1 v(t) \sin \omega_c t + a_3 v(t) \sin 3\omega_c t + \ldots \quad (6\text{-}2)$$

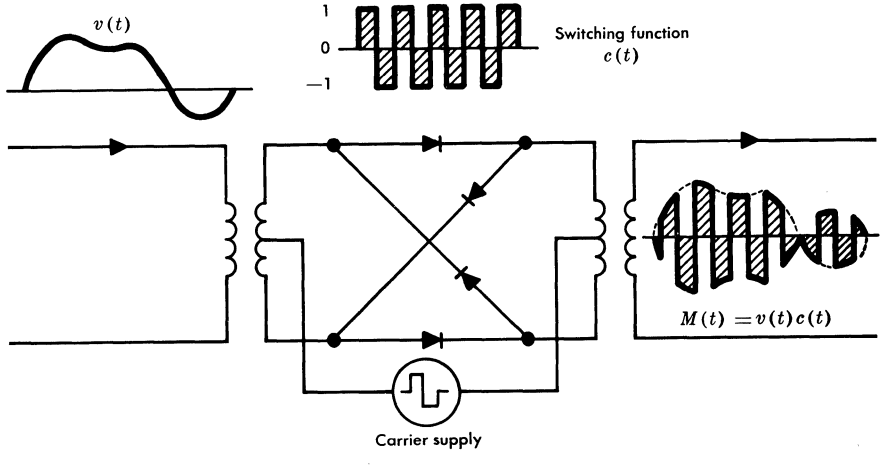


FIG. 6-1. Ring modulator.

Therefore, the baseband signal is multiplied by each of the odd harmonics to obtain the ideal frequency spectrum shown in Fig. 6-2, where the amplitudes of the sidebands around the harmonics fall off with frequency exactly as do the high-frequency components of the square wave series expansion. Either or both sidebands about the carrier frequency or any of its odd harmonics may be selected by filters for the desired signal. To prevent overlap, the carrier frequency must be at least equal to the highest frequency in the baseband signal.

In the ring modulator the switching function is assumed ideal. This means that the carrier, as it crosses the zero axis, switches the diodes instantaneously between zero and infinite impedance. This idealization produces two benefits.

The first benefit is that changes in carrier amplitude do not affect the switching function, and the modulator efficiency is therefore stabilized against carrier amplitude variations. In practice, a "stiffness" of 10 to 1 (0.1-dB increase in modulator loss for a 1-dB decrease in carrier power) can be easily achieved if desired. Actual message channel modulators are usually designed, however, to have a specified overload or limiting characteristic to protect the carrier system from loud-talker peaks and accidental tones. The desired overload characteristic is achieved by reducing the carrier so that the voice signal will control the diodes at a specified amplitude [1]. In typical cases this reduces stiffness to about 2 to 1.

The second benefit of an ideal switching function is that no products are produced other than those depicted in Fig. 6-2. These may be characterized as $n_o f_c \pm f_m$, where n_o is any odd integer, and f_m is any

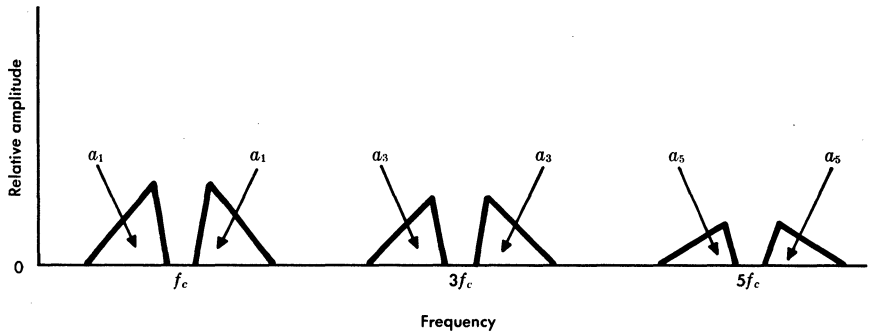


FIG. 6-2. Ring-type modulator frequency spectrum.

frequency component of the baseband signal. In practice, other products, represented by allowing an integer multiplier other than unity for f_m, are formed in the diodes during the finite switchover intervals. If the diodes of the lattice are identical, only odd harmonics of the baseband signal are produced. If the diodes are not identical, even harmonics of both the baseband signal and carrier will be produced. Diode balance giving 20 dB of suppression of these even harmonics is easy to obtain, and 40- or 50-dB suppression is possible by diode selection and is not unusual for telephone multiplex applications [2].

The ring modulator works equally well for the inverse process of demodulation. Consider application of an SSB signal occupying the band from $f_c + f_B$ to $f_c + f_T$ hertz applied to the ring modulator along with a carrier of f_c hertz. The resulting output spectrum contains frequencies that are sums and differences of the baseband signal ($f_c + f_B$ to $f_c + f_T$) and all odd harmonics of the carrier frequency, as shown in Fig. 6-3. Note that the demodulated signal is readily obtained by a low-pass filter that rejects all of the higher frequencies, which in this case consist of sidebands on each side of the *even* harmonics of the carrier frequency. Note also that the sideband components on either side of the even carrier harmonics are not of equal amplitude. This is denoted by the square wave series expansion coefficients (a_n) of Eq. (6-1), which are used to label the sidebands of Fig. 6-3.

Similarly, if a DSBSC signal with sidebands from f_B to f_T on each side of the carrier frequency, f_c, is applied to a ring modulator with an inserted carrier at f_c, the spectrum of Fig. 6-4 results. In

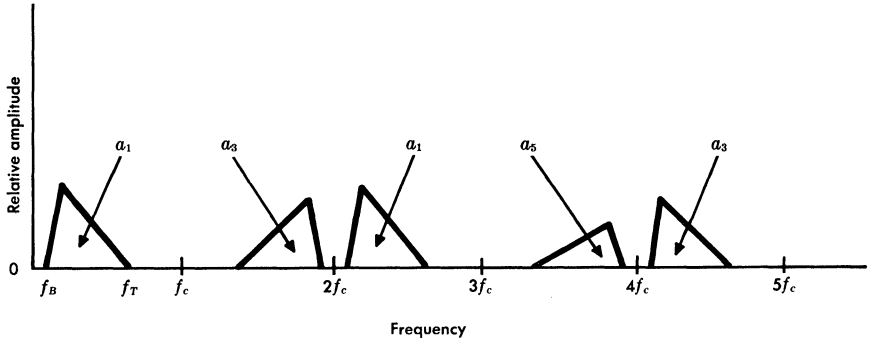


FIG. 6-3. Ring-type demodulator output spectrum—USB and f_c applied.

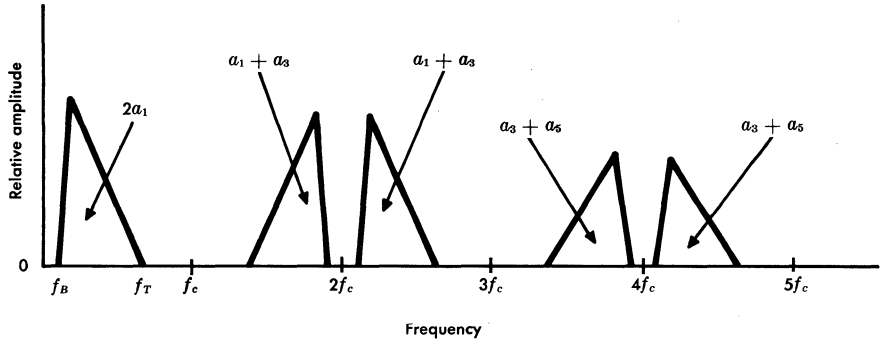


FIG. 6-4. Ring-type demodulator output spectrum—DSBSC and f_c applied.

this case the baseband component is twice the amplitude of the SSB case. This is a result of having two sidebands which (if the inserted carrier phase is identical to that of the original carrier) are coherent and thus will add in phase on a voltage basis. On the other hand, if f_c is 90 degrees out of phase with the original carrier, one baseband spectrum will be 180 degrees out of phase with the other, and the resulting baseband spectrum will be zero. This is the washout effect discussed in Chap. 5; it results in the demodulated baseband voltage being proportional to cos θ, where θ is the phase error.

Long-haul systems have standardized on SSB for FDM of baseband signals because it maximizes the channel capacity of the expensive long line. One advantage of carrier suppression is that intermodulation products of speech sidebands produced by repeater nonlinearity are unintelligible and noise-like in character, especially if large numbers of products occur. In contrast to this, transmitted carriers can produce intermodulation of carriers with an individual sideband to cause intelligible crosstalk in another channel. Since the long-haul systems are standardized as to signal format, a hierarchy of multiplexing terminals has evolved.

Bell System FDM Hierarchy

The practical implementation of frequency division multiplex may involve many steps of modulation and demodulation. Both

amplitude and angle modulation may be used within one complete system. Some of the steps commonly used and the reasons for them are discussed.

Figure 6-5 depicts the FDM transmission hierarchy for some of the common broadband transmission systems used in the Bell System. Only the transmitting terminals are shown, although the complete set of inverse operations must be performed at the receiving terminals.

The Message Channel. The basic building block of the hierarchical plan is the message channel. It is important from an administrative and maintenance standpoint, that a message channel may be considered an entity, characterized solely by measurements at terminal voice-frequency jacks made without detailed knowledge of the intervening transmission facilities. For example, an entire microwave transmission system may be replaced by a broadband coaxial cable system without seriously affecting the characteristics of the individual message channels.

The basic message channel, although originally intended for voice transmission, can be used for the transmission of data. Several narrowband data signals having components below 200 Hz can be multiplexed into a single message channel. For example, the 43-type multiplex terminal uses both FDM and FM to convert several narrowband data signals into a composite signal occupying a message channel [3]. This composite signal consists of several sinusoids in the voice-frequency range, each frequency modulated by one of the narrowband data signals. Since the composite signal has no components outside the 200- to 3400-Hz spectrum, it can be transmitted over any message channel. However, FM permits transmission of the narrowband signal components below 200 Hz.

Data signals requiring more bandwidth than that provided by the message channel are called wideband data. Some of the access points in the FDM hierarchy where wideband data can be conveniently inserted are shown in Fig. 6-5.

The Basic Group. The first multiplexing step for the message channels combines them into a set of 12 channels, called a group. The 12-channel modulating equipment is known as an A-type channel bank [4]. The 12-channel group output of the A-type bank is now a standard building block for most long-haul broadband systems. The channel bank has had a history of design improvements at about 5-year intervals, the latest version under development being the A6 channel bank.

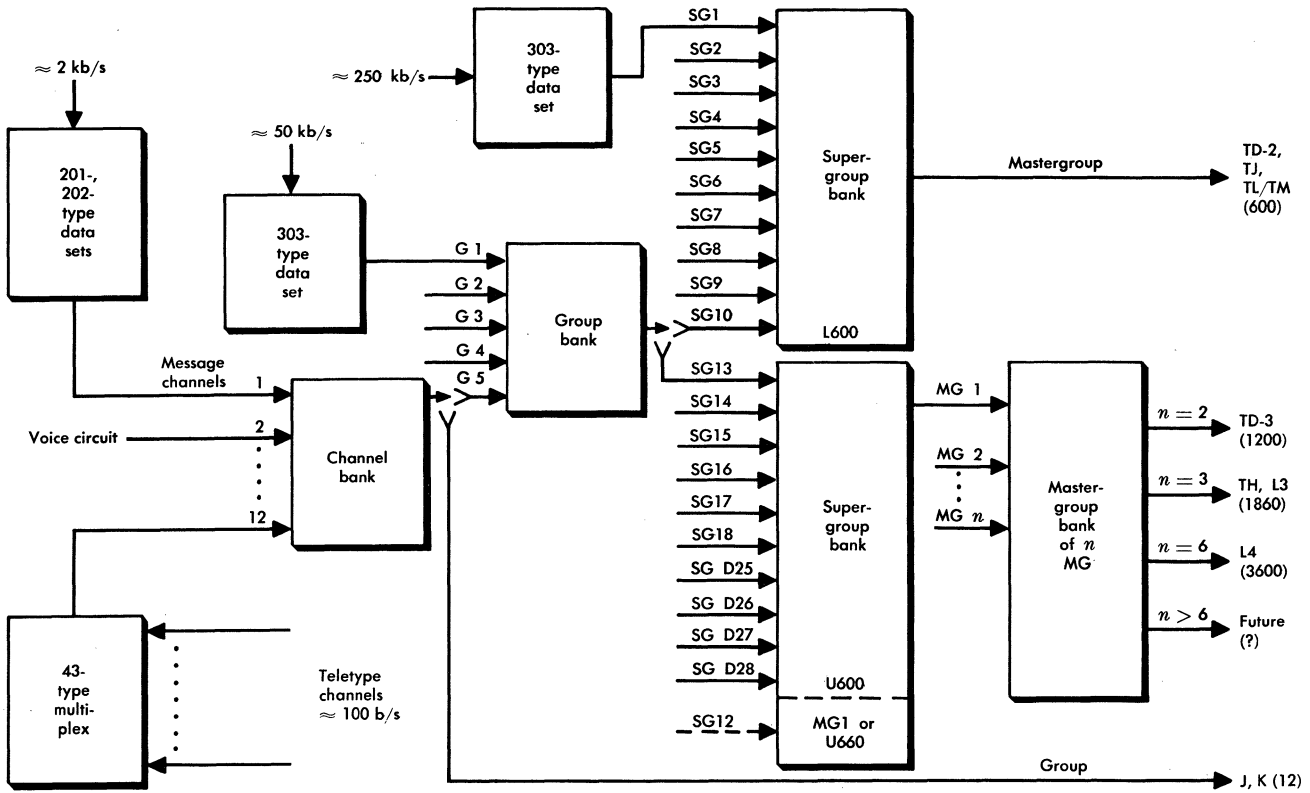


FIG. 6-5. Long-haul FDM hierarchy.

The composition of the 12-channel group is shown in Fig. 6-6. The slope of each message channel indicates sideband orientation with respect to audio frequency; i.e., highest amplitude represents the highest audio frequency. Note that the group is formed by taking channel n ($n = 1$ to 12) and having it modulate a carrier at a frequency of $112 - 4n$ kHz. The lower sidebands are then selected by filtering and combined. The result is a group of 12 inverted sidebands in the frequency range of 60 to 108 kHz. This frequency range was optimum for the crystal filters available at the time the original equipment was designed.

The group then defines another basic building block in the FDM scheme. Any other signal whose spectrum occupies the 48 kHz between 60 and 108 kHz could be treated as a group in further multiplexing steps. For example, several 303-type data sets have been developed that are capable of placing about 50 kilobits per second on a group facility. These have been used for encrypted voice transmission as well as other high-speed data applications. Some 303-type sets have been designed for half-group application and have a capability of about 20 kilobits per second.

Although the A-type channel bank group is most common, other group configurations are sometimes encountered. For example, the N3 short-haul carrier system forms groups of 12 message channels by modulating carriers of frequency $144 + 4n$ kHz and selecting the upper sideband [5]. The resulting group of 12 noninverted sidebands (with some transmitted carriers) occupies the 148 to 196 kHz portion of the spectrum. A single additional modulation step makes the N3 group compatible with the long-haul FDM hierarchy [6].

The Basic Supergroup. The next step in the FDM hierarchy shown in Fig. 6-5 is the combination of five groups into a 60-channel supergroup [7]. This is accomplished in a group bank as shown in Fig. 6-7 where the nth group ($n = 1$ to 5) modulates a $372 + 48n$ kHz carrier. A filter selects the lower (inverted) sideband and the five are combined to form a 240-kHz supergroup from 312 to 552 kHz. There are practical reasons for forming the supergroup in this manner rather than continuing the modulation of unique carriers by individual message channels. One reason is that economical filters of the required characteristics are only available over a limited frequency range. Another important consideration is the number of different filter designs and different carrier supplies needed. By using a group bank, five sets of the A-type channel banks can be used in parallel. In addition to making high production runs possible, the maintenance

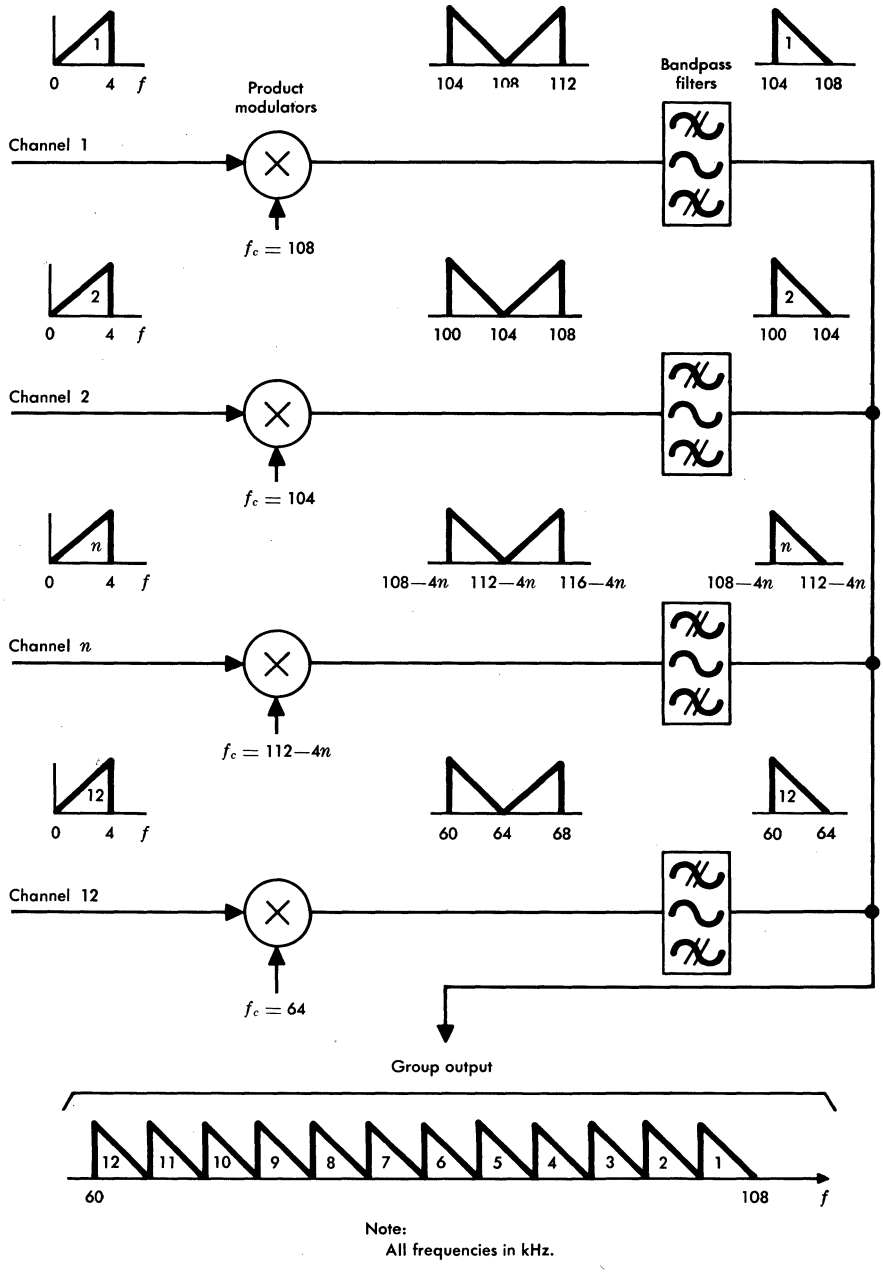


FIG. 6-6. Basic group with A-type channel bank.

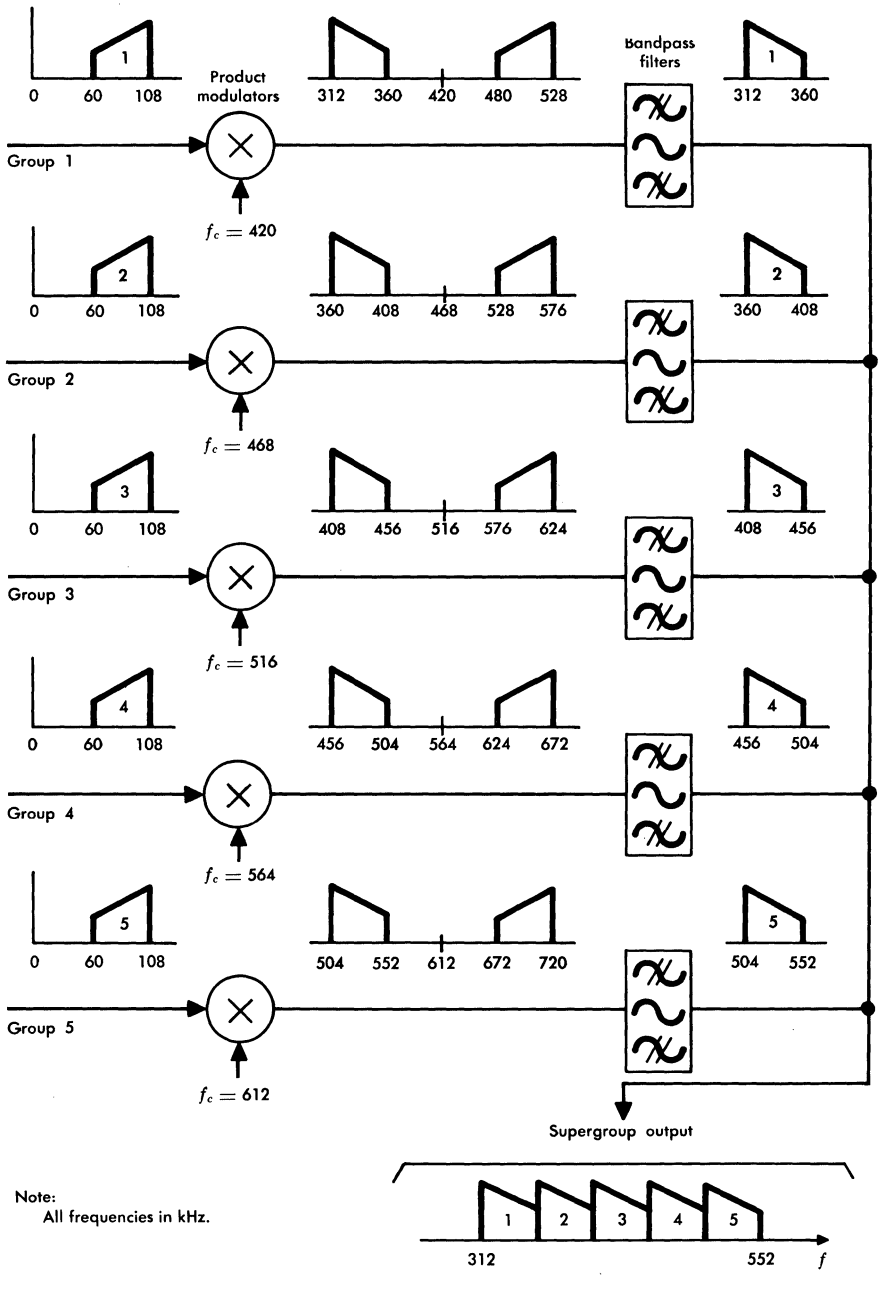


FIG. 6-7. Basic supergroup from group bank.

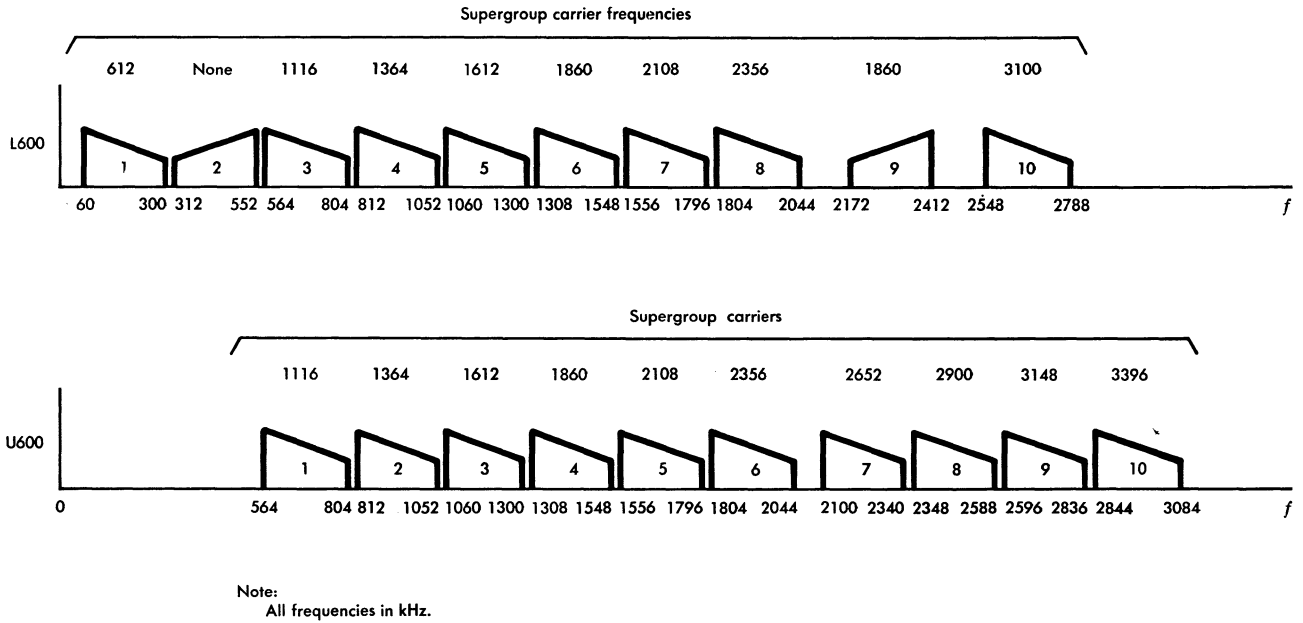
and spare parts problems in the field are simplified. Because of the two frequency inversions in supergroup generation, the channels in the basic supergroup have the same orientation as at audio frequencies. High-speed data of up to 250 kilobits per second can enter the FDM system through 303-type data sets in the basic supergroup spectrum as shown in Fig. 6-5.

The Basic Mastergroup. In a manner similar to that previously discussed, the combination of ten supergroups forms a 600-channel mastergroup [8]. The formation of a mastergroup by a supergroup bank is shown in Fig. 6-8. Note that two slightly different methods are shown. The L600 mastergroup occupies the 60 to 2788 kHz band of frequencies and is the broadband signal used on the L1 coaxial system and on the TD-2, TJ, TL, and TM radio systems [9-13]. The U600 mastergroup occupies slightly higher frequencies (564 to 3084 kHz) and is used as a building block for even larger groupings. Both L600 and U600 mastergroups have gaps between supergroups as a result of design considerations and limitations which were taken into account when the frequency allocations were adopted. Note that all sidebands used for the U600, and all but the second and ninth supergroups of L600, are lower sidebands and thus invert the basic supergroup frequencies.

Another signal that enters the FDM hierarchy above the supergroup level is commercial television. Although the television baseband spectrum has almost twice the bandwidth of a mastergroup, the radio systems are designed to accommodate it. The rough equivalence is a result of the differences in the frequency spectra. The spectrum of the mastergroup is basically flat to about 3 MHz, but the television spectrum rolls off with frequency.

Very Large Groupings. Modern broadband transmission systems are capable of even larger groupings than mastergroups. For L3 carrier and TH microwave, three mastergroups and one supergroup comprising 1860 message channels are combined as shown in Fig. 6-9 [14, 15]. The L4 system utilizes six U600 mastergroups multiplexed to form 3600 channels [16]. Since these larger groups are rather specialized and restricted to specific systems, there is no universal name for them.

Other Groupings. Although the message channel groupings discussed previously are almost always used in the Bell System for long-haul circuits, they are not universal standards. For example,



Note:
All frequencies in kHz.

FIG. 6-8. L600 and U600 mastergroups.

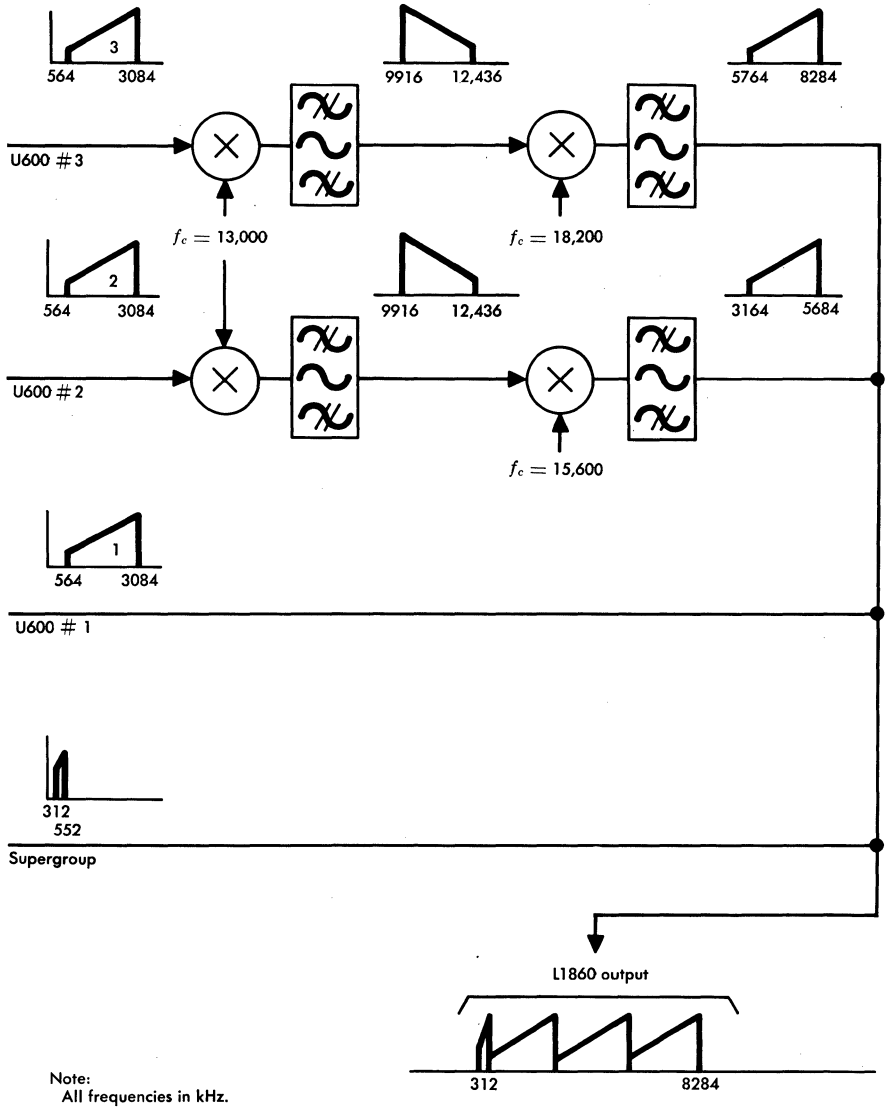


FIG. 6-9. Mastergroup multiplex.

submarine cable circuits utilize special multiplexing. Here, a more sophisticated terminal design is economically justifiable in order to obtain additional channels in the available spectrum. Channel banks have been developed which stack 16 channels in the standard group

frequency band of 60 to 108 kHz [17]. This is accomplished by spacing the channels at 3-kHz increments instead of 4 kHz. Although some of the "squeezing" is done in the bandwidth of the channel, two-stage modulation permits the effective bandwidth of the channels to be 200 to 3050 Hz so that most of the reduction is in the spacing between the channels in the group. This, of course, results in the need for more complex and expensive filters and becomes uneconomic for general application.

Terminal Carrier Supply. Because the channel carriers are not transmitted with the sidebands, the relative frequency accuracy requirement at opposite ends of a system is especially stringent. Frequency accuracy requirements on the order of one part in 10^8 are not unusual. This is stringent enough to preclude the use of separate free-running channel oscillators. Instead, a single, very accurate base frequency of 4 kHz is generated, and all carriers are derived as harmonics of this frequency. This requires carrier supply bandpass filters sharp enough to reject adjacent unwanted harmonics.

Stringent requirements make the carrier supply expensive, and its cost must be spread over many channels. Thus, long-haul terminal carrier supplies have evolved as highly centralized installations, common to as many as several thousand channels. This in turn has made necessary much redundancy and automatic protection switching, adding to the cost and complexity.

An oscillator of reasonable cost cannot be depended upon for the longtime stability required of the base frequency. Instead, each frequency supply is synchronized, or slaved, to a reference signal derived from an incoming line pilot or from another local frequency supply. The synchronizing pilots used include 64 kHz on the L1 coaxial and the TD-2, TJ, TL, and TM radio systems; 308 kHz on the L3 coaxial and the TH radio systems; and 512 kHz on the L4 coaxial system. The result of this synchronization is a configuration of frequency supplies connected by synchronizing paths originating with the Bell System reference frequency standard in New York.

The error is cumulative on these chains of supplies, and therefore a tight locking requirement is desirable. The latest frequency supplies are phase-locked, developing phase errors of only a few degrees at 4 kHz; thus the average frequency error approaches zero as the measuring interval increases [18]. The free-running stability of the supply is about one part in 10^8 per day which is adequate to weather a reasonable period of pilot outage [19].

Short-Haul Considerations

Multiplexed carrier systems economical for distances up to about 250 miles are called short-haul systems. The economics of such systems are much more sensitive to terminal costs than are those of long-haul systems. To keep short-haul costs down, the terminal design is usually specialized and integrated into the specific system design. In fact, short-haul systems have been developed that provide economical message circuits for distances as low as 15 miles, while performing satisfactorily for distances as great as 250 miles.

The N-type systems are designed for use on multipair cable and use a separate cable pair for each direction of transmission. Further electrical separation of the transmitted signals is obtained by using different frequency bands for each direction of transmission, i.e., 36 to 140 kHz (low group) for one direction on one pair, and 164 to 268 kHz (high group) for the other direction on the other pair. The interchange of the high and low groups at each repeater (called *frequency frogging*) is accomplished by modulation with a 304-kHz carrier. The resulting inversion provides first order equalization of line slope (increasing attenuation at high frequencies) while frequency frogging blocks a major circulating crosstalk path around each repeater.

The N1 and N2 systems employ DSBTC with low modulation index [20, 21]. Although this may appear uneconomical because of channel bandwidth and power inefficiency, overriding economies in the terminal design result from the DSB approach. Most important, requirements on channel bandpass filters are drastically relaxed because there is no unwanted sideband to be rejected at the transmitting end, and the carrier does not have to be suppressed by filters at either end. In addition, the transmitted channel carriers allow simple diode envelope detectors to be used at the receiving terminal and obviate synchronizing arrangements. Free-running crystal oscillators are adequate for generating carriers.

The possibility of additional channels becomes more attractive for the longer N systems, and the N3 terminal has been developed to place two groups of 12 SSB channels in the N-type frequency band. Some of the carriers are also transmitted at a low level to facilitate the demodulation process [22].

Another factor contributes to the desirability of using SSB on N3 in the short-haul carrier plant. Multiplexed circuits (including long-haul) are often dropped in small groups and extended on wire pairs. In many cases these may be short-haul carrier systems, which require

demultiplexing the long-haul group to voice frequency and remultiplexing on short-haul carrier terminals. If the long-haul and short-haul terminals had identical formats for a group of channels (12, for example), this group could be extended directly without going down to voice frequency. This is highly desirable, not only from the the standpoint of economy and maintenance, but also for the improvement of channel transmission characteristics. The N3 format, with 12 channel groups, permits compatibility with the basic long-haul group after slight processing.

The type-O short-haul open-wire system uses twin sideband multiplexing [23]. This is basically SSB, but two different channels share the same carrier, which is transmitted at reduced level. Equivalent four-wire performance is obtained on a single pair by using different frequencies for each direction of transmission. This gives maximum efficiency of the open-wire pairs which tend to be strung as required, one at a time, and so are not found in abundance. Also, the line capacity of a pole route is limited by the need for physical spacing of pairs to reduce crosstalk, especially when carrier frequencies are involved. The pair-at-a-time philosophy is consistent with the slow traffic growth pattern on open-wire routes. This factor is also recognized in the O-terminal design, whereby terminals can be installed economically in increments as low as four channels.

6.3 TIME DIVISION MULTIPLEX

Time division multiplex (TDM) is the third common type of multiplexing mentioned in the beginning of the chapter. As the name implies, it is simply the sharing of a common facility in time—an extension of the childhood principle of "taking turns." The basic idea, of course, is much older than that of FDM, but until recently most transmission systems used FDM almost exclusively. The fundamental reason for this condition is that state of the art limitations of switching devices delayed development of economical systems.

For signals that are not full time, TDM has often been used in telephone communications. For example, most telephones are in use only a small portion of the time; thus, several telephones can time share a common line to the nearest central office.

Speeding up the time scale, the same principle can be used to TDM several speech channels by taking advantage of pauses between words and statements. Utilizing this principle, a Time Assignment Speech

Interpolation (TASI) system is used on many overseas channels to effectively increase the capacity of the channels. Again, to avoid problems with more simultaneous talkers than available channels, the numbers of talkers and channels should be rather large (100 or so). Obviously, the switching in such a scheme must be very rapid, and the resulting complex equipment is not attractive except for use on expensive channels such as overseas applications.

So far, the discussion of time division multiplexing has considered signals not present at all times. The question arises as to possible ways of using TDM with signals having significant energy at all times. Since such signals do not exhibit times at which they can be ignored in favor of a different signal, some processing is necessary to break up the signal in time before multiplexing.

The fact that a bandlimited signal can be sampled at discrete times while preserving all of the information leads to the most popular method of processing for TDM. Before taking up this application of the sampling theorem, additional insight can be gained by considering another type of processing called time compression.

Time Compression

Consider the desirability of multiplexing N information channels, each with a top frequency of f_T hertz. By entering all N signals into parallel sections of a storage medium, and by reading sections of the storage medium serially through the channels at N times the input rate, a time-compressed output signal is obtained. To aid in visualizing the process, consider using tape recorders as the storage medium. By connecting each of N recorders to one of the N channels, a parallel recording function is accomplished. The serial readout is accomplished by moving a playback head N times as fast over a section of recorded tape, switching to the next machine and doing the same, etc. After reading part of all N machines, the playback head is switched back to the first tape and repeats the process. The frequencies of the baseband signal are all multiplied by N, resulting in a compressed time scale. A reverse procedure is necessary to demultiplex the time-compressed signals.

Since time compression results in multiplexing N channels of bandwidth f_T into a bandwidth of Nf_T, it is reasonable to question its desirability over standard SSB-FDM, which is just as conserving of bandwidth but much easier to implement. The answer lies in the signal distortion caused by nonlinear performance. In FDM systems, this nonlinear distortion results in spurious products being produced

which appear as noise in all of the demultiplexed channels. In broadband systems this intermodulation noise can become objectionably large. By using time compression, each signal appears on the broadband facility by itself in the times when it appears at all. Thus, although nonlinearities cause distortion of each individual baseband signal, there can be no intermodulation distortion between the different signals. This reduction of intermodulation noise can be an important consideration with some types of baseband signals (such as television) and warrants implementation of time compression or some other TDM scheme. No time-compression system is now in use, but its application has been seriously considered.

PCM Multiplexing

The time division multiplexing of four time-varying voltages into a PCM pulse stream is shown in Fig. 6-10. Each voltage (band-

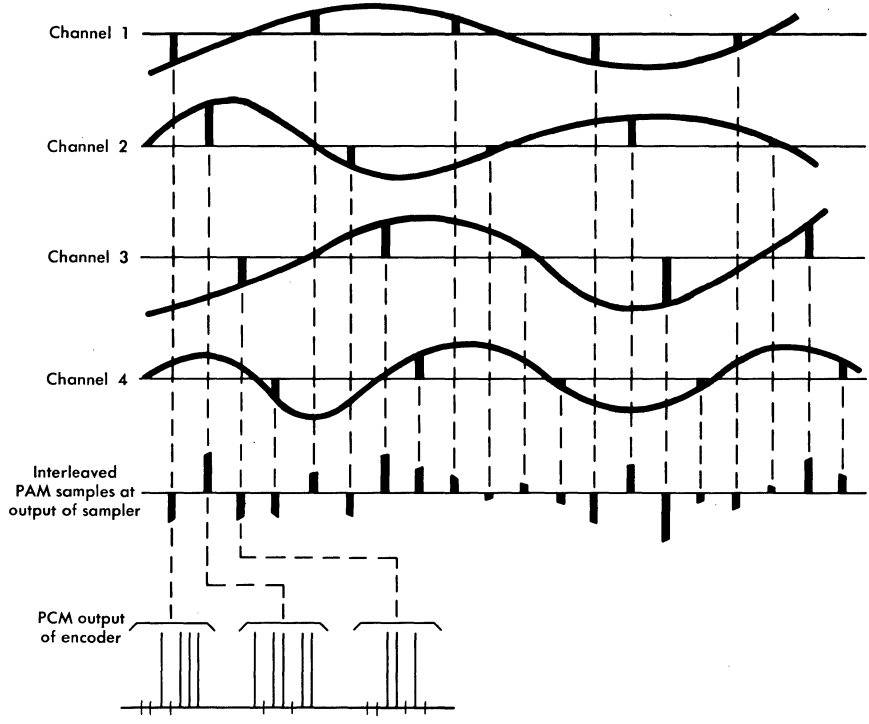


FIG. 6-10. Formation of PAM and PCM signals.

limited by a low-pass filter) is sampled; the resulting PAM signals are interleaved; and finally, each PAM sample is encoded into binary PCM. Although not shown, the inverse functions of decoding and demultiplexing must be performed at the receiving end. The *sampling interval, T,* is the time between successive samples of the message voltage in a channel. The time interval for each code word (representing a sample from a single message channel) must be equal to or less than $1/N$ times the sampling interval, if N message channels are to be accommodated.

Clearly, as the number of message channels is increased, the time interval that can be allotted to each must be reduced since all of them must be fitted into the sampling interval. The allowed duration of a coded pulse train representing an individual sample must be shortened, and the individual pulses moved closer together as the number of time division channels is increased. The more closely spaced pulses require greater bandwidth so that bandwidth limitations restrict the number of message channels.

A corollary to the sampling theorem shows that the maximum number of independent pulses per second that can be successfully transmitted through a bandwidth f_T hertz is $2f_T$ bauds. Thus, if f_T hertz is the highest frequency in the message and n is the number of pulses per code group, approximately nf_T hertz of bandwidth per message is required. This is n (typically 7 to 9) times the bandwidth required for direct transmission or for SSB. The increased bandwidth disadvantage is offset by the noise advantage resulting from the regeneration of a new, essentially noise-free symbol at each repeater.

When a number of wire communication routes converge on a single terminal, the ruggedness of the channels is a particularly important consideration. If the susceptibility of the channels to mutual interference is high, many separate FDM bands of frequencies may be required, and the total bandwidth required for the service will be large. Although PCM requires an initial increase of bandwidth for each channel, the resulting ruggedness usually permits many routes originating from, or converging toward, a single terminal to occupy the same frequency band. As a result, the occupancy of the cable facility by PCM is exceptionally good, and its other transmission advantages are then obtained with little, if any, increase in total bandwidth. In all, PCM is well suited for multiplex message circuits, where standard quality and high reliability are required.

Bell System PCM Hierarchy

Just as a frequency division multiplex hierarchy of group, supergroup, mastergroup, etc., has evolved, a similar plan is developing for time division multiplexing PCM systems as shown in Fig. 6-11 [24].

The D1 channel bank encodes 24 voice channels into 7-bit binary PCM for transmission over 1.544-Mb/s T1 repeatered lines. The D1 channel bank and T1 repeatered line are widely used in the Bell System as a short-haul carrier system strongly competitive with N carrier. The transmission medium is twisted pair cable of the same type typically used for voice-frequency (and N carrier) transmission. The T1 digital line is also capable of handling approximately 1.5 megabit-per-second data so it could also be fed by a wideband data terminal as shown in Fig. 6-11. This terminal could multiplex several slower data streams.

The D1 channel bank is being superseded by the D2 channel bank, which encodes 96 voice channels into 8-bit binary PCM, producing four lines at the T1 rate, and thus replacing four D1 banks.

The M12 digital multiplex combines four T1 pulse streams into a single 6.3-Mb/s pulse stream to be transmitted over the T2 repeatered line. Note that this is slightly higher than four times the T1 rate and allows for multiplexing unsynchronized lines through a technique called pulse stuffing described in Chap. 26. The T2 line is also designed to use conventional twisted pair cable but, because of the increased capacity, is more sensitive to repeater spacing for the smaller wire gauges. The 6.3-Mb/s T2 pulse stream appears capable of handling a 1-MHz PICTUREPHONE® signal if an appropriate coding method is used.

The M23 digital multiplex combines seven pulse streams at the 6.3-Mb/s T2 rate into a single 46.3-Mb/s pulse stream. At the present time this 46.3-Mb/s pulse stream is not transmitted over any system but can be used internally in the central office. The 46.3-Mb/s pulse stream can be generated from the binary words resulting from coding a U600 mastergroup. Although shown as a simple coder in Fig. 6-11, the mastergroup is first shifted in frequency before encoding into a 9-bit code. The 9-bit code is necessary in this case to meet noise objectives while allowing a number of digital encodings and decodings in tandem. An 8-bit code would have introduced too much noise.

Commercial color television signals are also sampled and encoded into 9-bit words resulting in a 92.6-Mb/s pulse stream which is equivalent to two 46.3-Mb/s streams.

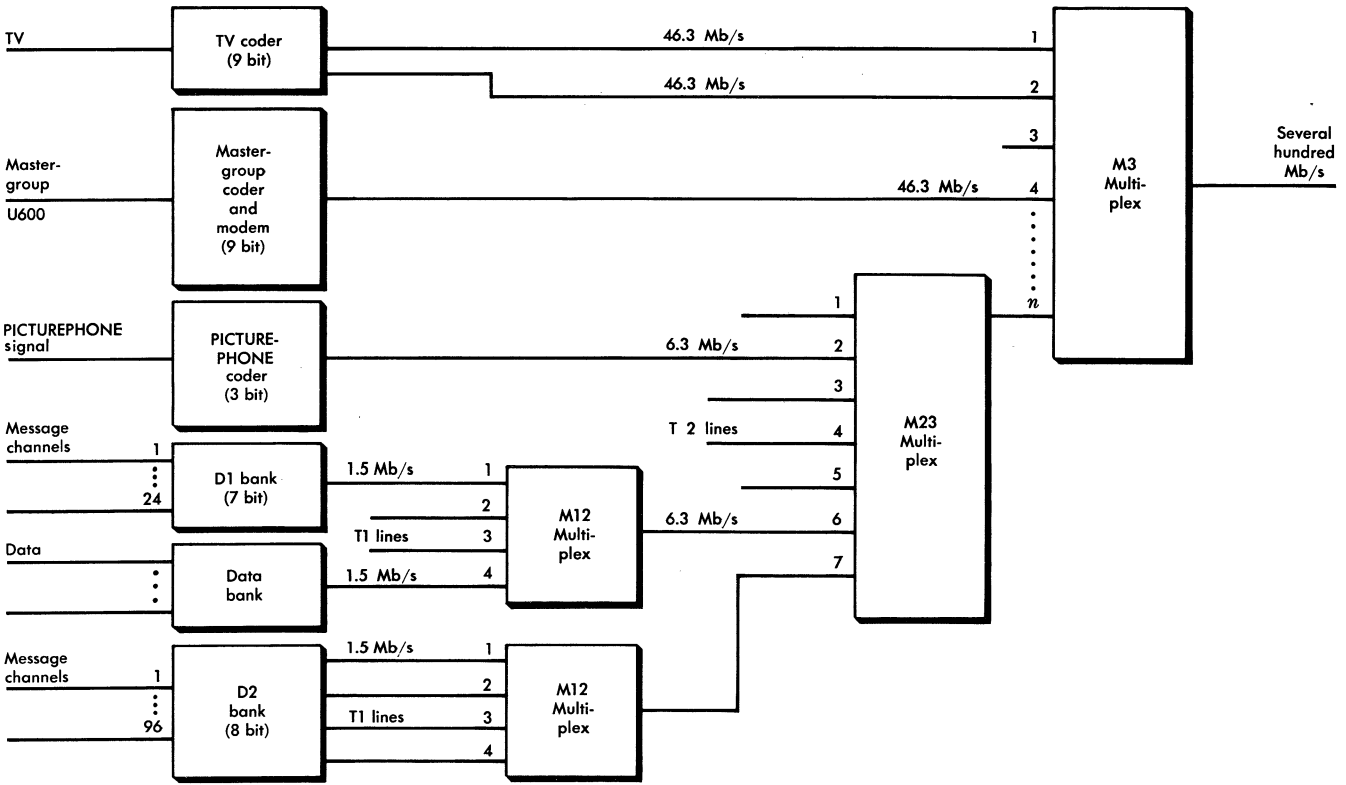


FIG. 6-11. PCM hierarchy.

Several of the 46.3-Mb/s pulse streams could be interleaved to form still larger systems. Experimental systems capable of several hundred megabauds over a coaxial cable facility have been built [25]. Still further multiplexing of these high-speed pulse streams has been proposed for waveguide or laser systems of the future.

REFERENCES

1. Albert, W. G., J. B. Evans, Jr., J. J. Ginty, and J. B. Harley. "Carrier Supplies for L-Type Multiplex," *Bell System Tech. J.*, vol. 42 (Mar. 1963), pp. 279-317.
2. Caruthers, R. S. "Copper Oxide Modulators in Carrier Telephone Systems," *Bell System Tech. J.*, vol. 18 (Apr. 1939), pp. 315-337.
3. Hysko, J. L., W. T. Rea, and L. C. Roberts. "A Carrier Telegraph System for Short-Haul Applications," *Bell System Tech. J.*, vol. 31 (July 1952), pp. 666-687.
4. Blecher, F. H. and F. J. Hallenbeck. "The Transistorized A5 Channel Bank for Broadband Systems," *Bell System Tech. J.*, vol. 41 (Jan. 1962), pp. 321-359.
5. Bleisch, G. W. "The N3 Carrier System Plan," *Bell Laboratories Record*, vol. 43 (Mar. 1965), pp. 77-83.
6. Haner, R. L., T. H. Simmonds, Jr., and L. H. Steiff. "A New Interface Between N3 and L Systems," *Bell Laboratories Record*, vol. 46 (Nov. 1968), p. 339.
7. Hallenbeck, F. J. and J. J. Mahoney, Jr. "The New L Multiplex—System Description and Design Objectives," *Bell System Tech. J.*, vol. 42 (Mar. 1963), pp. 207-221.
8. Graham, R. S., W. E. Adams, R. E. Powers, and F. R. Bies. "New Group and Supergroup Terminals for L-Multiplex," *Bell System Tech. J.*, vol. 42 (Mar. 1963), pp. 223-278.
9. Abraham, L. G. "Progress in Coaxial Telephone and Television Systems," *Trans. of the AIEE*, vol. 67 (1948), pp. 1520-1527.
10. Roetken, A. A., K. D. Smith, and R. W. Friis. "The TD-2 Microwave Radio Relay System," *Bell System Tech. J.*, vol. 30 (Oct. 1951), pp. 1041-1077.
11. Gammie, J. and S. D. Hathaway. "The TJ Radio Relay System," *Bell System Tech. J.*, vol. 39 (July 1960), pp. 821-877.
12. Hathaway, S. D., D. D. Sagaser, and J. A. Wood. "The TL Radio Relay System," *Bell System Tech. J.*, vol. 42 (Sept. 1963), pp. 2297-2353.
13. Friis, R. W., J. J. Jansen, R. M. Jensen, and H. T. King. "The TM-1/TL-2 Short-Haul Microwave Systems," *Bell System Tech. J.*, vol. 45 (Jan. 1966), pp. 1-95.
14. Ehrbar, R. D., C. H. Elmendorf, R. H. Klie, and A. J. Grossman. "The L3 Coaxial System," *Bell System Tech. J.*, vol. 32 (July 1953), pp. 781-1005.
15. Kinzer, J. P. and J. F. Laidig. "Engineering Aspects of the TH Microwave Relay System," *Bell System Tech. J.*, vol. 40 (Nov. 1961), pp. 1459-1494.
16. Klie, R. H. and R. E. Mosher. "The L-4 Coaxial Cable System," *Bell Laboratories Record*, vol. 45 (July-Aug. 1967), pp. 211-217.

17. Tucker, R. S. "Sixteen-Channel Banks for Submarine Cables," *Bell Laboratories Record,* vol. 38 (July 1960), pp. 248-252.
18. Clark, O. P., E. J. Drazy, and D. C. Weller. "A Phase-Locked Primary Frequency Supply for L-Multiplex," *Bell System Tech. J.,* vol. 42 (Mar. 1963), pp. 319-340.
19. Shields, J. G., D. C. Weller, and W. R. Wysocki. "A Phase-Locked Primary Frequency Supply," *Bell Laboratories Record,* vol. 44 (July-Aug. 1966), pp. 236-238.
20. Caruthers, R. S. "The Type N-1 Carrier Telephone System: Objectives and Transmission Features," *Bell System Tech. J.,* vol. 30 (Jan. 1951), pp. 1-32.
21. Boyd, R. C. and F. J. Herr. "The N2 Carrier Terminal—Objectives and Analysis," *Bell System Tech. J.,* vol. 44 (May-June 1965), pp. 731-759.
22. Bleisch, G. W. and C. W. Irby. "N3 Carrier System: Objectives and Transmission Features," *Bell System Tech. J.,* vol. 45 (July-Aug. 1966), pp. 767-799.
23. Fracassi, R. D. "Type-O Carrier: System Description," *Bell Laboratories Record,* vol. 32 (June 1954), pp. 215-220.
24. Hoth, D. F. "Digital Communication," *Bell Laboratories Record,* vol. 45 (Feb. 1967), pp. 39-43.
25. Mayo, J. S. "Experimental 224 Mb/s PCM Terminals," *Bell System Tech. J.,* vol. 44 (Nov. 1965), pp. 1813-1840.

Chapter 7

Noise and Its Measurement

The word *noise* is usually used to label those disagreeable or distracting sounds which one would rather not hear. As such, it is an acoustic term difficult to describe either qualitatively or quantitatively although descriptive words such as hiss, click, rumble, crash, pitch, loudness, etc. are often used. The total noise that reaches a listener's ears affects the degree of annoyance and the intelligibility of received speech. This total noise consists of room noise and circuit noise. Room noise reaches the subscriber's ear directly by leakage around the receiver cap, indirectly by way of the sidetone path through the transmitter and receiver of his telephone set, and over the normal transmission path from the far end. Control of such room noise can only be achieved through design of telephone sets and enclosures and will not be considered here.

There are many types of circuit noise. Actually, any interference to a communications channel can be considered noise. For ease in characterization of this circuit noise, all description will be of the electrical waves corresponding to noise at the system output. The disturbance associated with these generalized electrical noises includes a variety of end effects. In the case of television, for example, the ultimate effect produced by noise is against the eye rather than the ear, and terms such as snow or ghosts describe this effect subjectively. In the case of telegraph or data, the effect is not an aesthetic one but rather a threat to the accuracy of the received information.

There are many potential sources of noise or interference in a transmission system. Some of the most important ones warrant a chapter by themselves. For example, intermodulation noise resulting from nonlinearities in the transmission system is covered in Chap. 10. The interference produced by one transmission channel being coupled

to another (crosstalk) is covered in Chap. 11. The effects of thermal noise on complex networks are covered in Chap. 8. This chapter will attempt to describe the commonly encountered noise sources other than these.

7.1 COMMON TYPES OF NOISE

It is desirable to characterize electrical noise as accurately as possible. Unfortunately, the most common characteristic of noise is its nondeterministic nature; i.e., the exact waveform of the noise cannot be predicted. If prediction were possible, it would be an easy matter to generate a replica of the incoming noise and subtract it from the actual noise to achieve effective noise-free performance. Lack of accurate waveform information does not prevent placing a meter across a noise source to read the rms voltage (or current) produced by the noise and thus have a measure of the amount of noise. Similarly, the amount of electrical noise can be determined by the average, peak, or rectified average voltage (or current) measured on an appropriate meter. The relationships between these quantities are different for different types of noise. Furthermore, changing the frequency spectrum of the noise through filtering has an effect on the noise, depending on the type of noise.

A common method of characterizing many types of noise waveforms is through the use of probability; specifically, the probability distribution or probability density functions. The probability distribution, $P(V)$, gives the probability that the voltage is less than V. Plotted as a function of V, the probability distribution function goes from 0 for $V = -\infty$ to 1 for $V = \infty$. The slope of the distribution curve represents the probability density function, $p(V)$. Since the total area under the density function is equal to unity, the area under the $p(V)$ curve in the interval from $V = V_1$ to $V = V_2$ represents the fraction of time or probability of V being in that interval. The probability distribution function can be obtained from a noise wave by use of a level distribution recorder consisting of a bank of parallel threshold indicators feeding integrating output meters. Since the gathering of such data is rather laborious, it is fortunate that the probability function can usually be derived from a knowledge of the source of the electrical noise.

A probability function does not uniquely define random noise. The parameter missing is the time (or frequency) scale. This is commonly supplied by the frequency spectrum of the noise wave.

Since the wave is not time limited, the Fourier spectrum, which is an energy spectrum for all time, does not exist. Instead, the power spectrum is used, since this is measurable. Analytically, the mean square voltage spectrum, or spectral density, is commonly used and is identical with the power spectrum if the impedance is one ohm.

The time scale of a noise wave can also be characterized by the autocorrelation function. This is a function giving the degree of dependence between amplitudes of the wave at any two instants of time. It can be shown that the autocorrelation function and spectral density are Fourier transform pairs [1].

Single-Frequency Interference

It may seem strange to treat a sine wave of known amplitude, frequency, and phase as if it were a noise wave. From a strictly theoretical point of view, disturbance from such a wave can be eliminated by locally generating a sine wave 180 degrees out of phase to cancel the interference. In actual cases, however, an interfering sine wave may originate from an external uncontrolled source and its amplitude, frequency, and/or phase are subject to unpredictable variations. It is, therefore, sometimes convenient to treat single-frequency interference as if it were noise with a sinusoidal distribution of magnitude versus time.

If a sine-wave voltage has peak amplitude, A, the amplitude density function is given by [2]

$$\begin{aligned} p(V) &= \frac{1}{\pi\sqrt{A^2 - V^2}} && -A \leq V \leq A \\ &= 0 && |V| > A \end{aligned} \tag{7-1}$$

The corresponding distribution function is

$$\begin{aligned} P(V) &= \frac{1}{2} + \frac{1}{\pi}\arcsin\frac{V}{A} && -A \leq V \leq A \\ &= 0 && V < -A \\ &= 1 && V > A \end{aligned} \tag{7-2}$$

These are plotted in Fig. 7-1.

The peak value of this sine wave is given by A or 20 log A dBV. However, many voltmeters do not read peak values but instead rectify

the voltage and read the average. This average absolute value can be determined by doubling the density function of Eq. (7-1) for positive values of V, which corresponds to rectification, and by taking the expected value of V [2],

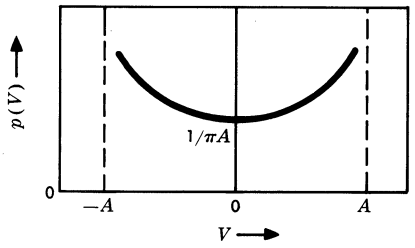
$$E[\,|\,V\,|\,] = \int_0^A \frac{2V}{\pi\sqrt{A^2 - V^2}}\,dV \quad = \frac{2A}{\pi} \tag{7-3}$$

Thus, the average absolute value is 20 log $\pi/2$ = 3.92 dB below the peak value.

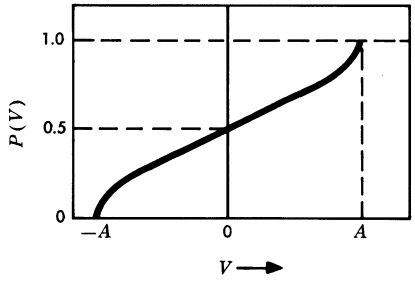
Similarly, the rms voltage is given by the square root of the expected value of V^2 or

$$V_{\text{rms}} = \sqrt{\int_{-A}^{A} \frac{V^2}{\pi\sqrt{A^2 - V^2}}\,dV} \quad = \frac{A}{\sqrt{2}} \tag{7-4}$$

The rms voltage is then 20 log $\sqrt{2}$ = 3.01 dB below the peak value. The ratio of rms voltage to average absolute voltage is called the form factor. For a sine wave this is 0.91 dB. This factor has significance because many meters are constructed to respond to the rectified average voltage and yet are calibrated in terms of rms sine-wave voltages (i.e., they read 0.91 dB high). Errors can result when noise voltages of other form factors are measured on such a meter and assumed to be the correct rms value.



(a) Density function



(b) Distribution function

FIG. 7-1. Probability density and distribution for a sine wave.

Pilots. Single-frequency sinusoids are purposely placed on a carrier facility as pilots for a number of reasons. Their use for synchronizing reinserted carriers has already been mentioned in the previous chapter. In addition, pilots are often employed for line regulation and maintenance. As a consequence, they can produce undesired interference which can be classed as single-frequency noise.

For example, nonlinearities in the system may produce harmonics of the pilots. If one of these harmonics falls into an FDM voice channel, the demodulated output for that channel will contain a single-frequency audio tone. Such effects can be minimized by careful selection of pilot frequencies. For example, by making all pilots multiples of 4 kHz and forcing all carriers to be at 4-kHz multiples, no intermodulation products of pilots can fall in a channel. The levels of pilots should also be carefully considered. Generally, the minimum level necessary for the pilot to perform its given function should be used. Higher level pilots not only increase the magnitude of the interference but can also use a significant amount of the allowed power on the system.

Supervision. The tones used for supervision (and occasionally signaling) are also sources of single-frequency interference. The most common type of SF supervision places a 2600-Hz sinusoid on all idle channels. This tone can crosstalk into adjacent busy voice-frequency circuits and appear as an undesired 2600-Hz noise tone. A more subtle means of interference can arise when FDM channels encounter a nonlinearity. In this case, the side frequency corresponding to the tone (2600 Hz from the carrier frequency) can interact with a pilot or other supervisory side frequency to produce spurious products that may disturb other channels in a multichannel system.

Thermal Noise

Thermal noise is a phenomenon associated with Brownian motion of electrons in a conductor. In accordance with the kinetic theory of heat, the electrons in a conductor are in continual random motion in thermal equilibrium with the molecules. The mean square velocity of the electrons is proportional to the absolute temperature. Since each electron carries a unit negative charge, each flight of an electron between collisions with molecules constitutes a short pulse of current. Because of the number of such randomly moving electrons and the frequency of collisions, some electrical manifestation of the behavior will be expected across the terminals of the conductor. Based on this model, it would be intuitively expected that the average voltage (d-c) is zero (otherwise, charges would pile up at one end of the conductor and stay there). But such random motion of charges would be expected to give rise to an a-c component. This a-c component of noise was first observed in 1927 by J. B. Johnson of Bell Telephone Laboratories [3]. A quantitative theoretical treatment was furnished by

H. Nyquist in 1928 [4]. The effect has been called Johnson noise, thermal noise, thermal agitation, and resistance noise.

The equipartition law of Boltzmann and Maxwell (and the works of Johnson and Nyquist) states that for a thermal noise source the available power in a 1-Hz bandwidth is given by

$$p_n(f) = kT \qquad \text{watts/Hz} \tag{7-5}$$

where k = Boltzmann's constant = $1.3805(10^{-23})$ joule/°K, and T is the absolute temperature of the thermal noise source in degrees Kelvin. At room temperature, 17°C or 290°K, the available power turns out to be $p_n(f) = 4.0(10^{-21})$ watts/Hz or −174.0 dBm/Hz.

The result given by the equipartition theory is one of a constant power density spectrum versus frequency. Because of this property, a thermal noise source is referred to as a white noise source—an analogy to white light which contains all visible wavelengths of light. Actually the analogy has not been correctly drawn since in optics the uniform distribution of white light is based on wavelength rather than frequency. The term white noise has become well established to mean uniform distribution with frequency and will be used here.

In all reported measurements, the available power of a thermal noise source has been found to be proportional to the bandwidth over any range from direct current to the highest microwave frequencies commonly used. If the bandwidths were unlimited, the results of the equipartition theory say that the available power of a thermal noise source would also be unlimited. This difficulty can be traced to shortcomings of the equipartition theory and can be resolved by applying a few principles of quantum mechanics to the problem. The result of this application is that kT must be replaced by $hf/[\exp(hf/kT) - 1]$ of quantum mechanics where h = Planck's constant = $6.625(10^{-34})$ joule-second. Applying this result to the expression for available power of a thermal noise source gives

$$p_n(f) = \frac{hf}{\exp(hf/kT) - 1} \qquad \text{watts/Hz} \tag{7-6}$$

Thus, at arbitrarily high frequencies the thermal noise spectrum eventually drops to zero. This does not mean that noiseless devices could be built at these frequencies. A quantum noise term equal to hf has to be added to Eq. (7-6) in this case. From Fig. 7-2 this transitional region is at about 40 GHz for $T = 2.9°K$, at 400 GHz for $T = 29°$K, and at 4000 GHz at room temperature.

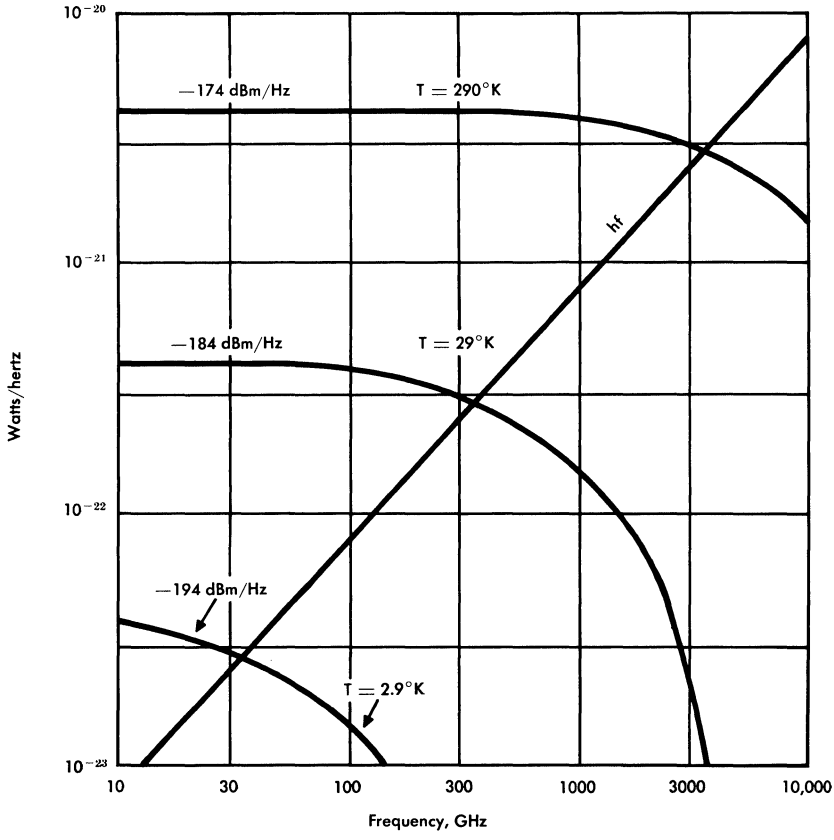


FIG. 7-2. Available thermal noise power at high frequencies.

For most practical purposes the available noise power of a thermal noise source is directly proportional to the product of the bandwidth of the system or detector and the absolute temperature of the source. Thus,

$$p_a = kTB_w \text{ watts} \tag{7-7}$$

where B_w is the noise bandwidth of the system or detector in hertz, and p_a is the available noise power in watts. Expressing the available noise power in dBm gives

$$P_a = -174 + 10 \log B_w \quad \text{dBm} \tag{7-8}$$

This represents a minimum amount of noise power which must ultimately limit the fidelity of amplification when the input signal is weak.

Gaussian Distribution. The gaussian distribution is the limiting form for the distribution function of the sum of a large number of independent quantities which individually may have a variety of different distributions [5]. This result is known in statistics as the *central limit theorem.* Thermal noise, which may be regarded as the superposition of an exceedingly large number of random, practically independent electronic contributions, satisfies the theoretical conditions for a gaussian distribution. The gaussian probability density function for zero mean is shown in Fig. 7-3, and its equation is

$$p(V) = \frac{1}{\sigma_n \sqrt{2\pi}} \exp\left(-V^2/2\sigma_n^2\right) \tag{7-9}$$

The distribution function is also shown in Fig. 7-3 and is given by the integral of Eq. (7-9)

$$P(V) = \frac{1}{\sigma_n \sqrt{2\pi}} \int_{-\infty}^{V} \exp\left(\frac{-x^2}{2\sigma_n^2}\right) dx \tag{7-10}$$

Values for this integral have been tabulated for various values of V/σ_n [6].

It can be easily shown that the mean square voltage (the expected value of V^2) is equal to the variance, σ_n^2. Thus, the rms voltage of a gaussian distributed noise source is given by σ_n, the standard deviation. The full-wave rectified average voltage can be obtained by taking the expected value of V using the one-sided "folded" density function

$$\begin{aligned} E(|V|) &= \frac{2}{\sigma_n \sqrt{2\pi}} \int_0^{\infty} V \exp\left(-\frac{V^2}{2\sigma_n^2}\right) dV \\ &= \sigma_n \sqrt{\frac{2}{\pi}} \end{aligned} \tag{7-11}$$

The form factor or ratio of rms to average absolute voltage is given by $\sqrt{\pi/2} = 1.253$ which is equivalent to 1.96 dB. From the previous analysis of a sine wave, the form factor for a sine wave is 0.91 dB.

This difference is of practical significance if a rectifying-type meter calibrated to read rms values for a sine-wave input is used to measure noise. The rms value indicated for thermal noise will be 1.05 dB too low.

Gaussian noise has a probability greater than zero of exceeding any finite magnitude no matter how large. Thus, the peak factor given by the ratio of peak to rms voltage does not exist for a thermal noise signal. For this particular case it is convenient to modify the definition of peak factor to be the ratio of the value exceeded by the noise a certain percentage of the time to the rms noise value. This percentage of time is commonly chosen to be 0.01 per cent. A table of the normal distribution shows that signal magnitudes greater than $3.89\sigma_n$ (i.e., $|V| > 3.89\sigma_n$) occur less than 0.01 per cent of the time. Since σ_n is the rms value of the noise signal, the peak factor for a thermal noise signal is 3.89, or 11.80 dB. Inclusion of 0.001 per cent peaks increases the peak factor by only 1.1 dB to 12.9 dB.

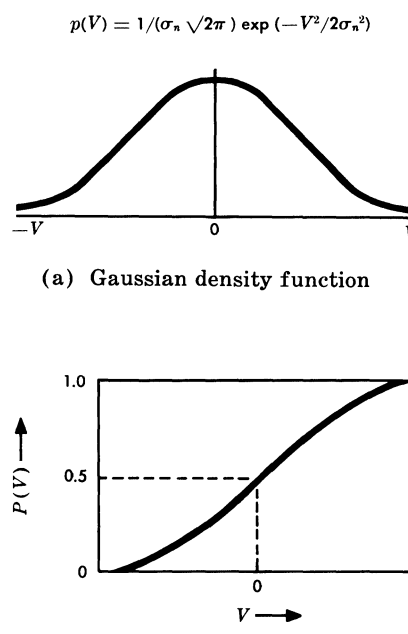


FIG. 7-3. Gaussian probability density and distribution functions.

This peak factor must be considered when making thermal noise loading tests on amplifiers and repeatered telephone systems. Consider a physical amplifier. This device can handle only a limited amplitude range before the signal is clipped or otherwise distorted. If the amplifier is to be used to amplify a sine-wave signal or a thermal noise signal without distorting the waveforms, the power handling capacity for thermal noise is 8.8 dB less than that for a sine wave.

The fact that thermal noise is white as well as gaussian has led many engineers into carelessly treating white and gaussian noise as synonymous. Such is not always the case. For example, passing gaussian noise through a linear network such as a filter will leave it gaussian but may drastically change the frequency spectrum. On

the other hand, a single impulse will not have a gaussian amplitude distribution but will have a flat or white frequency spectrum.

Simulation of White Noise by Randomly Phased Sine Waves. For purposes of analysis of communications systems, it is desirable to use alternate representations for the gaussian noise signal. One very useful representation is obtained by approximating the gaussian noise signal by a sum of a large number of sine waves of different frequencies having uniformly random phases. Thus,

$$e_n(t) = \sum_{k=1}^{k_N} A_k \cos\ (2\pi f_k t + \theta_k) \qquad (7\text{-}12)$$

where

k_N = the number of sine waves used to approximate the noise signal

A_k = the amplitude of the kth sine wave

f_k = the frequency of the kth sine wave

θ_k = the phase of the kth sine wave

In order that the combination of sine waves be a good representation of the noise signal, its statistics should approximate those of the noise in both the time and frequency domains.

The power spectrum of the noise signal for practical purposes is constant with frequency. In order to have the sine-wave representation simulate the power spectrum of the noise signal over a band of frequencies, f_1 to f_2, it is necessary to make the power spectrum of the sine-wave representation constant over this range of frequency. One convenient way of accomplishing this is by making the amplitudes of the various sine-wave components equal and by spacing the components equally between f_1 and f_2. The number of sine waves needed is determined by the required accuracy of the simulation. The probability distribution of the sum of randomly phased sine waves has been studied by several investigators and is not discussed here [7]. Figure 7-4 illustrates how the peaks are related to the rms value of the sine-wave sum in terms of the probability that peaks exceed a given level. The distribution applies to sine waves of the same or different frequencies so long as the phasing between them is uniformly random. Where different frequencies are involved, they should not be commensurably related.

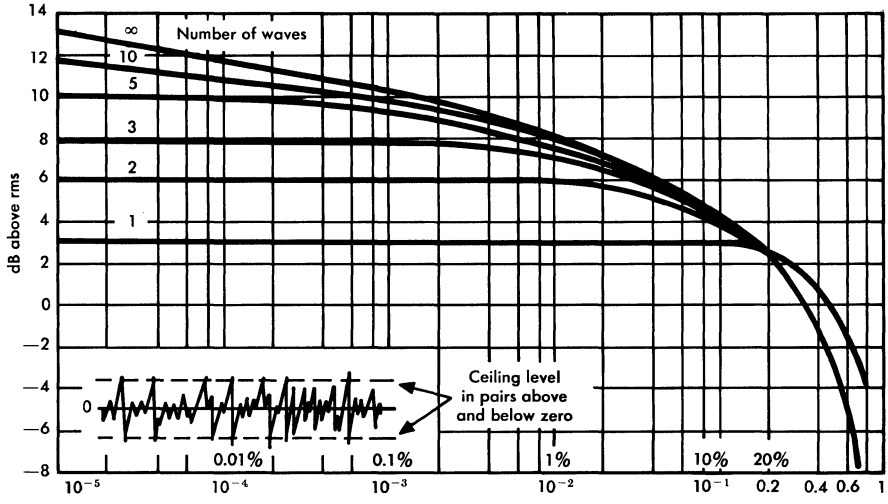


FIG. 7-4. Distribution of instantaneous amplitudes of randomly phased sine waves.

As to how many such sine waves would be required for a good approximation, an examination of Fig. 7-4 indicates that the distribution for ten sine waves closely approximates that of the gaussian distribution (infinite number of waves). From this it appears that ten or more sine waves are sufficient for engineering purposes.

Equivalent Circuits of Thermal Noise Sources. It was pointed out in Eq. (7-7) that for all practical purposes the available power of a thermal noise source at a temperature, $T°$K, in a bandwidth, B_w, is $p_a = kTB_w$. A good example of a thermal noise source is a resistor. A suitable noise equivalent circuit for a resistor is a noise voltage generator, e_n, connected in series with a hypothetically noiseless resistor having the same resistance, R. If this generator and resistor are connected to a load resistor having a resistance, R_1, as shown in Fig. 7-5,

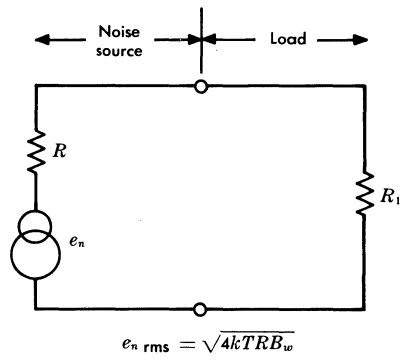


FIG. 7-5. Equivalent circuit of a noisy resistor.

the generator delivers power to the load resistor, R_1. It can be easily shown that the generator delivers its maximum power to the load resistor when $R_1 = R$. This maximum power is $p_a = (e_{n_{\text{rms}}})^2/4R$. The maximum power is said to be the available power of the Thévenin source shown in Fig. 7-5. The available power of a thermal noise source is $p_a = kTB_w$. Equating these two powers and solving for the rms value of voltage of the equivalent Thévenin generator,

$$e_{n_{\text{rms}}} = \sqrt{4kTB_wR} \tag{7-13}$$

A Norton equivalent circuit for a noisy resistor may be determined in a similar manner. In this case, Fig. 7-6, the rms current of the noise generator would be

$$i_{n_{\text{rms}}} = \sqrt{\frac{4kTB_w}{R}} \tag{7-14}$$

where R is the resistance of the original noisy resistor. Since a resistor is a thermal noise source, the temperature, T, is the actual physical temperature of the resistor in degrees Kelvin.

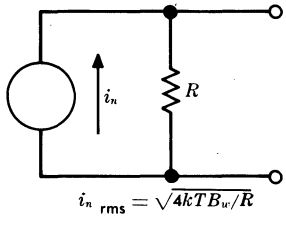


FIG. 7-6. Norton equivalent circuit of a noisy resistor.

If two noisy resistors, R_1 and R_2, are connected in series as shown in Fig. 7-7(a), the Thévenin equivalent of the resulting noise source can be determined. By combining the equivalent circuits of the two resistors shown in Fig. 7-7(b), the new equivalent circuit of Fig. 7-7(c) is obtained. The resistance, R, of the new equivalent circuit is $R_1 + R_2$, and the open circuit voltage, e_n, is

$$e_n = e_{n_1} + e_{n_2} \tag{7-15}$$

Since the voltage produced by these two generators is uncorrelated, the open circuit rms voltage produced by the two resistors is given

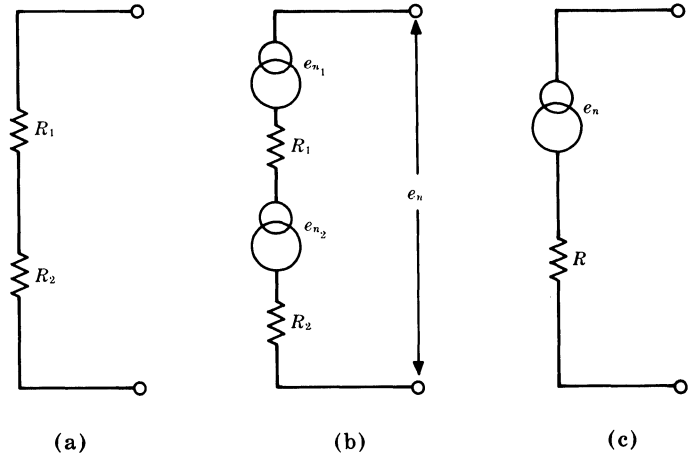


FIG. 7-7. Noisy resistors in series.

by a root sum square (rss) of the individual rms voltages. Since

$$e_{n_{1\text{rms}}} = \sqrt{4kTB_wR_1}$$

$$e_{n_{2\text{rms}}} = \sqrt{4kTB_wR_2}$$

it follows that

$$e_{n_{\text{rms}}} = \sqrt{4kTB_w\ (R_1 + R_2)} \tag{7-16}$$

It has been assumed that both resistors have the same physical temperature and that the rms open circuit voltage of the noise source is directly proportional to the square root of the internal resistance of the source.

In the case of a noise source consisting of two resistors, R_1 and R_2, connected in parallel, the equivalent source resistance is $R_1R_2/(R_1 + R_2)$. It can be shown that the equivalent rms open circuit voltage is

$$e_{n_{\text{rms}}} = \sqrt{4kTB_w \frac{R_1R_2}{R_1 + R_2}} \tag{7-17}$$

Again the open circuit rms voltage is directly proportional to the square root of the equivalent internal resistance of the noise source.

Care must be taken in deriving Eq. (7-17) if the Thévenin equivalent circuit is used. The expression for the instantaneous open circuit voltage should be calculated first, and from this, the rms circuit voltage should be computed. In general, it can be shown that for a two-terminal resistive network having all resistors at the same temperature, the rms open circuit noise voltage is directly proportional to the value of resistance seen looking into the resistive network, regardless of the interconnections existing among the resistors in the network.

Now the effect of reactive elements in connection with noisy resistors can be considered. If the circuit consists of a capacitor and resistor connected in parallel, the capacitor cannot dissipate any noise power from the resistor. If the capacitor generates noise, this power must be dissipated in the resistor. This would mean that the resistor would get hotter and the capacitor colder. This behavior is contrary to the entropy law of thermodynamics. Hence, the initial assumption that the capacitor generates noise is false. The same argument holds for inductive elements as well. Reactive elements do not contribute to noise in an RLC network, but since the impedance of a reactive element is frequency dependent, the noise appearing at the output of an RLC network can have a frequency shape. In general, for a two-terminal device consisting of passive linear elements having a driving point impedance, $Z(f) = R(f) + jX(f)$, the rms open circuit noise voltage in a small frequency band, df, is

$$e_{n_{\text{rms}}} = \sqrt{4kTR(f)\,df} \tag{7-18}$$

Noise Temperature. Since the available noise power of a thermal noise source is directly proportional to the absolute temperature of the source, it is said that the noise source has a noise temperature expressed in degrees Kelvin. In the case of a thermal noise source, the noise temperature is equal to the physical temperature of the source. The concept of noise temperature is extremely useful when characterizing the available power of other types of noise sources (such as noise diodes and microwave gas noise tubes). It can be said then that the noise temperature of such a device is the temperature of a thermal noise source which produces the same amount of available noise power as the device under consideration. That is, if a given noise source produces an available power of p_a watts in a small frequency interval of df hertz, the noise temperature of the noise

source is given by $T = p_a/kdf$. It should be emphasized that the noise temperature of a noise source does not have to equal the physical temperature of the source.

It should also be pointed out that the noise temperature of a noise source may be a function of frequency. In the definition of noise temperature of a noise source, the noise power is measured in a small frequency interval. Hence, the noise temperature of the noise source will be a function of frequency if the noise power spectrum of the source is not flat with frequency.

Note that the concept of noise temperature does not have to be restricted to noise sources alone. The noise power measured at the output of an amplifier can be expressed in terms of an equivalent noise temperature. The noise appearing at the output terminals of an antenna may also be expressed in terms of noise temperature. In this case the term *antenna noise temperature* is used. The more noise picked up by the antenna, the higher the antenna noise temperature. Such noise is due to radiation from objects on earth as well as objects in outer space such as the sun, moon, radio stars, and hot ionized interstellar gases.

Another noise temperature concept used in connection with one-port devices is that of excess noise temperature. Excess noise temperature of a noise source may be defined as the difference between the noise temperature of the source and the noise temperature of a thermal noise source at standard or room temperature (290°K). Thus,

$$T_x = T - T_0 \tag{7-19}$$

where

T_x = excessive noise temperature

T_0 = standard or room temperature

= 290°K = 17°C = 62.6°F

T = noise temperature of the noise source

With this definition it is possible to have noise sources having negative as well as positive excess noise temperatures, for example a resistor at a temperature less than T_0.

Many commercially available noise sources having resistive terminations are calibrated in terms of excess noise temperature. Excess noise temperature may be interpreted as the noise temperature of the

source in excess of that of the resistive termination in the source. If the termination is at standard temperature, then the noise source may be used as calibrated. If the termination resistor is not at standard temperature, then a suitable correction factor must be used, taking into account the difference between the actual termination temperature and the standard temperature. Care should therefore be taken when using the term excess noise temperature, especially in cases where such devices are used to make noise measurements on amplifiers.

Shot Noise

Shot noise is due to the discrete nature of electron flow and is found in most active devices. It was first observed in the anode current in vacuum-tube amplifiers and was described by W. Schottky in 1918. According to Schottky, the mean square noise current in a 1-Hz bandwidth is

$$i_{\text{rms}}^2 = 2qI \tag{7-20}$$

where q = charge of the electron = $1.6(10^{-19})$ coulombs, and I = direct current through the device in amperes.

Since shot noise is made up of a very large number of independent contributors, the central limit theorem implies that the amplitude distribution of shot noise would be gaussian, the same as thermal noise with the variance given by Eq. (7-20). Observations confirm this assumption. Similarly, over the frequency range of practical interest, it is often accurate to assume that each impulsive component of this noise contains frequency components uniformly distributed across all frequencies of interest, with the result being that shot noise (as well as thermal noise) is considered white noise. There are however two primary differences between shot noise and thermal noise.

1. The magnitude of thermal noise is proportional to absolute temperature, whereas shot noise is not directly affected by temperature.
2. The magnitude of shot noise is proportional to the square root of current. Thus, it is related to signal amplitude, whereas thermal noise is not.

Linear filtering or shaping of shot noise does not affect its gaussian properties but certainly does not leave it white. Again, the terms white and gaussian are not synonymous.

Low-Frequency ($1/f$) Noise

A third type of gaussian distributed noise is low-frequency noise, also called contact noise, excess noise, flicker noise, or $1/f$ noise because of its peculiar increase towards very low frequencies. This noise is associated with contact and surface irregularities in cathodes and semiconductors. It appears to be caused by fluctuations in the conductivity of the medium. Great advances have been made in the reduction of this effect by cleaning and passivating semiconductor surfaces. A good device may have negligible $1/f$ noise above about 1 kHz although the corner frequency can be a few decades higher in frequency for high-frequency low-noise transistors. The effect limits the performance of crystal video detectors for microwave frequencies. The law of variation for the spectral density of this noise is expressed by [2]

$$p(f) = \frac{K}{f^{\nu}} \quad \text{watts} \tag{7-21}$$

where ν ranges from about 0.8 to 1.5. It is interesting to note that if ν were exactly unity, the power in a band of frequencies from f_1 to f_2 would be given by

$$p = \int_{f_1}^{f_2} \frac{K}{f}\, df = K(\ln f_2 - \ln f_1) \tag{7-22}$$

This expression would give an infiite amount of noise power if the band extended down to zero frequency or up to infinite frequency. Since the actual noise power is finite, the exact $1/f$ law can only hold over a limited frequency band not including zero or infinity. It is remarkable that experimental observations have followed the $1/f$ law very closely over many frequency decades extending downward to a fraction of a hertz.

Reevaluating Eq. (7-22) for ν less than unity shows that the power would then remain finite for f_1 equal to zero but not for f_2 infinite. Similarly, if ν is greater than unity, the expression remains finite for infinite f_2 but not for f_1 equal to zero. Thus, no value of ν can give a law which is valid at both ends of the frequency spectrum. It is difficult to find a physical model that fits the experimental observations over a frequency band which is many octaves wide but does not include zero or infinite frequency.

Rayleigh Noise

Noise is considered to be of the narrowband type if the noise bandwidth is small compared with the midband frequency. Gaussian noise then assumes the characteristic appearance of a sinusoidal carrier at the midband frequency modulated in amplitude by a low-frequency wave whose highest frequency component is dependent upon the bandwidth of the noise.

The low-frequency envelope can be generated physically by applying the narrowband noise wave at high level to the well known envelope detector circuit whose output voltage represents the envelope and is a smooth curve through the positive peaks of the noise. When the noise is gaussian, the envelope has what is called the Rayleigh distribution. The probability density function is

$$p(V) = \frac{V}{\sigma^2} \exp\left(-\frac{V^2}{2\sigma^2}\right) \qquad V > 0 \tag{7-23}$$

and the distribution function is

$$P(V) = 1 - \exp\left(-\frac{V^2}{2\sigma^2}\right) \qquad V > 0 \tag{7-24}$$

Curves for these functions are shown in Fig. 7-8. Note that only positive values occur. The average value is not zero but is given by

$$E[V] = \int_0^\infty \frac{V^2}{\sigma^2} \exp\left(-\frac{V^2}{2\sigma^2}\right) dV = \sqrt{\frac{\pi}{2}}\,\sigma \tag{7-25}$$

The mean-square value is

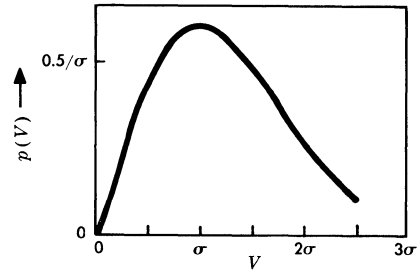
$$E[V^2] = \int_0^\infty \frac{V^3}{\sigma^2} \exp\left(-\frac{V^2}{2\sigma^2}\right) dV = 2\sigma^2 \tag{7-26}$$

That is, the mean square of the envelope is twice the mean square of the original noise wave. The mean-square a-c component is

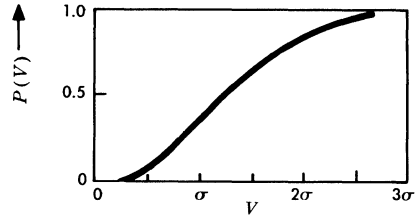
$$E[V^2] - \{E[V]\}^2 = \left(2 - \frac{\pi}{2}\right)\sigma^2 = 0.429\,\sigma^2 \tag{7-27}$$

The rms a-c component is the square root of Eq. (7-27), or 0.655 σ. The form factor of the complete Rayleigh noise wave is $2/\sqrt{\pi} = 1.128$, or 1.05 dB.

Defining peaks of Rayleigh noise as that value exceeded 0.01 per cent of the time, results in a peak factor of 9.64 dB, or over 2 dB less than that for gaussian noise. The Rayleigh distribution is important in the narrowband case in which an envelope detector is quite often a part of the receiving apparatus. Much confusion can result when the distinction between the Rayleigh and gaussian peak factors has not been considered.



(a) Rayleigh density function



(b) Rayleigh distribution function

FIG. 7-8. Rayleigh probability functions.

Impulse Noise

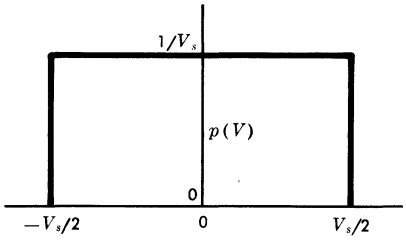
Impulse noise consists of short spikes of energy having an approximately flat frequency spectrum over the frequency range of interest. The noise arises from switching transients in central offices and from corona-type discharges that occur along a repeatered line. The human being appears to be reasonably tolerant of clicks and pops, i.e., impulses below levels which might cause hearing damage. However, PCM and data receivers are relatively intolerant of these impulses since they cannot distinguish between impulse noise and the pulses to be detected. Therefore, current study and control of impulse noise have emphasized effects on digital transmission.

If pulses occur independently at random times, the number arriving in any fixed interval follows a Poisson process. This process is characterized mathematically by [8]

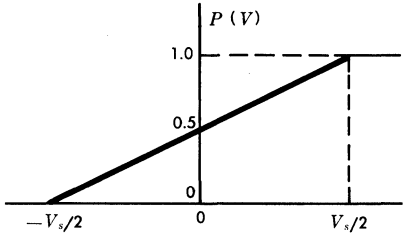
$$P(n) = \frac{(\nu T)^n e^{-\nu T}}{n!} \tag{7-28}$$

where $P(n)$ is the probability that exactly n pulses occur in a time interval of duration, T, and ν is the average number of pulses occurring in unit time. However, impulse noise on telephone channels does not follow a Poisson distribution. It has been found empirically that the number of arrivals per unit time can be approximated by a

log normal distribution. The important differences between noise impulses and steady noise is that the impulses are short relative to the time between them such that the receiving circuits resolve independent events. Narrowing the bandwidth would eventually cause the distinct pulses to merge into a steady noise wave. Before this merger takes place, however, the heights of the noise peaks tend to vary directly with bandwidth, whereas the rms noise follows the square root. This is because the isolated peaks represent addition of nearly equal in-phase components uniformly distributed in frequency. Band reduction cuts off a proportional number of equal contributors. The rms value is proportional to the square root of the average power which is directly proportional to bandwidth. It is thus possible to change the peak factor of impulse noise by filtering. Considerable reduction in the effects of impulse noise on a narrowband circuit can be achieved by preceding the bandlimiting part of the system by a wideband peak-clipping circuit. That is, it is better to clip while the peak factor is high than to wait until the pulses have been smeared over more time by bandlimiting.



(a) Rectangular density function



(b) Distribution function

FIG. 7-9. Probability functions for quantizing noise.

Quantizing Noise

The conversion of analog signals to digital form in PCM systems gives rise to round-off errors that result in what is called quantizing noise. The coded representation of the sample amplitude can be exactly right only when the sampled value corresponds exactly with one of the discrete code words. For all other values in a typical system, there is an error which can range from a negative half step to a positive half step. For a linear quantizer not subject to overload, the quantizing errors can be assumed equally likely, for if enough steps are used to make the quality acceptable, there is little tendency for the values to favor any region within a step. If the clipping caused by overload is ignored, the result is a noise which has a

uniform or rectangular density function throughout the range of minus half a step size ($-V_s/2$) to plus half a step size ($V_s/2$), as shown in Fig. 7-9 where V_s is the voltage difference between steps. The size of V_s can be made as small as desired by simply increasing the number of steps and hence the length of the code word per sample. However, this increases the required bit rate (thus required bandwidth) or decreases the capacity of a fixed bit-rate system so that it is advantageous to allow the quantizing noise to be as large as tolerable.

The probability density function, $p(V)$, for quantizing noise is constant and equal to $1/V_s$ throughout the range $-V_s/2$ to $V_s/2$, and zero outside this range as shown in Fig. 7-9. The distribution function, $P(V)$, is a ramp which represents the area of the density function up to the point, V, and thus starts from zero at $-V_s/2$ and increases to unity at $V_s/2$ as shown. The average value of V is zero since plus and minus values occur symmetrically. The average rectified value of V is obviously $V_s/4$. The mean-square value of V can be calculated as

$$E[V^2] = \int_{-V_s/2}^{V_s/2} \frac{V^2}{V_s}\, dV = \frac{V_s^2}{12} \tag{7-29}$$

The quantizing noise thus contributes an rms noise voltage equal to the step voltage divided by $\sqrt{12}$. The peak factor is $\sqrt{3}$, or 4.8 dB. The form factor is simply $2/\sqrt{3}$, or 1.25 dB.

It can be shown that the frequency spectrum of the quantizing noise is essentially flat over the range of interest [9]. Thus, if the sampling rate is twice the highest baseband frequency, the mean square noise at baseband is given by Eq. (7-29). If a narrower band of quantizing noise is selected, the gaussian form is approached with mean power proportional to bandwidth. By sampling at a higher rate than the minimum allowed, the quantizing noise performance can be improved by additional filtering. For example, if a baseband with highest frequency, f_T, is sampled at 4 f_T and all components above f_T are filtered out, the signal-to-quantizing noise ratio will be improved by 3 dB.

Defining the signal-to-noise ratio as the ratio of mean full-load sine-wave power to mean quantizing noise power results in the ratios shown in Fig. 7-10 for various numbers of quantizing levels. Note that each added binary digit improves the signal-to-noise ratio by 6 dB. Quadrupling the sampling rate and filtering would also improve the signal-to-noise ratio by 6 dB. Since adding a binary digit requires

Number of quantizing levels	Number of binary digits in coded representation	Signal-to-noise ratio, dB
8	3	20
16	4	26
32	5	32
64	6	38
128	7	44
256	8	50
512	9	56
1024	10	62

FIG. 7-10. Signal-to-noise ratios with various numbers of quantizing levels.

much less bandwidth increase than quadrupling the sampling rate, quantizing noise is made as small as necessary by utilizing enough digits in the code.

In the transmission of speech, the effects of quantizing noise can be reduced by making the quantizing steps large in the low probability amplitude ranges and making the steps smaller in the high probability ranges. For speech, this results in effective amplitude compression at the transmitting end with subsequent expansion at the receiving end and is called companding *(com*pressing—ex*panding)*. Companding can be performed on the analog signal before linear coding, or the same effect can be achieved with a nonlinear encoder. This is further discussed in Chap. 25.

One obvious characteristic of quantizing noise that is different from the other types discussed is that quantizing noise is only present when the signal is present. Technically, it is a form of distortion resembling in many respects the intermodulation noise discussed in Chap. 10. Analytic techniques are usually used to evaluate quantizing noise since it is very difficult to measure directly.

Summary

All of the form and peak factors of the types of noise amplitude distributions are summarized in the table of Fig. 7-11. In addition, this table lists the corrections necessary when measuring rms noise voltage with a meter responsive to the rectified average voltage. Again it should be emphasized that this table considers *amplitude* distributions in the time domain. The frequency spectral density may be flat, shaped, or bandlimited. Possible forms of the spectral density are discussed in the last column.

Distribution	Form factor, rms/rect. avg.	Peak factor (0.01% if necessary)	Correction for rms calibration of average reading sinusoid	Noise examples	Typical power spectrum
Sine	0.91 dB	3.01 dB	0 dB	Tones	Single frequency
Gaussian	1.96 dB	11.80 dB	+1.05 dB	Thermal noise	Flat over range of general interest
				Shot noise	
				1/f noise	Power proportional to wavelength
Rayleigh	1.05 dB	9.64 dB	+0.14 dB	Envelope of narrow-band gaussian noise	Bandlimited to the approximate passband of the gaussian noise
Poisson	Undefined	Undefined	Undefined	Impulse noise	Flat over range of usual interest
Rectangular	1.25 dB	4.77 dB	+0.34 dB	Quantizing noise (not directly measurable)	Approx. flat over band up to half the sampling frequency although signal dependent

FIG. 7-11. Summary of common noise distributions.

7.2 NOISE MEASUREMENT

The measurement of the amplitude of noise is made difficult by the nondeterministic nature of noise waveforms and by the amplitude dependence on bandwidth. Thus, it is usually necessary to average the noise amplitude over some time interval and to characterize the frequency response of the amplitude-indicating measuring device. This latter characterization is often performed by preceding a wideband measuring device with a filter described by a transmittance function. Transmittance, $H(f)$, of a two-port network is defined as the ratio of the output current or voltage to the input current or voltage. Transmittance can be dimensionless (voltage or current ratio), an admittance (output current to input voltage), or an impedance (output voltage to input current).

In many noise measurements, the interest is in the noise *power* rather than a voltage or current. As a consequence, the magnitude squared of the transmittance function, $|H(f)|^2$, is of more interest. Although $H(f)$ is complex in general, its phase characteristic is of no interest when dealing with average power. For noise analysis purposes, it is convenient to define an ideal bandpass transmittance, whose magnitude is shown in Fig. 7-12, to have a value of zero for all frequencies below f_1 and above f_2, and a constant value of H_0 from f_1 to f_2. The bandwidth of such a function is $f_2 - f_1$. If an ideal transmittance with $f_2 - f_1 = 1$ Hz is followed by a power meter (or mean-square voltmeter), the meter will read $|H_0|^2$ times the noise power density (or spectral density) at the input port of the transmittance averaged over this 1-Hz range.

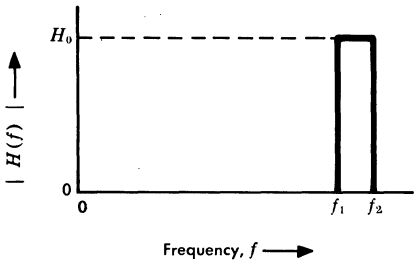


FIG. 7-12. Absolute value of ideal bandpass transmittance function.

The noise bandwidth of any transmittance function is defined as the width of an ideal bandpass filter which has an absolute transmittance value equal to the maximum absolute value of the transmittance function, and which delivers the same average power from

a white noise source as the given transmittance function. Thus, noise bandwidth, B_w, is given by

$$B_w = \frac{1}{|H_0|^2} \int_0^\infty |H(f)|^2 \, df \quad \text{Hz} \tag{7-30}$$

where H_0 = the maximum absolute value of $H(f)$.

This is illustrated in Fig. 7-13 showing a representative squared transmittance function and a rectangle, ABCD, having the same area. The width of the rectangle is the noise bandwidth, B_w. The determination of noise bandwidth reduces to the evaluation of an integral. For example, it can be shown that the noise bandwidth of a tuned RLC circuit is simply

$$B_w = \frac{\pi f_0}{2Q} \quad \text{Hz} \tag{7-31}$$

where f_0 is the resonant frequency, and Q is a measure of the selectivity given by $\omega_0 L/R$ or $\omega_0 C/G$.

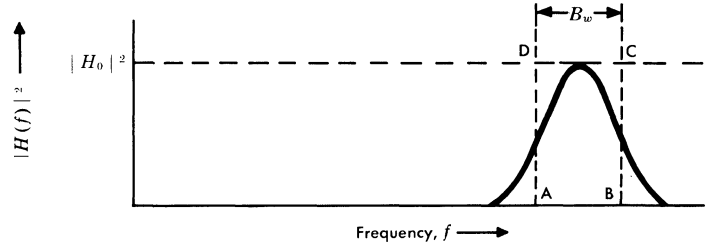


FIG. 7-13. Example of noise bandwidth.

Noise Measurement with a Voltmeter

Noise voltage can be determined under limited conditions with an a-c voltmeter. Since noise voltage is random, considerable fluctuation of a meter reading can be expected. Ideally, such fluctuation can only be eliminated by averaging the noise readings over an infinite time interval. More practically, a meter that integrates the reading over a time period which is long compared to the reciprocal of the bandwidth removes most of the fluctuation. For bandwidths greater

than several tens of kilohertz, the physical damping of most meter movements effectively performs the integration.

If the voltmeter reads true rms voltage and has a bandwidth (or frequency response) greater than the noise spectrum being measured, the voltmeter reading will be the total rms voltage of the noise. In the case of gaussian noise, this total rms voltage corresponds to σ_n, the standard deviation of the noise. If the rms voltmeter has a bandwidth less than that of the noise, the reading is proportional to that noise within the bandwidth of the meter. When the noise is shaped with frequency, the meter reading will, of course, be proportional to the average noise power in this bandwidth.

For example, if it is desired to measure the noise in a 4-kHz band within a broadband multiplex load, a filter having a 4-kHz noise bandwidth must be inserted between the measuring point and the voltmeter. If it is further desired to shape the band of noise with frequency, the filter transmittance function $|H(f)|^2$ must have the appropriate frequency shape. Such bandlimiting and shaping are often referred to as *noise weighting* and will be subsequently discussed.

When rectifying type a-c voltmeters are used to measure noise, correction factors must be applied to reflect the difference in form factors. Such correction requires knowledge of the character of the noise.

Noise Measurement with a Selective Detector

The determination of the power spectrum of a given noise source (such as the noise load applied to test a broadband transmission system) is often performed by the use of a selective detector. Actually, the selective detector is in principle an a-c voltmeter preceded by a filter as just discussed. However, most selective detectors are tunable (or switchable) over a range of frequencies much greater than their noise bandwidths. Most practical selective detectors have fixed input impedances and are often calibrated in terms of power rather than voltage. Average reading detectors must have form factor corrections applied.

For example, consider the case of obtaining a reading of N_x dBm of white noise known to be gaussian. If the selective detector used is sensitive to rectified average voltage, the form factor correction from

Fig. 7-11 amounts to 1.05 dB. Power spectral density at the detector input is given by

$$P_n = N_x - 10 \log B_w + 1.05 \quad \text{dBm/Hz} \tag{7-32}$$

where B_w is the noise bandwidth of the detector.

Noise Measurement on Telephone Channels

Although the previously discussed techniques are applicable to noise measurement in transmission channels, the near standardization of noise requirements for telephone channels has resulted in noise measuring meters specifically designed for such application.

Message Circuit Noise [10]. Noise measurement on message channels in the Bell System is characterized by an interest in how much the noise annoys the subscriber, rather than by the absolute magnitude of the average noise power. A meter which measures message circuit noise is essentially an electronic voltmeter with (1) frequency weighting, (2) an rms detector, and (3) a transient response resembling that of the human ear. These characteristics cause the noise measurement to approximate the interfering effect that the noise would create for the average telephone user.

Although other frequency weightings are used, as discussed in Chap. 2, the most common weighting is C-message weighting shown in Fig. 2-7. Quantitative effects of this and other weighting networks can be determined by integration of the appropriate transmittance function.

Let $W(f)$ represent the weighting of the noise shaping network in dB relative to the transmittance at 1 kHz. The squared transmittance function is given by

$$| H(f) |^2 = 10^{W(f)/10} \tag{7-33}$$

The total weighted noise power for noise of $p_i(f)$ watts/Hz is given by

$$p_T = \int_0^\infty | H(f) |^2 p_i(f)\, df \quad \text{watts} \tag{7-34}$$

The effect of the weighting over the frequency range from f_1 to f_2 is given by [11]

$$\chi = 10 \log \frac{\int_{f_1}^{f_2} p_i(f)\, df}{\int_{f_1}^{f_2} |H(f)|^2 p_i(f)\, df} \quad \text{dB} \qquad (7\text{-}35)$$

The weighting network attenuates the noise power by χ dB. For $f_1 = 0$ and $f_2 = 3$ kHz, $\chi \approx 2.0$ dB for flat $p_i(f)$ and C-message weighting. Thus, 0 dBm of flat noise from 0 to 3 kHz (with no power outside this range) is $90 - \chi = 88.0$ dBrnc. Similar data for any other weighting and/or noise shape can be obtained by evaluating Eq. (7-35).

Impulse Noise [12]. Digital signals such as data and PCM are not affected by noise in the same way as analog voice signals. For example, the annoying hiss due to thermal noise has no effect on digital signals unless its amplitude approaches the amplitude of the signals. On the other hand, impulses which cause tolerable clicks or pops on voice circuits result in almost certain errors because of their high amplitude. Specific counters have been designed to measure this impulse noise. Basically, an impulse counter consists of a weighting network, a rectifier, a threshold detector, and a counter of events above threshold. For ease of operation, the measuring sets include a timer which can be set to count automatically the events above threshold for a fixed time interval and then stop, holding the reading. Because of mechanical limitations of the counters used, the presently available counters can only resolve events separated by more than 7.5 milliseconds. Closer spaced impulses are counted as a single impulse, although they could be resolved by using electronic counters.

A typical impulse counter for use in the voice-frequency band has a threshold that is adjustable in 1-dB steps from 40 to 99 dBrn with a choice of terminating (600 ohm) input impedance or bridging (high) input impedance. A timer is capable of being set in 1-minute increments up to 15 minutes, although a 5-minute measuring interval is becoming standard for message circuits.

Several impulse counters can be used at different thresholds simultaneously to obtain information about the distribution of the magnitudes of the impulses. For efficiency, a four-threshold unit has been designed to facilitate such measurements, and it includes a timer capable of timing intervals up to one hour.

For wideband applications (such as wideband data or PICTUREPHONE service), special impulse measuring sets are being developed. These wideband noise measuring sets include wideband weighting networks and both average reading meters and impulse counters.

Psophometric Noise Weighting [13]. Although the dBrnc has become a standard unit of message circuit noise in the Bell System, it is not an international standard. The International Telegraph and Telephone Consultative Committee (CCITT) has defined noise as measured on a psophometer which includes a specified weighting that differs slightly from the C-message weighting used in the Bell System. For general conversion purposes, it is usually sufficient to assume that the psophometric weighting of 3-kHz white noise decreases the average power by about 2.5 dB (to be compared with the 2.0-dB factor for C-message weighting). The term *psophometric voltage* refers to the rms weighted noise voltage at a point and is usually expressed in millivolts.

It has become common in recent years to refer to average noise power (delivered to 600 ohms) rather than to noise voltage; this power is often expressed as picowatts psophometric (pWp). The relationship to psophometric millivolts is

$$\text{pWp} = \frac{(\text{psophometric mV})^2}{600} \times 10^6 \text{ picowatts} \tag{7-36}$$

or in dB quantities,

$$\text{dBp} = 10 \log (\text{pWp}) \tag{7-37}$$

For noise flat from 0 to 3 kHz, dBp can be related to dBrnc by

$$\text{dBp} = \text{dBrnc} - 0.5 \tag{7-38}$$

This relationship is not exact for other noise shapes because of the differences between psophometric and C-message weighting.

The results of the previously discussed noise units are summarized in Fig. 7-14. The data is particaularly useful when converting from one noise unit to another, since an estimate of the effects of frequency spectrum can be obtained by comparing the three conditions tabulated. The 1-kHz values are given for comparison of the various reference conditions used. The 1-kHz psophometric reading appears 1 dB high because the psophometric reference is 1 pW at 800 Hz. The 0- to

3-kHz band of white noise approximates the noise obtained from a message channel. The broadband white noise readings are proportional to the total area under the weighting curve and thus give significant information concerning the weighting function above 3 kHz. Similar data for other conditions or weightings can be obtained by integrating the appropriate weighting characteristic over the required frequency band.

Noise unit	Total power of 0 dBm		White noise of −4.8 dBm/kHz
	1000 Hz	0 to 3 kHz	
dBrnc	90.0 dBrnc	88.0 dBrnc	88.4 dBrnc
dBrn 3 KC FLAT	90.0 dBrn	88.8 dBrn	90.3 dBrn
dBrn 15 KC FLAT	90.0 dBrn	90.0 dBrn	97.3 dBrn
dBa*	85.0 dBa	82.0 dBa	82.0 dBa
Psophometric voltage (600 Ω)	870 mV	582 mV	604 mV
Psophometric emf	1740 mV	1164 mV	1208 mV
pWp	1.26×10^9 pWp	5.62×10^8 pWp	6.03×10^8 pWp
dBp	91.0 dBp	87.5 dBp	87.8 dBp

*The dBa is obsolete but is given here for reference.

FIG. 7-14. Comparison of various noise measurements.

REFERENCES

1. Lee, Y. W. *Statistical Theory of Communication* (New York: John Wiley and Sons, Inc., 1960).
2. Bennett, W. R. *Electrical Noise* (New York: McGraw-Hill Book Company, Inc., 1960).
3. Johnson, J. B. "Thermal Agitation of Electricity in Conductors," *Phys. Rev.*, vol. 32 (1928), pp. 97-109.
4. Nyquist, H. "Thermal Agitation of Electric Charge in Conductors," *Phys. Rev.*, vol. 32 (1928), pp. 110-113.

5. Fraser, D. A. S. *Statistics: An Introduction* (New York: John Wiley and Sons, Inc., 1958), p. 121.
6. *Mathematical Tables from Handbook of Chemistry and Physics* (Cleveland, Ohio: Chemical Rubber Publishing Co.).
7. Bennett, W. R. "Distribution of the Sum of Randomly Phased Components," *Quarterly of Applied Mathematics*, vol. 5 (Jan. 1948), pp. 385-393.
8. Davenport, W. B., Jr. and W. L. Root. *An Introduction to the Theory of Random Signals and Noise* (New York: McGraw-Hill Book Company, Inc., 1958).
9. Bennett, W. R. "Spectra of Quantized Signals," *Bell System Tech. J.*, vol. 27 (July 1948), pp. 446-472.
10. Cochran, W. T. and D. A. Lewinski. "A New Measuring Set for Message Circuit Noise," *Bell System Tech. J.*, vol. 39 (July 1960), pp. 911-932.
11. Aikens, A. J. and D. A. Lewinski. "Evaluation of Message Circuit Noise," *Bell System Tech. J.*, vol. 39 (July 1960), pp. 879-910.
12. Favin, D. L. "6A Impulse Counter," *Bell Laboratories Record*, vol. 41 (Mar. 1963), pp. 100-102.
13. The International Telegraph and Telephone Consultative Committee (CCITT). *Blue Book*, vol. 3 (International Telecommunication Union, 1965).

Chapter 8

Noise in Networks and Devices

The previous chapter discussed many of the types and general aspects of electrical noise affecting transmission systems. The discussion in this chapter will continue with the effects of noise on two-port networks and modulated signals. The most common types of noise are thermal noise (due to random motion of electrons) and shot noise (due to quantum flow of electricity in electronic devices). Since both are gaussian, most of the discussion will pertain to effects of gaussian noise. The results are sometimes applicable to other types of noise with the applicability justified by little more than the common attitude that if the noise cannot be predicted exactly, it may be assumed to be gaussian.

8.1 NOISE PRODUCED BY NETWORKS AND DEVICES

All transmission systems are made up of combinations of various electrical networks and devices which are themselves usually made up of individual components. In the following paragraphs the noise properties of these components are related to the terminal properties of specific, commonly used networks. The networks and devices will usually be characterized as two-port networks—a characterization which lends itself well to placing several networks in tandem to produce a working transmission system. Means of characterizing the terminal noise properties of two-port networks are discussed as are some of the means used to measure these properties. Most of the emphasis is on two-port networks in general, with specific results applicable to specific types of networks.

Calculation of Noise Output

Consider the common case of a two-port linear network being driven by a source of impedance, Z_s, as shown in Fig. 8-1. The open circuit output voltage of this network, V_0, can be related to the voltage across the input port, V_1, by the transmittance function, $H(f)$:

$$\frac{V_0}{V_1} = H(f) \tag{8-1}$$

Equation (8-1) is often awkward to use in practice. A knowledge of the source voltage, V_s, is not adequate to determine V_1; knowledge of the source impedance, Z_s, and network input impedance, Z_1, is also needed. Similarly, if the two-port is driving a load impedance, Z_L, measurement of the loaded output voltage will not determine the open circuit voltage unless Z_2 and Z_L are both known.

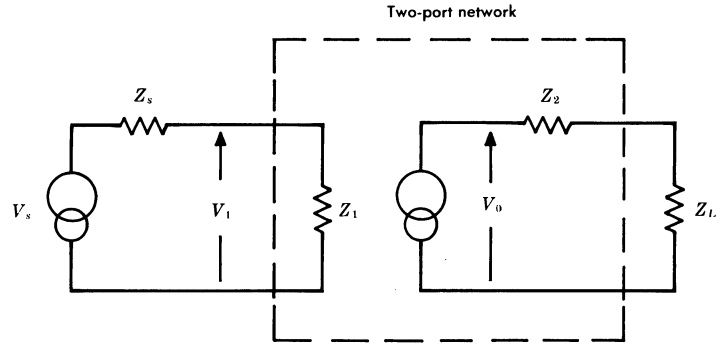


FIG. 8-1. A general two-port network with $H(f) = V_0/V_1$.

Some of these problems can be alleviated by defining the transfer characteristic of a two-port in terms of available powers. This was done in the previous chapter for a resistor where the available noise power is simply kTB_w. Note that this is independent of the resistance value. The available power gain of a two-port network is defined as the ratio of available signal power at the output terminals of the network, to the available signal power at the input terminals of the network. For the two-port shown in Fig. 8-1, the available source power is

$$\frac{|V_s|^2}{4R_s} \tag{8-2}$$

where $R_s + jX_s = Z_s$. The available output power is

$$\frac{|V_0|^2}{4R_2} \tag{8-3}$$

where $R_2 + jX_2 = Z_2$. Thus, from the above definition the available gain is given by

$$g_a = \frac{|V_0|^2}{|V_s|^2}\frac{R_s}{R_2} \tag{8-4}$$

From the network,

$$V_1 = \frac{Z_1}{Z_1 + Z_s} V_s$$

$$\frac{V_0}{V_s} = \frac{V_0}{V_1}\frac{Z_1}{Z_1 + Z_s} \tag{8-5}$$

Substituting Eq. (8-1) into (8-5), and (8-5) into (8-4) yields

$$g_a(f) = \left|\frac{H_1(f)\, Z_1}{Z_1 + Z_s}\right|^2 \frac{R_s}{R_2} \tag{8-6}$$

Note that the available gain is dependent on the source impedance, Z_s; the input impedance, Z_1; and the output resistance, R_2. However, the available gain is independent of the load impedance if the network is unilateral.

The utility of the concept of available gain is realized when cascading unilateral two-port networks. It can easily be shown that the available gain of unilateral networks in cascade is equal to the product of the available gains of the individual networks.

Consider a noiseless two-port network with available gain, $g_a(f)$, and with the input connected to a thermal noise source having a noise temperature, T. The available noise power from this source in a small band of frequencies, df, is given by

$$p_n = kTdf \text{ watts} \tag{8-7}$$

The available noise power at the output is given by

$$p_{no} = g_a(f)\, kTdf \text{ watts} \tag{8-8}$$

In general, this simple relationship between noise at the input and output of a network holds only for noiseless networks. If the network contains lossy elements or gain producing elements such as transistors, the network is not noiseless since these elements represent internal noise sources of the network. It is desirable then to have some means to characterize the amount of noise a network adds to a system by virtue of its internal noise sources. Two such means of characterization have been developed. These are the concepts of effective input noise temperature of a network and noise figure, or noise factor, of a network.

In the case of networks which are not unilateral, it is usually convenient to express the open circuit output noise voltage (or spectral density) in terms of short circuit transfer admittances, y_{0n} [1]. If each of the N uncorrelated noise generators in a network is considered as an individual port, by superposition, the output noise spectral density at port 0 is given by

$$S_0 = 2k \sum_{n=1}^{N} \left| \frac{Y_{0n}}{Y_{00}} \right|^2 T_n R_n \qquad \text{volts}^2/\text{Hz} \tag{8-9}$$

Since many practical transmission networks are unilateral and the available gain concept is intuitively more attractive than the transfer admittance approach, available gains are used in the following discussion. The discussion could have been carried out equally well by using the concept of actual power delivered to the load and the transducer gain of the network.

Effective Input Noise Temperature

Consider a two-port network having an available gain of $g_a(f)$. When it is connected to a noise source having a noise temperature of T, the available noise power in a small band, df, at the output of the network is p_{no}. This power is made up of two components: the power due to the external noise source, $g_a(f)kTdf$, and the power due to the internal noise sources of the network, p_{ne}. This can be expressed as

$$p_{no} = g_a(f)\,kTdf + p_{ne}$$

The power p_{ne} is the available noise power of the network when the input of the network is connected to a noise-free source. It is assumed that the noise voltages of the noise source driving the network and the noise sources internal to the network are uncorrelated.

If the noisy network is replaced by an equivalent noiseless one having two noise sources at its input, one of the noise sources will be the original external noise source having a noise temperature of T, and the other noise source will have a noise temperature which produces the noise power, p_{ne}. The effective noise temperature, T_e, of this equivalent representation of the internal noise source of the noisy network is

$$T_e = \frac{p_{ne}}{g_a(f)\,k\,df} \tag{8-10}$$

The available noise power at the output of the network in terms of the effective input noise temperature now becomes

$$p_{no} = g_a(f)\,k\,(T+T_e)\,df \tag{8-11}$$

The effective input noise temperature, T_e, can vary as a function of frequency, depending upon how g_a and p_{ne} vary. Furthermore, since the available gain of the network is a function of the manner in which the signal source is connected to the network, the effective input noise temperature will be a function of this as well. To reiterate, the effective input noise temperature of a two-port network is that input source noise temperature which, when connected to a noise-free equivalent to the network, results in an output noise power equal to that of the actual network when connected to a noise-free input source.

Noise Figure

The IRE definition of noise figure for a two-port network is as follows: "The noise figure (noise factor) at a specified input frequency is the ratio of (1) the total noise power per unit bandwidth at a corresponding output frequency available at the output when the noise temperature of the input source is standard (290°K) to (2) that portion of this output power engendered at the input frequency by the input source" [2]. The standard noise temperature of 290°K approximates the noise temperature of most input sources. In terms of previous definitions,

$$\text{Noise figure} = n_F = \frac{p_{no}}{g_a(f)\,kT_0 df} \tag{8-12}$$

where $T_0 = 290°\text{K}$. A noise figure such as this described for a narrow-band, df, is called a spot noise figure, which can vary as a function of frequency.

An alternative but equivalent manner of defining noise figure is in terms of the signal-to-noise degradation a network produces. The noise figure of a two-port network may be defined as the ratio of the available signal-to-noise power ratio at the input of the two-port network, to the available signal-to-noise power ratio at the output of the two-port network, when the temperature of the noise source at the input is standard. The following terms are defined:

p_{so} = the available signal power at the output of the two-port network.

p_{si} = the available signal power at the input of the two-port network.

p_{no} = the available noise power at the output of the two-port network in a small band, df.

p_{ni} = the available noise power at the input of the two-port network in a small band, df.

Since the temperature of the source is standard, $p_{ni} = kT_0df$. The definition of noise figure in terms of these symbols is then

$$n_F = \frac{p_{si}/p_{ni}}{p_{so}/p_{no}} = \frac{p_{si}/kT_0df}{p_{so}/p_{no}} \tag{8-13}$$

To show that this definition is equivalent to the IRE definition, it is only necessary to note that $g_a(f) = p_{so}/p_{si}$. Hence, Eq. (8-13) can be written as

$$n_F = \frac{p_{no}}{kT_0df} \cdot \frac{p_{si}}{p_{so}} = \frac{p_{no}}{g_a(f)\,kT_0df} \tag{8-14}$$

which is the expression resulting from the IRE noise figure definition.

In terms of the short circuit transfer admittances, the noise figure of any two-port can be determined from

$$n_F = \sum_{n=1}^{N} \left| \frac{y_{0n}}{y_{01}} \right|^2 \frac{R_nT_n}{R_sT_0} \tag{8-15}$$

where the y's are found by including the source resistance as port 1 of the $N+1$ port network.

Another useful noise figure concept is that of average noise figure. Average noise figure may be defined as the ratio of the total available

noise power at the output of a two-port network (when the noise temperature of the input source is standard), to that part of the total available output noise power due to the noise of the input source alone. Let p_{nt} be the total available noise power at the output of the two-port network. Then in terms of previous definitions,

$$\text{Average noise figure} = \bar{n}_F = \frac{p_{nt}}{g_0 k T_0 B_w} \tag{8-16}$$

where B_w = the noise bandwidth of the two-port network. Consider the relationship between spot noise figure and average noise figure. The available noise power, p_{no}, in a small band, df, at the output of a network is

$$p_{no} = n_F(f)\, g_a(f)\, kT_0 df \tag{8-17}$$

The total noise power is the integral over frequency of this power or

$$p_{nt} = \int_0^\infty p_{no} df = kT_0 \int_0^\infty g_a(f)\, n_F(f)\, df$$

Noise bandwidth, on the other hand, was defined in Eq. (7-30) as

$$B_w = \frac{1}{g_0} \int_0^\infty g_a(f)\, df$$

Substituting these expressions in the definition for average noise figure,

$$\bar{n}_F = \frac{\displaystyle\int_0^\infty g_a(f)\, n_F(f)\, df}{\displaystyle\int_0^\infty g_a(f)\, df} \tag{8-18}$$

This then is the quantitative relation between average and spot noise figure. Both of these noise figures may be expressed in dB by taking 10 log of the ratio.

Spot noise figure is useful when describing the noise behavior of a network as a function of frequency. In an FDM telephone system, channels are stacked in frequency; therefore, the spot noise figure of the repeaters is needed to describe the noise behavior from channel to channel. In an FM system it is necessary to know the total noise

across the bandwidth of the FM receiver in order to calculate *breaking* in such a system. Here the average noise figure of the FM receiver is a useful quantity.

Relation of Noise Figure to Effective Input Noise Temperature. The noise figure and effective input noise temperature of a two-port network are related analytically. It was shown in Eq. (8-11) that the available noise power appearing at the output of a network in a small band, df, is $p_{no} = g_a(f)k(T + T_e)\,df$, where T is the noise temperature of the input source, and T_e is the effective input noise temperature of the network. To make this output noise power conform with that required in the definition of noise figure, it is necessary to make the noise temperature of the input source equal to standard temperature, T_0. Hence, $p_{no} = g_a(f)k(T_0 + T_e)df$. The output noise in terms of noise figure is from Eq. (8-17) : $p_{no} = n_F g_a k T_0 df$. Equating these two powers, $n_F T_0 = T_0 + T_e$. Solving for n_F in one case and T_e in the other,

$$n_F = 1 + \frac{T_e}{T_0} \tag{8-19}$$

and

$$T_e = T_0(n_F - 1) \tag{8-20}$$

The concept of noise figure is most useful when the input source has a noise temperature approximately equal to standard temperature. Consider the expression for the noise power at the output of the network, $p_{no} = n_F g_a k T_0 df$. Rewriting this equation in terms of dBm (reference 1 mW),

$$\begin{aligned} P_{no} &= N_F + G_a + 10 \log kT_0 df + 30 \\ &= N_F + G_a - 174 \text{ dBm} + 10 \log df \qquad \text{dBm} \end{aligned} \tag{8-21}$$

The symbols are defined as follows:

$$P_{no} = 10 \log \frac{p_{no}}{10^{-3}} \qquad \text{dBm}$$

$$N_F = 10 \log n_F \qquad \text{dB}$$

$$G_a = 10 \log g_a \qquad \text{dB}$$

Hence, the available noise power of the two-port network in dBm can be written as the sum of the noise power of the thermal noise source

in dBm, the available gain of the network in dB, and the noise figure of the network in dB. The effects of internal noise sources of a two-port network can therefore be taken into account by adding the noise figure in dB to the available noise power of the source in dBm.

The concept of effective input noise temperature is useful when the source noise temperature differs from that of standard temperature. It has a distinct advantage when the noise performance of a complete communications system is being evaluated. Specific use of effective input noise temperature is discussed later with respect to low noise applications.

Cascaded Networks. The two networks connected in tandem as shown in Fig. 8-2 have effective input temperatures, T_{e_1} and T_{e_2}, and

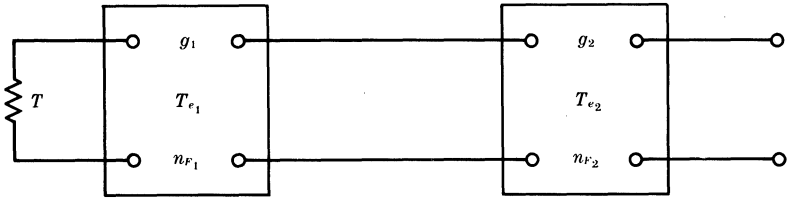


FIG. 8-2. Cascaded networks and noise.

available gains, g_1 and g_2. Suppose that these two tandem amplifiers are connected to a noise source having a noise temperature, T. In a small frequency band, df, the noise power due to the noise source only is g_1g_2kTdf. The noise power due to noise sources in the first network is $g_1g_2kT_{e_1}\,df$, and the noise power due to noise sources in the second network is $g_2kT_{e_2}\,df$. The total noise appearing at the output of the second network is $kg_2(g_1T+g_1T_{e_1}+T_{e_2})\,df$. The portion of this noise due to noise sources internal to the two networks is $kg_2(g_1T_{e_1}+T_{e_2})\,df$. The effective input temperature of the two networks in tandem is then

$$T_{e_{12}} = \frac{kg_2(g_1T_{e_1} + T_{e_2})\,df}{g_1g_2kdf}$$

$$= T_{e_1} + \frac{T_{e_2}}{g_1}$$

This result can be easily generalized to n networks in tandem. The

resulting effective input noise temperature is

$$T_{e_{1\ldots n}} = T_{e_1} + \frac{T_{e_2}}{g_1} + \cdots \frac{T_{e_n}}{g_1 g_2 \ldots g_{n-1}} \tag{8-22}$$

Using the relationship between noise figure and effective input noise temperature, it can be easily shown that the resulting noise figure of n stages in tandem is

$$n_{F_{1\ldots n}} = n_{F_1} + \frac{n_{F_2} - 1}{g_1} + \cdots \frac{n_{F_n} - 1}{g_1 g_2 \ldots g_{n-1}} \tag{8-23}$$

The significance of these two relationships becomes apparent when considering a multistage amplifier in which each stage has an available gain of at least 20 dB (ratio of 100). If each stage has the same effective input noise temperature, then the noise contribution of only the first stage is significant. Only noise sources occurring before or in the first stage of the amplifier need be considered so far as noise calculations are concerned. However, if the gain of the first stage is small or if the noise contribution of the second stage is large, then it is necessary to take these into account when making noise calculations.

Attenuators. Consider an attenuator inserted between a noise source and a load as shown in Fig. 8-3. Assume that the noise source has a noise temperature of T_s degrees and that the lossy elements of the attenuator are at a temperature of T degrees. Attenuators are usually calibrated in terms of insertion loss between fixed equal impedances. Such insertion loss (or gain) is equal to the more broadly defined available loss (or gain). However, attenuators are

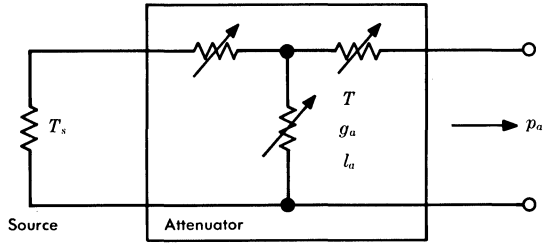


FIG. 8-3. Noise and the attenuator.

usually *not* unilateral, and available gains as used here are only generally applicable when the attenuator is properly terminated. At unity gain (unity loss, 0-dB loss, 0-dB gain), the available noise power in a band, B_w, at the output of the attenuator is kT_sB_w watts. At zero gain (infinite loss), the available noise power is kTB_w watts and is due to the lossy elements in the attenuator. If the noise source and the attenuator are at the same temperature, no change in available noise power at the output of the attenuator is noticed as the attenuator loss is varied from 0 dB to infinity. The noise produced by the source, which can be considered the signal, is proportional to the difference in temperature between the noise source and attenuator. This signal is attenuated by the lossy elements of the attenuator. The available signal power at the output of the attenuator is then $g_akB_w(T_s - T)$ for $0 \leq g_a \leq 1$. The total available power at the output of the attenuator is the sum of the signal and noise:

$$\begin{aligned} p_a &= g_akB_w(T_s - T) + kB_wT \\ &= kB_w[g_aT_s + (1 - g_a)T] \end{aligned}$$

Note that the available power given by this expression agrees with the values calculated for zero gain and unity gain. The effective input noise temperature of the attenuator is then

$$\begin{aligned} T_e &= \frac{p_a}{kB_wg_a} - T_s \\ &= \frac{kB_w[g_aT_s + (1 - g_a)T]}{kB_wg_a} - T_s \\ &= \frac{1 - g_a}{g_a}T \end{aligned} \tag{8-24}$$

If the loss ratio of the attenuator is l_a, then in terms of the gain, $l_a = 1/g_a$. The loss of the attenuator in dB is $L_a = 10 \log l_a$. The effective input noise temperature in terms of the attenuator loss ratio is then

$$T_e = T(l_a - 1) \tag{8-25}$$

and by Eq. (8-19) the noise figure of the attenuator is

$$n_F = 1 + \frac{T}{T_0}(l_a - 1) \tag{8-26}$$

If the lossy elements of the attenuator are at standard temperature, then the noise figure of the attenuator is equal to the loss of the attenuator; that is,

$$n_F = l_a$$

or

$$N_F = L_a \text{ dB} \tag{8-27}$$

This result can also be arrived at by using the signal-to-noise definition of noise figure.

The results given by Eqs. (8-25) and (8-26) apply only to those attenuating networks that achieve their attenuation through lossy elements such as resistors. Attenuating networks can be constructed using reactive elements. Such attenuators achieve their loss by reflecting power to the source. An attenuator constructed entirely of reactive devices will have an effective input noise temperature of zero degrees since pure reactive elements are noiseless. However, the loss of such attenuators must be included in Eqs. (8-22) and (8-23) when calculating the noise performance of a system containing them.

Semiconductor Noise. Semiconductor devices have, in addition to the usual thermal noise sources, other noise sources, which are d-c bias dependent. For clarity, all d-c bias currents and voltages are denoted by the upper case letters, I and V, respectively; and the noise currents and voltages have an rms value given by the lower case letters, i and e, respectively.

Diode Noise. The shot noise in a PN junction diode consists of the noise due to the minority carrier current, I_s (saturation current), and the majority carrier current, $I_s e^{qV/kT}$ (forward current). The two noise currents are statistically independent and add on a power basis. Therefore, from Eq. (7-20)

$$i_{rms}^2 = 2qI_s\,(1 + e^{qV/kT})\,B_w \tag{8-28}$$

The actual diode current is given by the familiar diode equation

$$I = I_s(e^{qV/kT} - 1) \tag{8-29}$$

Substituting into Eq. (8-28) yields

$$i_{rms}^2 = 2qB_w(I + 2I_s) \tag{8-30}$$

The available noise power from the diode would be delivered to a matched load of conductance,

$$G = \frac{dI}{dV} = \frac{qI_s}{kT} e^{qV/kT} = \frac{q(I + I_s)}{kT} \tag{8-31}$$

Thus, the available noise power from a PN junction is given by

$$P_n = \frac{i_{rms}^2}{4G} = \frac{kTB_w}{2} \frac{I + 2I_s}{I + I_s} \quad \text{watts} \tag{8-32}$$

For $I = 0$, or no current through the diode, there is thermal equilibrium, and the diode looks like a resistor at temperature T. For $I >> I_s$ there is strong forward conduction, and the diode looks only half as noisy as a resistor. For $I \rightarrow -I_s$ (the reverse bias condition), the diode looks like an extremely hot noise generator and in this condition is sometimes used as a noise source.

Transistor Noise. The important noise sources in a transistor can be obtained by combining two junctions and considering the effects of recombination in the base region. The easiest noise equivalent circuit to use for this application is the equivalent tee circuit. The noise sources include shot noise, arising from the random passage of carriers through the junctions in the transistor (diffusion fluctuations); partition noise, resulting from the random division of carriers between the base and collector (recombination fluctuations in the base region of the transistor); and thermal noise, arising from the resistive components in series with the base, collector, and emitter junction leads. (The most significant is the base resistance.)

For frequencies less than the α cutoff frequency, a simplified noise equivalent circuit as shown in Fig. 8-4 can be used. Since the emitter-base junction is usually forward-biased, noise in this area is just like that generated in the junction diode. Therefore a noise generator is placed in the emitter circuit to inject a noise current, i_{en}, as generally given by Eq. (8-30).

$$i_{en} = \sqrt{2qI_eB_w} \tag{8-33}$$

where I_e is the d-c emitter current in amperes.

The collector region provides three sources of noise. The first is due to the fraction of the emitter current reaching the collector. Hence a term given by $2q\alpha_0 I_e B_w$ or $2qI_cB_w$ is needed. The second is due to the collector saturation, I_{c0}, which flows when $I_e = 0$. This

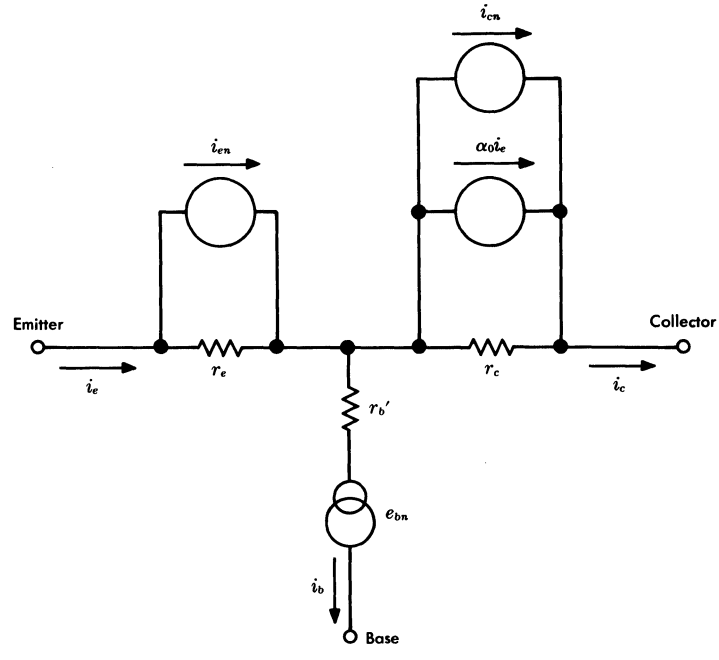


FIG. 8-4. Simple equivalent circuit for a noisy transistor.

term is usually small when compared to the first. The third is a partition noise term which results from the random fluctuations of the division of the emitter current between the base and collector. Combining these three effects, Van der Ziel [3] has shown that the total noise in the collector is

$$i_{cn} = \sqrt{2qB_w \left[I_{c0} + I_c (1 - \alpha_0)\right]} \tag{8-34}$$

In the base region a noise voltage is associated with the base spreading resistance, r_b', and is a thermal noise source with an rms voltage given by

$$e_{bn} = \sqrt{4kTB_w r_b'} \tag{8-35}$$

where T is the temperature of the base resistance in degrees Kelvin, and r_b' is the base resistance in ohms.

Several simplifying assumptions have been made to arrive at this equivalent circuit. First, the frequency dependence of the emitter

and collector noise generators has been neglected. Second, the emitter and collector noise generators have been assumed to be independent (not statistically correlated). Third, the equivalent tee circuit of the transistor has been simplified by neglecting the effect of space-charge layer widening, and by neglecting the frequency characteristics of the emitter and collector junctions. This equivalent circuit is valid up to the frequency $f_\alpha \sqrt{1 - \alpha_0}$, where f_α is the α cutoff frequency of the transistor. Above this frequency it is necessary to include the dependence of α upon frequency and the collector junction capacity into the equivalent circuit. More detailed analyses and equivalent circuits are given in the references listed at the end of this chapter.

The preceding model is also not accurate at very low frequencies. At these frequencies, $1/f$, or excess noise, dominates the noise picture. It is suspected that the $1/f$ noise is caused by fluctuations in the conductivity of the semiconductor medium; however, no physical model has yet been advanced to explain satisfactorily the observed $1/f$ noise spectral density.

It is instructive to obtain the noise figure of the above transistor model in the mid-frequency range ($1/f$ effects at low frequencies and α cutoff affects at high frequencies are ignored). To keep the mathematics simple, the following assumptions are made:

$$\begin{aligned} r_b' &<< \alpha_0 r_c \\ R_g + r_e &<< \alpha_0 r_c \\ r_e &= kT/qI_e \\ I_{c0} &<< I_c(1 - \alpha_0) \end{aligned} \tag{8-36}$$

Converting the current sources in Fig. 8-4 to voltage sources and driving the common base transistor with a source impedance of R_g results in the circuit of Fig. 8-5, with the values of the noise voltage sources given by

$$\begin{aligned} e_{gn} &= \sqrt{4kTR_gB_w} \\ e_{en} &= \sqrt{2kTr_eB_w} \\ e_{bn} &= \sqrt{4kTr_b'B_w} \\ e_{cn} &= \sqrt{2kT\frac{\alpha_0 r_c^2}{r_e}(1 - \alpha_0)B_w} \end{aligned} \tag{8-37}$$

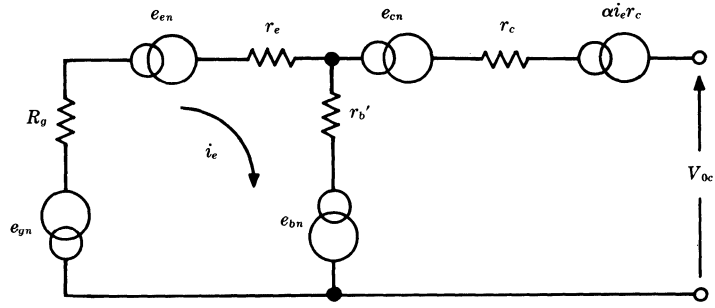


FIG. 8-5. Equivalent transistor circuit for calculation of noise figure.

The *ratio* of the available output noise power from all sources, to the available output noise power from the source resistance, R_g, gives the noise figure. This implies that the transistor is loaded with a load impedance, Z_L, equal to the conjugate of the impedance seen looking into the output port of the transistor. Computation of this impedance can be avoided by recalling that the available output power is equal to the square of the open circuit output voltage divided by $4R_L$. Since Z_L is constant for both conditions in determining the power ratio, the noise figure is also obtainable as the ratio of the mean squared open circuit output noise voltage of all sources, to the mean squared open circuit output noise voltage of the e_{gn} source.

The source, e_{gn} of Fig. 8-5, with all other noise sources shorted, will contribute to the loop current, i_e. By Ohm's law

$$i_e = \frac{e_{gn}}{R_g + r_e + r_b'} \tag{8-38}$$

yields an open circuit output voltage given by

$$\alpha_0 i_e r_c = \frac{\alpha_0 r_c e_{gn}}{R_g + r_e + r_b'} \tag{8-39}$$

where the portion of the output given by

$$\frac{r_b' e_{gn}}{R_g + r_e + r_b'}$$

can be ignored since $r_b' << \alpha_0 r_c$. Therefore, the mean squared open circuit output voltage due only to the source is given by

$$V_{g_{oc}}^2 = \frac{4kTB_w R_g \alpha_0^2 r_c^2}{(R_g + r_e + r_b')^2} \tag{8-40}$$

By a similar argument, the open circuit squared output voltage due to e_{en} can be obtained as

$$V_{e_{oc}}^2 = \frac{2kTB_w r_e \alpha_0^2 r_c^2}{(R_g + r_e + r_b')^2} \tag{8-41}$$

and that from e_{bn} (ignoring the direct term because $R_g + r_e << \alpha_0 r_c$) is

$$V_{b_{oc}}^2 = \frac{4kTB_w r_b' \alpha_0^2 r_c^2}{(R_g + r_e + r_b')^2} \tag{8-42}$$

The open circuited squared output voltage due to e_{cn} is simply e_{cn}^2 or

$$V_{c_{oc}}^2 = \frac{2kTB_w \alpha_0 r_c^2 (1 - \alpha_0)}{r_e} \tag{8-43}$$

The noise figure is given by

$$n_F = \frac{V_{g_{oc}}^2 + V_{e_{oc}}^2 + V_{b_{oc}}^2 + V_{c_{oc}}^2}{V_{g_{oc}}^2} \tag{8-44}$$

which for $T = T_0$ becomes [4]:

$$n_F = 1 + \frac{r_e}{2R_g} + \frac{r_b'}{R_g} + \frac{(1 - \alpha_0)\ (R_g + r_e + r_b')^2}{2\alpha_0 r_e R_g} \tag{8-45}$$

Note that the noise figure is not only dependent upon the transistor parameters, but also upon the source resistance, R_g which must be at standard temperature. In fact, there is an optimum value of R_g that will minimize the noise figure. Equation (8-45) suggests that a low noise figure can be obtained by making α_0 very close to unity (very high gain) and then driving from a source impedance which is large compared to r_e or r_b'. Small r_e implies that the d-c bias current, I_e, cannot be made too small. In practice, the minimum noise figure is usually broad enough so that R_g is often determined by considerations other than noise alone, such as impedance matching.

Actual noise figures in the range of 2 to 6 dB are obtainable in the mid-frequency region of operation (between excess noise and α cutoff effects) with source resistances, R_g, from 100 to 1000 ohms. It can be shown analytically that the noise figure obtained in this frequency region is practically independent of the circuit configuration (common base, common emitter, or common collector) that is used.

Effect of Terminations. Many situations arise in the design of broadband amplifiers where it is necessary to match the input impedance of a transistor amplifier to a constant real impedance over a large band of frequencies. For example, in order to avoid interaction effects between cable impedance and repeaters, it is often desirable to design the amplifier so that it matches the cable impedances. The value of noise figure for the amplifier will be affected differently, depending on the method used to obtain the desired input impedance.

One convenient way of accomplishing impedance matching is to use low impedance input circuits, such as common base stages or common emitter stages with shunt feedback, which can be easily matched to the driving source impedance by what is known as *brute force terminations.* This is simply a resistor placed in series with the input port to raise the low transistor impedance to the desired value. The dual network for this case can also be used where a shunt resistor is required to lower the impedance. This finds application in circuits with electron tubes, field effect devices, or series feedback transistor stages.

Consider the transistor input circuit shown in Fig. 8-6 with shunt feedback supplied from the feedback network denoted by β. If the amount of feedback is large, the impedance R_{in} will approach zero (an ohm or two is typical), and a good impedance match to the cable impedance, R_c, can be obtained by adding the series resistor, $R_t \approx R_c$. Note that the R_g used in previous transistor noise figure computations is equal to $R_c + R_t$ in parallel with the β circuit impedance. From

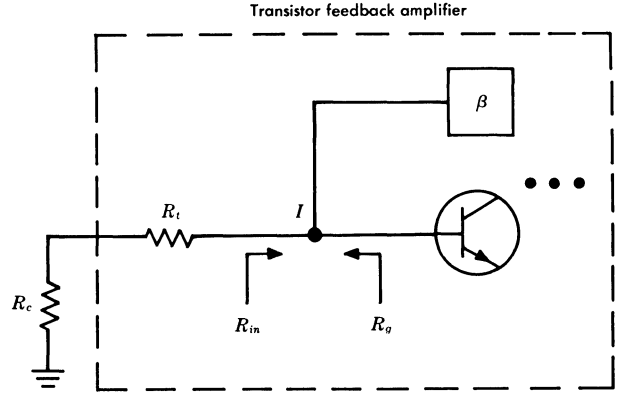


FIG. 8-6. Brute force input termination of shunt feedback amplifier.

Eq. (8-45), it can be seen that the transistor noise figure is a function of the source impedance, R_g. Changing R_g to R_c changes the noise figure of the transistor. However, it is convenient to compare the transistor noise figure with a source impedance of $R_c + R_t$, to the amplifier noise figure with a source impedance of R_c.

First, it must be recognized that feedback does not improve the noise figure of the amplifier. This can be seen by realizing that any improvement in noise performance afforded by the feedback is exactly compensated by the necessity for more gain in the active portion of the circuit. Signal levels at node I (the input) are lowered by feedback by the same amount as the noise, and thus, the signal-to-noise ratio (hence noise figure) is unchanged by feedback.

Next, it should be remembered that if $R_g = R_c + R_t$ and the shunting effect of the β circuit is ignored, this circuit is identical to that used for deriving the transistor noise figure. Thus, the total noise of the amplifier is directly proportional to the numerator of Eq. (8-44). However, the reference condition given previously by Eq. (8-40) is modified by replacing R_g with R_c. If the brute force termination, $R_t = R_c = \frac{1}{2}R_g$, the reference noise power of Eq. (8-40) is exactly half its previous value. As a consequence, the noise figure given by Eq. (8-44) is doubled. The result is a 3-dB increase in noise figure over that measured without a termination but with a new source impedance equal to twice the actual source impedance.

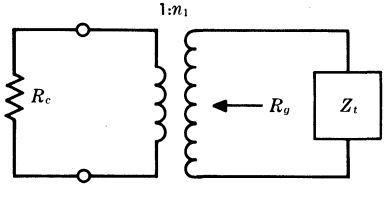


FIG. 8-7. Transformer input coupling network.

The 3-dB noise penalty associated with the brute force termination may be very undesirable. In some cases, this can be avoided by the use of transformer coupling as shown in Fig. 8-7. If Z_t is reasonably constant and real over the range of frequencies of interest, a transformer turns ratio of $1:n_1 = \sqrt{Z_t/R_c}$ will result in an input impedance to the low side of the transformer of R_c ohms. This also results in a high side impedance as seen by the transistor to be $R_g = Z_t$. If this value of R_g is not close to the optimum value for the transistor, a noise penalty results. The disadvantage of simple transformer coupling is its inability both to satisfy a good impedance match at

the input and to present the optimum source impedance to the transistor for minimum noise figure.

A method of termination that is acceptable from this standpoint results when the low side of the coupling network is connected as a hybrid. The circuit of such a coupling network is shown in Fig. 8-8. Circuit analysis of this network shows that the open circuit voltage gain is $n_2/2$, where n_2 is the turns ratio of the hybrid network, and the high side resistance in terms of the cable impedance, R_c, is

$$R_g = n_2^2 \frac{R_c}{2} \tag{8-46}$$

Thus, the turns ratio can be adjusted to optimize R_g for low transistor noise figure. It can be shown that the input impedance looking into the low side, or cable side, of the coupling network is exactly R_c when the impedance of the balancing network, Z_b, is adjusted to be

$$Z_b = \left(\frac{n_2 R_c}{2}\right) \frac{1}{Z_t} \tag{8-47}$$

where Z_t is the impedance of the load placed on the high side of the coupling network by the transistor amplifier and associated circuitry. Although the potential noise performance of the hybrid termination is ideal, such terminations are not universally used. This is primarily due to difficulties in controlling the characteristics of the hybrid over the wide frequency ranges of modern amplifiers. In the case of feedback amplifiers, the effects of the hybrid on the feedback characteristic may be important to frequencies many octaves above the highest frequency being amplified.

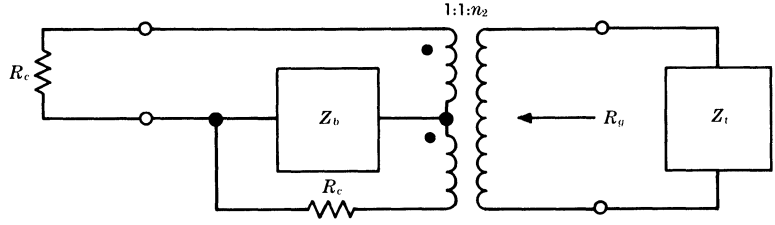


FIG. 8-8. Hybrid input coupling network.

The preceding methods of getting good input impedance do not exhaust the list of possibilities. Some of the other alternatives, such as the use of series and shunt feedback to obtain the correct terminating impedance and the use of bridge structures, are useful in certain cases. Hybrid feedback connections in particular appear to offer signal-to-noise advantages over purely passive structures. A complete examination of these alternatives is beyond the scope of this text.

Low Noise Applications. There are many applications in modern communications systems where the thermal input noise to the system may be much less than that referred to 290°K [5]. For example, the ground station antenna for a space communications system is normally pointed at high elevation angles where the atmospheric noise is only a few degrees Kelvin. This noise cannot be avoided, but it is important that the antenna does not pick up additional noise from the ground or the lower and warmer portions of the sky. If the side and back lobe characteristics are good, little noise above the minimum background level will be picked up. If this antenna is followed by an amplifier with the respectable noise figure of 1 dB, the amplifier noise will be much greater than the noise coming into the antenna, and the signal-to-noise degradation will be much more than 1 dB. This is true because the inescapable noise coming in on the antenna is much less than that due to a resistor at room temperature. The result is that the conventional definition of noise figure is not very convenient for low noise applications.

Instead, it has become common to use effective input noise temperature to characterize low noise devices. This was previously defined as that input source noise temperature which, when connected to a noise-free equivalent of the network, results in an output noise power equal to that of the actual network connected to a noise-free input source.

The available noise power appearing at the output of a network in a band, B_w, is $p_n = g_a k(T+T_e)B_w$. The noise temperature of the input source is T, and T_e is the effective input noise temperature of the network. For the case of a communications system consisting of an antenna and a receiver, the noise temperature of the input source would be the antenna temperature T_{ant}, and T_e would be the effective input noise temperature of the receiver. The latter includes the noise contribution due to loss of passive r-f components appearing between the receiver and the antenna. The sum of these two temperatures

is referred to as the total effective system noise temperature, or

$$T_{sys} = T_{ant} + T_e \tag{8-48}$$

This system noise temperature gives the total system noise power referred to the input of the receiver.

Consider a short piece of waveguide connecting a very low noise antenna to a very low noise amplifier such as a maser. From Eq. (8-25), the effective input noise temperature of an attenuator is given by

$$T_e = T(l_a - 1)$$

where T is the actual temperature of the loss elements in the attenuator. At room temperature (290°K), Eq. (8-25) reduces to the helpful approximation

$$T_e \approx 66.8L \qquad L < 0.5 \text{ dB} \tag{8-49}$$

where L is the attenuator loss in dB. For each tenth of a dB loss in the waveguide, the effective input noise temperature will increase by about 6.7°K.

Even small waveguide losses, if present between the antenna and a low noise receiver, can have drastic effects on the transmission quality of the system. Assume, for instance, a receiving system temperature of 25°K (all noise sources referred to the input of the low noise amplifier). The waveguide losses are then increased by 0.1 dB. This adds 6.7°K for a total receiver noise of 32.7°K, which is just 1 dB higher than 25°K. The signal-to-noise ratio of an incoming signal has therefore been reduced by 1.1 dB (1 dB due to noise and 0.1 dB due to signal loss). If the waveguide loss is higher than a few dB, the noise temperature of the attenuator reaches an asymptotic value of 290°K.

Example 8.1

Problem

Consider the satellite communication receiving system shown in Fig. 8-9. This system consists of a highly directional, high-gain

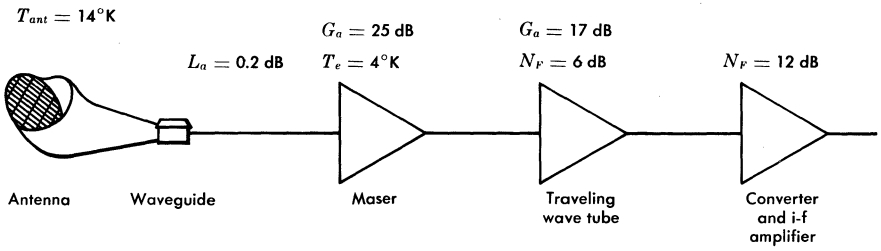


FIG. 8-9. An example of low-noise communications system and system noise calculations.

antenna followed by a low noise amplifier. Electrical properties of the various system components are:

Antenna
Noise temperature = 14°K

Waveguide (at room temperature)
Loss = 0.2 dB

Maser
Gain = 25 dB
Effective input noise temperature = 4°K

Traveling wave tube
Gain = 17 dB
Noise figure = 6 dB

Converter and i-f amplifier
Noise figure = 12 dB

Calculate the effective input noise temperature of the receiver and the system noise temperature.

Solution

First, the dB quantities should be converted into ratios.

Maser
Gain = 25 dB, gain ratio = 316

Traveling wave tube
Gain = 17 dB, gain ratio = 50
Noise figure = 6 dB, ratio = 4.0

Converter and i-f amplifier
Noise figure = 12 dB, ratio = 15.8

To compute the effective input noise temperature of the receiver, it is necessary in this case to use the formula for cascaded networks, Eq. (8-22),

$$T_e = T_{e_1} + \frac{T_{e_2}}{g_1} + \frac{T_{e_3}}{g_1 g_2} + \frac{T_{e_4}}{g_1 g_2 g_3}$$

Here T_{e_1}, T_{e_2}, T_{e_3}, and T_{e_4} are the effective input noise temperatures of the waveguide, maser, traveling wave tube, and converter, respectively; while g_1, g_2, and g_3 are the gains (available) of the waveguide, maser, and traveling wave tube, respectively. The relationship between noise figure and effective input noise temperature must be used in order to establish T_{e_3} and T_{e_4} for the traveling-wave tube and converter. This relation, Eq. (8-20), is

$$T_e = T_0(n_F - 1)$$

Then

$$T_{e_3} = 290\ (4 - 1) = 870°\text{K}$$

$$T_{e_4} = 290\ (15.8 - 1) = 4290°\text{K}$$

Equation (8-49) may be used to calculate the effective input temperature of the waveguide due to its losses.

$$T_e = 66.8L$$

$$T_{e_1} = (66.8)\ (0.2) = 13.6°\text{K}$$

The waveguide gain ratio, g_1, equals 0.955 for 0.2-dB loss.

The effective input noise temperature of the receiver may now be calculated using Eq. (8-22).

$$T_e = 13.6 + \frac{4.0}{0.955} + \frac{870}{(0.955)\ (316)} + \frac{4290}{(0.955)\ (316)\ (50)}$$

$$= 13.6 + 4.2 + 2.9 + 0.3$$

$$= 21.0°\text{K}$$

The system noise temperature is given by Eq. (8-48):

$$T_{sys} = T_{ant} + T_e$$

$$= 14° + 21° = 35°\text{K}$$

Measurement of Effective Input Noise Temperature and Noise Figure

In general, two methods of measuring noise figure exist [6]. One of these involves the use of a calibrated signal generator to characterize the gain, g, and the effective noise bandwidth, B_w, of the two-port network. The measured noise output power can then be compared to the theoretical value, given by gkT_0B_w, to obtain the noise figure. The second method utilizes a calibrated noise source placed at the input of the network. Both of these measurement techniques are similar and use the circuit arrangement shown in Fig. 8-10. The noise or signal source consists of a generator that can be turned off and on, and an internal source termination that supplies a fixed amount of thermal noise whether the generator is turned off or on. The source is connected to the input of the two-port network undergoing measurement. Note that the value of noise figure measured can be a function of the impedance match existing at this point. A power meter is connected at the output of the two-port network. In general, this power meter will read not the available output power of the two-port, but rather the actual power delivered to the meter. Therefore, in these calculations the transducer gain (g_t) of the two-port network is used.

In general, two power measurements are made at the output of the network: one with the generator in the signal source turned off (p_1), and the other with the generator in the signal source turned on (p_2).

Calibrated Noise Source. For the case of the calibrated noise source, the noise temperature of the source, T_1, is that of its internal termi-

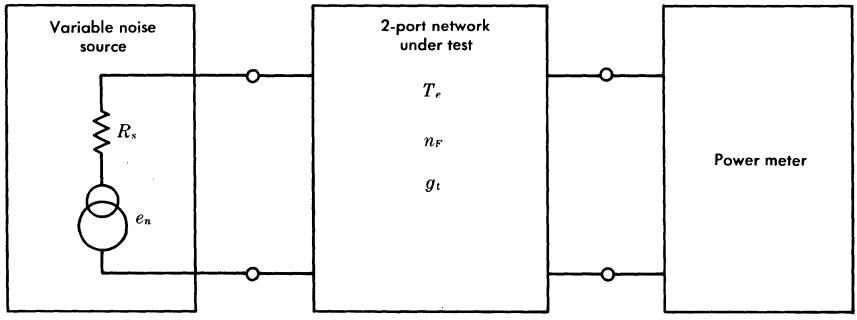


FIG. 8-10. Setup for measuring noise figure.

nating resistance when the generator in the noise source is turned off. With the generator on, the noise temperature of the source becomes T_2. The noise powers measured at the output of the two-port network under the generator off and on conditions are respectively

$$p_1 = g_t k B_w (T_e + T_1)$$

$$p_2 = g_t k B_w (T_e + T_2)$$

The unknowns are g_t, the transducer gain of the network; B_w, the noise bandwidth of the combined network and power meter; and T_e, the effective input temperature of the network. Of these, the latter is desired. Consider the ratio of these two equations. Let $y = p_2/p_1$; then,

$$y = \frac{p_2}{p_1} = \frac{g_t k B_w (T_e + T_2)}{g_t k B_w (T_e + T_1)} = \frac{T_e + T_2}{T_e + T_1}$$

Solving for T_e,

$$T_e = \frac{T_2 - yT_1}{y - 1} \tag{8-50}$$

The noise figure of the two-port network is given by

$$\begin{aligned} n_F &= 1 + \frac{T_e}{T_0} \\ &= \frac{\left(\frac{T_2}{T_0} - 1\right) + y\left(1 - \frac{T_1}{T_0}\right)}{y - 1} \end{aligned} \tag{8-51}$$

If the source impedance is at standard temperature (that is, $T_1 = T_0$), then this expression becomes:

$$n_F = \frac{(T_2/T_0) - 1}{y - 1} \tag{8-52}$$

Sources of white noise that can be used for such measurements are the temperature-limited noise diode, the gas tube noise generator, and resistors at different temperatures. The temperature-limited noise diode represents a variable noise source useful at frequencies less than 1 GHz. The noise output may be changed by simply varying the anode current of the diode. The gas tube noise generator

represents a source of noise power useful in the microwave frequency regions. Since the gas tube produces a constant available noise power, it is necessary to use a variable attenuator between it and the load if a variable source of noise power is desired. Resistors at different but precisely known temperatures may be used as very accurate noise sources for the measurement of noise figure. Accurate temperatures may be obtained by immersing resistors in ice baths, liquid nitrogen baths, etc.

Calibrated Signal Source. For the measurement of effective input noise temperature and noise figure using a calibrated signal source, it is assumed that the generator source impedance is at standard temperature. When the generator is on, the available power of the signal is p_c. The powers measured at the output of the two-port network, with the generator off and the generator on, are respectively

$$p_1 = kg_tB_w(T_e + T_0)$$

$$p_2 = kg_tB_w(T_e + T_0) + p_cg_t$$

Taking the ratio of these two powers,

$$y = \frac{p_2}{p_1} = \frac{g_tkB_w(T_e + T_0) + g_tp_c}{g_tkB_w(T_e + T_0)}$$

$$= 1 + \frac{p_c}{kB_w(T_e + T_0)}$$

Solving for T_e,

$$T_e = \frac{p_c}{kB_w(y - 1)} - T_0 \tag{8-53}$$

The expression for noise figure is

$$n_F = \frac{1}{y - 1} \cdot \frac{p_c}{kT_0B_w} \tag{8-54}$$

Bandwidth Effects. In order to complete the calculation of noise figure or effective input noise temperature, it is necessary to know the noise bandwidth, B_w. This quantity is not easy to measure accurately. The dispersed noise source method has the advantage in that it is not necessary to know the noise bandwidth.

If it is desired to measure *spot noise figure,* then the power detector must have a noise bandwidth considerably smaller than that of the network undergoing measurement. If this is the case, the noise bandwidth, B_w, used in the expressions given in this chapter is that of the power meter. If on the other hand it is desired to measure the average noise figure of the network, then the power meter should have a noise bandwidth considerably larger than that of the network. For this case the noise bandwidth, B_w, is the noise bandwidth of the network.

Example 8.2

Problem

A calibrated noise source is used to measure the noise figure of an amplifier. It is found that the excess noise temperature of the noise source required to double the noise power measured at the output of the amplifier is 1450°K. Assuming that the termination in the noise source is at standard temperature, what is the noise figure of the amplifier?

Solution

Since the termination is at standard temperature for one of the noise measurements, Eq. (8-52) may be used to compute the noise figure:

$$n_F = \frac{(T_2/T_0) - 1}{y - 1}$$

In this case T_2 is the sum of the excess noise temperature and standard temperature, i.e., $T_2 = T_x + T_0$. Hence, the expression for noise figure may be written

$$n_F = \frac{1}{y-1} \cdot \frac{T_x}{T_0}$$

Substituting the given information into this expression,

$$n_F = \frac{1}{2-1} \cdot \frac{1450}{290} = 5$$

or

$$N_F = 7.0 \text{ dB}$$

The ratio T_x/T_0 is sometimes referred to as the *excess noise ratio* of a noise source, and the ratio T_2/T_0, the noise ratio of a noise

source. By using a noise source whose termination is at standard temperature and by adjusting the excess noise temperature of the source so that the noise measured at the output of the amplifier doubles (increases by 3 dB), it is found that the noise figure is equal to the excess noise ratio of the noise source.

8.2 NOISE AND AMPLITUDE-MODULATED SIGNALS

Up to this point, this chapter has discussed the effects of noise when directly mixed with the desired signal. The remainder of the chapter is concerned with the effects of noise on a baseband signal when the noise is introduced after the baseband signal has been modulated and/or coded. As expected, some of the signal and noise relationships are complicated by the modulating process [7].

If an unmodulated carrier wave is linearly combined with a band of random noise having constant spectral density, the resultant wave is equivalent to a carrier wave that has been both amplitude- and phase-modulated by random noise. When the resultant wave is demodulated by either an ideal amplitude detector or an ideal phase detector, a random noise output is obtained.

SSB Modulated Wave

As noted in Chap. 5, the demodulation of an SSB wave consists of translating it in frequency with or without inversion. This is most easily accomplished with a product demodulator. If the SSB signal has random noise added to it, this noise will also be demodulated in the same way as the signal, i.e., a frequency translation.

For example, consider a 0-dBm SSB signal to be demodulated. To avoid demodulator constants, assume that the demodulated baseband signal is also 0 dBm (a condition easily achieved with suitable amplifiers and attenuators). Now consider adding at the input of the demodulator white noise with a power density of −100 dBm/kHz extending from f_a to f_b, as shown in Fig. 8-11(a). This assumes that the demodulator is preceded by a filter passing this band of frequencies. In general, this passband may be wider than that occupied by the signal. Upon demodulation by a product demodulator with carrier inserted at f_c, the noise spectrum will be translated by f_c to the baseband frequencies as shown in Fig. 8-11(b). The total noise power before demodulation also appears after demodulation with possible foldover at zero frequency. The "folded" noise power spectrums add on a power basis because the random noise is uncorrelated

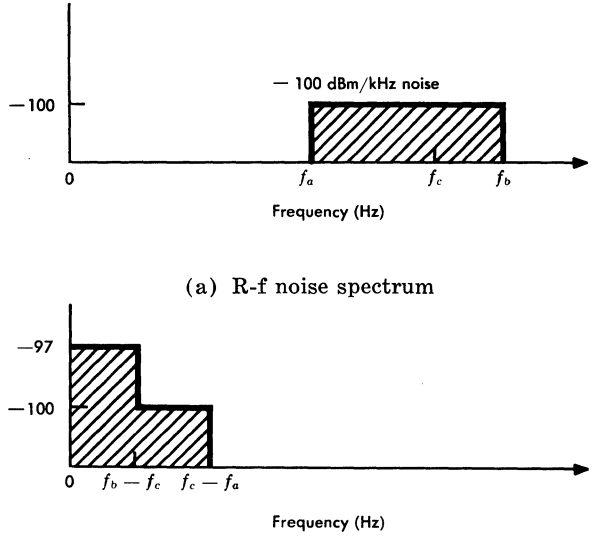


FIG. 8-11. Product demodulation of r-f noise.

between different frequencies. The result of all of this is the common statement that the signal-to-noise ratio of an SSB signal at r-f is not changed by the demodulation process. Because of the foldover effect, however, flat noise at r-f need not be flat under demodulation. For the best signal-to-noise ratio, of course, predetection filtering should remove all noise at frequencies not occupied by the modulation signal. Such filtering of an SSB signal removes any possible foldover of the noise in the demodulation process. In this case, the shape of the r-f power spectrum is maintained (with possible inversion) in the demodulation process.

The presence of quadrature distortion will not affect the noise in the demodulated output. Although quadrature distortion does affect the waveshape of the demodulated signal, it does not change the average signal power; therefore, it has no effect on the signal-to-noise ratio of a demodulated SSB signal.

DSBSC Modulated Wave

The noise performance of a product demodulator is the same for DSBSC as it is for SSB, except in the DSBSC case the minimum

bandwidth before demodulation is twice as much as for SSB. For flat noise at the input, this means that the noise output of the minimum bandwidth DSBSC demodulator is 3 dB greater than that from the minimum bandwidth SSB demodulator.

The demodulated signal consists of the two sidebands folded over each other around zero frequency. If the inserted carrier has no frequency or phase error, the components in each sideband are perfectly correlated so that the folded signal spectra add on a voltage basis. This means that the demodulated baseband voltage has doubled and that the power is 6 dB greater than that of either sideband. For example, a 0-dBm DSBSC signal (−3 dBm per sideband) would be demodulated by the previously discussed product demodulator to a +3-dBm baseband signal. If the input filter passes only those frequencies within ±4 kHz of the carrier frequency, the white noise with density of −100 dBm/kHz would be demodulated to a density of −97 dBm/kHz or a total noise of −91 dBm. Thus, the signal-to-noise ratio at r-f is $0 + 91 = 91$ dB, while the signal-to-noise ratio at baseband is $3 + 91 = 94$ dB. This 3-dB improvement in signal-to-noise ratio is characteristic of DSBSC demodulators with accurately inserted carriers.

If there is a phase error in the inserted carrier, the quadrature components in the sidebands cancel and only the in-phase components of the two sidebands add on a voltage basis. As a result, the demodulated baseband signal is reduced in amplitude by the factor $\cos \theta$, where θ is the phase error, and the signal-to-noise improvement is degraded.

DSBTC Modulated Wave

The demodulation of a conventional DSBTC wave is accomplished in the same manner as demodulation of a DSBSC wave except that a d-c baseband term results from the demodulation of the carrier. The chief advantage of transmitting the carrier is that it need not be added at the demodulator and results in no phase error (unless there is phase distortion in the r-f channel). The carrier is very wasteful of power, however. For example, a 0-dBm carrier can only support a −6-dBm sinusoid in each sideband. Demodulating with the previously discussed demodulator would result in a 0-dBm baseband signal.

Comparison of Amplitude Modulation Methods [8]

A comparison of the three amplitude modulation methods (SSB, DSBSC, and DSBTC) with respect to signal-to-noise ratio after de-

modulation depends upon the nature of the power limitation in the particular case. If average power density is limiting, as it tends to be in a multichannel situation, SSB is at a 3-dB disadvantage compared with DSBSC. Of course only half the number of channels can be carried by DSBSC in a given band. Adding the second sideband does not change the power density, but provides a 3-dB noise advantage due to coherency of signal sidebands as contrasted with noise sidebands.

A comparison of SSB with DSBTC depends on percentage modulation of the transmitted carrier. In single channel radio practice, an advantage of 9 dB is often quoted for SSB. This results from a peak-power comparison with DSBTC assuming 100 per cent tone modulation of the carrier. If the peak carrier power is 1 watt, the power on modulation peaks will then be 4 watts (peak envelope power), but the average power in each sideband will be only 1/4 watt. A single-sideband signal of 1/4 watt can be increased 16 to 1, or 12 dB, for the same 4-watt peak envelope power. However, due to the coherence of the two sidebands in the DSB wave, the net signal-to-noise advantage of SSB is reduced to 9 dB.

8.3 NOISE AND ANGLE-MODULATED SIGNALS

As noted previously, an unmodulated carrier and random noise applied to an ideal phase detector will produce a random noise output [9]. Since FM is the derivative of PM, an FM demodulator will also have a random noise output. The noise voltage at the output of the PM system is flat with frequency, whereas the noise voltage at the output of an FM system increases linearly with frequency. This is commonly referred to as the triangular noise spectrum of an FM system.

In the following discussion the carrier to which the interfering noise is added is assumed to be unmodulated. The results, however, are applicable to a low-index angle-modulated wave since most of the power is then in the carrier component. The effects of a band of random noise about the carrier will be characterized by (1) the total noise power appearing in the demodulated baseband signal, and (2) the spectral density of the demodulated noise at baseband.

It is shown in Chap. 20 that the mean square phase deviation, D_ϕ, due to adding p_n watts of noise power is given by

$$D_\phi = \frac{p_n}{2p_c} \qquad \text{rad}^2 \tag{8-55}$$

where p_c is the unmodulated carrier power. If, as is common, the noise power density at r-f is given by p_n watts/Hz, Eq. (8-55) gives the mean square phase deviation density in rad^2/Hz. Such a phase deviation spectrum has the same shape as that for the r-f noise spectrum shifted down in frequency by f_c. This results in a two-sided phase deviation spectrum (negative frequencies) which can be converted to the more practical one-sided spectrum by simply adding the mean square phase deviation spectrum for negative frequencies to that for positive frequencies. This corresponds to power addition. The actual noise voltage or power after demodulation requires knowledge of the type of demodulator (phase or frequency) and the necessary demodulator constants.

PM System Noise

The preceding analysis can now be applied to the specific problem of noise in the baseband of a PM system. It will be assumed that the random noise is flat versus frequency over the band from $f_c - f_1$ to $f_c + f_1$ with a power density of p_n watts/Hz as shown in Fig. 8-12(a), and that the carrier power is much greater than the noise power. To relate phase deviations to voltage, it is necessary to define a phase demodulator transfer constant. Since the constant, k, was used in Chap. 5 to relate baseband voltage to phase deviations, it is appropriate to use $1/k$ for the demodulator constant. Thus, a phase deviation of $1/k$ radian results in a 1-volt baseband output voltage. The *two-sided* baseband noise spectral density in volts^2/Hz can be derived from Eq. (8-55) as

$$\text{Noise spectral density} = \frac{1}{k^2}\frac{p_n}{2p_c} \quad \text{volts}^2/\text{Hz} \qquad (8\text{-}56)$$

On the more convenient one-sided basis obtained by folding negative frequencies over to positive frequencies, Eq. (8-56) becomes

$$\text{Baseband noise density} = \frac{p_n}{k^2 p_c} \quad \text{volts}^2/\text{Hz} \qquad (8\text{-}57)$$

This is illustrated in Fig. 8-12(b) showing the resulting flat spectrum. If telephone multiplex were transmitted over such a system, all channels would be equally noisy.

The total baseband noise is obtained by integrating Eq. (8-57) from 0 to f_1 Hz to obtain

$$\text{Total baseband noise power} = \frac{f_1 p_n}{k^2 p_c} \quad \text{volts}^2 \qquad (8\text{-}58)$$

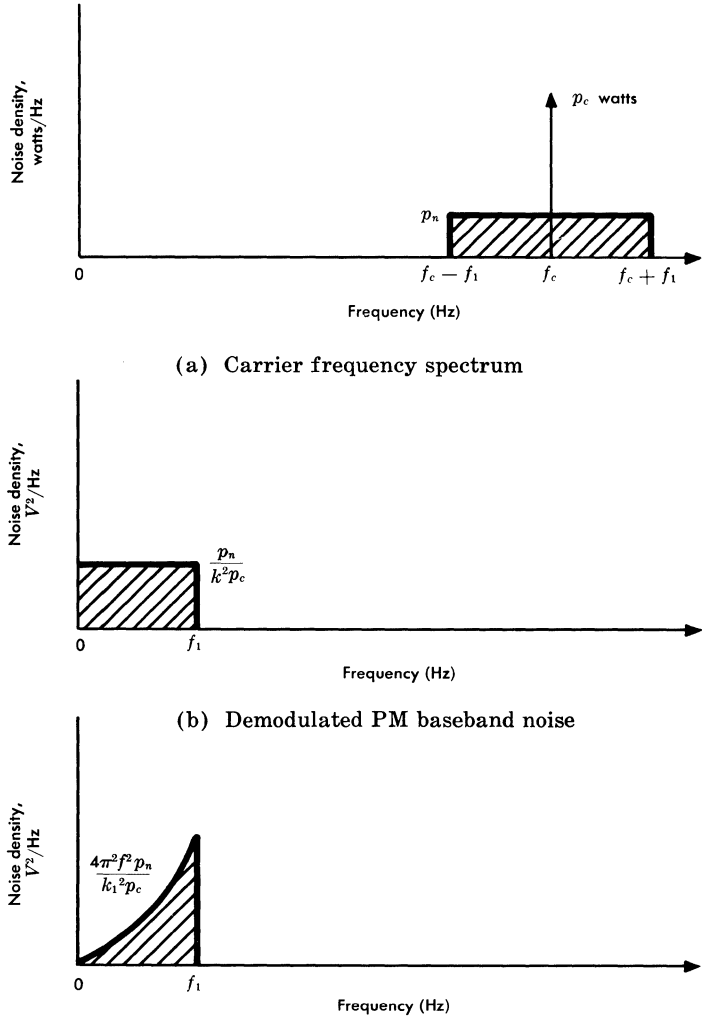


FIG. 8-12. Addition of a flat band of noise to a carrier and resultant baseband noise in PM and FM systems.

FM System Noise

As in the case of the PM system noise, it will be assumed that the random noise at r-f is flat with frequency over the band from $f_c - f_1$ to $f_c + f_1$ with a power density of p_n watts/Hz, as shown in Fig. 8-12(a), and that the carrier power is much greater than the

total noise power. The baseband voltage in an FM system is proportional to the frequency deviation of the carrier with the transfer constant given by $1/k_1$ volts/rad/sec.

The instantaneous frequency deviation caused by the noise can be obtained by differentiating the instantaneous phase deviation. Differentiation in the time domain corresponds to multiplication by ω in the frequency domain. The *two-sided* baseband noise spectral density can be derived from Eq. (8-55) as

$$\text{Noise spectral density} = \frac{\omega^2}{k_1^2}\frac{p_n}{2p_c} \qquad \text{volts}^2/\text{Hz} \qquad (8\text{-}59)$$

On a one-sided basis, this becomes

$$\text{Baseband noise density} = \frac{4\pi^2 f^2 p_n}{k_1^2 p_c} \qquad \text{volts}^2/\text{Hz} \qquad (8\text{-}60)$$

This is shown in Fig. 8-12(c) where the parabolic FM noise spectral density corresponds to the triangular *voltage* spectrum.

The total mean square noise voltage at the demodulator output can be obtained by integrating Eq. (8-60) from 0 to f_1 Hz to obtain

$$\text{Total baseband noise power} = \frac{4\pi^2 p_n f_1^3}{3k_1^2 p_c} \qquad \text{volts}^2 \qquad (8\text{-}61)$$

Comparison of FM and PM System Noise

The preceding analysis has shown that the random noise in a telephone channel at the output of an FM system is dependent on the frequency of the channel. Thus, if the top channel just meets noise requirements, the lower frequency channels are unnecessarily quiet. This is not very efficient.

In a PM system on the other hand, the noise is the same in all the telephone channels, since the phase modulation due to the signal and that due to the random noise are both flat with frequency. Here, the noise in a PM system is compared with the noise in the top channel of an FM system, under the condition of equal total baseband noise in each of the systems.

The total noise power in a PM system is given by Eq. (8-58) while that in an FM system is given by Eq. (8-61). Setting these two equal and solving for k^2 yields:

$$k^2 = \frac{3k_1^2}{4\pi^2 f_1^2} \qquad (8\text{-}62)$$

Substituting Eq. (8-62) into Eq. (8-57) gives the PM noise spectral density as

$$\text{PM noise spectral density} = \frac{4\pi^2 f_1^2 p_n}{3k_1^2 p_c} \quad \text{volts}^2/\text{Hz} \tag{8-63}$$

At the top end of the FM baseband, the noise density is given by substituting $f = f_1$ in Eq. (8-60):

$$\text{FM noise density at } f_1 = \frac{4\pi^2 f_1^2 p_n}{k_1^2 p_c} \tag{8-64}$$

The ratio of the top-of-band noise density for FM [Eq. (8-64)] to that for PM [Eq. (8-63)] is 3. Thus the top channel noise in an FM system is three times as high, or 4.8 dB greater, than the noise in each channel of a PM system.

FM Advantage with Respect to AM

It is possible to get better signal-to-noise performance in a frequency modulation system than in an AM system with the same transmitted power. To achieve this advantage, however, it is necessary to use large indices of modulation. Higher order sidebands become important, and a wider bandwidth is required than would be necessary for the corresponding AM system. The improvement in signal-to-noise performance which is obtained by using wider bandwidths is sometimes referred to as the FM advantage. To examine this quantitatively, a single-sideband AM system with suppressed carrier will be compared with an FM system.

Consider an SSB wave with peak power p_x watts at the low level point at the input to an amplifier. The noise at this point is assumed to be p_n watts/Hz and the signal has a bandwidth of f_1 hertz. The peak signal-to-average-noise density ratio of such a system is given by:

$$\left.\frac{\text{S}}{\text{N}}\right|_{\text{SSB}} = \frac{p_x}{p_n} \tag{8-65}$$

For comparison, consider an FM system with carrier power of $p_x/2$ watts (corresponding to a peak of p_x watts), a noise of p_n watts/Hz, and a peak frequency deviation of ΔF hertz. The peak demodulated signal voltage of such a system is given by:

$$\text{Peak signal voltage} = \frac{2\pi\Delta F}{k_1} \tag{8-66}$$

The top channel noise spectral density from the FM system is given by Eq. (8-64) as

$$\text{Noise spectral density} = \frac{8\pi^2 f_1^2 p_n}{k_1^2 p_x} \tag{8-67}$$

Squaring Eq. (8-66) and dividing by Eq. (8-67) gives the FM peak signal-to-average-noise density ratio:

$$\left.\frac{S}{N}\right|_{FM} = \frac{\Delta F^2 p_x}{2 f_1^2 p_n} \tag{8-68}$$

Comparing Eq. (8-68) with Eq. (8-65) shows that the noise performance of an FM system is $\Delta F^2/2f_1^2$ times better than an AM system. Thus,

$$\text{FM advantage} = 20 \log \frac{\Delta F}{f_1\sqrt{2}} \quad \text{dB} \tag{8-69}$$

Unless the peak frequency deviation is equal to or greater than $\sqrt{2}$ times the frequency of the top telephone channel, the FM advantage is negative. For an FM system where the peak frequency deviation is equal to the frequency of the top transmitted channel, the FM advantage is -3 dB, and the noise in the top channel will be 3 dB higher than in an SSB system. With pre-emphasis, the FM advantage can be raised to about 0 dB when the peak frequency deviation is equal to the frequency of the top channel.

By Carson's rule mentioned in Chap. 5, the bandwidth required to transmit an FM or PM signal is twice the sum of the peak frequency deviation and the top baseband frequency. When these are equal, the required bandwidth is therefore four times the top frequency. Since an SSB signal can be transmitted in a bandwidth equal to the top frequency, the pre-emphasized FM system requires at least four times the bandwidth of the AM system to achieve the same noise performance.

The advantage of SSB from the point of view of bandwidth utilization would appear to favor this form of modulation in microwave radio systems. As discussed in Chap. 10, the practical problems associated with obtaining linear amplifiers which have adequate gain and power output at microwave frequencies make the use of AM impractical at the present time. In applications where carrier power is limited, such as in space communications, the use of high-index FM allows enhancement of the signal-to-noise ratio at the expense of

bandwidth. In general, doubling the frequency deviation (or index) will improve the noise performance by 6 dB until breaking occurs (discussed in Chap. 20).

8.4 NOISE AND PCM SIGNALS

The deliberate quantization error or noise imparted to the PCM signal at the transmitting terminal has been discussed in Chap. 5. The error or noise produced by quantization is the major source of signal impairment and originates only at the transmitting (or coding) end of the system. It is shown in a later chapter that for a given bandwidth this type of noise can be minimized by nonuniform quantization, which results in effective amplitude compression of the baseband signal.

The other type of noise that characterizes a PCM system originates primarily at the receiving (or low level) points of the system. This is a false pulse noise which is incorrect interpretation of the intended amplitude of a pulse by the receiver; it is caused by noise spikes breaking through the receiver threshold or trigger level. As the signal power is increased above threshold, this noise decreases so rapidly that in any practical system it can be made negligible by design [10]. Aside from causing errors occasionally, random or fluctuation noise will have no other effect on the output signal. This is in contrast with the analog modulation systems previously considered, in which the noise affected the output signal continuously.

To detect the presence or absence of a pulse requires a certain minimum signal-to-noise ratio on the digital line. If the pulse power is too low compared to the noise, even the best possible detector will make mistakes and indicate an occasional pulse when there is none, or vice versa. Assume that the received pulses are of such a form that the pulse amplitude is nominally V_p with a pulse present, and zero with a pulse absent. This does not require the pulses to be flat-topped, but rather that they be zero at the sample times of adjacent pulses. A $(\sin x)/x$ shape meets these requirements as do many other possible shapes. If the signal plus noise exceeds $V_p/2$ at the sampling instant, it is interpreted as a pulse; if the signal plus noise is less than $V_p/2$, it is not a pulse. The result will be an error if the noise at the sampling instant exceeds $V_p/2$ in the proper direction.

If the noise has a gaussian distribution with an rms value of σ and the on-off binary pulses of amplitude V_p are used, it can be shown that

the probability of error is given by

$$p(\text{error}) = \frac{1}{\sqrt{2\pi}\,\sigma} \int_{\frac{1}{2}V_p}^{\infty} e^{-v^2/2\sigma^2}\, dv \tag{8-70}$$

As the signal voltage is increased, this function decreases very rapidly, so that if V_p/σ is large enough to make the signal intelligible at all, only a small increase will make the transmission practically error free. How rapidly this improvement occurs may be seen in Fig. 8-13.

Clearly, there is a distinct signal-to-noise threshold at about 20 dB; below this the interference is serious, and above it the interference is negligible. Comparing this with the ratio of approximately 60 dB required for high quality AM transmission of speech shows that PCM requires much less signal power at the expense of a much greater bandwidth.

The discussion of error probability and signal-to-noise ratio has been based on the transmission of binary (base 2) pulses. If pulses are used which have n different amplitude levels (i.e., a base n system), then a certain amplitude separation must exist between the adjacent levels to provide adequate noise margin. A greater total signal power is then needed to obtain the same (small) error probability if repeater spacing is the same for binary and n-base systems.

In the simple binary code, each digit is equally liable to error because of noise on the transmission path; however, the amounts of noise which these errors cause at the output of the decoder are unequal. Thus, an error in the most significant digit results in an error in the amplitude of the output signal much greater than that caused by an error in the least significant digit. The overall signal-to-noise ratio of the system could be improved by making the most significant digits less liable to error than those of less significance, so that they contribute equally to the output noise.

Several methods have been proposed for making the binary digits of a code group contribute equally to the output noise. For example, this can be accomplished by varying the pulse height or duration from digit to digit. It can be shown that by these methods an improvement in output signal-to-noise ratio is obtained over a wide range of input carrier-to-noise ratios. However, the proposed systems are much more complicated than a simple binary system, and in view

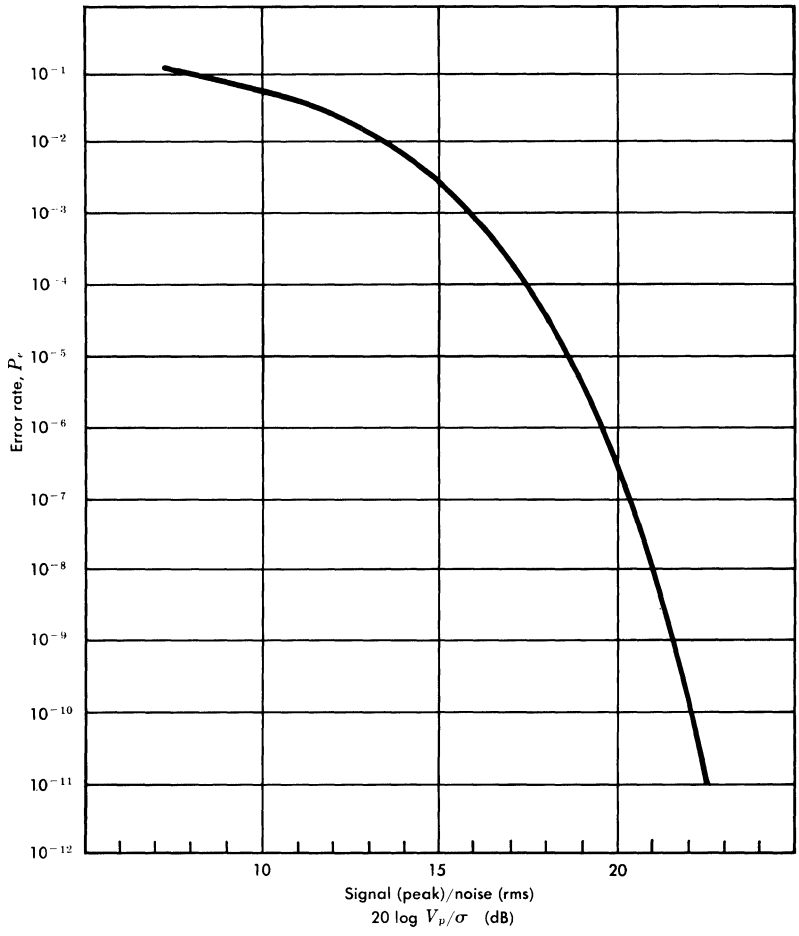


FIG. 8-13. Probability of pulse error versus peak signal to rms noise voltage ratio.

of the already excellent noise performance of a simple binary PCM system, the additional complexity is too high a price to pay for the gain in the signal-to-noise improvement.

It is of interest to compare the performance of a repeatered analog system with that of a PCM system. For example, if n analog repeater sections are to be used in tandem, the noise power added per link can only be $1/n$ of that which would be permissible in a single link. This is shown in the upper curve of Fig. 8-14 where it can be seen

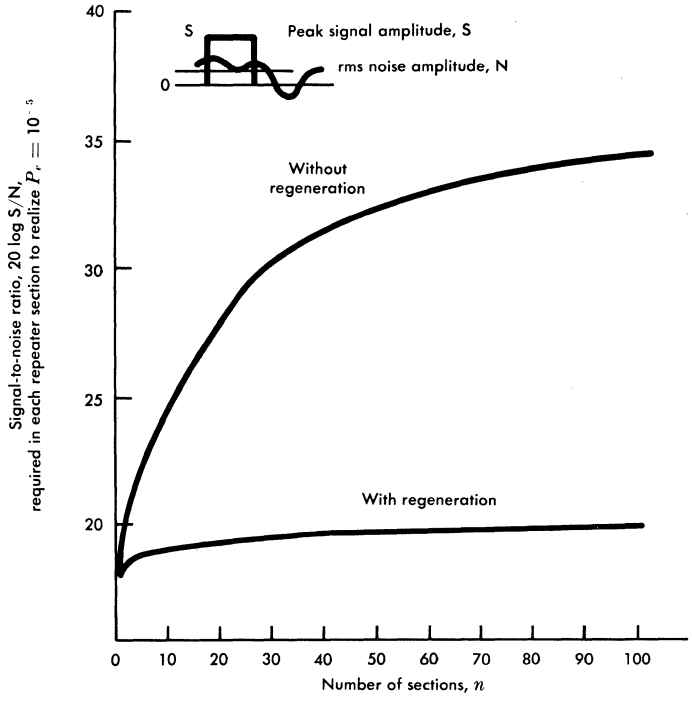


FIG. 8-14. Noise advantage of pulse regeneration.

that doubling the number of repeater sections requires a 3-dB increase in the per link S/N ratio. A binary PCM system, on the other hand, reduces the problem to recognition of the presence or absence of a pulse, with a low probability of error. If each link has an error probability, p, the error for n tandem links is given by

$$p_e = 1/2\ [1 - (1 - 2p)^n] \approx np \tag{8-71}$$

Thus, an n link PCM system requires the error probability per link to be $1/n$ times that of the whole system. From Fig. 8-13, it is apparent that a small S/N ratio improvement of a dB can result in an order of magnitude improvement in the error probability. The result is the nearly horizontal lower curve of Fig. 8-14 for regenerative repeaters. The advantage of regeneration indicated by Fig. 8-14 is obtained at the cost of a large increase in bandwidth over that required for analog systems.

It should be emphasized that the results shown in Figs. 8-13 and 8-14 were derived with ideal repeaters assumed. For several reasons, practical regenerative repeaters may require an increase in signal-to-noise ratio over the theoretical minimum. First, intersymbol interference between transmitted pulses reduces the amount of noise that can be tolerated in a repeater section. This interference is described as a noise impairment. Second, errors in pulse positions, called timing errors, can accumulate over a regenerative-repeatered medium. Third, imperfections and noise in repeaters can produce variations in pulse width and height which can also be translated directly into a noise impairment. Finally, external interference (crosstalk from other systems in the same environment), impulse noise (generated at central offices and picked up on a cable), and terminal noise require a considerable portion of the margin against error.

REFERENCES

1. Davenport, W. B., Jr. and W. L. Root. *Introduction to Random Signals and Noise* (New York: McGraw-Hill Book Company, Inc., 1958), Chapter 9.
2. American Standards Association. *Definition of Electrical Terms,* ASA-C42.65 (1957).
3. Van der Ziel, A. "The Theory of Shot Noise in Junction Diodes and Junction Transistors," *Proc. IRE,* vol. 43 (Nov. 1955), pp. 1639-1646.
4. Nielsen, E. G. "Behavior of Noise Figure in Junction Transistors," *Proc. IRE,* vol. 45 (July 1957), pp. 957-963.
5. Ko, H. C. "Temperature Concepts in Modern Radio," *The Microwave J.* (June 1961), pp. 60-65.
6. Haus, H. A., et al. "IRE Standards on Methods of Measuring Noise in Linear Twoports, 1959," *Proc. IRE,* vol. 48 (Jan. 1960), pp. 60-68.
7. Carlson, A. Bruce. *Communication Systems* (New York: McGraw-Hill Book Company, Inc., 1968), pp. 200-208.
8. Black, H. S. *Modulation Theory* (Princeton, N. J.: D. Van Nostrand Company, Inc., 1953), pp. 137-138.
9. Bennett, W. R. *Electrical Noise* (New York: McGraw-Hill Book Company, Inc., 1960), pp. 242-250.
10. Panter, P. F. *Modulation, Noise, and Spectral Analysis* (New York: McGraw-Hill Book Company, Inc., 1965), Chapter 21.

Chapter 9

Multichannel System Load

In this chapter the nature of the telephone speech load applied to multichannel telephone transmission systems is discussed. The characteristics of an FDM group, supergroup, mastergroup, or other ensemble of voice messages are discussed and related to the load carrying capacity of a multichannel system. This relationship is significant because the instantaneous system load fluctuates due to variations of signal amplitudes within the channel and variations in the number of active channels within the system. The message channel speech signal amplitude varies:

1. At an audio rate.
2. At a syllabic rate.
3. With talker volume.
4. With differences in loop and trunk losses in the circuits feeding the systems.

Ideally, the designer should have at his command all the field data on these considerations, measured in terms which will make it most usable for his purposes. Actually, gathering data on telephone speech load and subscriber reactions is a long and expensive process; therefore, whatever data is available must be used. This involves extrapolation, conversion from one type of measurement to another, and other sources of uncertainty.

9.1 SPEECH VOLUME CHARACTERISTICS

In any discussion of load carrying capacity, the first question must be: what is the magnitude of the signals applied to the system? More specifically, when overload is considered, the question takes the form: given a system equipped to carry N channels, what will

be the voltage-time functions at some chosen point in the system during the busy hour? A convenient point to choose is a zero transmission level point where the audio volume measurements are commonly referenced. From information about signal amplitudes at this point, it is possible to derive required load carrying capacities, required peak-frequency deviations, required voltage ranges of quantizers, etc.

The long-range average load objective per message channel has been established domestically in the Bell System at —16 dBm0. One of the purposes of this chapter is to illustrate the process used to establish this average. It is expected that future developments in telephone station sets and other station equipment will be compatible with this objective.

From an overload viewpoint, it is necessary to have knowledge of the peak power along with the average power value. In essence, the syllabic nature of speech results in much higher powers than —16 dBm0, provided that the duration of such power peak intervals is not too long. It is also the purpose of this chapter to explore the nature of such peaks.

Finally, it is important from an interference standpoint, that the average power spectrum of any message channel signal resemble that of the telephone speech signal. This limits the amount of single-frequency energy allowed. In summary, long-range objectives are based on message channel loads meeting the following characteristics:

1. A long-term average power of —16 dBm0 per channel.
2. A peak-to-rms ratio comparable to that for a telephone speech signal.
3. A frequency spectrum comparable to that for a telephone speech signal.

Since these load characteristics are closely associated with telephone speech signals, a detailed discussion of such signals is warranted.

Constant Volume Talkers

Although usually unrealistic, it is instructive to consider first the characteristics of talkers at a constant volume. A constant volume talker is one producing a speech signal fluctuating in amplitude at a syllabic rate but yielding a volume reading on a vu meter that is constant with time. Such control of volume is usually only found under specially controlled conditions. To illustrate the processes involved, it is instructive to consider the characteristics of a single speech channel.

Single Talker. It will be initially assumed that a single channel is carrying continuous speech as it would, for example, for a radio announcer giving a live description of a boxing match. If a vu meter were placed on the line (at 0 TLP), it would indicate a unique volume, V_{0c}, for this talker. If a very slow reading (heavily damped) power meter were placed in this line, it would indicate an average power being delivered to the load. It has been found empirically that for typical talkers of V_{0c}, the average power in dBm is 1.4 dB less than the vu reading, or

$$P_{0c} \approx V_{0c} - 1.4 \qquad \text{dBm0} \tag{9-1}$$

Thus, by definition, the power of a 0-vu talker is —1.4 dBm. The power P_{0c} is often referred to as the *long-term* average power of the talker. Long-term in this case means over a 10-second or longer sample of continuous speech, and the averaging process includes the time interval occupied by natural pauses such as interword and intersyllable, but not long pauses typically associated with marshalling thoughts or waiting for a reply.

The peak instantaneous power of this V_{0c}-vu talker could also be observed by use of a peak reading voltmeter. It has been observed that for a typical talker the peak instantaneous power is about 18.6 dB higher than the average power given by Eq. (9-1). This high peak factor means that a transmission system capable of handling this signal must be operated at a relatively low average power to avoid overload and distortion on peaks.

The speech waveform is often characterized by an audio-frequency carrier having a low-frequency envelope. Because the speech waveform is near zero for a portion of the time, it is also described as a series of talkspurts separated by gaps. The proportion of time occupied by the talkspurts is called the *activity*, τ. More precisely, activity is defined as the proportion of time that the rectified speech envelope exceeds some threshold. Although activity is threshold level dependent, it has been found by experiment that such dependency is relatively weak if the threshold is about 20 dB below the average power. Under such a condition, the activity, τ_c, of a continuous talker is found to be about 65 to 75 per cent. Since this corresponds to a talker whose average power is V_{0c}—1.4 dBm, the power of only the talkspurts in dBm is approximately equal to V_{0c} in vu. Thus, an alternative way of converting volumes in vu to average power in

dBm would be through the use of the full-time talker activity factor, τ_c, where τ_c excludes all gaps and pauses:

$$P_{0c} = V_{0c} + 10 \log \tau_c \quad \text{dBm0} \tag{9-2}$$

For $\tau_c = 0.725$, this equation gives the same result as Eq. (9-1).

Multiple Talkers. Next consider applying a second talker also of V_{0c} vu to the above transmission system. Up to this point, no restrictions have been made regarding the frequencies of the signal. This means that the average and peak power of the single talker could be measured after frequency translation without changing the power relationships. The addition of a second talker signal is only practical if it can be translated to a different segment of the multiplex band from the first talker signal. In such a case, the average power of the two talkers (assuming no correlation between them) is given by the sum of the average power of each. If both talkers have the same average power, the total average power is 3 dB higher than either. In general, N talker signals each at V_{0c} vu will have a total average power given by

$$P_{av} = V_{0c} - 1.4 + 10 \log N \quad \text{dBm0} \tag{9-3}$$

Note that P_{av} is the power in the frequency band occupied by all the talkers. Each talker may occupy a different portion of the frequency spectrum.

The peak instantaneous voltage of one talker could coincide with the peak instantaneous voltage of a second talker with these two voltages adding to increase the peak instantaneous power of the sum by 6 dB. The probability of the speech envelope attaining its highest peak at any given time is very low, and the probability of many uncorrelated speech patterns simultaneously reaching their peak values is much lower. The peak factor to consider for design purposes depends upon the risk or probability that the peak will be exceeded. The problem is similar to defining a peak factor for random noise.

Tests conducted by Holbrook and Dixon in 1938 showed that speech peaks could exceed the overload point of a vacuum tube amplifier about 0.1 per cent of the time without degrading the system performance [1]. To allow for possible different amplifier characteristics, it has become traditional to make the overload criterion the level exceeded 0.001 per cent of the time. With this criterion, Holbrook and Dixon found by experiment that the peak factor for two

equal level talkers is about the same as that for a single talker. The peak factor for many talkers tends to decrease as the number of talkers increases, approaching the characteristics of random noise for N greater than about 64. This is shown in Fig. 9-1 and corresponds to peaks being exceeded 0.001 per cent of the time. This peak factor has traditionally been called Δ_{c2}. It is standard practice to denote a system as overloaded when the overload point of the system is exceeded by peaks of a complex signal more than 0.001 per cent of the time. (It is *not* said to be overloaded 0.001 per cent of the time.)

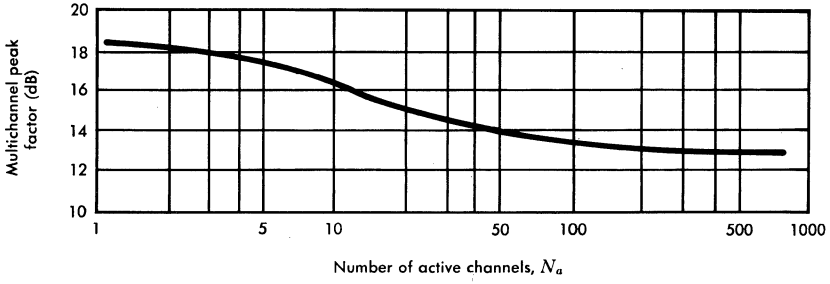


FIG. 9-1. Peak factor exceeded 0.001% for speech channels.

Up to this point, it has been assumed that a busy channel is being used by a continuous talker. However, this is not typical of a telephone conversation where, on the average, each party spends about equal time talking and listening. This talk-listen effect reduces the telephone speech activity by the factor $\tau_s \approx 0.5$. Thus, the talker power in each direction of a busy telephone circuit can be expected to be about half of what it is for a continuous talker.

The fact that all channels on a telephone system cannot be kept busy during the busy hour lowers the average power per circuit and is reflected in the *trunk efficiency factor*, τ_e. For example, an N channel system can never have more than N busy circuits because any additional callers would be turned away by a busy signal. To keep N circuits busy on such a system, it would be necessary to have a new call initiated at the same time that a busy call is terminated. Since call arrivals and departures are independent random variables, it is impossible to maintain an average of N busy circuits on such a system. In fact, to avoid excessive caller rejections, it is necessary to have considerably higher trunk capacity than the average busy hour load. In addition, each new call has a period of low activity at the beginning of the call while the connection is being set up.

These effects are characterized by a trunk efficiency factor, τ_e, which for domestic circuits is usually assumed to be about 0.70. For overseas circuits, where the channels are considerably more expensive and operators can hold some calls until a channel becomes available, the trunk efficiency factor is somewhat higher at around 0.85 or 0.90. There are other less important factors that may be considered. For example, in a normal telephone conversation there will be periods when neither party is talking while waiting for a reply or thinking of what to say. Such double pauses serve to lower the average power. On the other hand, there will also be some double talking (both parties talking simultaneously) which would tend to reduce the double pause effect. Since no accurate data is available on these two counteracting effects, they are often ignored. All of these effects can be combined into a *telephone load activity factor*, τ_L, defined as the ratio of average power of a telephone talker to average power of a continuous talker. If only τ_s and τ_e are considered, the telephone load activity for domestic circuits is

$$\tau_L = \tau_s \tau_e = 0.50 \times 0.70 = 0.35 \tag{9-4}$$

By taking into account other activity considerations, a lower value of 0.25 can be justified for τ_L. This value has traditionally been used in the design of telephone systems and has been adopted by the CCITT [2]. Note that τ_L is distinctly different from the speech activity factor, τ, given by $\tau_c \tau_s \tau_e$. This is because of the continuous talker reference condition used. However, speech activity and telephone load activity have not always been distinguished in the past with the assumption that a continuous talker is 100 per cent active, i.e., $\tau_c = 1$.

The average power of a single talker with a volume of V_{0c} on a telephone system with telephone load activity, τ_L, is given by modifying Eq. (9-1) to

$$P_{0c} = V_{0c} - 1.4 + 10 \log \tau_L \qquad \text{dBm0} \tag{9-5}$$

Now consider a system of N talkers of volume V_{0c} and load activity τ_L. The short-term average power of the ensemble of N talkers is proportional to the number actually talking. Unless the average is taken over a very long time interval, the power will fluctuate because the number of active channels at a given instant may be anything from zero to N. The probability that the number of talkers is n is

binomially distributed and given by

$$P(n) = \frac{N!}{n!(N-n)!} \tau_L{}^n (1 - \tau_L)^{N-n} \tag{9-6}$$

It has become standard practice to design systems in which the number of talkers, N_a, is assumed fixed and equal to the number of talkers exceeded one per cent of the time during the busy hour with all N channels busy. For a given N and τ_L, N_a can be determined from the following:

$$P(n \geq N_a) = 0.01 = \sum_{n=Na}^{N} \frac{N!}{n!(N-n)!} \tau_L{}^n (1 - \tau_L)^{N-n} \tag{9-7}$$

If N is large and τ_L is not too small ($N\tau_L \geq 5$), the binomial distribution is approximately normal with $\mu = N\tau_L$ and $\sigma^2 = N\tau_L(1-\tau_L)$. The number of talkers exceeded one per cent of the busy hour is given by

$$N_a \approx N\tau_L + 2.33\sqrt{N\tau_L\,(1 - \tau_L)} \tag{9-8}$$

Values of N_a are shown in Fig. 9-2 for various values of N and τ_L.

N	$\tau_L = 0.25$			$\tau_L = 0.35$		
	N_a	$N\tau_L$	Δ_{c1}	N_a	$N\tau_L$	Δ_{c1}
6	4.67	1.5	4.9 dB	5.35	2.1	4.1 dB
12	7.24	3.0	3.8 dB	8.68	4.2	3.1 dB
24	11.73	6.0	2.9 dB	14.50	8.4	2.4 dB
36	15.67	9.0	2.4 dB	19.77	12.6	2.0 dB
48	19.36	12.0	2.1 dB	24.72	16.8	1.7 dB
96	33.65	24.0	1.5 dB	44.06	33.6	1.2 dB
300	90.78	75.0	0.8 dB	122.20	105.0	0.7 dB
600	171.80	150.0	0.6 dB	232.30	210.0	0.4 dB
2000	528.40	500.0	0.2 dB	729.40	700.0	0.2 dB

FIG. 9-2. Number of active channels and Δ_{c1}.

The average talker load to be used during the busy hour is that corresponding to N_a talking channels or, rewriting Eq. (9-3):

$$P_{av} = V_{0c} - 1.4 + 10 \log N_a \qquad \text{dBm0} \tag{9-9}$$

From Fig. 9-2, it can be seen that N_a approaches the value of $N\tau_L$ for large N. Defining Δ_{c1} by

$$\Delta_{c1} \triangleq 10 \log \frac{N_a}{N\tau_L} \tag{9-10}$$

allows Eq. (9-9) to be written

$$P_{av} = V_{0c} - 1.4 + 10 \log \tau_L + 10 \log N + \Delta_{c1} \text{ dBm0} \tag{9-11}$$

This has the advantage of expressing the total average power as the product of the average power per channel and the total number of channels modified by the factor Δ_{c1} to account for load fluctuations due to talker activity. For large systems, Δ_{c1} approaches 0 dB. In review, Δ_{c1} is the correction for systems of a few channels where the central limit theorem is not applicable.

Another nearly equivalent approach is to derive Δ_{c1} in terms of speech activity, $\tau = \tau_c\tau_s\tau_e$. In such a case, Eq. (9-11) could be written

$$P_{av} = V_{0c} + 10 \log \tau + 10 \log N + \Delta_{c1} \text{ dBm0} \tag{9-12}$$

The addition of the τ_c factor is compensated by the omission of the 1.4 vu-to-dB conversion factor. Comparing Δ_{c1} for $\tau_L = 0.25$ and $\tau_L = 0.35$ in Fig. 9-2 reveals that Eq. (9-12) is practically identical to Eq. (9-11).

Talkers of Distributed Volume

Up to this point, it has been assumed that all talkers on a system are at a constant volume, V_{0c}. This is usually not true unless control is exercised over the channels with equipment such as voice operated gain adjustment devices. The distribution of talker volumes observed at 0 TLP on a message channel is a function of many parameters. The more obvious ones are:

1. Sex of subscriber.
2. Speech habits of subscriber.
3. Type of telephone.
4. Loss of toll connecting trunks.

5. Loss of subscriber loop.
6. Geographic area.
7. Distance to receiving party.

Latest observations have shown that men talk louder than women, and business calls are louder than social calls. People in large cities talk louder than those in small communities, and people talk louder over long distances than they do for local calls. A factor of 1-dB increase in volume for each 1000 miles of distance has been substantiated in spite of low loss and normal quality in most such connections. It has also been found that the observed variation in volumes is due more to the subscriber than to the telephone plant facilities [3].

Measured volumes of large populations of telephone talkers appear to be normally distributed in vu because of the combined effects listed in the preceding discussion. In all that follows it is assumed that the talker volumes are normally distributed with a mean of V_0 vu and a standard deviation of σ dB. The power corresponding to a continuous talker of mean volume, V_0 vu, is given simply by

$$P_0 = V_0 - 1.4 \quad \text{dBm0} \tag{9-13}$$

Of more interest is the average power per talker, P_{0p}, which is not given by Eq. (9-13) unless all volumes are constant as they were in previous discussions. This is true because talkers louder than average contribute much more power to the total load than is saved by those talkers softer than average.

For example, consider talker volumes normally distributed with a mean $V_0 = 1.4$ vu and $\sigma = 3$ dB. Although not typical numbers, they make the example easy to visualize. After conversion to dBm, talker powers are still normally distributed with a mean of 0 dBm and a standard deviation of 3 dB. This normal density function is shown in Fig. 9-3(a) plotted with dBm as the abscissa. If the dBm is converted to milliwatts, the density function of Fig. 9-3(b) results with milliwatts as the abscissa. This new function is called the log normal density function. Note that it is skewed to the left, and its mean (given by the expected value) is greater than 1 mW. Also, the difference between the power corresponding to the average volume talker and the average power per talker will increase as σ increases; the wider the normal dBm density function, the more skewed the log normal density function becomes. It can be shown that the necessary correction in dB is given by $0.115\,\sigma^2$, where σ is in dB [4].

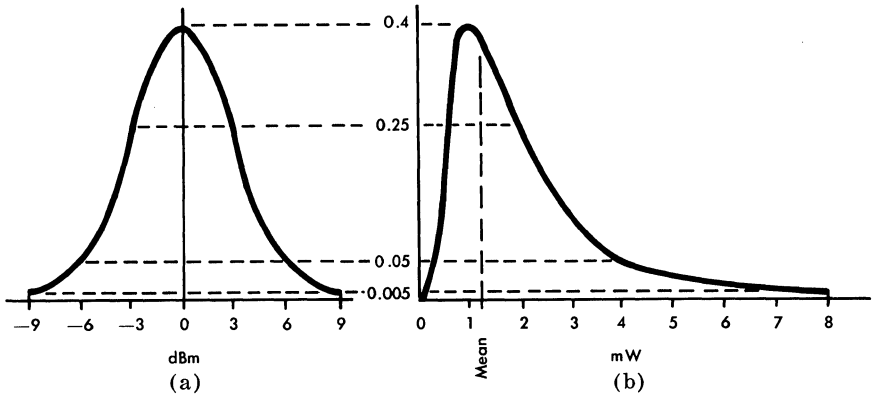


FIG. 9-3. The log normal density function from the normal distribution in dBm.

The average power per talker is then found to be 0.115 σ^2 dB higher than the power corresponding to the average volume talker. Thus, the average power per full-time talker for normally distributed talkers with mean V_0 vu and standard deviation σ dB is given by

$$P_{0p} = V_0 + 0.115\,\sigma^2 - 1.4 \qquad \text{dBm0} \tag{9-14}$$

For a combination of N telephone talkers with load activity τ_L, Eq. (9-11) must be modified to give an average power

$$P_{av} = V_0 + 0.115\,\sigma^2 - 1.4 + 10 \log \tau_L + 10 \log N + \Delta_{c1} \tag{9-15}$$

To be exact, the Δ_{c1} factor previously derived for constant volume talkers must be redefined to allow for talker volume variations. However, it is not radically different than previously defined. It can be shown that the proper value for Δ_{c1} with uncontrolled talker volumes is given by

$$\Delta_{c1} = 2.33\,\sigma_N - 0.115\,\sigma_N^2 \qquad \text{dB} \tag{9-16}$$

where σ_N is the standard deviation in dB of the approximately normal distribution of the total power in dBm and is given by

$$\sigma_N^2 = 43.43\,[\log\,(\tau_L N + b^2 - \tau_L) - \log \tau_L N] \tag{9-17}$$

where

$$10 \log b = 0.115\,\sigma^2 \tag{9-18}$$

The instantaneous peak factor for uncontrolled volumes is assumed the same as for controlled volumes and is given in Fig. 9-1 as Δ_{c2}. For all but very small systems, it is usually valid when using Fig. 9-1 to assume that $N_a = \tau_L N$ [1].

9.2 LOAD CAPACITY

The load capacity of a multichannel transmission system is the number of telephone channels the system can carry without undue distortion or noise. The system load will depend upon the talker volume, the distribution of talker volumes, and the activity of each talker. The load capacity of a system is determined by the maximum signal which can be impressed on the system without overload. This overload may be the result of the signal amplitude exceeding the dynamic range of an amplifier or repeater, of frequency deviations exceeding the bandwidth of an angle-modulated system, or of voltages exceeding the quantizing range of a digital quantizer.

Overload

Overload can be defined in many ways, depending upon the way in which the overload effect is observed when the system is subjected to an increasing signal amplitude. All criteria commonly used basically define the level at which the nearly linear performance of the system is no longer linear enough for satisfactory performance.

Although other definitions for overload may be used for specific systems, the general CCITT definition [2] is often applicable:

> *Overload point*—The overload point, or overload level, of a telephone transmission system is that average power in dBm0 of an applied sinusoid, at which the average power level of the third harmonic increases by 20 dB when the input signal is increased by 1 dB.

This definition is not applicable when the third harmonic falls outside the useful bandwidth of the system. The following alternate definition may then be used:

> *Overload point, alternate*—The overload point, or overload level, of a telephone transmission system is 6 dB higher than the average power in dBm0 of each of two applied sinusoids of equal amplitude and of frequencies α and β, when these input levels are so adjusted that an increase of 1 dB in both

their separate levels causes an increase, at the output, of 20 dB in the intermodulation product of frequency $2\alpha - \beta$.

The average power in dBm at 0 TLP of the single-frequency sinusoid causing overload is denoted by P_s. This represents the total maximum multichannel telephone load including pilots, carriers, and speech load. The peak instantaneous power of this sinusoid is $P_s + 3$.

It should be noted that most systems do not overload on average power but when instantaneous peaks exceed some threshold. As a consequence, a multichannel load with $P_{av} = P_s$ will severely overload a system because the peak factor for the multichannel load may be over 10 dB more than the 3-dB peak factor for a sinusoid. The load is obtained by setting the power of the multichannel load exceeded 0.001 per cent of the time equal to the peak power of the sinusoid ($P_s + 3$).

Multichannel Load Factor

Equation (9-15) gives the average load of a multichannel system which is exceeded only 1 per cent of the time at the busy hour. The maximum peak load is given by

$$\text{Maximum peak load} = P_{av} + \Delta_{c2} \qquad \text{dBm0} \tag{9-19}$$

where Δ_{c2} is the peak factor for $N_a \approx N\tau_L$ channels. This peak load can be equated to the peak sine wave power of $P_s + 3$ to yield

$$P_s = V_0 + 0.115\sigma^2 - 1.4 + 10 \log \tau_L + 10 \log N + \Delta_{c1} + \Delta_{c2} - 3 \qquad \text{dBm0} \tag{9-20}$$

For convenience, the last three terms of Eq. (9-20) have been combined and plotted as a function of N. The combination is known as Δ_c and is often referred to as the *multichannel load factor*. It is illustrated in Fig. 9-4 with curves given for three values of the standard deviation of the talker volume distribution. It has been assumed for the graph that the talker load activity, τ_L, is 0.25 since this is the activity commonly assumed. For activity or σ other than that shown in Fig. 9-4, it is often convenient to use the approximation given by

$$\Delta_c = 10.5 + \frac{40\sigma}{N\tau_L + 5\sqrt{2\sigma}} \qquad \text{dB} \tag{9-21}$$

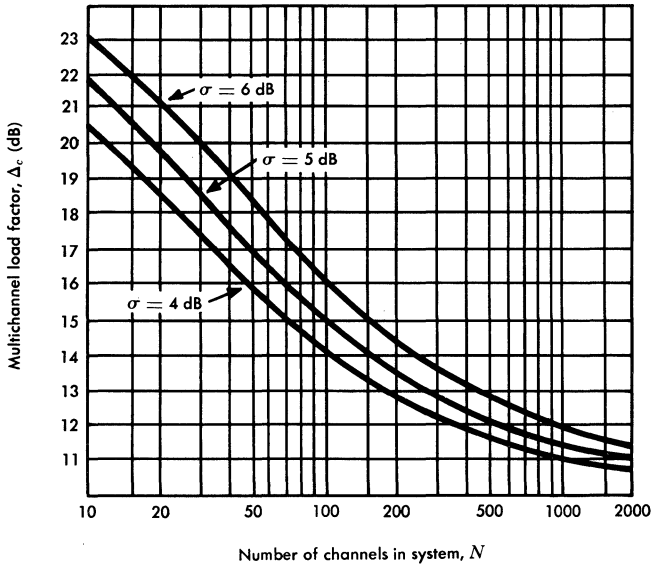


FIG. 9-4. The multichannel load factor, $\triangle_c$, for $\tau_L = 0.25$.

This empirical formula fits the results computed from theory within a fraction of a dB for a variety of parameters. In general, P_s can be simply expressed in terms of Δ_c by

$$P_s = V_0 + 0.115\sigma^2 - 1.4 + 10 \log \tau_L + 10 \log N + \Delta_c \quad \text{dBm0} \quad (9\text{-}22)$$

Load Simulation

It is often desirable in the design of a telephone transmission system to simulate the multichannel load of a band of FDM message channels. For a few channels, the simulation can consist of recordings of telephone talkers; however, for a simulation of several hundred channels, a different approach is needed. Fortunately, the characteristics of a large number of talkers approach that of gaussian noise. This is illustrated in Fig. 9-1 where the peak factor for many channels approaches the 13 dB expected for gaussian noise at the 0.001 per cent level. As a consequence, a band of FDM channels extending from f_B to f_T can be characterized by a bandlimited flat gaussian noise source over the same band with the same average power as the multiplexed load. Thus, the equivalent noise source should have an average power given by Eq. (9-15):

$$P_{av} = V_0 + 0.115\sigma^2 - 1.4 + 10 \log \tau_L + 10 \log N \quad \text{dBm0}$$

where Δ_{c1} has been ignored because hundreds of channels have been assumed. When the Bell System average load objective is used, this equation becomes.

$$P_{av} = -16 + 10 \log N \qquad \text{dBm0} \qquad (9\text{-}23)$$

A gaussian noise source with this average power is a good simulation when N is greater than several hundred channels. However, when N is less than this, the FDM baseband signal voltage amplitude is no longer gaussian distributed since the average number of active channels is less than about 64 and the average load per channel is increased by the Δ_{c1} factor. Hence, the noise source is not simulating the peaking characteristics of the signal. In these cases, judgment must be exercised in the load simulation. Increasing the average power of the noise source by Δ_{c1} would provide a better simulation of the average characteristic but would still be inadequate for the peaking characteristics. At this time, there is no hard and fast rule on how to simulate FDM signals with uniform-spectrum gaussian noise when N is less than a few hundred channels. The CCITT Recommendation G.223 for these cases appears to be a compromise between the two extremes of using the FDM average signal power and using a power which is artificially increased by Δ_c to reflect the higher multichannel load factors shown in Fig. 9-4 [2].

It should be noted that this simulation assumes that the spectrum of the multiplexed load is uniform across the frequency band. Of course, standard message channels have their energy shaped over a 3-kHz band but are spaced 4 kHz apart. The result is a spectrum which is only approximated by a uniform noise load. For characterizing the amplitude distribution of the load, such an approximation is sufficient for most engineering purposes. Any necessary allowance for such "bunching" effects is discussed in the next chapter.

9.3 TYPICAL DESIGN PARAMETERS

In spite of the extensive discussion on system load characterization, there remains a considerable range of possible solutions. This is partially due to the large number of parameters that must be considered and partially due to the uncertainties inherent in determining these parameters. Through the years, some of these parameters have been established in the Bell System using techniques that have proven successful in several different designs. It is of interest to review some of the more important of these.

Talker Volumes and Activity

The question of the best volume to use for average talker level, V_0, and load activity, τ_L, has been investigated by several people in the last 30 years [5]. The latest available study indicates that the distribution of all talker volumes on intertoll trunks in the DDD network has a mean of about −18 vu0 and a standard deviation of 6.5 dB [3]. Talker volumes on private networks tend to be somewhat higher. In the DDD network it also appears that the mean tends to increase about 1 dB per thousand miles of trunk length and the standard deviation may decrease to approximately 5 dB. For a transcontinental domestic system, the average talker volume and standard deviation will depend to some extent on the average trunk length and amount of private service. For a V_0 of −16 vu0 and a σ of 6 dB, the average power per channel (sometimes called rms load per channel) yields, for τ_L equal 0.35,

$$P_{op} = V_0 + 0.115\,\sigma^2 - 1.4 + 10 \log \tau_L = -18 \qquad \text{dBm0} \quad (9\text{-}24)$$

On the other hand, data, with its high activity, has a somewhat greater average power. The nominal data signal in the switched network is −13 dBm0. For full duplex operation (both directions transmitting simultaneously) a load activity factor of 0.70 can be assumed, resulting in an average channel load of −14.5 dBm0; for half duplex operation, the average channel load is −17.5 dBm0.

To accommodate a mix of both voice and data transmission, the Bell System has adopted a long-range objective of −16 dBm0 for the load per channel. This value provides for a reasonable mix of half duplex and full duplex data and allows some margin for supervisory tones and for increasing the efficiency of future telephone sets, particularly those on long loops.

Overseas message circuits generally have higher talker levels than domestic circuits. This is due to the greater distances spanned, the higher proportion of business calls, and a generally higher trunk efficiency factor. In addition, talker levels are higher in many foreign countries due to variations in plant and speech habits. An average volume of −12.5 vu with a σ of 5 dB has been used for overseas talkers. Assuming a 90 per cent trunk efficiency factor (making $\tau_L = 0.45$) results in an average power load of −14.5 dBm0 per channel.

The conventional value of average power per channel allowed by the CCITT is −15 dBm0. The CCITT also assumes a σ of 5.8 dB

and an activity factor of 0.25 from the Holbrook and Dixon results. Under these conditions, the —15 dBm0 per channel results in an average talker level of —11.5 vu which is 1 dB louder than previously assumed for overseas circuits. However, because of higher activity and the presence of data, the international load capacity of —15 dBm0 per channel may be inadequate in many applications, and some systems have been designed for higher loads.

Effects of Data and Tone Signals

Since modern message channel signals may be made up of data and supervisory tones instead of speech, the proper means of allowing for these in load capacity calculations can be very important. As previously mentioned, voiceband data transmission can usually be assumed to consist of a sinusoid or combination of sinusoids having a power of nominally —13 dBm0 when active. Similarly, domestic toll circuits usually have SF supervision such that idle toll circuits have placed on them a 2600-Hz tone at —20 dBm0.

The most common way of handling such signals is to add their powers to the average power load of the talkers and to use the same multichannel load factor as if all message channels had talkers. This method is generally applied to systems having large numbers of channels.

Another method that has been used is to break the system into two subsystems. All speech channels form one subsystem for which a P_s is determined. All sinusoidal sources form the second subsystem for which a P_s is determined with the peak factor analytically derived for the number of sinusoids in the subsystem. The value of P_s for the whole system is simply the power sum of the values of P_s found for each subsystem.

It is recognized that neither of these approaches is exact, and, as a consequence, conservative values of V_0 and τ_L are often assumed for additional protection. For example, there is evidence that the impulses introduced by hard clipping in a broadband system introduce high error rates in data signals long before they become audibly annoying in telephone circuits. More accurate load characterization in such systems must await further study on the effects of overload of modern transmission devices.

Effects of Shaped Levels

To this point, it has been assumed that the average signal power on all channels is identical. In many systems, a noise advantage can

be obtained if the multiplexed signal is frequency shaped. If this is the case, the signals at the overload point will not have identical volume distributions.

Obtaining the average power of such a shaped load requires modifying Eq. (9-15) by a shaping factor in a manner similar to that for noise weighting discussed in Chap. 7. Letting $C(f)$ represent the shaped gain in dB added to the flat signal, the new average power, P_{av}', is given in terms of the average power with flat levels by

$$P_{av}' = P_{av} + 10 \log \frac{1}{f_T - f_B} \int_{f_B}^{f_T} 10^{C(f)/10}\, df \qquad \text{dBm0}$$

where f_B and f_T are the bottom and top baseband frequencies, respectively.

Shaping of the talker levels affects the peak factor and hence Δ_c in a more complex manner. This problem has been investigated through the use of a computer for normally distributed talkers having a variety of signal shapes over the multiplexed band. For the same overload condition as already discussed, the results of the study indicated that the factor Δ_c is very well approximated by determining Δ_c for η channels instead of N channels. The symbol η represents the number of channels in the system having a level after shaping that is within 6 dB of the channel having the highest level.

REFERENCES

1. Holbrook, B. D. and J. T. Dixon. "Load Rating Theory for Multi-Channel Amplifiers," *Bell System Tech. J.*, vol. 18 (Oct. 1939), pp. 624-644.
2. The International Telegraph and Telephone Consultative Committee (CCITT). "Line Transmission," *Blue Book*, III (International Telecommunication Union, 1965).
3. McAdoo, Kathryn L. "Speech Volumes on Bell System Message Circuits—1960 Survey," *Bell System Tech. J.*, vol. 42 (Sept. 1963), pp. 1999-2012.
4. Bennett, W. R. "Cross-Modulation Requirements on Multichannel Amplifiers Below Overload," *Bell System Tech. J.*, vol. 19 (Oct. 1940), pp. 587-610.
5. Subrizi, V. "A Speech Volume Survey on Telephone Message Circuits," *Bell Laboratories Record*, vol. 31 (Aug. 1953), pp. 292-295.

Chapter 10

Nonlinearities

Like thermal noise, nonlinearities are present to some degree in all electrical networks. Nonlinearities fall into two basic classes. The first is the strong or intentional nonlinearity where the nonlinear performance is desired and controlled for some specific application. Examples include square and power law modulators, compandors, and limiters. The second class of nonlinearity is the weak or undesired nonlinearity where linear performance is desired, and any nonlinearities are considered parasitic in nature. In general, the effects of such weak nonlinearities limit the useful signal levels in a system and thus become an important design consideration. This chapter is primarily concerned with the characterization and system effects of weak nonlinearities.

The most important nonlinear elements commonly used in transmission systems are diodes, transistors or other active devices, and coils and transformers utilizing ferrous materials. Although the nonlinear effects of these devices are difficult to control in manufacture, they are somewhat reproducible so that results can be predicted with fair accuracy.

10.1 SERIES REPRESENTATION OF TRANSFER CHARACTERISTIC

Consider the voltage transfer characteristic of a general two-port which may be a device, network, or system. If a plot is made of the instantaneous output voltage versus the instantaneous input voltage, a graph similar to Fig. 10-1 would be obtained. For simplicity it is assumed that the two-port is memoryless; i.e., the output is an instantaneous function of the input voltage.

The transfer characteristic shown in Fig. 10-1 can be described by the Taylor series expansion

$$e_o = a_1 e_i + a_2 e_i^2 + a_3 e_i^3 + \cdots \tag{10-1}$$

For strong nonlinearities particular a_n's may be carefully controlled; e.g., a square law modulator requires a large a_2 with all other coefficients negligible. For a nearly linear two-port, a_1 is the voltage gain, and all higher order terms are relatively small.

For any physical network, Eq. (10-1) is only valid over a limited frequency range. If the voltage transfer characteristic is considered at discrete frequencies, the memoryless restriction can be relaxed slightly. Equation (10-1) will then be valid at a particular frequency if the phase (or delay) characteristic is ignored. It is common practice to characterize the nonlinear performance at points where the load impedance is flat with frequency which usually makes the frequency dependence of a_n small over the frequency range of interest. Although not treated here, a technique using Volterra kernels has also been used to characterize accurately the nonlinear performance of frequency dependent networks containing storage elements [1].

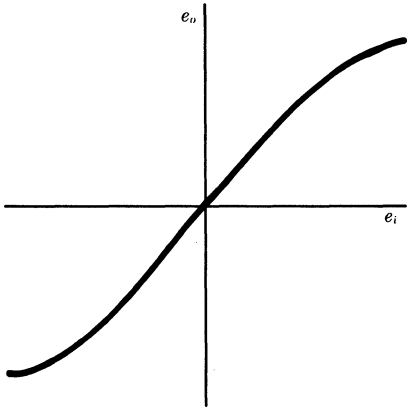


FIG. 10-1. Nonlinear voltage transfer characteristic of two-port.

Single-Frequency Input

Consider the application of a single-frequency sinusoid to a two-port whose transfer characteristic is given by Eq. (10-1) with $a_n = 0$ for $n > 3$. The input given by

$$e_i = A \cos \alpha t$$

results in an output

$$e_o = a_1 A \cos \alpha t + a_2 A^2 \cos^2 \alpha t + a_3 A^3 \cos^3 \alpha t$$

Applying the trigonometric identities

$$\cos^2\alpha t = \frac{1}{2} + \frac{1}{2}\cos 2\alpha t$$

$$\cos^3\alpha t = \frac{3}{4}\cos \alpha t + \frac{1}{4}\cos 3\alpha t$$

results in

$$e_o = \frac{1}{2}a_2A^2 + [a_1A + \frac{3}{4}a_3A^3]\cos \alpha t + \frac{1}{2}a_2A^2\cos 2\alpha t + \frac{1}{4}a_3A^3\cos 3\alpha t \qquad (10\text{-}2)$$

Examination of Eq. (10-2) reveals that the application of a single-frequency sinusoid to a nonlinear two-port results in an output consisting of components at the applied frequency and spurious components at zero frequency, the second harmonic, and the third harmonic of the applied frequency. Such spurious products are often called modulation products. If a bandpass filter passing α radians per second is placed at the output of this two-port, observation of the behavior of the fundamental would indicate that the apparent gain (or loss) is a function of the level of the applied signal. For the simple case given here, the change in gain is given by the factor $[1 + 3/4\ (a_3/a_1)\ A^2]$; 20 log of this factor is the *expansion* in dB; 20 log of the reciprocal of this factor is the *compression* in dB. Such compression or expansion is often used as an overload criterion for a two-port, with typical overload being a fraction of a dB.

Further examination of Eq. (10-2) shows that the amplitude of the second harmonic term is proportional to a_2 and the square of the amplitude of the applied signal. Thus, for a 1-dB increase in the power of the fundamental, a 2-dB increase in the second harmonic can be expected. Similarly, the third harmonic is proportional to a_3 and the cube of the amplitude of the fundamental. The 1-dB increase in the input power results in a 3-dB increase in the third harmonic output. These relationships, which provide a useful way of characterizing the nonlinear behavior of two-ports, are discussed later.

Three-Frequency Input

In many cases of practical interest, it is convenient to consider an input signal as a sum of individual sinusoids given by

$$e_i = A\cos(\alpha t + \phi_1) + B\cos(\beta t + \phi_2) + C\cos(\gamma t + \phi_3) + \ldots X\cos(\chi t + \phi_n)$$

The frequencies α, β, γ, etc. are not necessarily harmonically related. If $a_n = 0$ for $n > 3$ in Eq. (10-1), it can be shown that the *form* of all modulation products can be obtained by considering only three different sinusoids at the input. For mathematical simplicity, these sinusoids are assumed to have zero phase shift (an assumption that does not change the amplitude of the general random phase case) and can be represented by

$$e_i = A \cos \alpha t + B \cos \beta t + C \cos \gamma t \qquad (10\text{-}3)$$

Substituting this expression into Eq. (10-1) and applying trigonometric identities result in the terms listed in Fig. 10-2. For the nearly linear case the desired outputs are the first order products due to term 1. All other outputs are spurious and contribute to objectionable noise and interference.

The d-c term is generally of no interest because it is usually filtered out and thus causes no objectionable interference. The first order product due to the third term is a compression or expansion effect (depending upon the sign of a_3). This results in an apparent change in loss or gain and, since normally

$$a_1 >> \frac{3}{4} a_3 (A^2 + 2B^2 + 2C^2)$$

it will be ignored in this discussion.

There are two types of second order products due to a_2. The first of these is simply the second harmonic which is the same as that obtained if each input were applied separately. The other type of second order product consists of sum and difference frequencies of each pair of applied frequencies. The magnitude of the $\alpha \pm \beta$ product is a_2AB. Compared with the 2α product of 1/2 (a_2A^2), the $\alpha \pm \beta$ products are higher in level than the 2α products by the factor $2B/A$. Expressed in dBm (20 log because these are voltages),

$$P_{\alpha \pm \beta} = P_{2\alpha} + 6 + P_\beta - P_\alpha \qquad \text{dBm} \qquad (10\text{-}4)$$

Thus, if each of the input frequencies has the same amplitude, the power of the sum and difference products will be 6 dB higher than the power of the second harmonic products.

There are three different types of third order products due to a_3. The 3α, or third harmonic, terms are identical to those obtained when each of the input frequencies is applied individually and has an amplitude of 1/4 (a_3A^3). The $2\alpha \pm \beta$ products consist of the sum and difference of one frequency and the second harmonic of another and

Frequencies and relative magnitudes to be found in output, $e_o = a_1 e_i^1 + a_2 e_i^2 + a_3 e_i^3$, from applied signal, $e_i = A \cos \alpha t + B \cos \beta t + C \cos \gamma t$

	Term 1	Term 2	Term 3
d-c		$1/2\ a_2(A^2 + B^2 + C^2)$	
First order	$a_1 A\cos\alpha t + a_1 B\cos\beta t + a_1 C\cos\gamma t$		$3/4\ a_3 A(A^2 + 2B^2 + 2C^2)\ \cos\alpha t + 3/4\ a_3 B(B^2 + 2C^2 + 2A^2)\ \cos\beta t + 3/4\ a_3 C(C^2 + 2A^2 + 2B^2)\ \cos\gamma t$
Second order		$1/2\ a_2(A^2\cos 2\alpha t + B^2\cos 2\beta t + C^2\cos 2\gamma t) + a_2 AB[\cos(\alpha+\beta)t + \cos(\alpha-\beta)t] + a_2 BC[\cos(\beta+\gamma)t + \cos(\beta-\gamma)t] + a_2 AC[\cos(\alpha+\gamma)t + \cos(\alpha-\gamma)t]$	
Third order			$1/4\ a_3(A^3\cos 3\alpha t + B^3\cos 3\beta t + C^3\cos 3\gamma t) + 3/4\ a_3 \{A^2B[\cos(2\alpha+\beta)t + \cos(2\alpha-\beta)t]\ A^2C[\cos(2\alpha+\gamma)t + \cos(2\alpha-\gamma)t]\ B^2A[\cos(2\beta+\alpha)t + \cos(2\beta-\alpha)t]\ B^2C[\cos(2\beta+\gamma)t + \cos(2\beta-\gamma)t]\ C^2A[\cos(2\gamma+\alpha)t + \cos(2\gamma-\alpha)t]\ C^2B[\cos(2\gamma+\beta)t + \cos(2\gamma-\beta)t]\} + 3/2\ a_3 ABC[\cos(\alpha+\beta+\gamma)t + \cos(\alpha+\beta-\gamma)t + \cos(\alpha-\beta+\gamma)t + \cos(\alpha-\beta-\gamma)t]$

NOTE: Observe that if in the applied signal $A = B$, then the level of the $\alpha + \beta$ product, which is at the frequency $\alpha + \beta$, is 6 dB greater than the 2α product. Simliarly, $\alpha - \beta$ is 6 dB greater than the 2α product, and $2\alpha - \beta$ (and similar terms) are 9.6 dB greater than 3α. If $A = B = C$, then the $\alpha + \beta - \gamma$ term and similar terms (but do not confuse with $2\alpha - \beta$ type) are 15.6 dB greater than 3α. The compression, or first order component, arising from the e_i^3 term is at least 9.6 dB greater than 3α and may be much greater, depending on the number of signals applied; for the three-frequency input given above, it is 23.5 dB greater. If the a_n's are functions of frequency, the frequency effects must be added to the aforementioned level effects to determine the level differences between products.

FIG. 10-2. Expansion of power series for three-sinusoid input.

have magnitudes of 3/4 (a_3A^2B). Compared with the third harmonic amplitude, these products are larger by the factor $3B/A$. Expressed in dBm,

$$P_{2\alpha\pm\beta} = P_{3\alpha} + 9.6 + P_\beta - P_\alpha \quad \text{dBm} \tag{10-5}$$

Thus, if each of the input frequencies has the same amplitude, the power of the $2\alpha\pm\beta$ products will be 9.6 dB higher than that of the third harmonic product. The last type of third order product to be considered is the $\alpha\pm\beta\pm\gamma$, the sum and difference of each group of three frequencies. The magnitudes of the $\alpha\pm\beta\pm\gamma$ products are given by $3a_3ABC/2$ or are higher than the third harmonics by the factor $6BC/A^2$. Expressed in dBm,

$$P_{\alpha\pm\beta\pm\gamma} = P_{3\alpha} + 15.6 + P_\beta + P_\gamma - 2P_\alpha \quad \text{dBm} \tag{10-6}$$

If each of the input frequencies has the same magnitude, the power of the $\alpha\pm\beta\pm\gamma$ products will be 15.6 dB greater than that of the 3α harmonics.

In summary, a nonlinear network produces at its output additional frequencies that are various combinations of sum and difference frequencies of the input signals. It should be emphasized that the same generation of additional frequencies occurs with strong nonlinearities. Application of an α frequency carrier and an audio signal of β frequency to any power law device will result in an output consisting of the applied signals at α and β and many additional products at different frequencies. The output of this device could be passed through a bandpass filter centered near the frequency α to pass the $\alpha+\beta$ and $\alpha-\beta$ products and obtain a DSBTC signal. If instead the bandpass filter characteristic were centered at the frequency 2α, the 2α, $2\alpha+\beta$, and $2\alpha-\beta$ products would be passed to yield a DSBTC signal at the 2α frequency. The efficiency of such modulators depends, of course, on the magnitude of the coefficients of the power series. Modulator configurations such as the ring or lattice are chosen to produce the desired product with high efficiency while minimizing the other spurious products.

Demodulation can also be accomplished with a power law network. In this case, application of a carrier and its sidebands will result in $\alpha-\beta$ frequency products which will be the original baseband signal. It should be emphasized that power law modulators and demodulators always produce spurious products which must be filtered from the desired signal. It can be shown that such filtering will be successful

only if the carrier frequency, f_c, is greater than twice the highest baseband frequency, f_T; i.e., the condition $f_c > 2f_T$ must be satisfied.

Compensation of Nonlinear Characteristics

It might be expected that the nonlinear characteristic of a two-port could be compensated for by the addition of another two-port in tandem having a complementary nonlinear transfer characteristic. Such compensation cannot be extended over an arbitrarily wide dynamic range because of overload considerations; however, compensation over the normal working range may be desirable. Such compensation methods are generally impractical because of the great difficulty in controlling the nonlinearity to the degree required. In fact, achieving a desired nonlinear transfer characteristic is more difficult than achieving a linear one. Thus, except for special applications such as compandors, the compensation of a nonlinear characteristic by an inverse nonlinear characteristic is not usually attempted.

Even if compensating nonlinear characteristics could be achieved, care would be required in applying such compensation. Basically, *all* frequencies produced by the nonlinear two-port (and this can include those from zero frequency to very high harmonics) must be applied without frequency shaping to the compensating network. Such a constraint is usually rather difficult to meet in many practical applications.

10.2 EFFECT ON ANGLE-MODULATED WAVES

The effect of the nonlinear transfer characteristic given by Eq. (10-1) on an angle-modulated signal is now considered. The angle-modulated signal is represented by

$$e_i = A_c \cos [2\pi f_c t + \phi(t)]$$

Substituting in Eq. (10-1) and ignoring terms higher than third order gives

$$e_o = a_1 A_c \cos [2\pi f_c t + \phi(t)] + a_2 A_c^2 \cos^2 [2\pi f_c t + \phi(t)]$$
$$+ a_3 A_c^3 \cos^3 [2\pi f_c t + \phi(t)]$$

The terms may be expanded and collected:

$$e_o = \frac{1}{2}a_2A_c^2 + (a_1A_c + \frac{3}{4}a_3A_c^3)\cos[2\pi f_c t + \phi(t)]$$

$$+ \frac{1}{2}a_2A_c^2\cos[4\pi f_c t + 2\phi(t)]$$

$$+ \frac{1}{4}a_3A_c^3\cos[6\pi f_c t + 3\phi(t)]$$

The output wave consists of a d-c term and three angle-modulated waves centered respectively at frequencies f_c, $2f_c$, and $3f_c$. Assume for the moment that a filter can be used to extract the angle-modulated wave centered at f_c; the output becomes

$$e_o = (a_1A_c + \frac{3}{4}a_3A_c^3)\cos[2\pi f_c t + \phi(t)]$$

The nonlinear characteristic has done nothing more than modify the gain. This is an important difference between amplitude modulation and angle modulation and is the primary reason why angle modulation is used in microwave systems where nonlinear operation of amplifiers and other devices has thus far been unavoidable at the required output levels.

To achieve the desired output, it is necessary to separate the angle-modulated wave centered at f_c from the one centered at $2f_c$. Denote the peak frequency deviation by ΔF and the top baseband frequency by f_T hertz. Applying Carson's bandwidth rule mentioned in Chap. 5 and considering that the peak frequency deviation about the second harmonic of the carrier is doubled, the necessary condition for separation of the f_c and $2f_c$ waves requires that

$$f_c \geq 3\Delta F + 2f_T$$

The preceding analysis may seem to indicate that angle-modulated systems are free from distortion since they are insensitive to *amplitude* nonlinearities. However, angle-modulated systems are extremely sensitive to phase nonlinearities. Although phase nonlinearities are not as common as amplitude nonlinearities, they do exist and are often significant in angle-modulated systems.

A common type of phase nonlinearity is called AM to PM conversion. This is a result of the phase characteristic (or delay) of a two-port network being dependent upon the instantaneous amplitude of the signal. Thus, if the amplitude of the signal passed through such a network varies with time or is an amplitude-modulated sinusoid, the phase of the output will have a ripple around a linear (constant delay) characteristic. In many cases, the resulting phase ripple will have a "waveform" similar to the amplitude modulation. If an amplitude-modulated carrier having an index, m, is applied to a system, the observed peak phase deviation may be k_p radians. The ratio of k_p/m at the normal operating levels is the AM to PM conversion factor and is often expressed in dB by taking 20 log k_p/m. In devices having notoriously high AM to PM conversion, this factor may be as large as —6 dB. On the other hand, AM to PM conversion factors of —34 dB are not unusual for a well designed limiter. The AM to PM conversion in more linear circuits may be much less and can often be ignored.

In recent years it has become common to characterize AM to PM conversion in degrees of phase shift per dB change in amplitude [2]. Conceptually, such a conversion could be measured by applying a single sinusoid to the system and by observing at the output the phase shift due to a change in level of the applied signal. Unfortunately, such static measures are usually inadequate, and a dynamic measurement is required in which a "carrier" is modulated by a low level sinusoid. A good microwave repeater may have AM to PM conversion of less than 2 degrees per dB although higher values are usually encountered for traveling wave tubes.

These two methods of characterizing AM to PM conversion can be related by converting radians to degrees and fractional changes to dB. Thus,

$$\text{degrees/dB} \approx 6.6 \text{ antilog} \left[\frac{20 \log k_p/m}{20} \right]$$

$$\approx 6.6 \frac{k_p}{m}$$

Effects other than AM to PM conversion can also cause distortion in angle-modulated systems. It is shown in a later chapter that transmission deviations (i.e., nonuniform gain or nonlinear phase with frequency) in the r-f (or i-f) transmission portions of an angle-modu-

lated system can result in distortion of the baseband signal. This distortion is similar to that caused by passing the baseband signal through a two-port having an amplitude sensitive transfer characteristic. Indeed, such effects can be measured as nonlinearities in the baseband signal. However, they are not due to nonlinear two-ports and are not discussed in this chapter.

10.3 CHARACTERIZATION OF TWO-PORT NONLINEARITIES

The nonlinear or modulation distortion of any two-port *could* be characterized by the values of the a_n's of the voltage transfer characteristic. However, the direct measurement of the transfer characteristic to the accuracy required to determine the coefficients of the Taylor series is practically impossible. It is convenient to define another set of modulation coefficients more easily measured than the a_n's of the transfer characteristic. It has been pointed out that the application of a single-frequency input results in outputs at harmonic frequencies, and also that the amplitude of the harmonics is easily related to the amplitude of any other modulation product. Thus, the nonlinear performance of a two-port can be characterized by knowing the output voltage of an applied fundamental and the voltage of each harmonic component. Such data is obtainable with a sinusoidal source and a good frequency selective voltmeter.* Further simplification is possible because the amplitude of the second harmonic is proportional to the square of the amplitude of the applied fundamental. Thus, the ratio of the amplitude of the second harmonic to the square of the amplitude of the fundamental results in a modulation coefficient independent of the amplitude of the signal used to measure it. Similarly, the third order performance can be characterized by the ratio of the third harmonic to the cube of the amplitude of the fundamental [3].

The modulation coefficients obtained by these ratios are dependent upon the units used in the voltage readings. It is common to express all voltages in rms volts and to make the modulation coefficient dimensionless by multiplying the numerator by one volt raised to the appropriate power. The modulation coefficient in dB is then obtained by taking 20 log of this dimensionless ratio. For example,

*It is important that the sinusoidal source be relatively free of harmonics. The requirements on the voltmeter can be eased by suppressing the fundamental at the input to the meter.

the second order voltage modulation coefficient is obtained by multiplying the second harmonic rms voltage by one volt and dividing by the mean square voltage of the fundamental. In dB notation,

$$M_2 = 20 \log \frac{V_{2\alpha}\ (1 \text{ volt})}{V_\alpha{}^2}$$

Similarly, the third order voltage modulation coefficient is given by

$$M_3 = 20 \log \frac{V_{3\alpha}\ (1 \text{ volt})^2}{V_\alpha{}^3}$$

The chief interest in transmission systems is to determine the *power* produced by the spurious products of nonlinearities. Since selective power reading meters are readily available, the power of the fundamental and harmonic frequencies can be measured at the output, and these quantities can be used to define the modulation coefficients. It is convenient to define

$$m_2 \triangleq \frac{p_{2\alpha}\bar{p}}{p_\alpha{}^2} \tag{10-7}$$

$$m_3 \triangleq \frac{p_{3\alpha}\bar{p}^2}{p_\alpha{}^3} \tag{10-8}$$

where, for convenience, $\bar{p}$ is one milliwatt and all powers are in milliwatts. This insures that the modulation coefficients are dimensionless. Equivalent coefficients defined in dB (using 10 log because these are power ratios) are:

$$M_2 = 10 \log m_2 = 10 \log p_{2\alpha} + 0 \text{ dBm} - 20 \log p_\alpha$$

$$M_2 \triangleq P_{2\alpha} - 2P_\alpha \tag{10-9}$$

and

$$M_3 = 10 \log m_3 = 10 \log p_{3\alpha} + 0 \text{ dBm} - 30 \log p_\alpha$$

$$M_3 \triangleq P_{3\alpha} - 3P_\alpha \tag{10-10}$$

where all of the P's are in dBm. By these definitions, M_x is the xth harmonic power in dBm resulting from a 0 dBm fundamental. Note that the modulation coefficients defined in terms of power are *not* the same as those defined in terms of voltage unless the load is 1000 ohms.

If the two-port transfer characteristic follows a power series, the M's are independent of signal level. However, they are not independent of frequency if the load impedance or transfer characteristic is frequency dependent.

Relating m Coefficients to a Coefficients

From Fig. 10-2, with the compression term neglected, the rms value of the fundamental at the output is given by $a_1A/\sqrt{2}$, and the rms value of the second harmonic at the output is given by $a_2A^2/2\sqrt{2}$. Thus,

$$p_{2\alpha} = \frac{a_2^2 A^4}{8R_L} 1000 \quad \text{mW}$$

$$p_\alpha = \frac{a_1^2 A^2}{2R_L} 1000 \quad \text{mW}$$

$$\bar{p} = 1 \text{ mW}$$

$$m_2 = \frac{a_2^2}{2a_1^4}\left(\frac{R_L}{1000}\right) \tag{10-11}$$

Similarly, the rms value of the third harmonic at the output is given by $a_3A^3/4\sqrt{2}$ to yield

$$m_3 = \frac{a_3^2}{4a_1^6}\left(\frac{R_L}{1000}\right)^2 \tag{10-12}$$

In these equations, A is the peak voltage of the fundamental; a_1 is dimensionless; a_2 has units of reciprocal volts; a_3 has units of reciprocal volts squared; and R_L is the load resistance in ohms.

Determining the Output Power of a Specific Product

A common problem is to determine the expected power of a specific product when the modulation coefficients and the powers of the fundamentals are known. The necessary relationships are easily obtained by using Eqs. (10-4), (10-5), (10-6), (10-9), and (10-10):

$$\begin{aligned} P_{2\alpha} &= M_2 + 2P_\alpha \\ P_{3\alpha} &= M_3 + 3P_\alpha \\ P_{\alpha\pm\beta} &= M_2 + 6 + P_\alpha + P_\beta \\ P_{2\alpha\pm\beta} &= M_3 + 9.6 + 2P_\alpha + P_\beta \\ P_{\alpha\pm\beta\pm\gamma} &= M_3 + 15.6 + P_\alpha + P_\beta + P_\gamma \end{aligned} \tag{10-13}$$

Cascaded Two-ports

The nonlinear performance of a system consisting of cascaded two-ports is usually controlled by those having a combination of poor modulation performance and carrying relatively high signal levels. The quantitative relationships can be seen by considering a pair of two-ports in tandem as shown in Fig. 10-3. The first is assumed to have a power gain of G' when its load consists of the input impedance of the second. Under the same conditions, this first two-port has modulation coefficients given by M'_2 and M'_3. Similarly, the second two-port has a power gain of G'' when driven by the output impedance of the first and has modulation coefficients given by M_2'' and M_3'' under these same conditions.

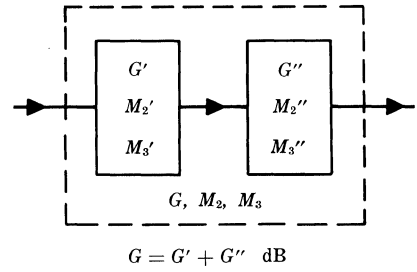


FIG. 10-3. Cascaded nonlinear two-ports.

For simplicity, it is assumed that the nonlinearities are small enough so that higher order effects can be ignored and superposition can be used. If the power of P_α at the output of the composite network is assumed to be 0 dBm, the resulting power of $P_{2\alpha}$ gives M_2 directly. Similarly, the power of $P_{3\alpha}$ under these same conditions gives M_3 directly. If P_α is 0 dBm at the output of the tandem two-ports, it is $-G''$ dBm at the output of the first. Thus, by the use of Eq. (10-13), the output of the first two-port consists of a second harmonic of power $P_{2\alpha}' = M_2' - 2G''$ which is amplified to a power of $M_2' - G''$ at the output of the composite. At the same time, the second two-port produces a second harmonic at its output at a power of M_2''. The total second harmonic output is then the sum of these two components at the same frequency. To sum these components accurately requires a knowledge of the phase characteristics of the two networks and their modulation products. Without this data, power addition is usually assumed to obtain

$$M_2 \approx (M_2' - G'') \text{ "+" } M_2'' \tag{10-14}$$

Using a similar procedure for the third harmonic results in

$$M_3 \approx (M_3' - 2G'') \text{ "+" } M_3'' \tag{10-15}$$

This approximation does not include the third harmonic component which results when the second harmonic from the first two-port sums with the fundamental in the second two-port. Such products can be ignored in most applications.

Examination of Eqs. (10-14) and (10-15) shows that if the power gain of the second two-port is sufficiently large, the modulation performance of the combination is almost entirely determined by the performance of the output or high level network. On the other hand, if the output network is lossy and linear (G'' negative and $M_2'' \to -\infty$), it can be seen that the modulation performance is determined primarily by the first network which in this case is handling the highest level signals. As a consequence the overall modulation performance of an amplifier is almost entirely determined by the modulation performance of its last active stage, and the effects of other stages can usually be ignored. This is in contrast to the random noise performance of cascaded networks (discussed in Chap. 8) where the lowest level stage determines the noise performance.

10.4 SYSTEM MODULATION PERFORMANCE

With the previously discussed relationships in mind, it is now appropriate to discuss the modulation behavior of a system at 0 TLP. It is convenient to define an index of system intermodulation, H_x, where x is the product under consideration (thus x may be 2α, $\alpha \pm \beta$, etc.).

> H_x is the power of the x-type product in dBm0, measured at the output of a system and formed by the intermodulation of fundamentals each of which is 0 dBm0.

Note that this definition is identical to that for the modulation coefficients except that all levels are referenced to 0 TLP rather than to the output of any particular network in the system. It is important to relate the modulation coefficients of a two-port in a system to 0 TLP; this requires that the level difference between the output at which the modulation coefficients are referenced and 0 TLP be known. The required factor is commonly denoted C.

> C is the level difference between 0 TLP and the point where the modulation coefficients are defined; C is positive when this point is below 0 TLP.

Stated another way, if a 0 dBm signal appears at the output where the M's are defined, this signal will be C dBm0. The factor C may be a function of frequency, although in the initial stages of design, it is often assumed to be flat with frequency. It may be carefully shaped

in the later, more refined stages of design. However, if the reference point for defining the modulation coefficients is carefully chosen, the assumption that C is flat with frequency can often be justified.

With knowledge of C, the modulation coefficients, and the power of the fundamentals at 0 TLP, the expected magnitude of any modulation product can be computed by the use of Eq. (10-13). For example, consider a 0 dBm0 fundamental. The power of this fundamental at the modulation reference is $-C$ dBm, and the power of the second harmonic at this same point is $M_2 - 2C$ dBm. At 0 TLP, the power of this second harmonic is $M_2 - C$ dBm. If there is only *one* nonlinear source in the system, the above value of the second harmonic corresponds to the value of $H_{2\alpha}$. By similar reasoning, the powers of all modulation products at 0 TLP with fundamental powers in dBm0 can be tabulated in terms of H_x:

$$\begin{aligned}
P_{2\alpha} &= H_{2\alpha} + 2P_\alpha \quad \text{dBm0} \\
P_{3\alpha} &= H_{3\alpha} + 3P_\alpha \quad \text{dBm0} \\
P_{\alpha\pm\beta} &= H_{2\alpha} + 6 + P_\alpha + P_\beta \quad \text{dBm0} \\
P_{2\alpha\pm\beta} &= H_{3\alpha} + 9.6 + 2P_\alpha + P_\beta \quad \text{dBm0} \\
P_{\alpha\pm\beta\pm\gamma} &= H_{3\alpha} + 15.6 + P_\alpha + P_\beta + P_\gamma \quad \text{dBm0}
\end{aligned} \tag{10-16}$$

Choice of a Modulation Reference Point

Up to this point, the discussion of nonlinearities has been general and has assumed that the reference point for defining the modulation coefficients is at the output of some general two-port. It has also been implied that a wise choice for this reference point may tend to keep C reasonably flat with frequency and thus simplify some of the system calculations. It is also desirable to choose the point such that the resulting modulation coefficients are reasonably independent of frequency, which implies a point with an impedance nearly constant for all frequencies of interest. With these restrictions it may not always be possible to avoid frequency shapes entirely. In a repeatered analog system the output of the repeater is usually working into the impedance of the transmission cable which is reasonably flat with frequency. As a consequence, this repeater output is often chosen as the reference point for the M's. The chief drawback in using this reference point is the extra manipulation necessary to relate these M's to the modulation coefficients of the nonlinear device which is rarely directly at the output of the repeater. Also, such a reference point does not guarantee a C that is flat with frequency.

A second point sometimes used for the modulation reference point is the collector of the output stage transistor. The arguments favoring this point are: (1) it relates directly to device measurements; (2) the effects of feedback are easily seen; (3) overload considerations may result in a flat signal spectrum at this point, thus making C flat with frequency. Unless the output network is fairly simple, however, keeping the load impedance constant with frequency may be difficult.

Additional Considerations with Transistors

It may be impossible to choose a modulation reference point that makes both C and the M's flat with frequency. In such a case, H_x will not be flat with frequency. A preliminary approach for initial calculations is to make a worst case computation assuming H_x flat with frequency. It must be remembered that some advantage may be obtained by allowing for the frequency shape.

A point that must be checked using this worst case analysis is whether or not the modulation indices, M_2 and M_3, are constants with varying signal level and thus follow power-law relationships. This may be accomplished by varying the amplitude of the fundamental 1 dB and noting whether or not there are 2- and 3-dB changes in the second and third harmonic, respectively. When modulation indices are not constant, allowances must be made in the system analysis for their variations. For example, an $\alpha \pm \beta$ product may not be 6 dB greater than a 2α product as previously indicated [4].

The allowance that can be made in the system analysis is to measure $M_{\alpha\pm\beta}$, $M_{\alpha\pm\beta\pm\gamma}$, or any other modulation index directly in terms of the powers at the reference point. That is, M_2 is measured as the ratio of the second harmonic power to the square of the fundamental power, both being measured at the reference point chosen. Thus, the following relations hold where all powers are measured in dBm at the modulation reference point:

$$\begin{aligned} M_{2\alpha} &= P_{2\alpha} - 2P_\alpha \quad \text{dB} \\ M_{3\alpha} &= P_{3\alpha} - 3P_\alpha \quad \text{dB} \\ M_{\alpha\pm\beta} &= P_{\alpha\pm\beta} - P_\alpha - P_\beta \quad \text{dB} \\ M_{2\alpha\pm\beta} &= P_{2\alpha\pm\beta} - 2P_\alpha - P_\beta \quad \text{dB} \\ M_{\alpha\pm\beta\pm\gamma} &= P_{\alpha\pm\beta\pm\gamma} - P_\alpha - P_\beta - P_\gamma \quad \text{dB} \end{aligned} \tag{10-17}$$

10.5 NONLINEAR EFFECTS ON MULTIPLEXED TALKERS

It has been shown that if single frequencies are present at the input of a system, the system nonlinearities will cause the generation of undesired tones which will be present at the output. The levels of these undesired tones are a function of the levels of the single frequencies at the input. Since the telephone transmission system usually carries speech signals rather than single-frequency tones, it is important to relate the measured nonlinearity of a system to the expected intermodulation noise due to talkers. Several techniques are available for this purpose.

For narrowband low capacity systems, it is often practical to characterize the talker load by a set of appropriately spaced sinusoids. Since analysis of such a case is merely an extension of the single-frequency performance already discussed, it is not treated further.

Three techniques for characterizing intermodulation noise in multiplexed telephone systems are discussed in some detail. The first, based on methods developed by W. R. Bennett, relates intermodulation noise to single-frequency modulation parameters of the system. Although still very useful, this technique is being replaced in broadband applications by the second technique called noise loading. With the third technique, intermodulation noise can be computed using spectral densities and autocorrelation functions.

Bennett's Method

This method for calculating the modulation noise seeks answers to the following questions:

1. What is the amplitude of one modulation product arising from talkers of known constant volume in systems of known nonlinearity?
2. How many such products will fall in the channel of interest? This question faces the fact that not all of the channels are carrying active speech at the same time. The number sought is N_x, the probable number of x-type products that will fall into the channel of interest during the busy hour.
3. What is the effect on the fact that talker volumes are not all the same, but distributed as shown in Chap. 9?
4. What is the total modulation noise for a system of given linearity? Or, what must the linearity be in order to limit the intermodulation noise to some given value?

The answers to 1, 2, and 3 will be different for each type of product ($\alpha \pm \beta$, $2\alpha - \beta$, etc.) and must be found for each before the last question can be answered.

This method does not take into consideration the special cases in which linearity requirements may be set by factors besides intermodulation noise, e.g., intelligible crosstalk. The magnitude of any particular intermodulation product will be a function of the nonlinearity of the system, H_x. The number of possible intermodulation products in a given channel is a function of the frequency allocation of the FDM message load [5].

It can be assumed that by the time the probable number of superimposed products exceeds 20 or 30 they become indistinguishable from random noise. This would appear to be particularly valid if the masking effect of an equal or greater magnitude of truly random noise power is expected in the disturbed channel.

Product Amplitude. Assume then that each message channel of the system is loaded by a band of gaussian noise having a flat spectral density from 0 to 4 kHz with a total power of 0 dBm0. Recall that H_x was defined in terms of single-frequency signals applied to a system. It is interesting to determine if the same H_x factors would apply for bands of gaussian noise. It can be easily shown that, as long as each fundamental band "enters" the product only once, the same modulation index applies for both single frequencies and bands of noise. However, when there are multiple "entries" such as 2α, 3α, or $2\alpha \pm \beta$, there is a distinct difference between the two cases. For example, with sinusoids the 2α product is simply the second harmonic. However, with bands of noise, the 2α product corresponds to the sum of the second harmonics of all frequency components plus many $\alpha_1 \pm \alpha_2$ products. As with sinusoids, the $\alpha_1 \pm \alpha_2$ products are 6 dB "hotter" than the second harmonic. The net result is that the $H_{2\alpha}$ and $H_{2\alpha \pm \beta}$ are each 3 dB larger for bands of noise than they are for single-frequency signals. Similarly, $H_{3\alpha}$ is 7.8 dB larger for noise than it is for single frequencies. These relationships can be shown more analytically by convolving the appropriate spectral densities. The system intermodulation index for bands of noise can be defined as:

> H_x^* is the power of the x-type product in dBm0, measured at the output of a system and formed by the intermodulation of 4-kHz bands of gaussian noise each of which is 0 dBm0.

From the preceding discussion,

$$H_x^* = H_x + \chi_x \tag{10-18}$$

where

$$\chi_x = 3 \qquad \text{for } x = 2\alpha \text{ and } 2\alpha \pm \beta$$

$$\chi_x = 7.8 \qquad \text{for } x = 3\alpha$$

$$\chi_x = 0 \qquad \text{for all other } x$$

Note that the products due to bands of noise will be spread over an 8-kHz band for the second order products and a 12-kHz band for the third order products.

Talker Amplitude Distributions. As has been shown in Chap. 9, the telephone load does not consist of 0 dBm0 talkers; real talkers form a normal distribution with a median, V_0 vu or $P_0 = V_0 - 1.4$ dBm0, and a standard deviation of σ dB. The following steps compute the modulation noise from such a distribution, relative to the noise from 0 dBm0 talkers.

The x-type products arising from fundamentals which are normally distributed in dB will also be normal distributions in dB. The average value of the product distribution is the sum of the dB (or vu) values of the averages of the fundamentals. The standard deviation of the product magnitudes will be a function of the standard deviations of the fundamentals, the number of fundamentals which form each product, and the number of times they "enter."

The statistics of the product distribution are based on the formation of the product by multiplying the fundamentals (adding dB). The $2\alpha - \beta$ product, for example, is a third order product. Thus the average value of the distribution of $2\alpha - \beta$ products will be $3P_0$ dB higher than the reference product, i.e., a product of the same type formed by 0 dBm0 talkers. If η_x is defined as the order of the product, then in general it can be said that the average of the distribution of products from real talkers will be $\eta_x P_0$ dB with respect to the reference product of H_x^* dBm0. The familiar formula for the standard deviation of a distribution formed by adding a number of normal distributions is

$$\sigma_s = \sqrt{\sigma_1^2 + \sigma_2^2 + \sigma_3^2 + \dots} \tag{10-19}$$

Therefore, the expression for the standard deviation of the distribution of products formed by real talkers can be written

$$\sigma_s = \sigma \sqrt{\lambda_x} \tag{10-20}$$

where σ is the standard deviation of the talker volumes, and

$$\lambda_x = r_1^2 + r_2^2 + \cdots \tag{10-21}$$

for a product of the type $r_1\alpha \pm r_2\beta \pm \ldots$ For example, λ_x is 5 for $2\alpha-\beta$ products, so that the standard deviation of the $2\alpha-\beta$ product distribution will be 11.2 dB if σ is 5 dB.

The relations between average and standard deviation for the modulation products, and the corresponding quantities, P_0 and σ, for the fundamental are shown in Fig. 10-4. Thus, the power of the x-type product caused by loading the channels with flat gaussian noise of P_0 dBm0 is given by $\eta_x P_0 + H_x^*$.

Modulation product	Average product in dB referred to product from 0-dBm0 talkers ($\eta_x P_0$)	Standard deviation in dB ($\sigma\sqrt{\lambda_x}$)
2α	$2P_0$	$2\ \sigma$
$\alpha \pm \beta$	$2P_0$	$\sqrt{2}\ \sigma$
3α	$3P_0$	$3\ \sigma$
$\pm 2\alpha \pm \beta$	$3P_0$	$\sqrt{5}\ \sigma$
$\alpha \pm \beta \pm \gamma$	$3P_0$	$\sqrt{3}\ \sigma$

FIG. 10-4. Average and standard deviation for volume distribution of intermodulated talkers.

If the channels are loaded with flat noise whose amplitude distribution is normal in dB, the average power of the resulting x-type product must be corrected by the $0.115\sigma_s^2$ term discussed in Chap. 9. Since $\sigma_s^2 = \lambda_x\sigma^2$ and $P_0 = V_0 - 1.4$, the average power of an x-type product can be written

$$R_x = \eta_x V_0 + 0.115\ \lambda_x\sigma^2 - 1.4\ \eta_x + H_x^* \qquad \text{dBm0} \tag{10-22}$$

The Product Count. In a typical broadband system, many intermodulation products of a given type can fall into a channel. To find the total intermodulation noise, it is necessary to know the number of products present.

Bennett provides formulas from which the total number of x-type products which can possibly fall in the channel of interest can be computed. These formulas assume that the frequency allocation is as indicated in Fig. 10-5. Carrier slots are located at intervals of f_0

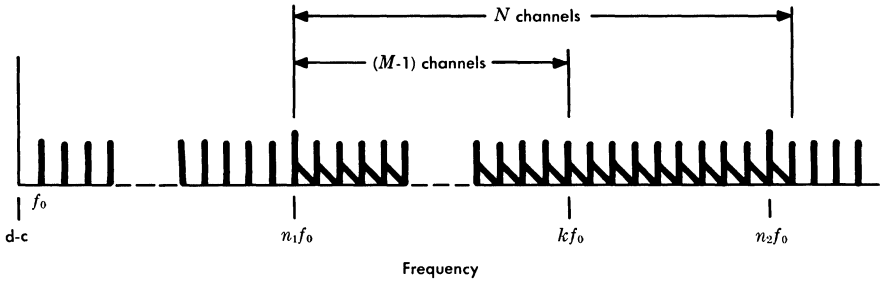


FIG. 10-5. Frequency allocation for product count.

hertz from zero frequency. The system transmits N channels with carrier frequencies from $n_1 f_0$ to $n_2 f_0$ inclusive. Channels are identified by the carrier with which they are associated; the channel of interest is identified by the kth carrier (from zero frequency), M carriers from the lower band edge, such that

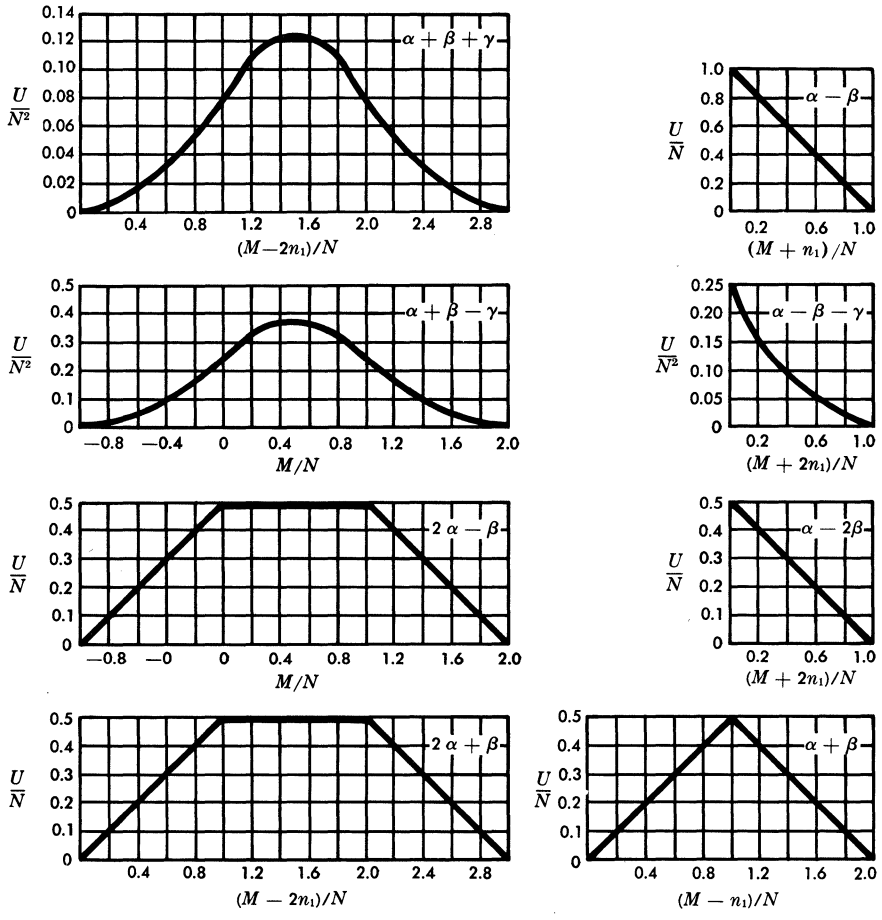
$$M = k - n_1 + 1 \tag{10-23}$$

The 1 occurs because the first transmitted channel is identified by the carrier, $M = 1$, $k = n_1$. The terminology is generalized so that the kth carrier may lie outside the transmitted band. Since the total possible number of products that can fall in the channel of interest is being found, it is assumed that every channel in the system (except the channel of interest) is transmitting energy.

Although Bennett's exact equations are straightforward to understand and use, they are rather long, and repetitive computations become laborious. Variations in bandwidth and therefore in the number of channels transmitted are commonplace during the preliminary design stage of a new system. Counting products in these circumstances can be a time-consuming effort.

A useful set of approximations to Bennett's formulas is given by the graphs in Fig. 10-6. The accuracy of these graphs is poor for narrowband systems. Errors of 50 to 100 per cent may be expected for a ten-channel system. The accuracy improves rapidly with bandwidth and for a 100-channel system is somewhat better than ± 5 per cent. For a 1000-channel system the error is negligible.

The quantities used for the abscissa of these graphs call for comment. When the abscissa is M/N, the number of products that fall in any channel is independent of the position in the frequency spectrum



Legend

U = Total possible products of a given type.

N = Number of channels transmitted with carriers $n_1 f_0$ to $n_2 f_0$ inclusive, where f_0 is the base frequency in hertz and n_1 and n_2 are integers with $n_2 > n_1$.

k = Channel of interest associated with carrier kf_0 within the fundamental band $n_1 < k < n_2$. When values of $k < n_1$ or $k > n_2$ are of interest, the relationships are equally valid.

M = Channel of interest associated with carrier Mf_0 where $M = k - n_1 + 1$.

Note: The number of products is zero outside limits of curves.

FIG. 10-6. Approximate modulation product count for multichannel telephone systems.

occupied by the system. In the case of other products, for example $\alpha + \beta$, where the abscissa is $(M - n_1)/N$, particular frequency allocations will minimize the number of products. Suppose that for any value of N, n_1 is made equal to $n_2/2$. In this case, no $\alpha + \beta$ products will fall inband; however, there is no such escape from $2\alpha - \beta$ or $\alpha + \beta - \gamma$ products, where the abscissa is M/N.

To evaluate the noise arising from modulation products, the number of talker products that will probably fall in the channel of interest during the busy hour must be determined. If the probability that a channel is carrying speech is τ and the total possible number of x-type products that can fall into the channel of interest is U_x, then the probable number of products is

$$N_x = U_x \tau^{\mu_x} \tag{10-24}$$

where μ_x is the number of fundamentals needed to form an x-type product (e.g., $\mu_{2\alpha-\beta} = 2$, and $\mu_{\alpha+\beta-\gamma} = 3$).

Thus, if the total number of $2\alpha - \beta$ products in the kth channel is found to be 160 and the value of τ is 0.25, then the probability that any given possible product is present is $(0.25)^2$, and the probable number of products, $N_{2\alpha-\beta}$, is 10. Since the average number of probable products falling in a given channel is N_x, the total average intermodulation power for the x-type products is

$$R_{x_T} = R_x + 10 \log N_x \qquad \text{dBm0} \tag{10-25}$$

or, taking all factors into account,

$$R_{x_T} = \eta_x V_0 + 0.115\lambda_x \sigma^2 + H_x + \chi_x + 10 \log U_x \tau^{\mu_x} - 1.4\eta_x \qquad \text{dBm0} \tag{10-26}$$

The power computed by the above equation will not be confined to a 4-kHz interval since each second order product occupies an 8-kHz band and each third order product occupies a 12-kHz band. However, the amount of power extending into the adjacent channels is the same as the adjacent channel intermodulation power extending into the channel of interest. As a result, Eq. (10-26) does indeed give the total average intermodulation noise power in a 4-kHz channel at 0 TLP.

Correction Factors for Talkers Instead of Noise. The difference between talkers and flat noise can be taken into account to arrive at a correction factor consisting of three components.

1. *The noise bandwidth factor* accounts for the difference between an effective message channel bandwidth of 3 kHz and a noise bandwidth of 4 kHz. Therefore, a noise bandwidth factor of 10 log 4/3, or 1.25 dB, must be subtracted to obtain the noise power in 3 kHz.
2. *The spectrum shape factor* relates the spectral density of the signal to the annoyance of the intermodulation noise. Since talker power tends to be concentrated between 500 and 1000 Hz, the intermodulation power generated is not flat with frequency. The important second order products produced by talkers have a spectral shape such that they are about 1 dB less annoying than those produced by flat noise. Similarly, shaped third order products are about 1.5 dB less annoying than those of flat noise.
3. *The product amplitude factor* is required because the speech spectrum is not as flat as white noise nor as concentrated as a sine wave. Therefore, the χ_x correction on H_x should only be about half as much (in dB) for talkers as it is for flat noise.

The effects of these three considerations are combined into a single correction factor, C_{w_x}, as tabulated in Fig. 10-7. To convert the units of Eq. (10-26) to dBrnc0, it is necessary to add 88 dB for flat noise, to use the single-frequency value for H_x, and to subtract the correction factor, C_{w_x}. The final expression for the x-type intermodulation noise is given by

$$W_x = H_x + \eta_x V_0 + 0.115\,\lambda_x \sigma^2 + 10 \log U_x \tau^{\mu_x} - 1.4\eta_x - C_{w_x} + 88 \quad \text{dBrnc0} \quad (10\text{-}27)$$

x	Factors (dB)			C_{w_x} (dB)
	Bandwidth	Spectrum shape	Product amplitude	
$\alpha + \beta$	1.25	1.0	0	2.25
$\alpha - \beta$	1.25	1.0	0	2.25
$2\alpha - \beta$	1.25	1.5	−1.5	1.25
$\alpha + \beta - \gamma$	1.25	1.5	0	2.75

FIG. 10-7. Factors comprising the speech-noise correction factor, C_{w_x}.

Using Eq. (10-27) for a given frequency allocation and talker population, it is possible to express the total second order intermodulation noise as

$$W_2 = H_{2\alpha} + K_2 \qquad \text{dBrnc0}$$

where K_2 converts $H_{2\alpha}$ dBm0 to W_2 dBrnc0. Similar relationships exist for third order intermodulation noise.

Effects of Shaped Signal Levels. The nonlinearity computations were based on the assumption that all effects are flat with frequency. Although the multiplex band is flat with frequency at 0 TLP, signal levels in a transmission system will often be shaped to optimize the noise performance of the system. This results in C being shaped with frequency. Thus, even if the modulation coefficients, M_x, are flat with frequency, the system intermodulation coefficients, H_x, will be frequency dependent. This complicates the computation of intermodulation noise because the levels of the intermodulation products will vary with the frequency making up the product. As a result, the total noise in a particular channel must be obtained by summing all products rather than by multiplying by a product count. For example, Eq. (10-27) is evaluated for $x = \alpha - \beta$ with the assumption that C (thus H_x) is shaped with frequency. To do so, the terms $H_x + 10 \log N_x$ must be replaced by a more accurate summation. For $x = \alpha - \beta$,

$$H_{\alpha-\beta} + 10 \log N_{\alpha-\beta} = M_2 + 6 + 10 \log \tau^2 + 10 \log \sum_{1}^{U_{\alpha-\beta}} \text{antilog } (C_{\alpha-\beta} - C_\alpha - C_\beta)/10 \qquad (10\text{-}28)$$

where $C_{\alpha-\beta}$ is the value of C at the frequency $\alpha - \beta$; C_α is the value of C at the frequency α, etc. The summation signs must extend over all possible values of α and β which can produce an $\alpha - \beta$ product in the channel of interest. From Fig. 10-6 the channel in the bottom of the band would have the most products, while none would appear in the top channel of the band.

Similarly, for $x = \alpha + \beta - \gamma$, Eq. (10-27) must be modified by replacing $H_{\alpha+\beta-\gamma} + 10 \log N_{\alpha+\beta-\gamma}$ by

$$M_3 + 15.6 + 10\tau^3 + 10 \log \sum_{1}^{U_{\alpha+\beta-\gamma}} \text{antilog } (C_{\alpha+\beta-\gamma} - C_\alpha - C_\beta - C_\gamma)/10 \qquad (10\text{-}29)$$

Similar equations could, of course, be developed for all other possible products.

This discussion indicates the need to compute and add hundreds of thousands or even millions of products to obtain the intermodulation noise in only one channel. Although computer programs are available, even a high-speed computer may not be able to do this job efficiently. To save machine time, a good approximation can be achieved by scaling, i.e., dividing all channel numbers (and therefore the total number of channels) by a factor, k. Intermodulation noise of each type is computed for this scaled down version and the answers increased by the factor $10y \log k$ where y is one less than the number of independent fundamentals (e.g., $y = 1$ for $\alpha - \beta$, 1 for $2\alpha - \beta$, 2 for $\alpha + \beta - \gamma$, etc.). This correction factor can be justified by inspection of Fig. 10-6.

If M_x is not flat with frequency, i.e., the power series transfer characteristic of Eq. (10-1) is frequency dependent, the preceding procedure must be followed with slight modifications. One approach is to redefine the reference point to some fictitious point that makes it flat with frequency and to modify C for the correct results. The same result can be obtained by computing the amplitude of each possible product individually. In effect, both M_x and C are brought within the summation signs of Eqs. (10-28) and (10-29).

The final complication in some practical problems is the discovery that the M_x coefficients are a function of signal amplitude. This comes about if the nonlinearity cannot be represented by a power series. The most probable cause of this is the presence of more than one nonlinearity in the two-port under measurement. For example, transistors are subject to both input and output distortion as well as transfer distortion and thus do not always closely follow the power law. It is often practical in such cases to measure the powers of the various products while applying fundamentals with powers very close to that of the normal signals. Also, it is often found that the behavior with amplitude can be represented by a very simple function such as a straight line over the range of interest.

Measurement of Intermodulation by Noise Loading

In modern broadband systems the accuracy of converting single-frequency modulation measurements into intermodulation performance of thousands of products becomes questionable. A technique called noise loading is now commonly used for measuring the nonlinearity of very broadband systems. Noise loading consists of

applying a band of flat gaussian noise to the system at a point where the signal is already multiplexed. It will be shown that if the noise extends from the lowest to the highest frequencies in the multiplex band and is of the proper level, it will nearly simulate the normal multichannel system load with each channel intermittently active. In order to measure the intermodulation noise (which will be indistinguishable from the applied noise), it is necessary to maintain some of the channels quiet. This is generally accomplished by following the noise generator with some sharp band-rejection filters before applying the noise to the system. To create quiet channels, notches can be placed within the multiplex band, as long as they do not occupy a significant portion of the total bandwidth. Assume for this discussion that each notch is 4 kHz wide. To measure the total intermodulation noise, the total power in each quiet channel is monitored at the output of the system. The total noise power observed is a combination of intermodulation noise and thermal noise. The two components of the total noise can be easily separated by disabling the noise source and noting the thermal noise. From these two quantities, the intermodulation noise can be readily determined. To determine whether the intermodulation noise is second order or third order (or higher), the effect of noise loading power on intermodulation noise must be noted. For example, if the intermodulation noise changes 2 dB for every dB change of input power, it is second order noise; a 3-dB change per dB input change signifies third order noise. Finally, the nature of the dominant modulation products can be determined by noting the frequency shape of the intermodulation noise (assuming that the modulation performance is flat with frequency) and using Fig. 10-6. For example, second order noise in the lower end of the band is $\alpha - \beta$ type noise, and that in the upper part of the band is $\alpha + \beta$ type noise. Third order noise increasing with decreasing frequency is probably $\alpha - \beta - \gamma$. Note from Fig. 10-6 that third order $\alpha - 2\beta$ type products also increase with decreasing frequency, but the ordinate in this case is U/N rather than U/N^2 as for the $\alpha - \beta - \gamma$ type products. If N is large, the $\alpha - \beta - \gamma$ type products will predominate.

Analysis. In Chap. 9 the average busy hour talker load on a system was established as

$$P_{av} = V_0 + 10 \log \tau_L N + 0.115\sigma^2 - 1.4 = P_s - \Delta_c \quad \text{dBm0} \quad (10\text{-}30)$$

For equivalent average power, the applied noise load on the system should then be as given, with equal noise power applied to each

channel (i.e., the noise applied at 0 TLP is flat over the multiplex band). The noise power applied to each 4-kHz channel is then given by

$$N_c = P_{av} - 10 \log N$$

$$= V_0 + 0.115\sigma^2 - 1.4 + 10 \log \tau_L \qquad (10\text{-}31)$$

The power of a single x-type intermodulation product is given by

$$R'_x = H_x^* + \eta_x N_c \qquad \text{dBm0} \qquad (10\text{-}32)$$

In this case, unlike the previous ones, all channels are active so that the number of x-type products falling in a channel is the total possible number, U_x. Therefore,

$$R'_{x_T} = H_x^* + \eta_x N_c + 10 \log U_x \qquad \text{dBm0} \qquad (10\text{-}33)$$

Combining all of these results yields

$$R'_{x_T} = \eta_x V_0 + 0.115\eta_x\sigma^2 - 1.4\eta_x + 10 \log \tau_L^{\eta_x}$$
$$+ H_x + \chi_x + 10 \log U_x \qquad \text{dBm0} \qquad (10\text{-}34)$$

Comparing Eq. (10-34) with Eq. (10-26) reveals any differences between the analytical computations and the noise loading technique.

Subtracting Eq. (10-34) from Eq. (10-26) and assuming $\tau_L = \tau$ reveals

$$R_{x_T} - R'_{x_T} = 0.115(\lambda_x - \eta_x)\sigma^2 - 10(\eta_x - \mu_x) \log \tau \qquad \text{dB} \qquad (10\text{-}35)$$

The values of λ_x, η_x, and μ_x are tabulated in Fig. 10-8 for the normally important products. For all but the $2\alpha - \beta$ products the difference given by Eq. (10-35) is zero since $\lambda = \eta = \mu$. Evaluating Eq. (10-35) for $2\alpha - \beta$ products shows

$$R_{2\alpha-\beta_T} - R'_{2\alpha-\beta_T} = 0.23\sigma^2 - 10 \log \tau \qquad \text{dB} \qquad (10\text{-}36)$$

For a typical case of $\sigma = 5$ and $\tau = 0.25$, this error amounts to 11.75 dB, which seems to be significant. However, by referring again to Fig. 10-6, it becomes apparent that $2\alpha - \beta$ products are quickly outnumbered by many of the other third order products because of the

x	λ_x	η_x	μ_x
$\alpha - \beta$	2	2	2
$\alpha + \beta$	2	2	2
$\alpha + \beta - \gamma$	3	3	3
$2\alpha - \beta$	5	3	2

FIG. 10-8. Factors for important x-type modulation products.

N rather than the N^2 factor in the denominator of the ordinate. Therefore, the error in the $2\alpha - \beta$ product will usually be insignificant if the number of channels is large. Since noise loading is usually only used on broadband systems, the results should be compatible with calculations and this difference can be ignored.

Finally, to convert the noise loading intermodulation powers in dBm0, to noise in dBrnc0, use must be made of the correction factor, C_{w_x}, previously discussed. There may be some doubt as to which x-type product applies. If the intermodulation appears to be second order, $C_{w_2} = 2.25$ dB. If the intermodulation noise is third order and the system has a large number of channels, the correction for $\alpha + \beta - \gamma$ is $C_{w_3} = 2.75$ dB. If second and third order cannot be distinguished, little significant error results from assuming $C_w = 2.5$ dB.

Intermodulation noise of R'_t dBm0 can be converted to dBrnc0 by:

$$\begin{aligned} W_2 &= R'_t + 88 - 2.25 \quad \text{dBrnc0} \\ W_3 &= R'_t + 88 - 2.75 \quad \text{dBrnc0} \\ W_t &= R'_t + 88 - 2.5 \quad \text{dBrnc0} \end{aligned} \tag{10-37}$$

Measurement. The preceding analysis assumed the presence of ideal 4-kHz notch and bandpass filters. Such filters are not practical, so it is of interest to investigate effects of using more practical filter shapes. Since the function of the notch filter at the noise source output is to remove the applied noise from the channel to be measured, the only penalty in making the notch wider than required is the increased error in the assumption that all channels are loaded. The test of whether the notch filter at the source is wide enough and the bandpass filter at the detector narrow enough is to connect them in tandem

and observe the residual noise reading. Practical noise loading equipment usually shows the measured noise in the notch to be at least 80 dB below the applied signal power measured with the notch filter removed. If the equivalent noise bandwidth of the detector is 4 kHz, all of the previous noise loading results can be applied directly.

It is of interest to consider means of obtaining useful noise loading results with detectors having equivalent noise bandwidth of other than 4 kHz. Of course, if the equivalent noise bandwidth, B_w, of the detector is known, the necessary correction for a 4-kHz bandwidth can be obtained by adding the correction factor, 10 log $4000/B_w$, to the detector reading. Note that if the detector is not a true rms reading instrument, a correction factor for gaussian noise must be used as described in Chap. 7.

The use of these detector correction factors can be avoided by the use of *noise power ratio* (NPR). This technique is particularly useful when the multiplex band is shaped with frequency since conversion of the results to 0 TLP is simple. Noise loading measurements require that the noise load applied to the system be N_c dBm0 per 4 kHz. At the output of the system, which need not be at 0 TLP and can be shaped with frequency, it is necessary to take three readings with the detector:

1. The noise load on the system, i.e., with no rejection filter after the loading source.
2. The noise appearing in a quiet slot, i.e., with a slot rejection filter between the source and input.
3. The noise with drive removed, i.e., with the noise source turned off.

The ratio of reading (1) to reading (2) expressed in dB is called NPR_1. This ratio corresponds to thermal plus intermodulation noise and is given by

$$P_N \text{ "+" } R'_t = N_c - \text{NPR}_1 \qquad \text{dBm0} \tag{10-38}$$

Similarly, the ratio of reading (1) to reading (3) is called NPR_2 and is a measure of thermal noise per channel by the relationship

$$P_N = N_c - \text{NPR}_2 \qquad \text{dBm0} \tag{10-39}$$

Using Eqs. (10-38) and (10-39), the value of intermodulation noise, R'_t, in dBm0 per 4-kHz channel can be determined. To convert to

dBrnc0, it is only necessary to make use of Eq. (10-37). Note that since ratios are used, neither the bandwidth nor type of detector is important.

Another approach, which is often more easily understood and gives equivalent results, consists of converting the units for the load to dBrnc0. The channel load, N_c, can be converted to dBrnc0 by adding 88 dB and the bandwidth factor of 1.25 dB. For example, if $N_c = -16$ dBm0 per 4 kHz, as is standard for most domestic systems, a signal load of approximately 71 dBrnc0 is obtained. The thermal noise per channel is then given by

$$W_N = 71 - \text{NPR}_2 \qquad \text{dBrnc0} \tag{10-40}$$

If the spectrum shape factor is ignored, the total noise per channel is given by

$$W_t = 71 - \text{NPR}_1 \qquad \text{dBrnc0} \tag{10-41}$$

The spectrum shape factor can be applied by extracting the components of the intermodulation noise and by lowering the second order noise 1 dB and the third order noise 1.5 dB. However, many designers choose to ignore this effect.

Intermodulation Noise Computed from Spectral Densities

If the multiplexed signal is assumed to be represented by a band of noise as is done in the noise loading technique, it is possible to compute directly the intermodulation spectrum generated by passing this noise loading signal through a nonlinear device. The advantage of this technique is that a single expression can be derived for all second order intermodulation noise. Likewise, one can be derived for all third order intermodulation noise. More important, in some cases the intermodulation performance can be computed analytically even if the signal is shaped with frequency. As a result, this technique is becoming very popular.

Consider again the nonlinear voltage transfer characteristic given by Eq. (10-1), ignoring terms higher than third order:

$$e_o = a_1 e_i + a_2 e_i^2 + a_3 e_i^3 \tag{10-42}$$

In this case, the input signal is a band of noise whose spectral density is denoted by $S_1(f)$ which represents the multiplex load applied to the system. In the notation commonly used, it is assumed that $S_1(f)$

is characterized by volts squared per hertz at the *input* port of the nonlinear device having the nonlinear transfer characteristic. The desired spectrum at the output of the nonlinear two-port will be given by

$$a_1^2 S_1(f) \tag{10-43}$$

The spectrum at the output corresponding to the e_i^2 term will correspond to all of the second order intermodulation noise and will be denoted by $a_2^2 S_2(f)$. Similarly, third order intermodulation noise due to the cubed input voltage will be denoted by $a_3^2 S_3(f)$.

By the use of autocorrelation functions, $S_2(f)$ and $S_3(f)$ can be determined. If it is assumed that $e_i(t)$ is gaussian, it can be shown that the autocorrelation functions of $e_i^2(t)$ and $e_i^3(t)$ can be written in terms of the autocorrelation function $\mathcal{R}_1(\tau)$ of $e_i(t)$. For example, the autocorrelation function $\mathcal{R}_2(\tau)$ of $e_i^2(t)$ is given by [6]

$$\mathcal{R}_2(\tau) = 2\ \mathcal{R}_1^2(\tau) + \mathcal{R}_1^2(0) \tag{10-44}$$

Since spectral density and the autocorrelation function are a Fourier transform pair,

$$\begin{aligned} S_n(f) &= \int_{-\infty}^{\infty} \mathcal{R}_n(\tau)\ e^{-j2\pi f\tau}\, d\tau \\ \mathcal{R}_n(\tau) &= \int_{-\infty}^{\infty} S_n(f)\ e^{j2\pi f\tau}\, df \end{aligned} \tag{10-45}$$

Equation (10-45) can be used to express Eq. (10-44) in the frequency domain as

$$\begin{aligned} S_2(f) &= \int_{-\infty}^{\infty} [\,\mathcal{R}_1^2(0) + 2\,\mathcal{R}_1^2(\tau)\,]\ e^{-j2\pi f\tau}\, d\tau \\ &= \left[\int_{-\infty}^{\infty} S_1(f)\ df\right]^2 \delta(f) + 2\int_{-\infty}^{\infty} S_1(f')S_1(f-f')\ df \end{aligned} \tag{10-46}$$

Use has been made of the fact that the transform of $\mathcal{R}_1^2(\tau)$ may be expressed as the convolution of $S_1(f)$ with itself. Note also that the mean square voltage of $e_i(t)$ is given by

$$V^2_{\text{rms}} = \int_{-\infty}^{\infty} S_1(f)\ df \tag{10-47}$$

To illustrate a specific case, assume that the input spectral density, $S_1(f)$, is flat from 0 to f_T hertz as shown in Fig. 10-9(a). The total area under the $S_1(f)$ curve is equal to V^2_{rms}.

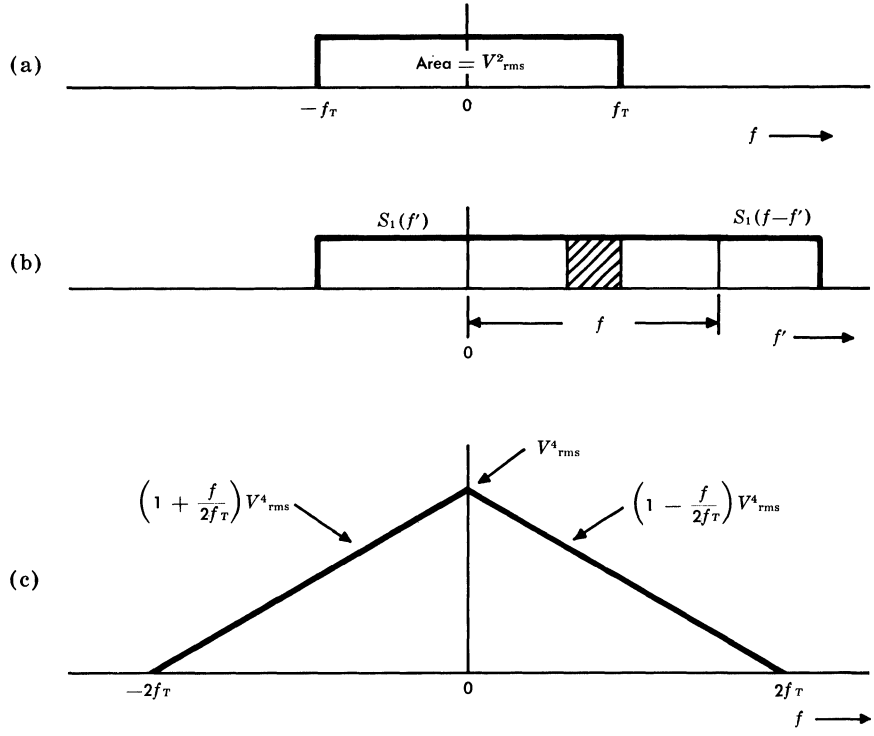


FIG. 10-9. Convolution of a flat spectrum with itself.

Evaluation of the convolution integral of Eq. (10-46) is most easily accomplished graphically as shown in Fig. 10-9(b). The value of the convolution integral at any particular value of f is simply proportional to the overlapping or shaded area in Fig. 10-9(b) [7]. For $|f| > 2f_T$, the value is zero since $S(f')$ and $S(f-f')$ do not overlap in this condition and the product is zero. At $f = 0$, the convolution integral is given by the density squared times the total width, or is simply $V^4_{\text{rms}}/2f_T$. From $f = 0$ to $|f| = 2f_T$, the overlapping area will decrease linearly to zero so

$$\int_{-\infty}^{\infty} S_1(f')\, S_1(f-f')\, df' = 0 \qquad |f| > 2f_T$$

$$= \left(1 - \frac{|f|}{2f_T}\right) V^4_{\text{rms}}/2f_T \qquad |f| \leq 2f_T$$

This is shown in Fig. 10-9(c). Thus, Eq. (10-46) can be rewritten for $|f| \leq 2f_T$

$$S_2(f) = V^4_{\text{rms}}\,\delta(f) + 2\left(1 - \frac{|f|}{2f_T}\right) V^4_{\text{rms}}/2f_T \qquad (10\text{-}48)$$

To be consistent with previous discussions, the signal and intermodulation power can be referred to the output of the network. If the total output power of the desired spectrum is p_f milliwatts,

$$a_1^2 S_1(f) = \frac{p_f R_L}{2f_T\,(1000)} = \frac{a_1^2 V^2_{\text{rms}}}{2f_T} \qquad (10\text{-}49)$$

The ratio of the distortion to the desired spectral density at the output (ignoring the d-c term of the distortion) is

$$\frac{a_2^2 S_2(f)}{a_1^2 S_1(f)} = \frac{2a_2^2\left(1 - \dfrac{|f|}{2f_T}\right)\left(\dfrac{p_f R_L}{1000\,a_1^2}\right)^2\left(\dfrac{1}{2f_T}\right)}{\dfrac{p_f R_L}{2f_T\,(1000)}}$$

$$= \left[\frac{a_2^2}{2a_1^4}\frac{R_L}{1000}\right] 4p_f\left(1 - \frac{|f|}{2f_T}\right) \qquad |f| \leq f_T \qquad (10\text{-}50)$$

Although the intermodulation noise spectrum and the desired signal spectrum in Eq. (10-50) are two-sided, the *ratio* is equally valid for one-sided spectra, which permits dropping the absolute value signs for frequency. Also, since the frequencies of interest are inband ($f < f_T$), the following equations are given only for the inband frequency range. From Eq. (10-11), the term in brackets can be seen to be simply m_2, so

$$\frac{a_2^2 S_2(f)}{a_1^2 S_1(f)} = 4m_2 p_f\left(1 - \frac{f}{2f_T}\right) \qquad (10\text{-}51)$$

Therefore at 0 TLP for a single nonlinear characteristic, the ratio of the second order interference to the noise load on a per hertz basis is, in dB,

$$M_2 + 6 + P_f + 10\log\left(1 - \frac{f}{2f_T}\right) \qquad f \leq f_T \qquad (10\text{-}52)$$

This is identical to the noise power ratio previously discussed. The total second order noise power in a 4-kHz band at 0 TLP due to a

single second order nonlinearity is

$$R'_2 = N_C + M_2 + 6 + P_f + 10 \log \left(1 - \frac{f}{2f_T}\right) \quad \text{dBm0} \tag{10-53}$$

where N_C is the noise load per channel in dBm0.

Since P_f is at the modulation reference point, it can be referred to 0 TLP by the factor C previously defined so that

$$P_{av} = P_f + C \tag{10-54}$$

Since all N channels are noise loaded, N_C is related to P_{av} by the number of channels, N:

$$N_C = P_{av} - 10 \log N \tag{10-55}$$

Thus,

$$R'_2 = P_{av} - 10 \log N + M_2 + 6 + P_{av} - C + 10 \log \left(1 - \frac{f}{2f_T}\right)$$

Since $M_2 - C$ refers the modulation performance to 0 TLP and can be replaced by $H_{2\alpha}$,

$$R'_2 = 2P_{av} + 6 + H_{2\alpha} - 10 \log N + 10 \log \left(1 - \frac{f}{2f_T}\right)$$

From Eq. (10-30)

$$P_{av} = V_0 + 0.115\sigma^2 - 1.4 + 10 \log \tau_L N \quad \text{dBm}$$

Thus,

$$R'_2 = 2(V_0 + 0.115\sigma^2 - 1.4) + H_{2\alpha} + 6 + 10 \log N + 10 \log \left(1 - \frac{f}{2f_T}\right) + 20 \log \tau_L \quad \text{dBm} \tag{10-56}$$

This expression gives the noise power in a 4-kHz channel for all second order intermodulation. To convert to annoyance in dBrnc0 at the system output:

$$W_2 = R'_2 + 88 - C_{w_2}$$

$$W_2 = R'_2 + 88 - 2.25 \quad \text{dBrnc0} \tag{10-57}$$

It should be noted that the product amplitude factor of C_{w_x} does not apply when spectral analysis is used. Since this factor is zero for the second order products, the previously determined value for C_{w_x} is correct.

By reasoning similar to that done for second order noise, the relations for third order noise can also be developed. In this case, the autocorrelation function corresponding to cubing a gaussian voltage is given by

$$\Re_3(\tau) = 9\, V^4_{\text{rms}}\, \Re_1(\tau) + 6\Re_1{}^3(\tau) \qquad (10\text{-}58)$$

The total third order noise due to noise loading the system can be related to the other system parameters by

$$R'_3 = 3\,(V_0 + 0.115\sigma^2 - 1.4) + H_{3\alpha} + 12.6$$
$$+ 20 \log N + 30 \log \tau_L + 10 \log\left(1 - \frac{f^2}{3f_T{}^2}\right) \quad \text{dBm0} \qquad (10\text{-}59)$$

Or,

$$W_3 = R_3 + 88 - C_{w_3} \qquad \text{dBrnc0} \qquad (10\text{-}60)$$

where $C_{w_3} = 2.7$ dB.

Differential Gain and Phase

Another common means of denoting nonlinearity of a two-port is by differential gain and phase. This is often desirable in situations where the signal is relatively "fragile" and occupies a wide frequency range in octaves such as video or very high quality audio facilities.

Differential gain is defined as the difference in gain encountered by a low-level, high-frequency sinusoid at two stated instantaneous amplitudes of a superimposed low-frequency signal.

The two stated amplitudes of the low-frequency signal are usually zero and the peak signal normally handled by the network. The low-level, high-frequency signal is usually much lower in level (by tens of dB) than the high-level, low-frequency signal.

Differential phase is defined as the difference in phase shift encountered by a low-level, high-frequency sinusoid at two stated instantaneous amplitudes of a low-frequency signal.

The levels and frequencies applicable are the same as those used for differential gain.

To illustrate a possible cause of differential gain (or phase), consider application of both a low-frequency, high-level and a high-frequency, low-level sinusoid to a two-port network. If the two-port exhibits significant compression at the peaks of the high-level, low-frequency signal, it is apparent that the high-frequency sinusoid

will exhibit less gain at times corresponding to these peaks than it will when the low-frequency signal is passing through zero amplitude. If the high-frequency signal is filtered from the low-frequency component, the waveform of Fig. 10-10 results.

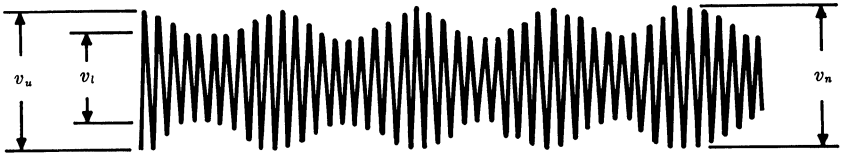


FIG. 10-10. Envelope of high-frequency signal subjected to differential gain from a low-frequency sinusoid.

This is simply the waveform of an amplitude-modulated wave. Examination of the figure reveals that if the compression is symmetrical for both the positive and negative peaks of the low-frequency signal, the sidebands will be at twice the low frequency. If the compression is not symmetrical, sideband components equal to the low frequency will also be present. A quantitative measure of the differential gain may be obtained from Fig. 10-10 where v_n is the normal peak-to-peak voltage of the high-frequency sinusoid with no low-frequency signal applied. Application of the low-frequency signal will result in a minimum value of the envelope given by v_l and a maximum value given by v_u. In this simple case, $v_u = v_n$ since all distortion is due to compression of the high-amplitude signal. Differential gain is quantitatively defined as follows:

$$\text{Positive diff gain} = 20 \log \frac{v_u}{v_n}$$
$$\text{Negative diff gain} = 20 \log \frac{v_n}{v_l} \tag{10-61}$$

The larger of these two quantities is the differential gain of the two-port.

Letting β represent the low frequency and α the high frequency, the spectral diagram of Fig. 10-11 can be drawn. All of the components near the frequency α can be drawn as phasors in a manner similar to that used in Chap. 5. Such a diagram is shown in Fig. 10-12. The phasor (1) corresponds to the component at α. Note that because of compression, the amplitude of this phasor is not identical to the amplitude of the α component when β is not applied.

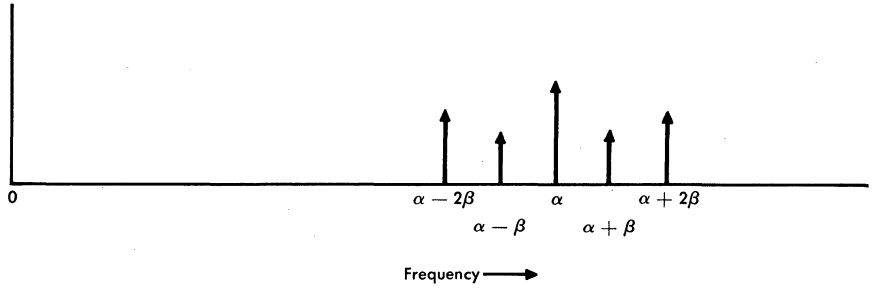


FIG. 10-11. Spectrum of the waveform shown in Fig. 10-10.

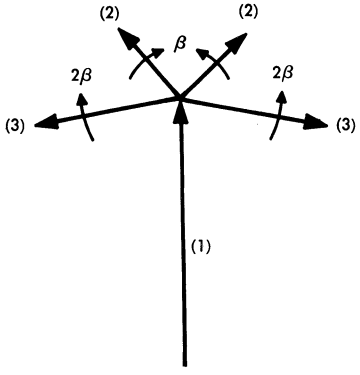


FIG. 10-12. Phasor diagram of waveform shown in Fig. 10-10 for all differential gain.

The phasors (2) are counter-rotating at the frequency β and correspond to $\alpha \pm \beta$ products of the α and β frequencies. The vectors (3) are also counter-rotating but at the frequency 2β and correspond to $\alpha \pm 2\beta$ products of the α and β frequencies. Note that in this diagram, the phases are such that the resultant maintains constant phase. This is not true in general, but in this case it means that all distortion is due to differential gain, and there is no differential phase. Also, the phase relationship between the (2) and (3) sets of phasors is such that they all add directly when pointing upward but not at any other time (including pointing downward). The maximum value of the resultant phasor corresponds to v_u and the minimum value to v_l. From these, the differential gain can be determined if the amplitudes and phases of all of these components are known.

The phasor diagram of Fig. 10-12 could be redrawn with different phase relationships as shown in Fig. 10-13. In this case, almost all of the distortion shows up as differential phase instead of gain since the resultant of these five phasors has a near constant amplitude but has a varying phase. The resultant of phasor (2) and phasor (3) is always perpendicular to phasor (1). The phase deviation in radians is given approximately by this resultant phasor

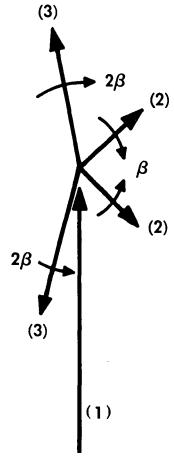


FIG. 10-13. Effect of shifting phases of Fig. 10-12 to obtain differential phase.

divided by phasor (1). The maximum possible value of phase deviation is the differential phase in radians. Multiplying by 57.3 gives the differential phase in degrees.

Example 10.1

Problem

Consider a television transmission system that must meet certain differential gain and phase requirements. Assume the index of system modulation is given by $H_{2\alpha} = -60$ dB and $H_{3\alpha} = -40$ dB, both flat with frequency over the video spectrum and measured across 100 ohms at the system output. The problem is to estimate the differential gain and phase of this system, which would be realized by a 1.4 volt peak-to-peak 15 kHz low-frequency sinusoid and a —20 dBV 3.6 MHz high-frequency sinusoid. Since the indices of modulation do not give phase information, the approach will be to assume worst case conditions for both gain and phase.

Solution

The first step is to compute the magnitudes of components ± 15 kHz and ± 30 kHz from 3.6 MHz. Each of the ± 15 kHz components will

have a power given by $P_{\alpha\pm\beta}$ of Eq. (10-13)

$$P_{\alpha\pm\beta} = M_2 + 6 + P_\alpha + P_\beta$$

P_β is due to 1.4 volt peak-to-peak or 0.5 volt rms across 100 ohms, i.e., $P_\beta = 4$ dBm. Similarly, P_α is due to -20 dBV $= -10$ dBm. Thus, $P_{\alpha\pm\beta} = -60 + 6 - 10 + 4 = -60$ dBm. Each of the ± 30 kHz components is due to $\alpha \pm 2\beta$ and has a power given by:

$$P_{\alpha\pm2\beta} = M_3 + 9.6 + P_\alpha + 2P_\beta$$

or

$$P_{\alpha\pm2\beta} = -40 + 9.6 - 10 + 8 = -32.4 \quad \text{dBm}$$

Since the $P_{\alpha\pm2\beta}$ components are much larger than the $P_{\alpha\pm\beta}$, the latter can be ignored. The resulting phasor diagram of the voltages (assuming maximum differential gain) is shown in Fig. 10-14.

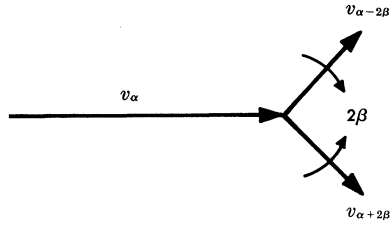


FIG. 10-14. Phasor diagram for computing differential gain.

The maximum value of this vector is given by

$$P_\alpha \quad \text{"+"} \quad P_{\alpha+2\beta} \quad \text{"+"} \quad P_{\alpha-2\beta}$$

Since $P_{\alpha-2\beta} = P_{\alpha+2\beta} = -32.4$ dBm, the last two terms are easily combined by adding 6 dB. If compression is ignored,

$$\text{Diff gain} = (P_\alpha \quad \text{"+"} \quad -26.4) - P_\alpha$$

where

$$P_\alpha = -10 \text{ dBm}$$

Using the voltage addition charts (p. 36) results in:

$$\text{Diff gain} = -8.8 + 10 = 1.2 \text{ dB}$$

If compression cannot be ignored, the differential gain could be as much as twice this number, or 2.4 dB.

If instead, only differential phase is assumed, the phasor diagram of Fig. 10-15 applies. In this case, the maximum phase shift is equal to

$$\arctan\left(\frac{v_{\alpha-2\beta} + v_{\alpha+2\beta}}{v_\alpha}\right)$$

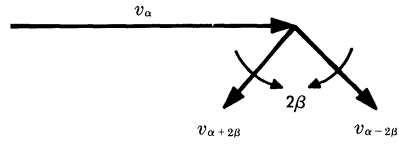


FIG. 10-15. Phasor diagram for computing differential phase.

But $(v_{\alpha-2\beta} + v_{\alpha+2\beta})$ corresponds to $(P_{\alpha-2\beta}$ "+" $P_{\alpha+2\beta})$ or −26.5 dBm, and v_α corresponds to $P_\alpha = -10$ dBm. Thus, the above ratio can be rewritten as

$$\frac{v_{\alpha-2\beta} + v_{\alpha+2\beta}}{v_\alpha} = \text{antilog}\left(\frac{-26.5 + 10}{20}\right) = 0.151$$

The differential phase is thus

$$\arctan\,(0.151) = 8.5 \text{ degrees}$$

Based on the modulation indices, the above system could have differential gain of as much as 2.4 dB *or* differential phase of up to 8.5 degrees.

Measurement. A typical scheme for measuring differential gain or phase of a commercial video system is shown in Fig. 10-16. The amplitude of the 15-kHz signal is adjusted to swing from the reference white to black levels to cover the full signal swing and produce maximum differential gain and phase. The waveform presented on the oscilloscope gives a direct reading of phase shift or gain deviation of the 3.6-MHz signal as a function of the instantaneous amplitude of the 15-kHz signal which drives the horizontal deflection of the oscilloscope. The differential gain or phase can be computed from this reading and knowledge of the detector parameters.

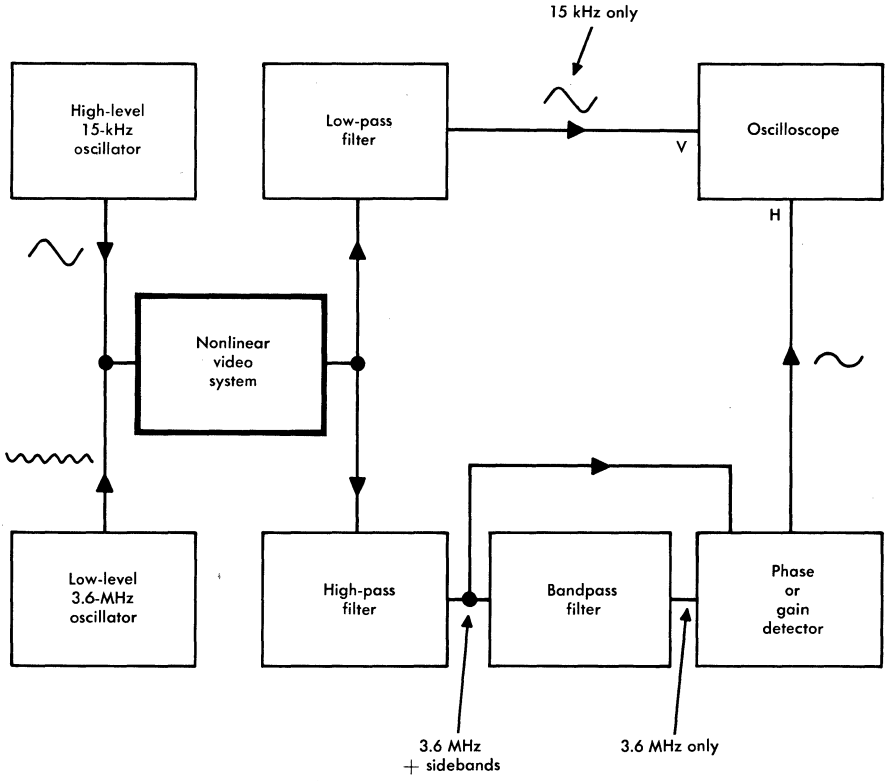


FIG. 10-16. Measurement of differential gain and phase of a video transmission system.

REFERENCES

1. Maurer, R. E. and S. Narayanan. "Noise Loading Analysis of a Third-Order Nonlinear System with Memory," *IEEE Trans. on Comm. Tech.*, vol. COM-16 (Oct. 1968), p. 701.
2. Cross, T. G. "Intermodulation Noise in FM Systems Due to Transmission Deviations and AM/PM Conversion," *Bell System Tech. J.*, vol. 45 (Dec. 1966), pp. 1749-1773.
3. Ketchledge, R. W. "Distortion in Feedback Amplifiers," *Bell System Tech. J.*, vol. 34 (Nov. 1955), pp. 1265-1285.
4. Mulligan, J. H. "Amplitude Distortion in Transistor Feedback Amplifiers," *Trans. AIEE, Communications and Electronics*, part I, vol. 80 (July 1961), pp. 326-335.
5. Bennett, W. R. "Cross Modulation Requirements on Multichannel Amplifiers Below Overload," *Bell System Tech. J.*, vol. 19 (Oct. 1940), pp. 587-610.
6. Laning, J. H., Jr. and R. H. Battin. *Random Processes in Automatic Control* (New York: McGraw-Hill Book Company, Inc., 1956), pp. 166-167.
7. Bracewell, R. M. *The Fourier Transform and Its Applications* (New York: McGraw-Hill Book Company, Inc., 1965), Chapter 3.

Chapter 11

Crosstalk

Crosstalk is defined as the disturbance created in one communication circuit by the signals in other communication circuits. It is properly, therefore, a subdivision of the general subject of interference. Because of its importance, crosstalk has been singled out for separate and more detailed discussion in this chapter.

The term *crosstalk* was originally coined to indicate the presence in a telephone receiver of unwanted speech sounds from another telephone conversation [1]. The methods which were developed for computing and measuring crosstalk quantitatively were soon found to be useful in studying interference between nontelephone circuits. Therefore, the term has been gradually broadened to apply to interference between any kinds of communication circuits.

The nature of the interfering crosstalk is often described as either intelligible or unintelligible. Crosstalk between unlike channels (for example, different types of carrier facilities) is usually unintelligible because of frequency inversion, frequency displacement, or digital encoding. Such crosstalk often retains the syllabic pattern of speech and is more annoying than a steadier noise (such as thermal noise) of the same average power. Also, undesired intermodulation products in an FDM multichannel telephone system are a form of interchannel crosstalk that is usually unintelligible but sometimes has a recognizable syllabic pattern. In most of these cases of unintelligible crosstalk, the interference is generally grouped with other noise-type interferences.

When crosstalk interference is intelligible (or nearly intelligible), it is particularly annoying and also objectionable because of a real or fancied loss of privacy. Stringent requirements are necessary

to maintain a low probability that a customer will hear a "foreign" conversation. Crosstalk objectives are discussed in Chap. 3.

The use of the words *intelligible* and *unintelligible* can also be applied to nonvoice circuits. In such cases, intelligible implies that the crosstalk interference is of the same type as the desired signal and thus could be amplified and decoded if the desired signal were absent. Unintelligible crosstalk for nonvoice circuits results from crosstalk between different types of systems. As a consequence, it is usually considered noise rather than unintelligible crosstalk.

There are three basic causes of crosstalk in communications systems. One, of course, is nonlinear performance within an FDM system. Thus, intermodulation products can often be considered crosstalk. A second cause is poor control of frequency response. This type can result from poor design of the filters in FDM terminals or from a poor match between pulse shapes and frequency-response in TDM systems. The third and most obvious crosstalk cause is electrical coupling between various transmission media. For example, crosstalk coupling will exist between various twisted pairs of a multipair telephone cable or between coaxials of a multicoaxial cable. Although the emphasis in this chapter is on coupling crosstalk, a brief discussion of the other two causes of crosstalk is in order.

11.1 NONLINEAR CROSSTALK

As the name implies, nonlinear crosstalk is a direct result of the presence of nonlinearities in an analog system. Such interference was discussed in detail in the previous chapter as intermodulation noise and is usually considered as such rather than as unintelligible crosstalk. However, when the undesired intermodulation product is intelligible, it is often referred to as crosstalk and must meet more stringent requirements than unintelligible intermodulation noise. It is of interest to discuss briefly some of the possible ways that nonlinearities may cause intelligible crosstalk.

Pilots consisting of single-frequency sinusoids are often sent along with the frequency division multiplexed signal for a number of purposes as discussed in Chap. 6. If the frequency allocation of the pilots is not carefully engineered, nonlinearities produce intermodulation products between pilots and speech signals, which may be intelligible. If the spectrum of a particular channel is translated to coincide with another channel without inversion, intelligible crosstalk will result

in the demodulated output. If both the pilots and the carrier frequencies are multiples of 4 kHz, this type of intelligible crosstalk can occur.

A more subtle nonlinear crosstalk path can result from a single-frequency tone applied to one of the channels of the multiplexed load. For example, consider a 2-kHz tone applied to a channel which is subsequently multiplexed with a carrier which is a multiple of 4 kHz. The resulting single-frequency signal will be at a frequency of $\alpha = 4n \pm 2$ kHz. This signal can interact with a speech signal of the multiplexed load to form a $2\alpha \pm \beta$ product which may be intelligible in another channel.

If the nonlinear crosstalk coincides with another channel but is inverted in frequency, the crosstalk is unintelligible. However, subjective tests have shown that inverted crosstalk, although not as annoying as intelligible crosstalk, is more annoying than random intermodulation noise. This is due to the syllabic pattern of such unintelligible crosstalk. Results of the subjective tests have shown such inverted crosstalk spectrums to be between 6 and 12 dB more annoying than random noise of the same power, depending upon the level of the crosstalk (the greater difference corresponding to the higher crosstalk powers). In large systems this effect can be masked by random noise and the random nature of many components of intermodulation noise.

If the pilots and frequency allocations of an FDM signal are judiciously chosen, the possibility of intelligible crosstalk will be reduced. However, unintelligible crosstalk due to nonlinearities is very likely to appear as otherwise intelligible products shifted by a fixed frequency, e.g., 1 kHz. Such a shift in frequency of the crosstalk spectrum is called *staggering* and, for a given disturbing talker volume, may reduce the effective crosstalk output power in the disturbed channel because not all of the disturbing speech energy frequency spectrum will fall into the disturbed channel. The reduction in power, in dB, is called the staggering advantage and may be between 0 and 10 dB in practical systems.

In summary, intermodulation products in frequency division multiplexed analog systems can give rise to crosstalk coupling between channels of the multiplex group. It can be distinguished from the other types of crosstalk because the crosstalk coupling (ratio of disturbing channel power to interference power in the disturbed channel) measured for nonlinear crosstalk will be a function of signal

level in the disturbing channel. This means can also be used to determine the order of nonlinearity causing an observed crosstalk interaction.

11.2 TRANSMITTANCE CROSSTALK

Transmittance crosstalk is that interference caused by inadequate control of the frequency response of various elements of a transmission system. An obvious example is in the design of FDM equipment, i.e., combinations of modulators and filters. If the filters do not adequately reject undesired products from the modulators, these products may produce crosstalk in other channels of the multiplex load. Such crosstalk may be either intelligible or unintelligible and may affect individual channels or large numbers of channels. Like nonlinear crosstalk, the crosstalk spectrum may be inverted and/or shifted in frequency depending upon the design of the modem. The control of such crosstalk paths determines the stop band loss requirements of many of the filters used in modems. Because it results from inadequate control of the transfer characteristic or the transmittance of networks, the crosstalk is called transmittance crosstalk.

Another form of transmittance crosstalk occurs as interchannel crosstalk in TDM systems when each sample is not confined to its assigned time slot. This results in what is often referred to as intersymbol interference. Insufficient bandwidth of the common transmission path will cause each sample pulse of the TDM signal to smear into neighboring time slots, rather than be confined to its assigned time slot. The magnitude of the resulting interchannel crosstalk is dependent on the upper and lower cutoff frequencies of the transmission path.

Attenuation and phase distortion at high frequencies will prevent a pulse of one channel from decaying to zero before the time slot of the next channel is ready to receive its pulse. On the other hand, distortion at low frequencies causes d-c level variations which may result in crosstalk from one channel into all the other channels. It is therefore necessary for the crosstalk caused by low-frequency distortion to be controlled more than that caused by high-frequency distortion, usually appreciable only between adjacent channels. The low-frequency cutoff of the common TDM transmission path must therefore be considerably lower than the lowest modulating frequency employed.

It should be noted that PCM systems are much less susceptible to such crosstalk than other TDM systems because a code word, corresponding to a single-channel sample, is sent as a sequence of pulses. Thus, it is not as probable that adjacent pulses belong to different channels as is the case in other TDM systems. Intersymbol interference in PCM systems (after coding) will not result in intelligible crosstalk unless the individual pulses of the code are interleaved with similar pulses from other channels. Further discussion of intersymbol interference in digital systems is in later chapters.

11.3 COUPLING CROSSTALK

Coupling crosstalk is caused by electromagnetic coupling between physically isolated circuits. The most common coupling is due to near-field effects and can usually be characterized by mutual inductance and direct capacitance. This can best be illustrated by considering two parallel balanced transmission paths as shown in Fig. 11-1. It will be initially assumed that the transmission paths are short in terms of the wavelengths of the frequencies of interest. This allows the coupling effects to be characterized by lumped element capacitors and transformers as shown and allows the use of simple

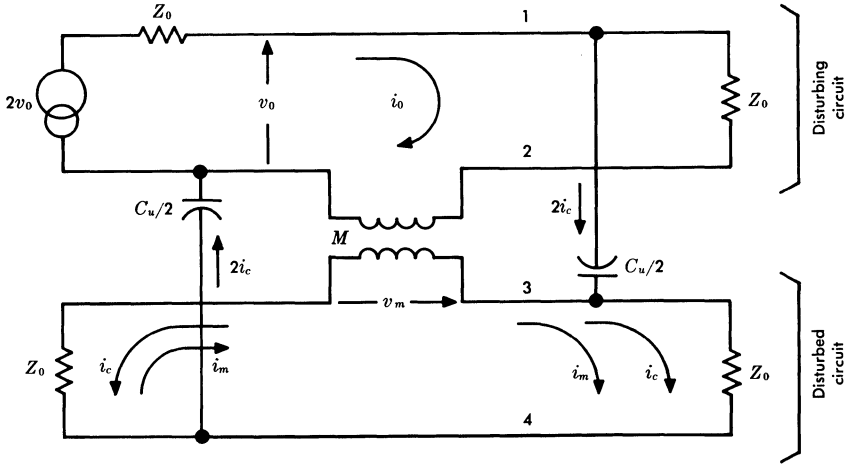


FIG. 11-1. Inductive and capacitive coupling between two circuits.

lumped element network equations. The two capacitors, shown as $C_u/2$, characterize the unbalance capacitance, C_u, between the circuits [2]. The components from which C_u is derived include all the direct capacitances between the wires of Fig. 11-1. Letting C_{ij} represent the direct capacitance between wires i and j results in

$$C_u = C_{13} + C_{24} - C_{14} - C_{23} \tag{11-1}$$

Equation (11-1) ignores all higher order terms in C_u, which result from capacitances to ground, etc.

The current i_0 in the disturbing circuit causes a voltage, $v_0 = i_0 Z_0$, to appear across this circuit. This voltage, $i_0 Z_0$, across the disturbing circuit drives a current, $i_0 Z_0 j\omega C_u/4$, through the two capacitors in series. (This assumes that $Z_0/2$ is much less than the impedance of the two capacitors in series and that the mutual inductance effect is ignored.) Half of this current appears at each end of the disturbed circuit as shown by i_c where

$$i_c = \frac{i_0 Z_0 j\omega C_u}{8} \tag{11-2}$$

The effects of the mutual inductance, M, can be easily seen by ignoring the capacitance, C_u. Current i_0 in the disturbing circuit results in a voltage, $v_m = i_0 j\omega M$, in the disturbed circuit producing a current given by

$$i_m = -\frac{i_0 j\omega M}{2Z_0} \tag{11-3}$$

The shielding effect of intervening conductors or shields may reduce the unbalance capacitance, C_u, and the mutual inductance, M, and thus reduce the current coupled into the disturbed circuit.

A quantitative measure of the coupling loss between circuits is the equal level coupling loss (ELCL). The ELCL at some point is defined as the ratio of the power passing through this point in the disturbing circuit to the induced power passing through the same location on the disturbed circuit. Note that ELCL is independent of signal levels.

Consider now the termination on the left end of the disturbed circuit. Because this termination is located at the same end as the

source of the disturbing circuit, the ELCL at the left end of the circuits is a measure of the near-end crosstalk. A more exact definition of near-end crosstalk, allowing for crosstalk between circuits that are not coterminous, results from considering directions of energy flow in the circuits. Thus, the disturbing circuit in the preceding example has energy flowing from left to right. The current in the left resistor of the disturbed circuit is due to the energy flow from right to left.

Near-end crosstalk (NEXT) is crosstalk whose energy travels in the opposite direction to that of the signal in the disturbing channel.

In accordance with Fig. 11-1, the NEXT current is given by $i_n = i_c - i_m$. The ELCL, or NEXT ratio, is

$$\frac{i_n}{i_0} = \frac{i_c - i_m}{i_0} = \frac{j\omega C_u Z_0}{8} + \frac{j\omega M}{2Z_0} \tag{11-4}$$

Similarly, the power at the far end of the disturbed circuit can be considered far-end crosstalk defined by:

Far-end crosstalk (FEXT) is crosstalk whose energy travels in the same direction as the signal in the disturbing channel.

The FEXT current in Fig. 11-1 is given by $i_f = i_c + i_m$. The FEXT ratio is:

$$\frac{i_f}{i_0} = \frac{i_c + i_m}{i_0} = \frac{j\omega C_u Z_0}{8} - \frac{j\omega M}{2Z_0} \tag{11-5}$$

Note that the capacitive crosstalk and inductive crosstalk oppose each other for FEXT and add for NEXT. This makes NEXT the more serious problem in the case considered. The degree of cancellation in the far-end case, Eq. (11-5), depends upon the relative magnitudes of the capacitive and inductive components as well as Z_0. Examination of Eqs. (11-4) and (11-5) shows that for high impedance circuits, capacitive coupling is the most significant mode of crosstalk interference. On the other hand, for low impedance circuits, capacitive effects are minimized and inductive coupling is more significant.

This procedure will not be applicable if the transmission line is long compared to a wavelength. In this case, the solution is obtained by considering short sections of line and then summing the result, taking into account attenuation and phase shift as characterized by the propagation constant, γ. Let the mutual inductance unbalance

per unit length at a distance, x, from the near end be $M(x)$ and the capacitance unbalance per unit length be $C_u(x)$. To a first order approximation, $M(x)$ and $C_u(x)$ are independent of frequency.

To illustrate crosstalk relationships in a simplified manner, it is convenient to assume that the mutual impedance and admittance per unit length are constant. Although crosstalk relationships have been derived without this constraint, the somewhat more involved mathematics (utilizing the autocorrelation in x of the coupling) gives basically similar results.

Near-End Crosstalk

Consider two parallel cable circuits of length L as shown in Fig. 11-2. Assume that the mutual admittance per unit length is given by $j\omega C_u$ and the mutual impedance per unit length by $j\omega M$. The component of crosstalk current, di_n, induced along dx at a distance, x, from the source is given from Eq. (11-4) as

$$\frac{di_{nx}}{i_{0x}} = j\omega\left(\frac{C_u Z_0}{8} + \frac{M}{2Z_0}\right)dx \tag{11-6}$$

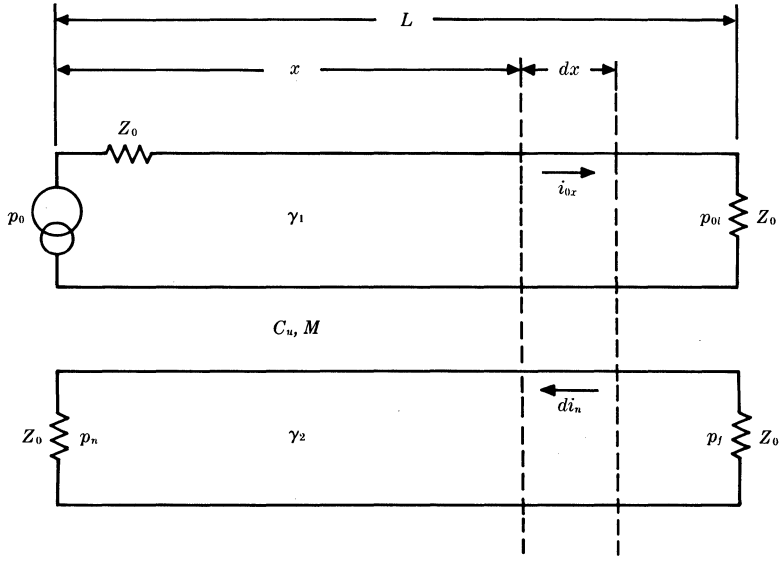


FIG. 11-2. Differential crosstalk coupling between two circuits.

but

$$i_{0x} = i_0 e^{-\gamma_1 x} \tag{11-7}$$

where γ_1 is the propagation constant of the disturbing line and may be written

$$\gamma_1 = \alpha_1 + j\beta_1 \tag{11-8}$$

Referring the crosstalk current to the near end by γ_2, the propagation constant of the disturbed line, gives

$$\frac{di_n}{i_0} = j\omega\left(\frac{C_u Z_0}{8} + \frac{M}{2Z_0}\right) e^{-(\gamma_1+\gamma_2)x}\, dx \tag{11-9}$$

Letting

$$\left(\frac{C_u Z_0}{8} + \frac{M}{2Z_0}\right) = C_n \tag{11-10}$$

allows Eq. (11-9) to be written

$$\frac{di_n}{i_0} = j\omega C_n\, e^{-(\gamma_1+\gamma_2)x}\, dx \tag{11-11}$$

Assuming uncorrelated crosstalk coupling with length, the ELCL or NEXT power ratio is obtained by squaring each differential current element given by Eq. (11-11) and integrating over the length of the parallel cables [3]:

$$\frac{p_n}{p_0} = \left|\frac{i_n}{i_0}\right|^2 = \omega^2 C_n^2 \int_0^L e^{-(\gamma_1+\gamma_2)x} e^{-(\gamma_1^*+\gamma_2^*)x}\, dx$$

$$= \omega^2 C_n^2 \int_0^L e^{-2(\alpha_1+\alpha_2)x}\, dx \tag{11-12}$$

Or,

$$\frac{p_n}{p_0} = \frac{C_n^2\omega^2}{2(\alpha_1+\alpha_2)}\left[1 - e^{-2(\alpha_1+\alpha_2)L}\right] \tag{11-13}$$

Skin effect causes the attenuation of transmission lines to be proportional to the square-root of frequency for the frequencies of usual interest. Thus,

$$\alpha_1 = K_1\sqrt{f}$$

$$\alpha_2 = K_2\sqrt{f} \tag{11-14}$$

and Eq. (11-13) can be rewritten as

$$\frac{p_n}{p_0} = \frac{4\pi^2 C_n^2}{(2K_1 + K_2)} f^{3/2} \left[1 - e^{-2(\alpha_1 + \alpha_2)L}\right] \tag{11-15}$$

or

$$\frac{p_n}{p_0} = k_n f^{3/2} \left[1 - e^{-2(\alpha_1 + \alpha_2)L}\right] \tag{11-16}$$

Examination of this equation for long lengths of line, L, shows that the term in brackets goes to unity, and the NEXT ratio is given by

$$\frac{p_n}{p_0} \approx k_n f^{3/2} \tag{11-17}$$

or,

$$\text{NEXT} = 10 \log k_n + 15 \log f \quad \text{dB} \tag{11-18}$$

where 10 log k_n is the NEXT loss at unit frequency (hertz, kilohertz, or megahertz—the convenient units for f). Note that the NEXT is independent of length for long transmission lines and increases with frequency at the well-known 4.5 dB per octave rate. If the NEXT loss for a long line is K_0 dB at a frequency f_0, it can be found for any other frequency, f, by

$$\text{NEXT loss} = K_0 - 15 \log f/f_0 \quad \text{dB} \tag{11-19}$$

Equation (11-17) is valid when $2(\alpha_1 + \alpha_2)L >> 1$. Since $\alpha_1 L$ represents the total loss in nepers through the disturbing circuit and $\alpha_2 L$ is the total loss in the disturbed path, it can be shown that Eq. (11-17) is less than 10 per cent in error if the total loss in the two parallel lines is greater than 10 dB.

In summary, NEXT is independent of path lengths if the sum of the loss in the two paths is greater than 10 dB. NEXT is frequency dependent and usually increases with frequency at a 4.5 dB per octave rate.

Far-End Crosstalk

The FEXT can be analyzed in a manner similar to that for NEXT. Again, reference can be made to Fig. 11-2. In this case, Eq. (11-5)

can be expressed in differential form:

$$\frac{di_{fx}}{i_{0x}} = j\omega\left(\frac{M}{2Z_0} - \frac{C_u Z_0}{8}\right) dx \tag{11-20}$$

Letting

$$C_f = \left(\frac{M}{2Z_0} - \frac{C_u Z_0}{8}\right) \tag{11-21}$$

and referring di_{fx} to the far end ($x = L$) results in

$$\frac{di_f}{i_0} = j\omega C_f \left[e^{-\gamma_1 x} e^{-\gamma_2 (L-x)}\right] dx \tag{11-22}$$

Squaring each differential current element and integrating over the length of the parallel cables gives:

$$\frac{p_f}{p_0} = \left|\frac{i_f}{i_0}\right|^2 = \omega^2 C_f^2 e^{-2\alpha} 2L \int_0^L e^{-2(\alpha_1 - \alpha_2)x} dx$$

$$\frac{p_f}{p_0} = \frac{\omega^2 C_f^2 e^{-2\alpha} 2L}{2(\alpha_2 - \alpha_1)} \left[e^{-2(\alpha_1 - \alpha_2)L} - 1\right] \tag{11-23}$$

For reasons that will become obvious when levels are discussed, it is often convenient to express the ELCL at the far end of the disturbing circuit. Since $p_{0l} = p_0 e^{-2\alpha_1 L}$, Eq. (11-23) may be rewritten in terms of equal level crosstalk as

$$\frac{p_f}{p_{0l}} = \frac{p_f}{p_0 e^{-2\alpha_1 L}} = \omega^2 C_f^2 e^{-2(\alpha_2 - \alpha_1)L} \left[\frac{e^{-2(\alpha_1 - \alpha_2)L} - 1}{2(\alpha_2 - \alpha_1)}\right] \tag{11-24}$$

This expression is not homogeneous in α_2 and α_1, and therefore the magnitude of the crosstalk depends upon which is the disturbing circuit.

A case of special interest is that of FEXT between two like pairs. In this case, α_1 equals α_2 and Eq. (11-24) becomes

$$\frac{p_f}{p_{0l}} = \omega^2 C_f^2 L = k_f f^2 L \tag{11-25}$$

or,

$$\text{FEXT (equal level)} = 10 \log k_f + 20 \log f + 10 \log L \quad \text{dB} \tag{11-26}$$

where 10 log k_f is the FEXT loss of a unit length system at unit frequency. Note that the FEXT power is directly proportional to the length of the crosstalk path and increases with frequency at the well known 6 dB per octave rate. If the FEXT loss is K_f dB at a frequency f_0 for a length L_0, it can be determined for any other length by:

$$\text{FEXT loss} = K_f - 20 \log \frac{f}{f_0} - 10 \log \frac{L}{L_0} \quad \text{dB} \qquad (11\text{-}27)$$

For like pairs ($\gamma_1 = \gamma_2$), all crosstalk paths connecting p_0 and p_f of Fig. 11-2 through various unbalances have the same time of propagation. Crosstalk currents due to capacitance unbalances at any two points combine in the same or opposite sign. One correcting unbalance of suitable sign connected between the pairs at *any* point may be used to annul the distributed unbalances [4]. Likewise, the distributed mutual inductance may be nearly balanced by a single lumped mutual inductor.

Effects of Systematic Coupling

In the derivation of the crosstalk equations [Eqs. (11-16) and (11-25)], power addition of the differential contributors was assumed; i.e., correlation between them was disregarded. This is not valid for very short lengths of transmission line, and a more realistic result can be obtained by integrating the current, taking phase shifts into account, to arrive at a magnitude of current to be squared for the power relationships. If phase shifts due to both length and random fluctuations in the coupling are ignored, the crosstalk equations can be easily rederived by assuming voltage (or current) addition in the integration before squaring.

If this is done for NEXT on short exposure lengths (those with negligible phase shift), the ELCL for NEXT becomes

$$\frac{p_n}{p_0} \approx \omega^2 C_n^2 L^2 \qquad (11\text{-}28)$$

Similarly, the FEXT between *identical* pairs with *uniform* coupling is incorrectly given by Eq. (11-25) which should be rederived with voltage addition to yield

$$\frac{p_f}{p_0} = \omega^2 C_f^2 L^2 \qquad (11\text{-}29)$$

with the result that the 10 log L term of Eq. (11-26) becomes 20 log L. If the lines are truly identical and uniform, Eq. (11-29) is applicable regardless of the lengths of the parallel lines; however, in practical lines, the propagation constants are not identically matched or uniform. As a result, for long exposure paths the crosstalk components have different phases and thus add on a power basis. Therefore Eq. (11-25) often gives a better approximation under many practical conditions.

There are some important practical situations where the FEXT calculations are more accurate with Eq. (11-29) than with Eq. (11-25). One is the crosstalk between adjacent coaxials used in the L-type carrier systems. The physical structure of each coaxial is relatively uniform along its length as is the physical location of adjacent coaxials. As a consequence, both the coupling between adjacent coaxials and the propagation velocity along the coaxials are extremely uniform. This *systematic coupling* results in the crosstalk adding on a voltage basis and the crosstalk power ratio increasing as the square of the length [5]. Another example is when color-to-color splicing is used with multipair plastic-insulated cables. It has been found that crosstalk measurements on reel lengths cannot be scaled to longer lengths by a 10 log L factor but instead require close to 20 log L. Fortunately, in systems subject to such coupling, methods of minimizing systematic crosstalk (such as transposition and splicing rules) can be found. The remaining random coupling is then the only one to consider.

Indirect Crosstalk

Up to this point, consideration has been given to direct crosstalk coupling paths between circuits. This crosstalk is transverse in nature. In each elementary length along two parallel circuits, the currents and charges in the disturbing circuit induce a voltage in the disturbed circuit in that elementary length. This voltage results in a crosstalk current at a circuit terminal. The total transverse crosstalk current at the terminal is the vector sum of the transverse contributions of all the elementary lengths.

Crosstalk by way of other (tertiary) circuits is called indirect crosstalk. For engineering purposes it is convenient to divide indirect crosstalk into two components called transverse indirect and interaction indirect crosstalk. Thus, transverse crosstalk may be both direct and indirect, but interaction crosstalk occurs only indirectly via tertiary circuits.

Transverse Crosstalk. This type of indirect crosstalk results when tertiary circuits (including phantoms or ground-return circuits) change the capacitance unbalance between two circuits. It should be noted that tertiary circuits do not change the mutual inductance between two circuits, but only the capacitance unbalance. Thus, if the disturbing circuit is a coaxial or shielded pair, it has no external electric field (barring holes in the shield), and there is no transverse indirect crosstalk [6]. In balanced systems, any unbalance to tertiary or to ground results in transverse indirect crosstalk. Examples include resistance unbalance caused by unequal wire diameters or poor joints and inductance unbalance caused by one conductor being unsymmetrically wrapped around its mate. Although transverse indirect crosstalk is via a tertiary path, the crosstalk current does not propagate along the tertiary path.

By utilizing the mutual capacitance in the previously given direct crosstalk equations rather than the direct or unbalance capacitance, both direct and indirect components of transverse crosstalk can be treated together. This is a technique which is frequently used.

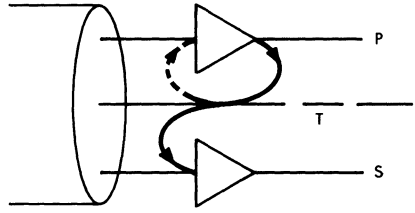
Interaction Crosstalk. Crosstalk as a result of direct coupling to a tertiary circuit, propagation along this circuit, and coupling into the disturbed circuit is called interaction crosstalk (IXT). Figure 11-3 shows four kinds of interaction crosstalk. In the order of their importance they are:

1. Near-end near-end interaction crosstalk (NE-NE-IXT).
2. Near-end far-end (NE-FE-IXT) or far-end near-end interaction crosstalk (FE-NE-IXT).
3. Reflected near-end crosstalk.
4. Far-end far-end interaction crosstalk (FE-FE-IXT).

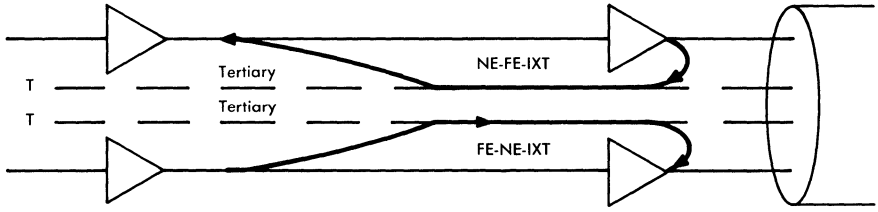
NE-NE-IXT results because a primary circuit, P, may crosstalk into a tertiary circuit, T, in an elementary length, as shown in Fig. 11-3(a). The crosstalk current in the tertiary circuit is then propagated toward the sending end of the primary circuit and crosstalks into the secondary circuit, S (the ultimately disturbed circuit), at some other point. The crosstalk current in the secondary circuit is propagated to the far end of that circuit. Therefore, NE-NE-IXT is a component of the total observed far-end crosstalk between a primary and a secondary circuit. NE-FE-IXT (or FE-NE-IXT) is a component of near-end crosstalk and sometimes causes concern; although with the repeater configuration shown in Fig. 11-3(b), such paths are blocked. Reflected NEXT is a consequence of a mismatch

between repeater output impedance and cable impedance. Figure 11-3(c) shows the path taken by the reflected NEXT. If necessary, a better impedance match between cable and repeater output could be used to reduce this kind of interference. FE-FE-IXT is usually an unimportant component of far-end crosstalk.

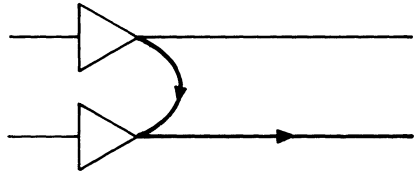
In many repeatered transmission systems, direct NEXT is minimized by separating each direction of transmission (for example, by using different cables). In such cases, other cable paths not



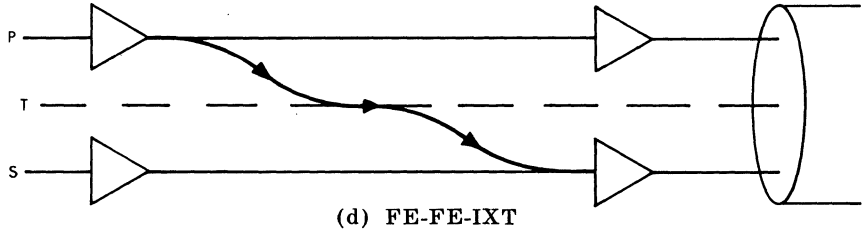
(a) NE-NE-IXT



(b) NE-FE-IXT or FE-NE-IXT



(c) Reflected NEXT



(d) FE-FE-IXT

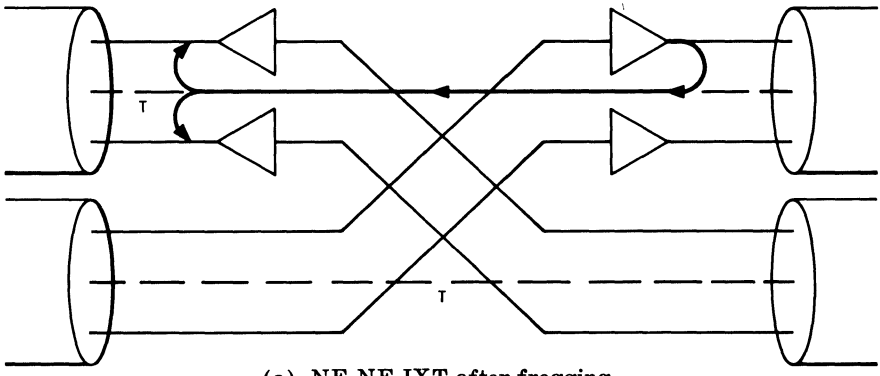
FIG. 11-3. Interaction crosstalk paths.

interrupted at a repeater site provide a crosstalk path around the repeaters. An example is the sharing of a cable between nonrepeatered voice-frequency circuits and repeatered carrier circuits. In such cases the NE-NE-IXT paths shown in Fig. 11-3(a) often become predominant. Circuits P and S represent two one-way branches transmitting the same carrier frequencies in the same direction. Circuit T represents a tertiary circuit not repeatered at the carrier repeater points. The solid line connecting P and S indicates an interaction crosstalk path between the output of a repeater in circuit P and the input of a repeater in circuit S. This path involves the sum of (1) the near-end crosstalk loss between P and T in the second repeater section and (2) the near-end crosstalk loss between T and S in the first repeater section. The path indicated by the solid line is much more serious than that indicated by the dotted line since the former is at a higher power level than the latter by the gain of one repeater.

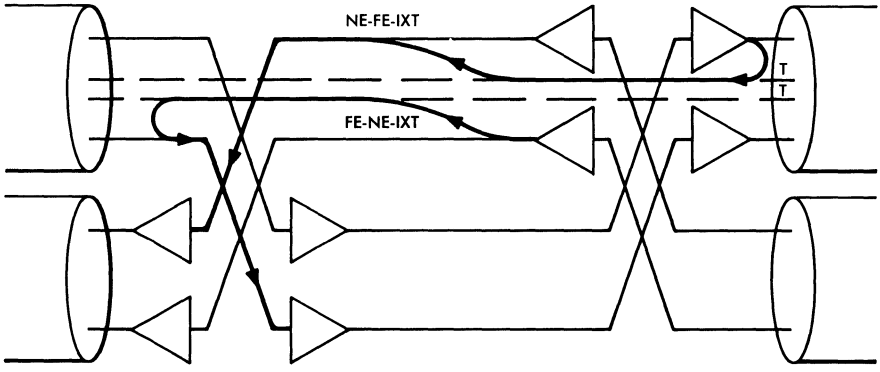
Since it is usually economical to make the repeater gain large, means must be found for greatly increasing the loss in the crosstalk path shown by the solid line. A similar path may connect input and output of the same repeater, as indicated by the dashed line, and thus endanger echo and singing margins.

Interaction crosstalk at repeater points is often minimized by "frogging," that is, interchanging the east- and west-bound branches of the four-wire system between the two different cables ordinarily used with such systems. This scheme is illustrated in Fig. 11-4. NE-NE-IXT paths are connected between repeater outputs, and crosstalk currents through such paths cannot reach a subscriber. However, NE-FE-IXT paths or FE-NE-IXT paths, as shown in Fig. 11-4(b), transmit crosstalk currents from repeater outputs to repeater inputs, and such currents do propagate through the system. Such paths are usually secondary contributors to total crosstalk since they are superimposed on direct FEXT of greater magnitude.

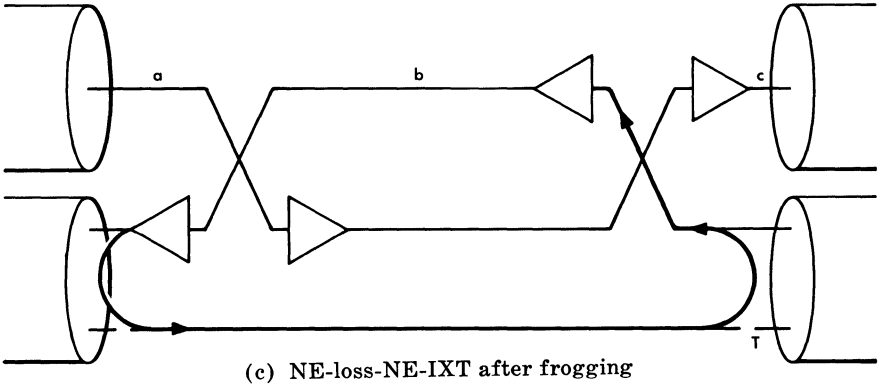
The effects of interaction paths after frogging are dependent on variations in spacing between repeaters. For example, the FE-NE-IXT path shown in Fig. 11-4(b) is attenuated along the tertiary path. If the loss along this path is less than the repeater gain, the crosstalk is enhanced by this gain difference. Normally, with uniform repeater spacings the cable loss and amplifier gains are complementary. However, if spacing is not uniform, differences in gains between sections may enhance the potential crosstalk. This same variation in repeater spacings can make significant an NE-loss-NE-IXT coupling path, as shown in Fig. 11-4(c). The first near-end coupling originates in



(a) NE-NE-IXT after frogging



(b) NE-FE-IXT and FE-NE-IXT after frogging



(c) NE-loss-NE-IXT after frogging

FIG. 11-4. Effects of frogging.

section a, proceeds on the tertiary path past section b, and couples into section c (repeater input) via a second near-end path. The crosstalk will be reduced by the NEXT coupling losses and the attenuation of length b, but the disturbed signal will also be attenuated by length c. If c = b, the crosstalk is the same as the ELCL of NE-NE-IXT. If, however, length c is much greater than b, the crosstalk is enhanced by the difference in attenuation. The remedy is to set a limit on the difference in section lengths.

With the type N cable carrier system, NE-NE-IXT is prevented by shifting the carrier frequency of each channel at each repeater point as indicated in Fig. 11-5. This shifting of frequencies to block crosstalk paths is called frequency frogging. The crosstalk current through an interaction path such as $n_2 n_1$ is prevented from reaching a subscriber set by frequency selectivity. A talker on a channel with a

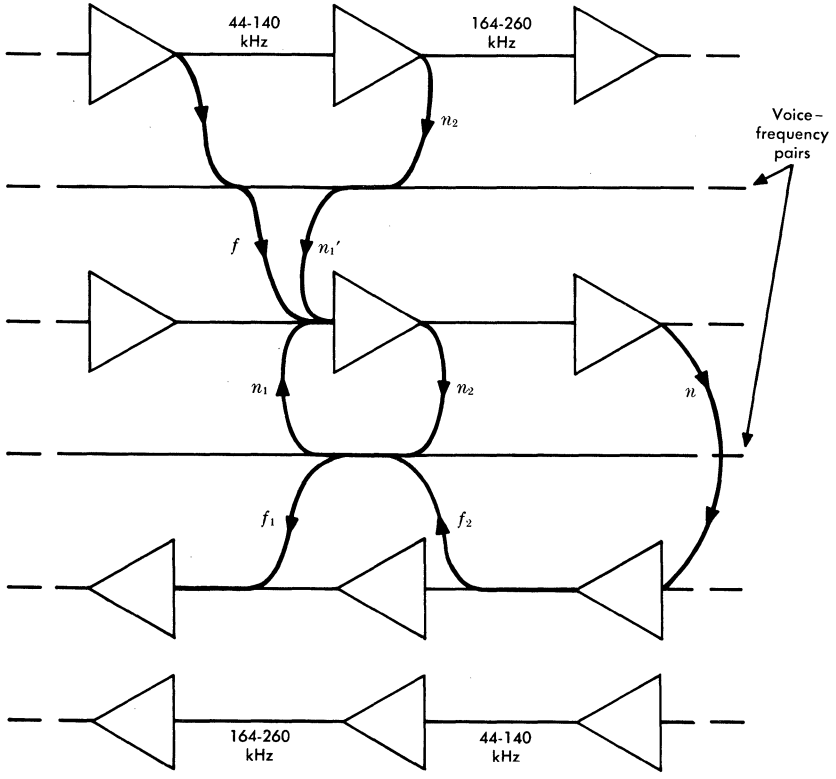


FIG. 11-5. N carrier system crosstalk paths.

carrier frequency of 256 kHz would, via n_2n_1, cause a crosstalk current at the input of a repeater passing only the band of 44 to 140 kHz. He would also cause a crosstalk current via a path such as n_2f_1 to reach the input of a repeater passing 256 kHz. Therefore, NE-FE- and FE-NE-IXT were given consideration in the design of the N carrier system, and mild restrictions were set on variations in length of repeater sections.

If 44 to 140 kHz were transmitted in both directions in the first repeater section, and 164 to 260 kHz were used in the second section, etc., interaction crosstalk paths such as n_2f_1 and f_2n_1 would be blocked by frequency selectivity. However, such a frequency allocation would result in serious NEXT through the n paths since these paths would no longer be blocked by frequency selectivity. FEXT paths, denoted by f, cannot be economically blocked by frequency frogging.

Effects of Transmission Levels

The observed crosstalk between two circuits at their terminals is a function not only of the coupling loss between circuits, but also of the relative transmission levels of the circuits. When the circuits contain gain, as do most communication circuits, the transmission level differences and their effects may be greatly increased.

For example, consider the near-end and far-end crosstalk between the two circuits shown in Fig. 11-6, which are typical of short two-wire voice-frequency toll cable circuits. The symbols at A, B, and C are conventional representations of two-way repeaters with gain in both directions. The transmission level diagrams show the levels of the speech signal at each point for the two directions of transmission. Since the circuits are assumed to be alike, the same level diagrams apply to both circuits. The gain or loss between any two points is easily obtained from the level diagrams by finding the difference between levels at those points.

Now consider at the A terminals of the upper and the lower circuits the NEXT resulting from the 60-dB crosstalk path shown at B. This crosstalk path might be a lumped coupling at the point indicated, or it might represent the NEXT measured at B on that section of line between B and C.

The crosstalk loss between the two circuits at A is evidently equal to the loss in the upper circuit from its terminal at A to the point of coupling, plus the 60-dB loss between the circuits, plus the loss of the lower circuit from the coupling to its terminal at A. From the level diagrams, there is evidently a gain of 3 dB (or a loss of −3 dB)

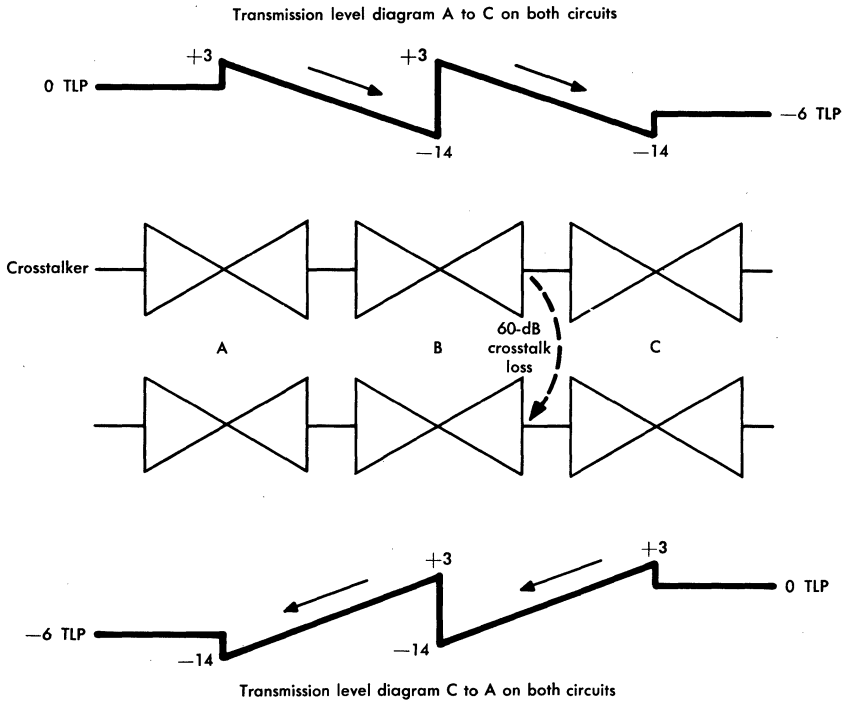


FIG. 11-6. Crosstalk on repeatered circuits.

from A to the coupling, and a gain of 14 — 6, or 8 dB (or a loss of —8 dB), on the lower circuit from the coupling to its A terminal. Therefore, the near-end crosstalk loss between the circuits at A is —3 + 60 — 8, or 49 dB. Evidently, a crosstalk loss of 60 dB at B appears as a crosstalk loss of 49 dB measured from A. The apparent gain of 11 dB is called the *crosstalk amplification.* If there is a known crosstalk loss between two circuits in a particular section or piece of equipment, its importance cannot be judged without knowing the crosstalk amplification. Before different crosstalk couplings along a circuit can be compared or combined, they must be reduced to a common base by correcting for the crosstalk amplification.

If the gain of the terminal repeater of the disturbed circuit in Fig. 11-6 is increased by 6 dB, the output of this circuit would be a zero level point, and the near-end crosstalk loss would be 49 — 6, or 43 dB. This would be the *equal-level crosstalk loss,* i.e., the coupling that would be measured between points of the same transmission level on the disturbing and disturbed circuits.

Measurements and Units

The magnitude of the crosstalk coupling between two circuits is fundamentally a matter of how well speech energy is transferred from one circuit to the other; a single-frequency measurement is inadequate unless the coupling is flat with frequency. This condition is often satisfied over the bandwidth of any particular channel in FDM carrier circuits; but at voice frequencies, where capacitive coupling is usually the dominant mechanism, the coupling loss has an average slope of 6 dB per octave. Furthermore, the discontinuities caused by splices and different gauges of wire cause the coupling loss to vary as much as 6 to 8 dB around this average slope.

For laboratory measurements of the effects of crosstalk, speech and listeners are used; the criterion of intelligibility is taken as the ability to understand four words during a 7-second time interval. For field measurements of voice-frequency circuits, thermal noise, shaped to have the same power spectrum as speech, is often used as the input to the disturbing circuit. The output of the disturbed circuit can then be measured using a noise measuring set with C-message weighting. (A 0-vu speech volume applied to a noise meter thus weighted gives a reading of about 88 dBrn.) For smooth 6 dB per octave capacitive coupling, this method gives a coupling loss 2.8 dB less than the 1-kHz coupling loss. If the coupling has a smooth 6 dB per octave slope, 1-kHz measurements can be corrected. The noise measurement provides protection against errors which would be involved in this slope assumption, and gives a single integrated value for the jagged curve of coupling loss versus frequency.

A unit often used in crosstalk computations is the dBx, which is equal to the dB difference between 90-dB loss and the transmission loss of the coupling path. For example, if the coupling loss between two circuits is 60 dB, this fact is expressed by saying that the coupling is 30 dBx. The dBx was introduced in order that the number of dB would increase as couplings become tighter, rather than decrease as in the case of coupling loss. Thus, as crosstalk becomes worse, the numbers in dBx get larger.

Summation of Many Crosstalk Components

The total crosstalk loss between two circuits may be computed or estimated if the crosstalk losses for the various parts are known. The individual losses must be corrected for their individual crosstalk

amplifications and then combined. The method of combination depends on the nature of the coupling between the disturbing and disturbed circuits.

When two crosstalk losses are combined, the total effective crosstalk loss is smaller than the loss of either component because the power in the disturbed circuit increases. To avoid confusion, the expression of crosstalk coupling in dBx will be assumed in all that follows.

There are many valid reasons for assuming that differential crosstalk paths have characteristics randomly distributed. Irregularities in construction are important factors in this variation. Some irregularities, such as twisted pairs with different twist lengths, can be purposely introduced to force the crosstalk coupling to be random. Since random crosstalk is a matter of chance, some consideration of probability is necessary for analytical purposes.

It has been observed that crosstalk among different cable pairs is usually normally distributed in dB. The resulting log normal distribution in power is the same as discussed in Chap. 9. With n crosstalk exposures having a mean of C_m dBx and a standard deviation of σ_c dB, the average crosstalk coupling is given by

$$C_x = C_m + 10 \log n + 0.115\,\sigma_c^2 \qquad \text{dBx} \qquad (11\text{-}30)$$

This is often called the rms crosstalk in dB since it is 20 log the rms current or voltage ratios.

A more difficult problem is the determination of the distribution of the total crosstalk coupling due to a large number of exposures log normally distributed. The most obvious need for such information is in the determination of the expected total crosstalk coupling exceeded with some given probability such as 1 per cent. The determination of the distribution of a sum of log normally distributed parameters has been studied in detail, and straightforward computations can be used to determine the new distribution [7]. When the number of contributors, n, is large and their variance is not too great ($\sigma_c < 10$ dB), the distribution of the sum approaches a normal distribution with a mean and standard deviation given approximately by

$$\mu \approx C_m + 0.1\,\sigma_c + 10 \log N \qquad \text{dB}$$

$$\sigma \approx \sigma_c/\sqrt{n}$$

Another interesting approach that is often applicable to analog systems is to consider the total crosstalk as a time-varying voltage (or current) made up of a large number of components. In such a case, the instantaneous crosstalk voltage (or current) waveform will have a gaussian distribution with a zero mean and a standard deviation given by the rms value corresponding to Eq. (11-30). This magnitude exceeded 1 per cent of the time is then readily determined by the 2.33 σ point, which is 20 log 2.33 = 7.3 dB greater than the rms value given by Eq. (11-30). Of course, the crosstalk exceeded 1 per cent of the time is only the same as the crosstalk exceeded in 1 per cent of the cases if the process is ergodic so that the ensemble and time statistics are identical. This is usually true for analog systems where crosstalk of tandem repeaters adds, but not true for digital systems where each regenerative repeater removes the crosstalk of the previous section.

Crosstalk Example

The practical application of many of the above considerations is often required in the short-haul exchange area. Extensive use is made of ordinary voice-frequency grade unshielded wire pairs for higher capacity and higher frequency transmission systems. These systems may be either analog (such as N carrier) or digital (such as T carrier). For such an environment, the most serious limitation is often crosstalk between pairs bundled in the same cable. This problem is compounded when many two-way systems are sharing the same cable.

These multipair exchange area cables were originally placed with no provision for repeaters; the voice-frequency loss over long distances is reduced by inserting loading coils at regular intervals, normally 6000 feet. As a consequence, access to the cable is available at these points (usually via a manhole). The natural extension to using these cables for higher frequency repeatered systems results in fixed repeater spacing at the old loading coil spacing of 6000 feet. The crosstalk paths affecting a typical repeater-to-repeater 6000-foot link are depicted in Fig. 11-7. It can be seen that for two-way (single cable) operation, only the NEXT is significant since the FEXT paths include the loss of the cable section. If the cable were used for one direction only, the NEXT paths would be suppressed and the FEXT would become significant. The NEXT path is between pairs operating in different directions of transmission. Figure 11-8 shows the

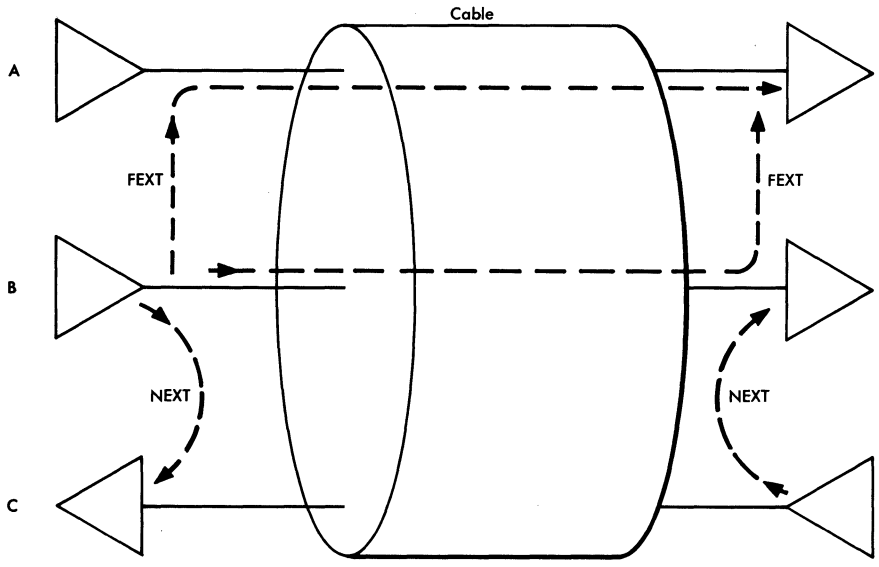


FIG. 11-7. Crosstalk in repeatered multipair cable.

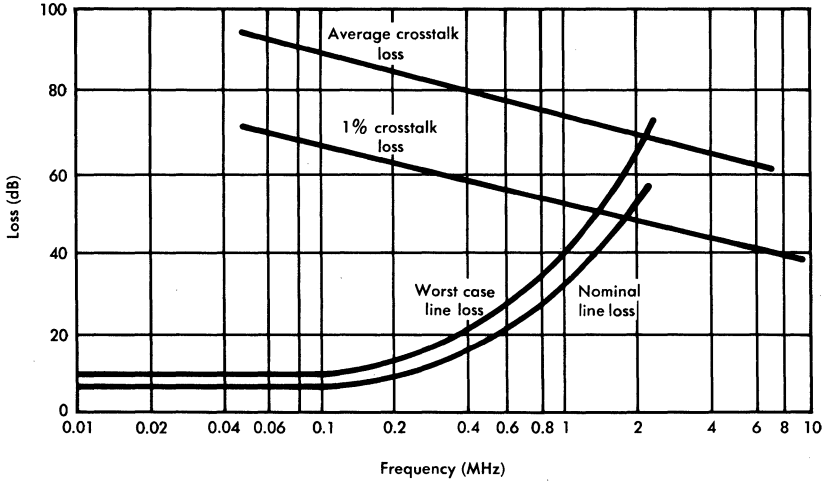


FIG. 11-8. Line loss and crosstalk loss in 22-gauge exchange area cable.

line loss of a 6000-foot, 22-gauge cable and the distribution of NEXT. The equivalent NEXT path has a slope of —4.5 dB per octave as expected. If it is assumed that the crosstalk loss is normally distributed in dB, the 21-dB difference between the average and the 1 per cent crosstalk loss represents 2.33 σ; therefore σ equals 9 dB.

Consider the NEXT between circuit B and C of Fig. 11-7 at a frequency of 0.1 MHz. From Fig. 11-8, assuming worst case (highest loss) lines, the gain of each repeater at 0.1 MHz is about 10 dB. The mean crosstalk loss between the output of the A repeater and the input of the B repeater is shown in Fig. 11-8 as 89 dB; i.e., the NEXT coupling is 1 dBx. If the outputs of all repeaters are assumed to be at equal levels, the equal level crosstalk coupling will be increased to 11 dBx by the gain of the repeater. However, there is a 1 per cent chance that the crosstalk coupling could be more than 21 dB higher than this, or the 1 per cent equal level crosstalk coupling is 32 dBx for this single exposure.

To illustrate the addition of many crosstalk paths, it will be assumed in this example that the total system consists of 20 links in tandem so that there will be 20 NEXT exposures between the B and C circuit. In addition, it will be assumed that a total of 25 two-way systems share the same cable. As a consequence, each of the other circuits will have 20 crosstalk exposures into the B circuit giving a total of 20 $\times$ 25 or 500 NEXT exposures. The previously computed mean crosstalk per exposure (11 dBx) is slightly conservative since worst case line loss was used, and for this many exposures the nominal loss would be more appropriate. The result is a mean equal level NEXT of about 9 dBx per exposure. It should be noted that a digital transmission system with regeneration would only be susceptible to the 25 exposures since regeneration of pulses breaks the tandem effects. This is another illustration of the "ruggedness" of digital systems.

It can be assumed, usually quite validly, that the individual crosstalk paths are uncorrelated. The average crosstalk can be found by use of Eq. (11-30) with $C_m = 9$ dBx and $\sigma_c = 9$ dB:

$$C_x = 9 + 10 \log 500 + 0.115\ (81) = 45.3 \text{ dBx} \qquad (11\text{-}31)$$

For this many contributors, it can be assumed that the 1 per cent crosstalk coupling is 7.3 dB higher than the average value or

$$C_{1\%} \approx 45.3 + 7.3 = 52.6 \text{ dBx} \qquad (11\text{-}32)$$

Since the preceding example has assumed that all of the systems are carrying the same signal levels at the output of all of the repeaters, the average signal-to-crosstalk noise ratio is given by

$$S/N = 90 - 45.3 = 44.7 \text{ dB} \tag{11-33}$$

The ratio exceeded less than 1 per cent of the time is given by

$$S/N_{1\%} = 90 - 52.6 = 37.4 \text{ dB} \tag{11-34}$$

It is obvious from Fig. 11-8 that at sufficiently high frequencies, the NEXT loss may be less than the line loss. In such a case, a repeatered system would "sing" and thus be unusable. If this exchange area cable is used at high frequencies, it is necessary either to space the repeaters more closely or to segregate the two directions of transmission into separate cables.

REFERENCES

1. Weaver, M. A. "The Long Struggle Against Cable Crosstalk," *Bell Telephone Quarterly* (Jan. 1935).
2. "Dr. G. A. Campbell's Memoranda of 1907 and 1912," *Bell System Tech. J.*, vol. 14 (Oct. 1935), pp. 558-572.
3. Cravis, H. and T. V. Crater. "Engineering of T1 Carrier System Repeatered Line," *Bell System Tech. J.*, vol. 42 (Mar. 1963), pp. 431-486.
4. Weaver, M. A., R. T. Tucker, and P. S. Darnell. "Crosstalk and Noise Features of Cable Carrier Telephone System," *Bell System Tech. J.*, vol. 17 (Jan. 1938), pp. 137-161.
5. Booth, R. P. and T. M. Odarenko. "Crosstalk Between Coaxial Conductors in Cable," *Bell System Tech. J.*, vol. 19 (July 1940), pp. 358-384.
6. Schelkunoff, S. A. "The Electromagnetic Theory of Coaxial Transmission Lines and Cylindrical Shields," *Bell System Tech. J.*, vol. 13 (Oct. 1934), pp. 532-579.
7. Nasell, I. "Some Properties of Power Sums of Truncated Normal Random Variables," *Bell System Tech. J.*, vol. 46 (Nov. 1967), pp. 2091-2110.

Chapter 12

Introduction to Analog Cable Systems

The analog cable system derives its name from both the medium used and the properties of the repeaters required to compensate for the loss of the medium. In contrast to the regenerative repeater of the digital system, the repeater of the analog system is intended to reproduce at its output an exact, linearly scaled version of the input signal. Inevitably, in real repeaters there are accumulations of thermal noise and amplitude distortion that ultimately determine the quality of performance achieved by the total system. The design of analog cable systems involves the analysis and control of these factors and the ways in which they interact with the medium to establish the final properties of the system. The media commonly used in cable systems are discussed in Chap. 2.

Specifying an analog system does not limit the types of signals to be transmitted by that system. In general, the system load will include a mixture of voice, digital, video, and supervisory signals. It is only necessary that the interference requirements for each type of signal be compatible with system performance and that the power and bandwidth capacities of the system be consistent with the resultant combination of signals. Since analog cable systems are designed to operate over distances at which it becomes uneconomic to provide each message circuit with its own wire pair, frequency division multiplex is used to assemble the mixture of signals eventually applied to the line. These distances run from about 10 miles in exchange area systems to 4000 miles in the long-haul coaxial systems.

12.1 GENERAL SYSTEM FEATURES

Analog cable systems at this time constitute a large portion of the total carrier facility within the telephone plant. Alternatives

available at the present are digital cable systems in the exchange area, and analog microwave radio systems for short- and long-haul use. The systems in place have evolved since the mid-1930's, with the most recently installed new long-haul designs being the L4, a buried cable system placed in service in 1967, and the SF, a submarine cable system in service since 1968. The most recently designed short-haul analog cable system is the N2, which has been in use since 1962. In some ways, the thrust of continuing development over the years has been quite different in the short-haul and the long-haul areas. Furthermore, within the long-haul area, a different emphasis is applied to the design of land cable systems than to submarine cable systems. The contrasts in each case are primarily traceable to economic factors.

The short-haul application normally involves the use of wire pairs in a multipair cable for the interconnecting medium. These pairs are relatively inexpensive and one result of this and the relatively short length of these systems is a sensitivity to terminal costs that is much greater than that of long-haul systems. This is one of the important reasons that little emphasis has been placed on attaining major increases in the bandwidth of these systems. The emphasis, rather, has been on the application of new art toward the reduction of repeater and terminal costs. Increased circuit needs are accommodated by using more pairs of an already installed cable, or by installing new multipair cables and equipping pairs as required.

In comparison with short-haul systems, the more stringent transmission objectives of long-haul systems of high capacity dictate the use of a more uniform, lower loss, and higher quality cable, such as the coaxial cables used in the L-type and undersea systems. As a result, the medium for long-haul systems is the most important cost factor; therefore, there is much greater motivation for achieving wider bandwidths on existing coaxial cable facilities or on new installations of the same kind of cable. At this time, for example, the installed cost of the coaxial cable used in the L-type systems makes up 60 to 90 per cent of the total system cost, including necessary electronics, buildings, and terminal gear. Thus, the history of long-haul cable systems has been one of striving for increased channel capacity by use of the advancing technology.

After the 12-channel K system came the 600-channel L1 system, the 1860-channel L3 system, and the 3600-channel L4 system. Presently in development is the L5 system which will have 9000 channels per coaxial line. The signal-to-noise analysis of Chap. 13 shows why

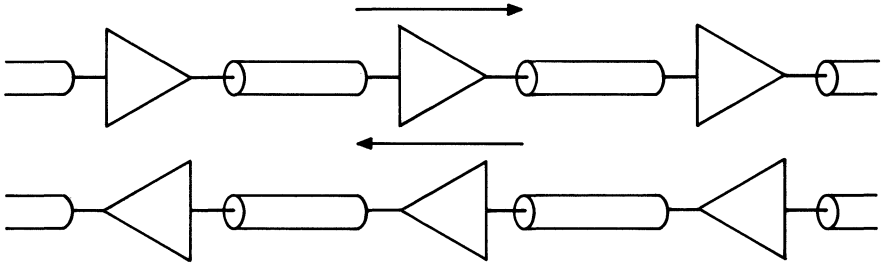
the increasing capacity of these systems has led to an eight-fold increase in the number of repeaters required for a given span (in going from L1 to L5). The treatment of misalignment and equalization in Chapters 14 and 15 considers the effects of this increase in the number of repeaters on such things as S/N penalties and transmission deviations.

Due to the high cost of replacing a failed repeater in a submarine cable system, the repeater designs are undertaken with extremely high reliability as a main objective. Although reliability is naturally a desirable feature for the repeaters of the land cable systems, their accessibility makes reliability of the kind called for in submarine systems unnecessary. In the high capacity long-haul land systems, an added degree of operational reliability is achieved by allocating a pair of the coaxials in a multiline cable to provide automatic switched protection of the remaining working lines. Installing and equipping a submarine cable for spare use is, of course, not economically feasible. The level of reliability required in the submarine cable application is achieved through a series of manufacturing controls, aging processes, and component selection, which result in a very expensive repeater (perhaps two orders of magnitude more expensive than a land system repeater). Consequently, the cost of the electronics in a submarine system is not a minor fraction of the total cost. The resulting facility is a particularly expensive commodity (costs per channel-mile are about an order of magnitude above comparable land systems), and special techniques, such as TASI (Chap. 28), are used at the terminals to maximize utilization.

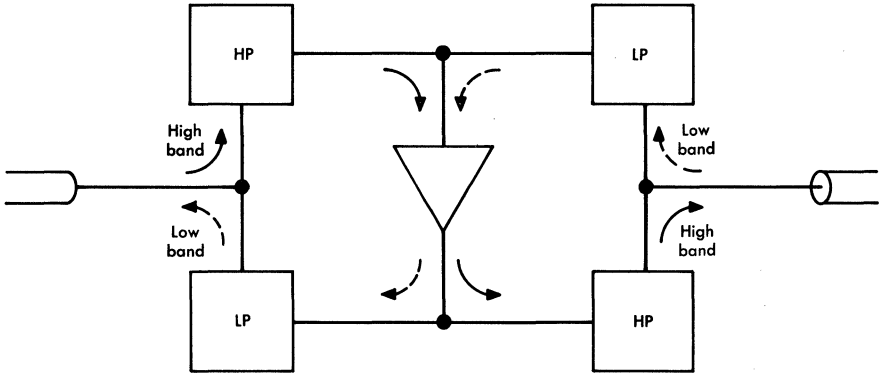
For the short-haul land systems, additional wire pairs tend to be both readily available and inexpensive. For long-haul land cable systems, there is always the possibility of converting an older system to a new one by adding the required intermediate repeater stations. Increasing the capacity of an existing submarine cable plant is, however, not an economically attractive alternative. Thus, since a new submarine cable system will always involve the laying of a new cable, the possibility of a new design of cable for that system receives much more attention than in land cable systems.

The analog cable systems of interest today are generally four-wire systems—either physical four-wire or equivalent four-wire. Typical of the former are the L-type coaxial cable systems, which use a separate coaxial line for each direction of transmission. Probably the most important examples of equivalent four-wire operation

are the submarine cable systems in which a single coaxial cable is used with frequency diversity employed to achieve the two directions of transmission. Figure 12-1 illustrates the layout for these two modes of transmission.



(a) Physical four-wire system



(b) Equivalent four-wire system

FIG. 12-1. Physical and equivalent four-wire transmission.

12.2 TRANSMISSION CONSIDERATIONS

In order to achieve a high quality of transmission, it is important to maintain as large a signal-to-noise ratio as possible. When the interference is independent of signal magnitude, as is the case for thermal noise, this objective is achieved by transmitting the signal at the maximum level which will not overload the line repeaters. Such a system is designated a *thermal noise-overload limited* system. On the other hand, when the nonlinearities of the repeaters produce significant interference due to the intermodulation of the broadband

signal itself, the optimum operating level for the line repeaters is a value lower than that established by the repeater overload point. Systems where the optimum repeater transmission level is below that which would be set on the basis of repeater overload are *thermal noise-intermodulation limited* systems. Long-haul broadband systems tend to fall in this category, whereas short-haul smaller capacity systems are usually overload limited.

In either case a predictable and stable line repeater gain is required to compensate for the loss of the associated section of cable. This is usually achieved by using negative feedback amplifiers (Chap. 16). The use of negative feedback has the additional merit of reducing the effect of the initially small amplifier nonlinearities. This improvement is usually essential to long-haul system realization. Short-haul systems often use techniques such as companding to achieve signal-to-noise advantages (Chap. 28).

An important aspect of any analog cable system is the equalization of the end-to-end transmission response. The combination of signals to be carried is applied at the system input at specified transmission levels. To permit the interconnection of systems, these signals must be delivered at the system output at specified levels, which may be different from the input levels. Consequently, in traversing the system, all of the signals must experience a specified constant amount of gain or loss independent of the transmitted frequency. The provision of the desired flat transmission response requires the correction, or equalization, of whatever deviations from nominal may exist in the response of the many elements making up the system. These deviations may or may not be time-varying and can often be associated with a specific cause. Furthermore, unless all line repeaters are at the same transmission level (i.e., the gain of the repeater exactly compensates for the loss of the associated wire pairs or cable at all frequencies), a signal-to-noise penalty results. This is because the net gain or loss along the repeatered line makes it impossible to operate all repeaters at the same optimum transmission level. These two considerations, system interconnection and optimum level control, dictate equalization along the length of the system and at the terminals to hold transmission deviations from the ideal condition within acceptable bounds. Some of these transmission deviations can be reasonably predicted in advance and thus can be compensated by fixed equalizers; other deviations, however, require adjustable equalizers for proper correction. While some adjustable equalizers may be set manually at the time of installation or

as part of regular maintenance, others (which correct for relatively rapid gain changes) may utilize some form of automatic adjustment.

In Chap. 13 the effects of thermal noise, intermodulation noise, and repeater load-carrying capacity on system performance and design are examined. That discussion assumes a system that is ideal in the sense that repeater gain exactly matches the loss of the associated cable. Chapters 14 and 15 discuss quantitatively the effects of deviations from this ideal condition and the control of these deviations by equalization. Finally, the effects of device parameters and circuit configurations on repeater performance indices are considered in Chap. 16.

Chapter 13

Analysis and Design of Analog Cable Systems

The analysis and design of analog cable systems involve a collection of analytical and empirical techniques leading to a system realization in which a particular signal-to-noise ratio is achieved over a specified length of system. The layouts of different systems are determined by the properties of the interconnecting cable, the amplifiers which compensate for the cable loss, the channel capacity specified for the system, and the system noise requirements.

The transmitted signal in all analog systems is corrupted by thermal noise, and a system design is required that limits the degree to which this occurs. In some systems, repeater nonlinearities cause additional distortion which must be considered. This chapter first considers the thermal noise in analog cable systems and the design of systems in which this is the only significant kind of interference. Later considered is the effect of intermodulation distortion on signal transmission.

13.1 THERMAL NOISE IN THE SYSTEM

The origin and nature of thermal noise have been considered in Chapters 7 and 8. It is of interest here to treat the manner in which this noise accumulates and can eventually determine the quality of transmission on the facility in question. As stated previously, it is assumed throughout these early system considerations that the system consists of a series of identical, uniformly spaced amplifiers (repeaters), connected by identical lengths of cable, and that the gains of the repeaters exactly match the loss of the intervening cable sections. The equivalent amplifier circuit for noise shown in Fig. 13-1 is used, where N_R is the equivalent noise power

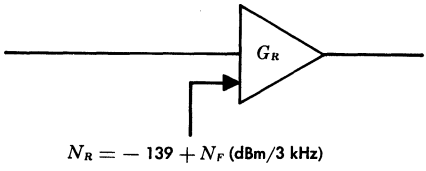


FIG. 13-1. Equivalent amplifier circuit for noise.

(3-kHz band) originating in the actual amplifier and the amplifier as shown is assumed to be noise-free. From the definition of noise figure, N_F,

$$N_R = -174 + 10 \log B_w + N_F \quad \text{dBm}$$
$$= -139 + N_F \quad \text{dBm/3 kHz}$$

Given a series of repeaters of insertion gain, G_R dB*, and noise figure, N_F, separated by cable sections of loss G_R, as shown in Fig. 13-2, then the noise at the output of the first repeater is

$$N_R + G_R \quad \text{dBm}$$

and is due only to the first repeater. Let p be the same power in milliwatts, i.e., $N_R + G_R = 10 \log p$. At the output of the second repeater, this noise appears at its original value of p. It has been attenuated by the loss of the intervening cable section and amplified by the gain of the second repeater (both $= G_R$).

At that point it joins the noise due to the second repeater, which is also p milliwatts. Thus, the noise power at the output of the second repeater is $2p$, and the corresponding noise power at the output of the nth repeater is np milliwatts. The total noise in dBm can be written 10 log np, or,

$$N_R + G_R + 10 \log n \quad \text{dBm}$$

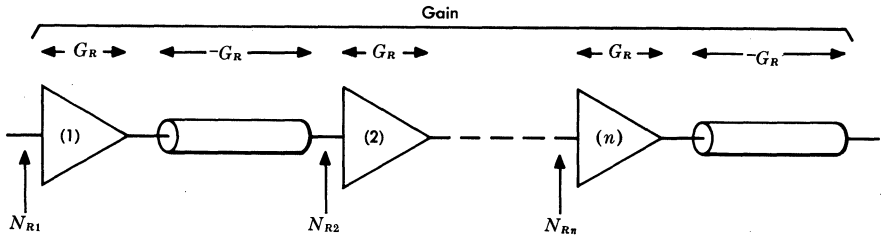


FIG. 13-2. Equivalent system for noise calculation.

*Insertion gain is applicable here because the repeater input impedance is usually conjugate to the cable impedance.

Define the transmission level at the repeater output to be C dB below zero level (Fig. 13-3).* The definition of system noise performance at a reference point of zero dB transmission level was discussed in Chap. 7, and it is of interest here to determine the noise powers associated with n repeater sections at that point. Furthermore, the noise power at zero level due to the n repeaters can be expressed as an annoyance, W_{n0} dBrnc0. From Chap. 7, 0 dBm of white noise in a 3-kHz band corresponds to 88 dBrnc as measured by a noise meter. Thus,

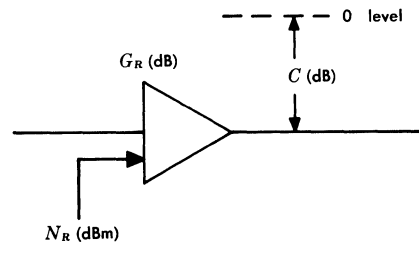


FIG. 13-3. Definition of reference point.

$$W_{n0} = N_R + G_R + 10 \log n + C + 88 \qquad \text{dBrnc0}$$

In the general case, N_R, G_R, and C may be functions of frequency. Therefore, the noise in a channel, W_{n0}, may depend on the line frequency of the channel in the system.

It is now possible to write the first of the system equations on which the remainder of this and subsequent chapters are constructed. The system thermal noise requirement, expressed as an annoyance, is defined to be W_{NS}. (The subscript N here implies thermal noise, as opposed to modulation noise, and does not refer to the number of repeaters.) Simply stated, in a system consisting of n repeaters, the thermal noise of the system must be equal to or less than the system requirement, or $W_{n0} \leq W_{NS}$. If this condition must be met initially with some margin, A_N dB, then

$$W_{n0} + A_N \leq W_{NS}$$

Thus,

$$N_R + G_R + 10 \log n + C + 88 + A_N \leq W_{NS} \qquad (13\text{-}1)$$

The allowance for margin, A_N, must take into account performance uncertainties, misalignment, and aging effects. The magnitudes of

*Exactly what point in the repeater (input, output, or somewhere in between) is used as a reference depends on overload and intermodulation considerations. It turns out frequently that the output of the repeater is a suitable reference point. For the purposes of this discussion, it will be assumed that this is the case here and that the output of every repeater is C dB below zero transmission level.

these margins are based on judgment, past experience, and additional analytical studies. The margins allowed may change as the design progresses from initial studies toward final development, but seldom will all factors be so precisely known that computed performance can be allowed to equal exactly the system requirement. In general, then, a system is designed so that its computed thermal noise, W_{n0}, is less than the system thermal noise requirement, W_{NS}, by at least some margin, A_N.

As an example of the use of Eq. (13-1), consider a system of 100 repeaters separated by cable sections each having 44-dB loss at the frequency of interest. Assume that the repeaters may operate at a transmission level no higher than −10 dB (i.e., $C = 10$ dB) because of limitations on their load-carrying capability. Assume further that no noise margin is to be allowed and that the zero level annoyance is not to exceed 40 dBrnc0. The maximum allowable repeater noise figure is to be determined.

Equation (13-1) shows that the noise figure may be increased dB-for-dB as C is decreased ($N_R = -139 + N_F$). That is, the highest noise figure is permissible if the signals are transmitted at the highest possible level. In the example this means $C = 10$ dB. Substituting into Eq. (13-1) yields

$$-139 + N_F + 44 + 10 \log 100 + 10 + 88 \leq 40$$

or

$$N_F \leq 17 \text{ dB}$$

Usually margins must be allowed in a design. If the system may become misaligned, the nominal repeater levels must be lowered by increasing C to avoid overloading the repeaters when the misalignment is positive (i.e., of a sort which increases signal amplitudes relative to the aligned condition). The system must meet the noise objective even if negatively misaligned. Finally, there may be some degradation of noise figure with time. The maximum allowable noise figure of 17 dB calculated above must be reduced by an amount equal to the sum of these factors.

A number of other interesting relationships exist among the terms in Eq. (13-1). For example, an improvement in repeater design which reduces the noise figure reduces system noise dB-for-dB. Similarly, if for any reason it is possible to reduce C, the system zero level thermal noise is reduced dB-for-dB. If the length

of the system is doubled, the $10 \log n$ term shows that the longer system is 3 dB noisier than the shorter. The effect of changing repeater spacing can also be seen. Suppose, in the preceding example, the repeater spacing is halved. The value of G_R would be reduced by 22 dB, and the $10 \log n$ term would increase by 3 dB for a fixed length of system. The net effect would be to reduce system noise by 19 dB (for the same repeater noise figure and C), to allow a 19-dB reduction in the load capacity of the repeater (for the same noise figure and system noise), or to allow a 19-dB increase in noise figure (for the same system noise and C).

Another interesting example is to evaluate the effect of changing the cable diameter, a possibility often considered in the design of submarine cable systems. Assume a 44-dB repeater gain and a 25 per cent increase in cable diameter. Since the loss of the cable is approximately inversely proportional to cable diameter, the loss of a cable section is $[1/(1 + 0.25)]\ 44 = 35$ dB. This corresponds to a 9-dB reduction in G_R and therefore a 9-dB improvement in system thermal noise performance.

13.2 LOAD CAPACITY

The calculation of total power associated with a multichannel message load is developed in Chap. 9 where it is shown that the result can be expressed in terms of the power of a single sinusoid. From Chap. 9,

$$P_S = V_0 + 0.115\,\sigma^2 + 10 \log N\tau_L - 1.4 + \Delta_C \qquad \text{dBm0}$$

where all but the last term of the right side constitute the average power of the multichannel speech load at zero level. The term P_S is defined as the power of the equivalent sine wave which the system must be able to sustain *at zero level* given the statistical structure assumed in the derivation of the multichannel load factor, Δ_C. Of prime practical interest is the corresponding load capacity, P_R, of the *repeater,* whose output is not generally 0 TLP. This is the maximum single-frequency power which the repeater can maintain at its output without overloading.

There are several possible overload phenomena, one of which is usually controlling in a particular system. These include excessive gain compression or expansion, significant degradation of modulation coefficients, and damage to the repeater. The validity of equating

the power of a broadband signal with the power, P_S, of a single-frequency signal requires that overload be independent of frequency. This is one of the considerations mentioned previously, which determines the selection of the reference point in the repeater used to define C. For the purposes of this discussion, the assumption is continued that the repeater output satisfies this condition. Accordingly, it is necessary that the power which the repeater can maintain at its output, referred to zero level, must at least equal P_S. Thus,

$$P_R + C \geq P_S$$

As with the thermal noise analysis, various uncertainties and deviations from ideal make it desirable to allow some margin, A_P, against overload. Then,

$$P_R + C - A_P \geq P_S \tag{13-2}$$

This is the second of the system equations with which the basic system analysis is carried out. Using Eqs. (13-1) and (13-2), it is possible to proceed with the design of a thermal- and overload-limited system. The analysis of such a system assumes that modulation noise, which results from the repeater nonlinearities is negligible compared to thermal noise. In most practical situations it will be necessary to make at least rough computations along the lines described in Section 13.7 to verify that a design on the basis of the overload-limited assumption is valid.

13.3 THERMAL NOISE- AND OVERLOAD-LIMITED SYSTEMS

For the analysis of a thermal noise- and overload-limited system, Eqs. (13-1) and (13-2) are used and are repeated here for convenience:

$$N_R + G_R + 10 \log n + C + 88 + A_N \leq W_{NS}$$

$$P_R + C - A_P \geq P_S$$

The total cable loss (L_S) at the frequency of interest, usually that of the top channel at this point in a design, must be compensated by the gains of the n repeaters. Thus,

$$L_S = nG_R \qquad \text{dB}$$

and

$$G_R = L_S/n \qquad \text{dB} \tag{13-3}$$

Combining Eqs. (13-3) and (13-1) to eliminate G_R,

$$N_R + \frac{L_S}{n} + 10 \log n + C + A_N + 88 \leq W_{NS} \tag{13-4}$$

Solving Eq. (13-2) for C,

$$C \geq P_S - P_R + A_P \tag{13-5}$$

Solving Eq. (13-4) for C,

$$C \leq W_{NS} - N_R - \frac{L_S}{n} - 10 \log n - A_N - 88 \tag{13-6}$$

Combining Eqs. (13-5) and (13-6) to eliminate C,

$$\frac{L_S}{n} + 10 \log n \leq W_{NS} - N_R - (A_N + A_P) - (P_S - P_R) - 88 \tag{13-7}$$

Equation (13-7) defines for the several system parameters the basic relationship which must be preserved throughout the design of a thermal noise- and overload-limited system. Which of the parameters are known and which are unknown depends upon the application of the equation. For example, it will often be the case that the medium to be used in a particular system design is specified in advance. This may result from a need to make more efficient use of existing facilities. It may also be that the cost of producing a new cable design specifically for the system is prohibitive. Thus, the cable loss per unit length is frequently not a system design variable, and L_S is a function only of frequency and system length. The problems in achieving the necessary performance are usually greatest at the highest frequencies where the cable attenuation is maximum, and it is common in the early design stages to concentrate attention on the highest transmitted frequency (top channel). As a result, if the top channel is made to satisfy the thermal noise requirement, with transmission levels and amplifier noise figures which are the same at all frequencies, then the lower frequency channels will be quieter than required. Signal shaping, which is discussed later, provides a means of improving performance at the top channels at the expense of the better-than-necessary lower frequency performance.

The number of repeaters, n, is often the quantity to be determined. In this connection it must be observed that the left side of Eq. (13-7) has a minimum. This means that it is possible that no choice of n will satisfy the conditions. Under these circumstances no satisfactory system is possible under the assumptions that have been used.

The system thermal noise requirement, W_{NS}, is usually a known function of length and is the same for all channels. It is based on some overall Bell System plan for achieving a particular grade of service. Depending upon the application, repeater performance parameters, N_R and P_R, may be the quantities to be evaluated and given to the repeater designer as requirements, or they may be taken as estimates of achievable repeater performance. Since talker statistics are usually given, P_S will be the load capacity required by the assumed bandwidth. The margin terms, A_N and A_P, will reflect the effect of uncertainties, misalignment, and such refinements as signal shaping. Which of these factors are fixed and which must be established as the analysis evolves will usually differ from system to system. Further clarification of the use of Eq. (13-7) during a system design is best achieved by some examples.

Illustrative Designs

Consider a short-haul system to be designed for 600-channel capacity which will use repeaters for which a +10 dBm load capacity can probably be achieved. For the present example it is postulated that the system is thermal- and overload-limited, a fact not usually known before the design begins. The system is to use a cable for which the loss at the top channel is 3 dB per mile and is to operate over distances up to 250 miles. On the basis of laboratory studies and similar amplifiers used elsewhere, a repeater noise figure of about 8 dB can be assumed achievable. Margins of 3 dB on both noise and load capacity will be provided initially. From Fig. 3-6 the system noise objective is 34 dBrnc0.

First calculate P_S:

$$P_S = V_0 + 0.115\,\sigma^2 - 1.4 + 10 \log \tau_L N + \Delta_C \quad \text{dBm0}$$

For $V_0 = -12.5$ vu, $\sigma = 5$, $\tau_L = 0.25$, and $N = 600$,

$$\Delta_C \approx 12$$

and

$$P_S \approx 23 \text{ dBm0}$$

Also,

$$L_S = (250 \text{ miles}) \times (3 \text{ dB/mile}) = 750 \text{ dB}$$

Substituting the known quantities into Eq. (13-7),

$$\frac{750}{n} + 10 \log n \leq 34 - (-139+8) - (3+3) - (23-10) - 88$$

or

$$\frac{750}{n} + 10 \log n \leq 58$$

Solving for the smallest integer which satisfies the inequality yields $n = 17$. The corresponding repeater spacing will be $250/17 = 14.7$ miles, and the repeater gain must be $750/17 = 44.1$ dB. The transmission level is established by calculating C from either Eq. (13-1) or (13-2), recognizing that slightly different values may result since an integral value of n was specified. From Eq. (13-2)

$$C = P_S - P_R + A_P$$

$$= 23 - 10 + 3 = 16 \text{ dB}$$

Thus, the repeater output is established as a −16 TLP. The important characteristics can be summarized:

600-channel capacity;
14.7-mile repeater spacing;
Initial noise performance of about 31 dBrnc0, which allows 3-dB margin with respect to requirement;
8-dB repeater noise figure;
7-dBm repeater load, allowing 3-dB margin with respect to the repeater capacity;
−16 TLP at repeater output;
44.1 dB repeater gain at top channel.

It may be of interest to determine the effect of increasing the channel capacity by 50 per cent to 900 channels. In the expression for P_S, 10 log $\tau_L N$ is increased by 1.8 dB, and Δ_C is decreased by 0.4 dB; therefore, there is a net increase of 1.4 dB in P_S. If it is assumed that attenuation is proportional to the square root of frequency, the new top frequency loss $L_S = 750 \sqrt{1.5} = 917$. It is assumed that repeater noise figure and load-carrying capacity remain unchanged for the broader band repeater. Margins are also held constant. Then,

$$\frac{917}{n} + 10 \log n \leq 56.6$$

and

$$n = 22$$

$$G_R = 41.7 \text{ dB}$$

$$\text{spacing} = 11.4 \text{ miles}$$

$$C = P_S - P_R + A_P$$

Since P_R and A_P are assumed unchanged, the value of C increases 1.4 dB due to the increase in P_S. This illustrates the fact that as bandwidth increases, repeater spacing must be decreased such that the repeater gain decreases slightly. Increase in bandwidth also makes it necessary to operate repeaters at a lower transmission level as shown by the increase in C. These effects are even more marked when the effects of modulation noise and increasing bandwidth on repeater performance are incorporated.

Another use of Eq. (13-7) indicates the range of repeater parameters, N_F and P_R, which satisfies a 20-mile spacing requirement. It could be, for example, that at this interval there are existing facilities in large quantities whose continued use is worthwhile. Leaving the other parameters unchanged, Eq. (13-7) yields

$$P_R - N_F \geq 12.8$$

Thus any combination of amplifier noise figure (in dB) and load capacity (in dBm) for which this relation is preserved will permit a 20-mile spacing to be realized with satisfactory noise performance. For the originally specified 8-dB amplifier noise figure, the required P_R now is 21 dBm. This is 11 dB higher than for the 14.7-mile spacing of the first part of this example. This is the increased cable loss between repeaters, reduced by a factor reflecting the smaller number of repeaters (noise sources).

13.4 INTERMODULATION DISTORTION

A detailed discussion of the nature and origin of the nonlinearities in analog systems is included in Chap. 10. The effect of repeater nonlinearities on system design is now considered. The deviation from perfect linearity in a repeater is usually very small by normal standards; however, when a signal passes through many repeaters in tandem, the effect of nonlinear distortion can accumulate to a

significant value. It results in modulation noise where the intermodulation of many signals produces an interference subjectively indistinguishable from white noise. For this modulation noise to be acceptable, it may be necessary to impose surprisingly stringent linearity requirements on each repeater. In the case of long-haul high-capacity systems, for example, the permissible third harmonic distortion at the output of a repeater, given a milliwatt of fundamental signal power, may be on the order of 10^{-10} to 10^{-13} milliwatts. This degree of linearity is usually difficult to achieve, and its relation to repeater design is considered at greater length in Chap. 16. It is the system aspect of nonlinear distortion that is of immediate interest.

Accumulation of Modulation Noise

Consider the repeaters of Fig. 13-4. Of interest is the way in which the nonlinear distortion originating in one repeater combines

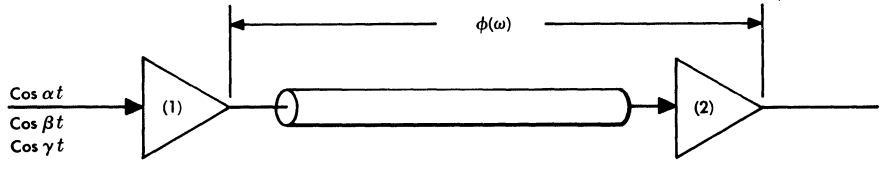


FIG. 13-4. Section of repeatered line.

with the distortion in subsequent repeaters. Associated with any repeater section will be a phase characteristic like that of Fig. 13-5 in which it is possible to describe the phase shift by

$$\phi(\omega) = m\omega + b$$

The assumption of a phase shift which varies linearly with frequency is equivalent to an assumption of no delay distortion. In the calculation of overall intermodulation distortion, this seems to be a good approximation for most long-haul coaxial systems.* This does not mean that the phase linearity is adequate without further

*The directional filters required by equivalent four-wire systems do introduce sufficient delay distortion to make modulation computation results based on the assumption of linear phase overly pessimistic.

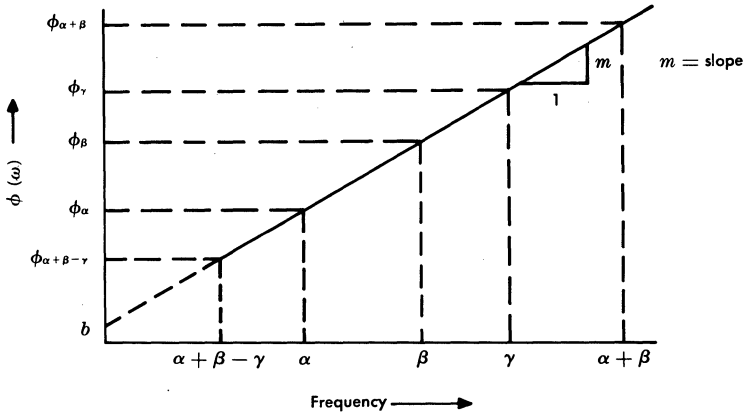


FIG. 13-5. Ideal phase-frequency characteristic of a repeater section.

equalization for the transmission of such signals as high speed digital or video signals which are very vulnerable to even small amounts of delay distortion.

To determine how the various products formed in one repeater combine with those formed in subsequent repeaters, it is convenient to compare at the second repeater of Fig. 13-4 the phase of the locally generated product with the phase of the product generated in the preceding repeater.

With the fundamental signals proportional to cos αt and cos βt at the input to the first repeater, there will be, among others, a component at the output proportional to $\cos(\alpha+\beta)t$ resulting from the second order nonlinearity of the repeater. From Fig. 13-5, the fundamentals, in traversing to the output of the second repeater, will undergo phase shifts of ϕ_α and ϕ_β, while the product frequency signal will undergo a phase shift of $\phi_{\alpha+\beta}$. The fundamental signals arriving at the second repeater are proportional to $\cos(\alpha t+\phi_\alpha)$ and $\cos(\beta t+\phi_\beta)$. They generate in the second repeater an $\alpha+\beta$ product proportional to cos $[(\alpha+\beta)t+\phi_\alpha+\phi_\beta]$. At the second repeater output, the $\alpha+\beta$ product from the first repeater is proportional to cos $[(\alpha+\beta)t+\phi_{\alpha+\beta}]$. The manner of combination of the two products depends on the relationship between $\phi_{\alpha+\beta}$ and $\phi_\alpha+\phi_\beta$. From Fig. 13-5,

$$\phi_\alpha = m\alpha + b$$

$$\phi_\beta = m\beta + b$$

thus,

$$\phi_\alpha + \phi_\beta = m(\alpha+\beta) + 2b$$

and finally,

$$\phi_{\alpha+\beta} = m(\alpha+\beta) + b$$

The difference in phase between the products generated by the consecutive repeaters is seen to be equal to b. It should be clear that these two products reinforce each other if b is an integral multiple of 2π. On the other hand, they cancel if b is an odd integral multiple of π. For other values of b, the products combine to a resultant between these two extremes.

In reality, things are not quite as neat as in this analytical model. There is usually at least some curvature in the phase characteristic over a broadband. As shown in Fig. 13-6, this leads to a change in the effective b as a function of frequency. Furthermore, the transmission gain and phase of one repeater and cable section will not be mathematically identical to all the others. As a result, if the $\alpha+\beta$ product were measured on a particular system, while varying the frequency of the product and one of the fundamentals,

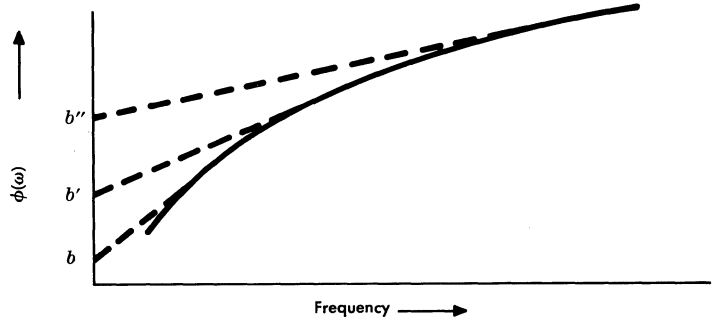


FIG. 13-6. Example of an actual phase-frequency characteristic of a repeater section.

a characteristic similar to that of Fig. 13-7 might be observed. If the object of the computation is to compute the single-frequency interference due to the intermodulation of discrete frequencies (e.g., pilots), it is therefore necessary to know the transmission characteristic of each repeater section rather precisely. However, the real

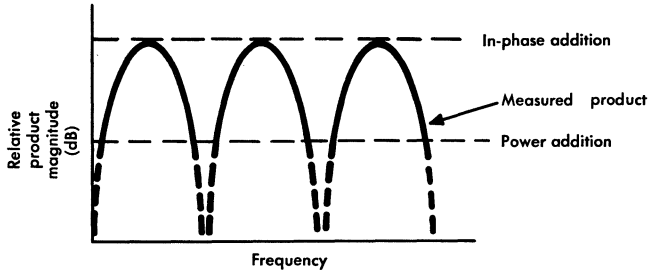


FIG. 13-7. Possible second order product addition.

object of the computations in this chapter is to estimate the intermodulation noise produced by the intermodulation of many channels. Under these circumstances the total interference falling on a particular channel will be the resultant of many products whose fundamentals appear at different points in the transmission band. These products will correspond to varying degrees of reinforcement and cancellation. The total interference from a large number of channels and repeaters will tend to average the accumulation from repeater to repeater somewhere between systematic in-phase addition and perfect cancellation. An assumption of random addition is usually satisfactory for such systems. It shall therefore be assumed that $\alpha + \beta$ and $\alpha - \beta$ products accumulate on a random basis from repeater to repeater, which is equivalent to power addition.

Next, the $\alpha + \beta - \gamma$ type distortion is considered in terms of the ideal model. At the output of the first repeater, there will be signals proportional to cos αt, cos βt, cos γt, and cos $(\alpha + \beta - \gamma)t$, each of which will traverse the system to the output of the second repeater. They undergo phase shift of ϕ_α, ϕ_β, ϕ_γ, and $\phi_{\alpha+\beta-\gamma}$, respectively. At the output of the second repeater, the locally generated $\alpha+\beta-\gamma$ product is proportional to cos $[(\alpha+\beta-\gamma)t + \phi_\alpha+\phi_\beta-\phi_\gamma]$, while the product originating in the first repeater is proportional to cos $[(\alpha + \beta - \gamma)t + \phi_{\alpha+\beta-\gamma}]$. Again, the manner of accumulation depends on the relationship between $\phi_{\alpha+\beta-\gamma}$ and $\phi_\alpha + \phi_\beta - \phi_\gamma$. From Fig. 13-5

$$\phi_\alpha = m\alpha + b$$

$$\phi_\beta = m\beta + b$$

$$\phi_\gamma = m\gamma + b$$

therefore,

$$\phi_\alpha + \phi_\beta - \phi_\gamma = m\ (\alpha + \beta - \gamma) + b$$

Finally,

$$\phi_{\alpha+\beta-\gamma} = m(\alpha + \beta - \gamma) + b$$

which is identical to $\phi_\alpha + \phi_\beta - \phi_\gamma$, regardless of the value of b. As a result, $\alpha + \beta - \gamma$ products will add in phase for any value of b, not just for special values as is the case with the $\alpha + \beta$ product. It therefore can be concluded that the realization of a linear phase characteristic in a system, which is often desirable from other viewpoints, has an undesirable side effect of in-phase addition of $\alpha + \beta - \gamma$ products. As discussed previously, a real system will deviate somewhat from the ideal model used in the preceding derivation; however, unlike the $\alpha + \beta$ product, the $\alpha + \beta - \gamma$ product is of a form where disturbed and disturbing channels can all come from the same portion of the band. Thus, even a delay characteristic with curvature can approach in-phase addition as long as the actual characteristic falls reasonably close to a piece-wise linear approximation. Therefore, the assumption of in-phase addition deduced from the ideal model has not proved to be excessively pessimistic when compared to measurements on actual systems.

The discussion of repeater-to-repeater product accumulation can now be summarized. The possible intermodulation products up to third order can be expressed as $k_a\alpha + k_b\beta + k_c\gamma$, where $k_x = 0, \pm 1, \pm 2$, or ± 3, subject to the constraint that

$$| k_a | + | k_b | + | k_c | \leq 3$$

The products will accumulate in phase (voltage addition) for those cases where

$$k_a + k_b + k_c = 1$$

Otherwise, the products will accumulate randomly (power addition). Therefore, power or random addition is assumed for 2α, 3α, $\alpha + \beta$, $2\alpha + \beta$, $\alpha + \beta + \gamma$, and $\alpha - \beta - \gamma$; in-phase or voltage addition for $\alpha + \beta - \gamma$ and $2\alpha - \beta$. The term representing the cumulation of power addition products generated by n repeaters will be $10 \log n$; for voltage addition products, $20 \log n$.

System Modulation Requirements

In Chap. 10, the coefficients M_2 and M_3 were defined as the power of the second and third harmonics corresponding to a 0 dBm fundamental. This assumes that the nonlinear behavior under consideration is adequately described by adding a quadratic and a cubic term to the linear term which represents the desired amplification. On the basis of this power series model, it was shown that the second harmonic changes 2 dB and the third harmonic 3 dB for each 1-dB change in fundamental. The power series representation also predicts behavior independent of frequency. In Chap. 16 it is pointed out that, approximately at least, the harmonic product is reduced by the magnitude of feedback at the product frequency. Since feedback may be a function of frequency, the M coefficient will also tend to be a function of product frequency but independent of fundamental frequency (e.g., fundamentals at 1 and 3 MHz produce the same 4-MHz power as fundamentals at 1.9 and 2.1 MHz).

Another consequence of the power series model is the simple relationship between the harmonic coefficients and the sum and difference coefficients:

$$M_{\alpha\pm\beta} = M_{2\alpha} + 6$$

$$M_{\alpha\pm\beta\pm\gamma} = M_{3\alpha} + 15.6$$

$$M_{2\alpha\pm\beta} = M_{3\alpha} + 9.6, \text{ etc.}$$

Therefore, at least to the degree that the model is valid, if the harmonic performance of an amplifier is known, all other modulation product magnitudes are also determined.

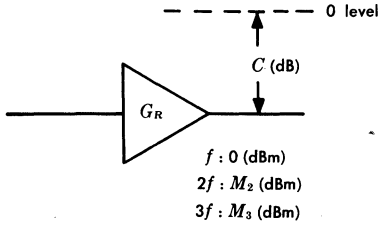


FIG. 13-8. Fundamental and harmonic magnitude relations at repeater output.

The next step in the derivation of system modulation equations is to define modulation coefficients for the system as a whole, which are analogous to the M coefficients for individual repeaters.

From Fig. 13-8, it is apparent that a 0 dBm0 signal appears at the output of each repeater at $-C$ dBm. By definition, 0 dBm fundamentals at the repeater output produce M_2 dBm second harmonics and M_3 dBm third harmonics at

the repeater output. Therefore, on the basis of the power series model, $-C$ dBm fundamentals produce $(M_2 - 2C)$ dBm second harmonics and $(M_3 - 3C)$ third harmonics at the repeater output. These products experience C dB of gain between repeater output and zero level; consequently, the harmonic power at zero level generated in one repeater by 0 dBm0 fundamentals will be $(M_2 - C)$ dBm and $(M_3 - 2C)$ dBm.

Next the distortion products are calculated as they exist at the output of the system as the result of the contributions of the many repeaters involved. They will be defined at zero level, and use will be made of the earlier results showing how the different kinds of products add from repeater to repeater. Chapter 10 defines H_x as the power of the x-type product at the system output, referred to zero level, which results from the application to the system of 0 dBm0 fundamental signals. Consequently, for second order products,

$$H_x = M_x - C + 10 \log n \qquad \text{dBm0}$$

For power-adding third order products,

$$H_x = M_x - 2C + 10 \log n \qquad \text{dBm0}$$

For voltage-adding third order products,

$$H_x = M_x - 2C + 20 \log n \qquad \text{dBm0}$$

Specifying a few products most commonly of interest and using the power series relationships between the cross-product amplitudes and the amplitude of the harmonic products,

$$H_{\alpha \pm \beta} = H_{2\alpha} + 6 = M_2 - C + 10 \log n + 6 \qquad \text{dBm0}$$

$$H_{2\alpha+\beta} = H_{3\alpha} + 9.6 = M_3 - 2C + 10 \log n + 9.6 \qquad \text{dBm0}$$

$$H_{2\alpha-\beta} = M_3 - 2C + 20 \log n + 9.6 \qquad \text{dBm0}$$

$$H_{\alpha+\beta-\gamma} = M_3 - 2C + 20 \log n + 15.6 \qquad \text{dBm0}$$

It is now possible to derive for system modulation distortion some relationships similar to those for thermal noise and overload performance. These equations permit the analysis of the general case in which it is not initially known whether the system will be limited by overload or modulation distortion factors.

It is shown later that the annoyance associated with a multichannel speech signal can be related to the single-frequency distortion by the terms K_2 and K_3 as follows:

$$W_2 = M_2 - C + 10 \log n + K_2 \qquad \text{dBrnc0} \qquad (13\text{-}8)$$

$$W_3 = M_3 - 2C + 20 \log n + K_3 \qquad \text{dBrnc0} \qquad (13\text{-}9)$$

where W_2 and W_3 are the system noise due to the totality of second and third order products, respectively. It is implied by the use of $20 \log n$ term in Eq. (13-9) that the third order distortion performance is determined by those third order products which add voltage-wise across the system. The terms K_x depend on such things as talker statistics and system load. Designating W_{2S} and W_{3S} as the second and third order system noise requirements and allowing margins A_2 and A_3 results in:

$$M_2 - C + 10 \log n + K_2 + A_2 \leq W_{2S} \qquad (13\text{-}10)$$

$$M_3 - 2C + 20 \log n + K_3 + A_3 \leq W_{3S} \qquad (13\text{-}11)$$

Equations (13-10) and (13-11) are the system equations for second and third order modulation noise corresponding to Eqs. (13-1) and (13-2) for thermal noise and overload. In general, W_2, K_2, and M_2 are functions of frequency so that different channels of a given system will tend to have different amounts of noise. As mentioned in connection with thermal noise, the technique of signal shaping offers a method of making noise in all channels approximately equal. Again, the technique in laying out a new system is to compute modulation noise in the top channel for flat levels (C constant) and to include the estimated effect of signal shaping in the margin terms. When the system has been completely defined, it is possible to use the equations of Section 13.2 in conjunction with the methods developed later in this chapter to optimize the shaping and to compute expected performance as a function of frequency.

The validity of the system equations depends on how well the M coefficients actually describe the nonlinear behavior of the repeater. The accuracy with which the M coefficients (and power series approach on which they are based) represent actual nonlinear behavior in a particular case depends on various factors. A single dominant source of modulation within each repeater and negligible higher

order terms in the power series are some important conditions which must be satisfied. If the assumption that product power is a function only of product frequency and not of fundamental frequency is to be satisfied, the reference location in the repeater to which fundamental and product powers are referred must be carefully selected. In general, a suitable choice is the actual location where the modulation occurs or any other point that differs in level from this point by a constant amount at all frequencies. In what follows, it will again be necessary to relate the modulation reference point to zero transmission level, and it will be desirable to make this level difference C. For the moment, it will be assumed that the reference point used previously, the repeater output, is a satisfactory one for the definition of the M coefficients.

Intermodulation with Speech Load

So far in this chapter the system relationships have been derived without considering the nature of the signal being transmitted. The factors K_2 and K_3 were used to relate the harmonic distortion coefficients M_2 and M_3 to the noise generated by a multichannel message load. For present purposes, the message load is assumed to be a multichannel speech signal, having the properties described in Chap. 10. The relationship between the disturbance (annoyance) associated with the intermodulation of a multichannel speech load and the single-frequency distortion indices, M_2 and M_3, is also derived in Chap. 10. Equation (10-27) is repeated for convenience:

$$W_x = H_x + \eta_x V_0 + 0.115\lambda_x\sigma^2 + 10 \log U_x \tau^{\mu_x}$$

$$- (1.4\eta_x + C_{w_x}) + 88 \qquad \text{dBrnc0}$$

This equation defines the annoyance at zero level due to the x-type product only; Eqs. (13-8) and (13-9) define the *total* second and *total* third order noise, respectively. In the case where only $\alpha + \beta$ type second order distortion is important,

$$W_{\alpha+\beta} = W_2 = H_{\alpha+\beta} + \eta_{\alpha+\beta} V_0 + 0.115\lambda_{\alpha+\beta}\sigma^2$$

$$+ 10 \log U_{\alpha+\beta}\, \tau^{\mu_{\alpha+\beta}} + 88 - (1.4\eta_{\alpha+\beta} + C_{w_{\alpha+\beta}})$$

Since $H_{\alpha+\beta} = H_{2\alpha} + 6$, and Eq. (13-8) defines $W_2 = H_{2\alpha} + K_2$, then, for this case,

$$K_2 = K_{\alpha+\beta} = \eta_{\alpha+\beta}V_0 + 0.115\lambda_{\alpha+\beta}\sigma^2$$

$$+ 10 \log U_{\alpha+\beta}\, \tau^{\mu_{\alpha+\beta}} + 88$$

$$- (1.4\eta_{\alpha+\beta} + C_{w_{\alpha+\beta}}) + 6$$

If the channel of interest incurs significant distortion from both $\alpha+\beta$ and $\alpha-\beta$ type products, which is not the case for a top channel analysis,

$$K_2 = K_{\alpha+\beta} \text{ "+" } K_{\alpha-\beta}$$

In the case of third-order distortion, it is likely that only the voltage-adding products of the $2\alpha - \beta$, $\alpha + \beta - \gamma$ type need be considered. Accordingly,

$$K_3 = K_{2\alpha-\beta} \text{ "+" } K_{\alpha+\beta-\gamma}$$

where

$$K_{2\alpha-\beta} = \eta_{2\alpha-\beta}V_0 + 0.115\lambda_{2\alpha-\beta}\sigma^2 + 10 \log U_{2\alpha-\beta}\, \tau^{\mu_{(2\alpha-\beta)}}$$

$$+ 88 - (1.4\eta_{2\alpha-\beta} + C_{w_{2\alpha-\beta}}) + 9.6$$

and

$$K_{\alpha+\beta-\gamma} = \eta_{\alpha+\beta-\gamma}V_0 + 0.115\lambda_{\alpha+\beta-\gamma}\sigma^2 + 10 \log U_{\alpha+\beta-\gamma}\, \tau^{\mu_{\alpha+\beta-\gamma}}$$

$$+ 88 - (1.4\eta_{\alpha+\beta-\gamma} + C_{w_{\alpha+\beta-\gamma}}) + 15.6$$

In Eq. (10-27), x then takes on the value of any product type such as $\alpha + \beta$ or $2\alpha - \beta$. The terms V_0, σ, and τ are the talker statistics, and the terms C_{w_x}, λ_x, η_x, and μ_x depend only on x as discussed in Chap. 10. The term U_x is the total number of products of type x, originating anywhere in the transmission band, that can fall into a channel at the line frequency under study. It is a function of both x and line frequency. As mentioned previously, in preliminary system layouts the top frequency is usually used. The term $U_x \tau^{\mu_x}$ represents the average number of products. This way of multiplying the power of one product by the average number of products is only valid if all products of a particular type falling at a particular frequency have the same power. This condition is met only if C is independent

of frequency and M_2 and M_3 have been defined at a location which makes them independent of fundamental frequencies. These conditions have indeed been assumed, and the expressions for K_x can be evaluated for each type of product.

In the initial stages of system design, only the total noise requirement is specified. The allocation of this total requirement to W_{NS}, W_{2S}, and W_{3S} is discussed in the following.

13.5 ALLOCATION OF TOTAL SYSTEM NOISE TO THE POSSIBLE CONTRIBUTORS

For a system in which only one of the noise sources is important, it is a straightforward matter to allocate the total system noise requirement, W_{TS}, to that source; however, more than one source may be significant. This may be determined by trial solutions to the modulation noise system equations after a tentative design based on thermal noise and overload. In such cases, it is important to allocate the total to the separate sources in such a way as to optimize overall performance for a specified repeater performance and spacing. The optimum allocation will depend on the relative importance of the thermal and the modulation noise and whether the dominant modulation noise is second or third order. The optimization process amounts to selecting the best transmission level for a given set of conditions. The variations of the several types of noise with transmission level, as indicated by C, are shown in Fig. 13-9.

Consider first the case where only thermal noise and second order modulation noise are significant and where W_T is the total noise in dBrnc0.

$$W_T = W_2 \text{ "+" } W_N$$

Define $W_{x0} = 10 \log p_{x0}$, where the subscript zero indicates that this is the value of noise from the appropriate equation with C set equal to zero.

Then

$$W_2 = W_{20} - C = 10 \log p_{20} + 10 \log \left[\log^{-1} \left(\frac{-C}{10} \right) \right]$$

$$W_2 = 10 \log p_2 = 10 \log \left[p_{20} \log^{-1} \left(\frac{-C}{10} \right) \right]$$

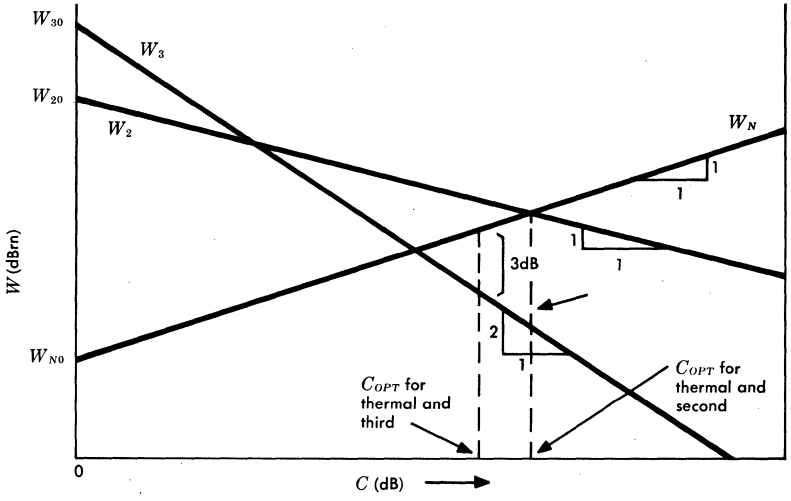


FIG. 13-9. Determination of optimum C.

therefore,

$$p_2 = p_{20} \log^{-1}\left(\frac{-C}{10}\right) = p_{20} e^{-\frac{(\ln 10)C}{10}} = p_{20} e^{-y}$$

where

$$y = \frac{(\ln 10)\, C}{10}$$

Similarly, since

$$W_N = W_{N0} + C,$$

$$p_N = p_{N0} e^{y}$$

and

$$p_T = p_2 + p_N = p_{N0} e^{y} + p_{20} e^{-y}$$

To minimize p_T with respect to $y\,(C)$, dp_T/dy is set equal to zero. This leads to

$$p_{N0} e^{y} - p_{20} e^{-y} = 0$$

which indicates that the total noise in a thermal and second order modulation noise limited system is minimized if the transmission level is adjusted to make the zero level contribution from each of the sources equal, and 3 dB less than the total.

In a similar way it can be shown that the total noise in a thermal and third order modulation noise limited system is minimized if

$$p_{N0}e^{y} - 2p_{30}e^{-2y} = 0$$

The total noise in this case is therefore minimized if C is adjusted to make the zero level contribution of the thermal noise twice that of the third order noise (i.e., $W_{3S} = W_{NS} - 3$). This means $W_{NS} = W_{TS} - 1.8$, and $W_{3S} = W_{TS} - 4.8$. If both second and third order modulation distortion must be considered, the optimum relationship becomes

$$p_N = p_2 + 2p_3$$

However, this is not usually the case.

In summary, the allocations will be as follows:

System description	Allocation to thermal noise	Allocation to modulation noise
Thermal and second limited	$W_{TS} - 3$	$W_{TS} - 3$
Thermal and third limited	$W_{TS} - 1.8$	$W_{TS} - 4.8$
Overload limited	W_{TS}	negligible

The parameter C is selected to realize these allocations for the repeater characteristics and spacing given.

13.6 SUMMARY OF BASIC SYSTEM RELATIONS

It is convenient at this point to review the several basic relationships developed thus far and to combine them in ways which will aid in the analysis of any kind of analog system, whatever the limitations.

The thermal noise requirement [Eq. (13-1)] is

$$N_R + G_R + C + 10 \log n + 88 + A_N \leq W_{NS}$$

The overload requirement [Eq. (13-2)] is

$$P_R + C - A_P \geq P_S$$

The second-order requirement [Eq. (13-10)] is

$$M_2 - C + 10 \log n + K_2 + A_2 \leq W_{2S}$$

The third-order requirement [Eq. (13-11)] is:

$$M_3 - 2C + 20 \log n + K_3 + A_3 \leq W_{3S}$$

The cable loss—repeater gain requirement [Eq. (13-3)] is:

$$L_S = nG_R; \; G_R = \frac{L_S}{n}$$

Previously in this chapter, Eqs. (13-1), (13-2), and (13-3) were combined to eliminate C and G_R and obtain the fundamental *thermal- and overload-limited system equation,* **S1:**

$$\frac{L_S}{n} + 10 \log n \leq (W_{NS}-N_R) - (A_N+A_P) - (P_S-P_R) - 88 \quad \textbf{(S1)}$$

Equations (13-1), (13-10), and (13-3) can be combined to eliminate C and G_R and obtain the *thermal and second order limited system equation,* **S2:**

$$\frac{L_S}{n} + 20 \log n \leq (W_{NS}+W_{2S}) - (N_R+M_2) - (A_N+A_2) - (88+K_2) \quad \textbf{(S2)}$$

Likewise, Eqs. (13-1), (13-11), and (13-3) can be combined to obtain the *thermal and third order limited system equation,* **S3:**

$$\frac{L_S}{n} + 20 \log n \leq \left(W_{NS} + \frac{W_{3S}}{2}\right) - \left(N_R + \frac{M_3}{2}\right) - \left(A_N + \frac{A_3}{2}\right) - \left(88 + \frac{K_3}{2}\right) \quad \textbf{(S3)}$$

If the equations are used to determine n for a particular set of parameters (bandwidth, total noise, talker statistics, margins, and repeater performance), it will not usually be known beforehand which of the three equations **(S1, S2, or S3)** is the limiting one. The question, then, is to find the minimum n for each case. The largest of these minima will be controlling and will identfy whether overload, second order modulation, or third order modulation is determining the choice of C. In carrying out this procedure, the results

of the discussion in Section 13.5 require that:

In equation **S1**, the overload-limited case,

$$W_{NS} = W_{TS}$$

In equation **S2**, the second order modulation-limited case,

$$W_{NS} = W_{2S} = W_{TS} - 3$$

In equation **S3**, the third order modulation-limited case,

$$W_{NS} = W_{TS} - 1.8; \; W_{3S} = W_{TS} - 4.8$$

If **S2** and **S3** give similar values of n, it will be necessary to reoptimize and increase n somewhat to reflect the contribution of both types of modulation to the total noise.

In the derivation resulting in the system equations, it has been assumed that the repeater output was a suitable reference point in defining C. This presumed that the overload characteristic defined there was flat with frequency and that modulation products referred to that point were independent of fundamental frequencies (though they could be functions of product frequency). Even if this is true of the basic amplifier making up the repeater, in those cases where there is a significant shaped loss between the amplifier output and the repeater output, it is not true of the overall repeater. It is interesting to see what happens to the equation if some other reference location within the repeater turns out to be better—for example, the collector of the output transistor. In this case, C can be defined as the power in dBm at zero transmission level corresponding to 1 volt at the collector. The M coefficients are defined in terms of the modulation product voltage at the collector resulting from a 1-volt fundamental at the collector. The overload point of the repeater would be defined as the collector signal voltage at which overload occurs. With these modified definitions, Eqs. (13-2), (13-3), (13-10), and (13-11) can be derived exactly as they have been stated for the case where repeater output is used as reference. The thermal noise equation must, however, be modified. Define Q as the signal power at the repeater output corresponding to 1 volt at the collector. Note that if Q is a constant, the repeater output is as good a reference point as the collector, but this is not true if Q is a function of frequency. Previously, it was shown that the noise at the output of the last repeater is $N_R + G_R + 10 \log n$ dBm. By the definition of Q, the voltage at the collector is then $N_R + G_R + 10 \log n - Q$ dBV. Using the new definition of C, the noise at zero level will then be $N_R - Q + G_R + C + 10 \log n + 88$ dBrnc0.

Comparing this to the previous result, it can be seen that this more general case will result in all equations having exactly the same form as previously if a new effective N'_R is defined such that $N'_R = N_R - Q$. The quantities P_R, M_2, M_3, and C will appear as before but will have different numerical values for a given case because of the change in their definitions. Since this approach is only necessary if Q is a function of frequency, N'_R will also be a function of frequency, even if N_R was not. However, N_R was never constrained to be constant, and, in general, will not be so; therefore, this does not introduce any new complexity.

13.7 DESIGN OF AN ANALOG CABLE SYSTEM

At this point, some examples will be helpful to illustrate the significance and use of the derived system equations. Equations **S1, S2,** and **S3,** in conjunction with Eq. (13-3), provide the basis for the analysis and design of analog systems. They can be used, for example, to establish repeater requirements, repeater spacing, noise performance, and maximum capacity solutions.

Example 13.1

Consider first a situation in which some exploratory repeater design work has provided a good idea of the load capacity, noise figure and linearity that can be achieved in an amplifier of roughly the required bandwidth.

Problem 1

Let the system under consideration be required to have a channel capacity of at least 3000 and be required to meet a 40 dBrnc0 noise objective for 4000 miles. The coaxial cable has a loss of 14 dB per mile in the vicinity of the proposed top channel. Let the estimates of repeater performance be:

$$P_R = 23 \text{ dBm}$$
$$N_F = 6 \text{ dB}$$
$$M_2 = -80 \text{ dB}$$
$$M_3 = -100 \text{ dB}$$

In summary, the other factors of interest are:

$$W_{TS} = 40 \text{ dBrnc0}$$
$$N = 3000 \text{ channels}$$
$$L_S = 56{,}000 \text{ dB (4000 miles at 14 dB/mile)}$$

Let

$$V_0 = -15 \text{ vu}$$
$$\sigma = 6$$
$$\tau = 0.25$$
$$A_N = A_P = A_2 = A_3 = 3 \text{ dB}$$

Among the questions to be answered are:

1. What repeater spacing, if any, will permit the system noise objective be met?
2. Will the system be overload- or modulation-limited and, if modulation-limited, by second order or by third order distortion?
3. At what transmission level (C) should the system operate to provide performance?

Solution

First assume that thermal noise and overload are controlling, and calculate by equation **S1** the permissible repeater spacing (i.e., find n). The calculation will be repeated, using equations **S2** and **S3**. The solution corresponding to the largest n will satisfy the requirements of the others and, in so doing, will indicate which of the noise sources is dominant and thereby limits system performance.

Equation **S1** is repeated here for convenience:

$$\frac{L_S}{n} + 10 \log n \leq (W_{NS}-N_R) - (A_N+A_P) - (P_S-P_R) - 88$$

Calculate P_S:

$$P_S = V_0 + 0.115\sigma^2 - 1.4 + 10 \log N\tau_L + \Delta_C$$

For the specified values, $P_S \approx 27.0$ dBm.

Thus,

$$\frac{56{,}000}{n} + 10 \log n \leq [40-(-139+6)] - (3+3) - (27-23) - 88$$
$$\leq 75$$

A quick check on the existence of a solution can be made by calculating the minimum value of the left side. Differentiating the left side with respect to n and equating to zero yields the minimum at

$$n = \frac{L_S}{4.34}$$

For such n,

$$\frac{56{,}000}{n} + 10 \log n = 45.4$$

and since this is less than 75, it is clear that a solution exists. Proceeding either graphically or by cut-and-try will show that the n for which the inequality is just satisfied is 1275. (A larger n will provide increased margin.) Therefore overload limitations are satisfied if

$$n = 1275$$

Equation S2 is repeated here for convenience:

$$\frac{L_S}{n} + 20 \log n \leq (W_{NS}+W_{2S}) - (A_N+A_2) - (N_R+M_2) - (88+K_2)$$

Allocate W_{TS} to W_{NS} and W_{2S}:

$$W_{NS} = W_{TS} - 3 = 37 \text{ dBrnc0}$$

$$W_{2S} = W_{TS} - 3 = 37 \text{ dBrnc0}$$

Calculate K_2:

Since the top channel is being studied, the only second order product of present interest is the $\alpha + \beta$ type.

Therefore,

$$K_2 = K_{\alpha+\beta} = \eta_{\alpha+\beta} V_0 + 0.115\lambda_{\alpha+\beta}\sigma^2 + 10 \log \left(U_{\alpha+\beta}\, \tau^{\mu}{}_{\alpha+\beta}\right)$$

$$+ 88 - (1.4\eta_{\alpha+\beta} + C_{w_{\alpha+\beta}}) + 6$$

Substituting values,

$$K_2 = 87$$

Thus,

$$\frac{56{,}000}{n} + 20 \log n \leq (37+37) - (3+3) - (-139+6-80) - (88+87)$$

$$\leq 106$$

As before, a quick check on the existence of a solution results from determining that the left side is minimized for

$$n = \frac{L_S}{8.68}$$

For such n,

$$\frac{56{,}000}{n} + 20 \log n = 85$$

and a solution does exist. Proceeding with the solution shows that the n for which the inequality is just satisfied is 1280. Therefore, second order modulation limitations are satisfied if:

$$n = 1280$$

Next, reexamine equation **S3:**

$$\frac{L_S}{n} + 20 \log n \leq \left(W_{NS} + \frac{W_{3S}}{2}\right) - \left(N_R + \frac{M_3}{2}\right) - \left(A_N + \frac{A_3}{2}\right) - \left(88 + \frac{K_3}{2}\right)$$

Allocate W_{TS} to W_{NS} and W_{3S}:

$$W_{NS} = W_{TS} - 1.8 = 38.2 \text{ dBrnc0}$$

$$W_{3S} = W_{TS} - 4.8 = 35.2 \text{ dBrnc0}$$

Calculate K_3:

Because of the large number of channels, it is likely that the third-order distortion in the top channels will be predominantly of the $\alpha + \beta - \gamma$ type. Proceeding with that assumption,

$$K_3 = K_{\alpha+\beta-\gamma} = \eta_{\alpha+\beta-\gamma} V_0 + 0.115 \lambda_{\alpha+\beta-\gamma} \sigma^2 + 10 \log \left(U_{\alpha+\beta-\gamma} \tau^{\mu_{\alpha+\beta-\gamma}}\right) + 88 - \left(1.4\eta + C_w\right)_{\alpha+\beta-\gamma} + 15.6$$

Substituting values,

$$K_3 = 109.6$$

Thus,

$$\frac{L_S}{n} + 20 \log n \leq \left(38.2 + \frac{35.2}{2}\right) - \left(-139 + 6 - \frac{100}{2}\right) - \left(3 + \frac{3}{2}\right) - \left(88 + \frac{109.6}{2}\right)$$

$$\leq 91.5$$

Proceeding with a solution, the value of n for which the inequality is just satisfied is 2310. Therefore, third order modulation limitations are satisfied if:

$$n = 2310$$

A comparison of the solutions to **S1**, **S2**, and **S3** shows that third order modulation imposes the limiting requirement on repeater spacing, and thus the system is thermal and third order limited, although it is not obvious at this point that the second order noise is entirely negligible.

For $n = 2310$, the repeater spacing is about 1.7 miles, and the corresponding top channel repeater gain is 24 dB. The repeater output transmission level at which the system should operate at the top channel frequency can be calculated from Eqs. (13-11) and (13-1). Repeating Eq. (13-11),

$$M_3 - 2C + 20 \log n + K_3 + A_3 \leq W_{3S}$$

Solving for C,

$$C \geq (M_3 + 20 \log n + K_3 + A_3 - W_{3S}) \left(\frac{1}{2}\right)$$

From Eq. (13-1):

$$C \leq W_{NS} - (N_R + G_R + 10 \log n + 88 + A_N)$$

Substituting values leads to

$$22.4 > C > 22.3$$

Given this transmission level ($C = 22.3$), the $\alpha + \beta$ modulation noise can be calculated from the left side of Eq. (13-10):

$$M_2 - C + 10 \log n + K_2 + A_2 = -80 - 22.3 + 33.6 + 87 + 3$$
$$= 21.3 \text{ dBrnc0}$$

Thus, the second order distortion noise is more than 13 dB less than the allocation to third order noise, and setting system transmission levels based on the third order modulation noise alone will not introduce appreciable departures from optimum S/N performance in the top channel.

The use of Eq. (13-2) will show how close to the repeater-overload point the system would be operating. Repeating Eq. (13-2):

$$P_R + C - A_P \geq P_S$$

Substituting, where A'_P is the actual margin,

$$23 + 22.3 - A'_P \geq 27$$

Thus, $A'_P = 18.3$ dB, the *actual* operating margin against repeater overload where the transmission levels have been set by third order modulation considerations.

Problem 2

Another problem based on the same data would be to determine the combination of repeater parameters which would permit a two-mile repeater spacing to be used. Assuming that the repeater load capacity would still be about +23 dBm, it seems reasonable to begin with the working assumption that the top channel performance will be limited by thermal noise and third order modulation distortion of the $(\alpha + \beta - \gamma)$ type. Second order modulation noise can then be set at a value which insures the validity of this stipulation.

Solution

Equation **S3** is used in rearranged form as follows:

$$\frac{L_S}{n} + 20 \log n \leq \left(W_{NS} + \frac{W_{3S}}{2}\right) + 139 - \left(N_F + \frac{M_3}{2}\right) - \left(A_N + \frac{A_3}{2}\right) - \left(88 + \frac{K_3}{2}\right)$$

Of interest is the term $(N_F + M_3/2)$.

$$\left(N_F + \frac{M_3}{2}\right) \leq \left(W_{NS} + \frac{W_{3S}}{2}\right) + 139 - \left(A_N + \frac{A_3}{2}\right) - \left(88 + \frac{K_3}{2}\right) - \left(\frac{L_S}{n} + 20 \log n\right)$$

For $n = 2000$ (two-mile spacing),

$$\left(N_F + \frac{M_3}{2}\right) \leq -47.3$$

Rearranging,

$$M_3 \leq 2\ (-47.3 - N_F)$$

This expression points out the relationship between noise figure and third order modulation index for equivalent noise performance. Under the constraints imposed, each dB of improvement in repeater

noise figure will reduce the requirement on the repeater M_3 by 2 dB. For the 6-dB noise figure used in problem 1 of the example, it can be seen that an increase in repeater spacing from 1.7 to 2.0 miles would increase the requirement on M_3 from −100 to −106.6 dB. To check the impact on the required M_2, a slightly rearranged form of equation S2 is useful:

$$(N_F + M_2) \leq W_{NS} + W_{2S}) + 139 - (A_N + A_2) - (88 + K_2) - \left(\frac{L_S}{n} + 20 \log n\right)$$

This leads to

$$(N_F + M_2) \leq -73.8$$

It can be seen that there exists a one-for-one relationship between N_F and M_2. Finally, from equation S1

$$(N_F - P_R) \leq W_{NS} + 139 - (A_N + A_P) - P_S - 88 - \left(\frac{L_S}{n} 10 \log n\right)$$

which leads to

$$P_R - N_F \geq 3$$

Again a one-for-one relationship is evident between the repeater noise figure and the required repeater load capacity. Each improvement (reduction) in noise figure of a dB will permit the signal levels to be dropped a dB lower while still resulting in equivalent zero level thermal noise. Consequently, the repeater output load would correspondingly diminish by a dB.

Problem 3

It is also interesting to investigate the upper limit on achievable channel capacity for a given set of repeater performance parameters and system requirements. As discussed in Chap. 16, a given level of repeater performance becomes more difficult to achieve as top frequency is increased. For example, the modulation distortion will be greater for a given state of the device and circuit art; the repeater linearity also may interact with any change in insertion gain. For the purposes of this example, however, these effects are ignored.

Solution

Consider again equation **S3**:

$$\frac{L_S}{n} + 20 \log n \leq \left(W_{NS} + \frac{W_{3S}}{2}\right) - \left(N_R + \frac{M_3}{2}\right) - \left(A_N + \frac{A_3}{2}\right) - 88 - \frac{K_3}{2}$$

For a particular set of noise requirements, repeater parameters, etc., all terms of the right side are independent of the number of channels except the last term, $-K_3/2$. Reviewing the definition of K_3 reveals that it is a monotonically increasing function of N, the rate depending only on the type of third order product involved. Accordingly, it is useful to modify **S3** as follows:

$$\frac{L_S}{n} + 20 \log n \leq -\frac{K_3}{2} + \text{Constant}$$

or

$$K_3 \leq 2\left[\text{Constant} - \left(\frac{L_S}{n} + 20 \log n\right)\right] \qquad (13\text{-}12)$$

By maximizing the right side of Eq. (13-12), K_3 is maximized, and therefore N, the channel capacity, is also maximized. This, in turn, is done by minimizing

$$\frac{L_S}{n} + 20 \log n$$

While checking the existence of solutions to **S2** and **S3** in a preceding problem, it was found that the left side of **S2** or **S3**, $L_S/n + 20 \log n$, is minimized for $L_S/n = 8.68$ dB. This corresponds to the "1 Neper" solution to a modulation-limited system problem.

Since L_S is the total cable loss at the frequency of the top channel, it is a function of the system channel capacity, N. To express L_S as a function of N, let L_0 be the 1-MHz loss of some length of coaxial cable. Then the loss in dB at other frequencies is given by $L_0 \sqrt{f}$, where f is in MHz. When $f = f_T$, the frequency of the top channel, and L_0 is the 1-MHz loss for the total system length for which the design applies, then $L(f) \triangleq L_S = L_0 \sqrt{f_T}$.

For 4 kHz per channel,

$$f_T = 4 \times 10^3 \times N \quad \text{Hz}$$
$$= 4\,N' \text{ MHz}$$

where N' is the channel capacity in thousands of channels.

Thus,

$$L_S = 2\,L_0 \sqrt{N'} \tag{13-13}$$

If

$$\frac{L_S}{n} = 8.68 \text{ dB}$$

the number of repeaters is given by [using Eq. (13-13)]

$$n = \frac{L_S}{8.68} = \frac{2\,L_0 \sqrt{N'}}{8.68}$$

Therefore, in terms of the channel capacity,

$$\text{MIN}\left(\frac{L_S}{n} + 20 \log n\right) = 8.68 + 20 \log \frac{2\,L_0}{8.68} + 10 \log N' \tag{13-14}$$

With the resulting Eq. (13-14) and the original repeater and system specifications of the preceding example, it is possible to determine the maximum channel capacity of the system which could be realized with these constraints. It will be recalled that the original problem specified a particular channel capacity (3000) as a requirement, which, together with other requirements and parameters, resulted in a repeater spacing of about 1.7 miles. Although not explicitly stated in the example, the assumed cable loss was that correpsonding to a 3/8 inch coaxial cable. For such a cable, the loss at 1 MHz is approximately 4 dB per mile. Thus $L_0 = 4 \times 4000$ dB, and from Eq. (13-14),

$$\text{MIN}\left(\frac{L_S}{n} + 20 \log n\right) = 8.68 + 20 \log \frac{32{,}000}{8.68} + 10 \log N'$$
$$= 80.4 + 10 \log N'$$

Substituting MIN $(L_S/N + 20 \log n)$ along with the other specified constants of the example into Eq. (13-12) yields

$$K_3 \leq 131.8 - 20 \log N' \tag{13-15}$$

Assuming that $K_3 = K_{\alpha+\beta-\gamma}$ for this calculation and noting that for the top channel of the system $U_{\alpha+\beta-\gamma} \approx N^2/4$ (from Fig. 10-6), K_3 is also given by

$$K_3 = 40.1 + 20 \log (N' \times 10^3)$$
$$= 100.1 + 20 \log N' \qquad (13\text{-}16)$$

Combining Eqs. (13-15) and (13-16) gives

$$40 \log N' \leq 31.8$$

For the equality,

$$N' = \log^{-1}\left(\frac{31.8}{40}\right) = 6.24 \text{ (thousand channels)}$$

Thus,

$$f_T = 4\,N' \approx 25 \text{ MHz}$$

To determine the repeater spacing implied by this solution, where the cable loss is given to be $4\sqrt{f_T}$ dB/mile at f_T (in MHz), and k is the repeater spacing in miles, then

$$(4\sqrt{f_T})\ (k) = 8.68$$

and

$$k = \frac{8.68}{4\sqrt{f_T}} = 0.434 \text{ miles}$$

Consequently, if the repeater parameters of the 3000-channel 1.7-mile spacing example could be achieved at 25 MHz, a system of about 6240 channels with 0.434-mile spacing would be possible.

A maximum capacity solution such as the one just obtained is not generally the best solution. Unless the traffic the system is to carry can justify the channel capacity, it is uneconomic to pay the cost for the increased number of repeaters required as compared to some smaller capacity system. Furthermore, minimum cost per channel is usually achieved at somewhat lower than maximum bandwidths. From a reliability point of view it is also better to have several smaller capacity systems rather than one large capacity system since the number of channels lost during a failure is less.

The previous examples should have made clear that the system equations by themselves do not result in determining a unique best system. Many factors, quantitative and qualitative, which are not

included in their derivations must be considered. In general, a system design is a circular process. Some estimates of achievable repeater performance are made on the basis of past experience and any new art that has become available. The equations are then made to generate repeater spacing as a function of channel capacity and any other parameters that are not constrained. Any performance parameters that are highly uncertain might also be varied to study their impact on the system. Cost and traffic growth estimates are combined with the repeater spacing results to permit economic comparisons. These will reflect in a quantitative way that a smaller capacity system, even though its per channel cost might be higher, could be more attractive if traffic growth rates are low. This is true because the smaller system will more closely provide channel capacity as it is needed and avoid tying up capital on channel capacity that will not be needed for a long time. Finally, reliability and compatibility with existing plant must be considered. All of these considerations will permit defining a tentative bandwidth, repeater spacing, and insertion gain required of the repeater. Different repeater circuits and devices can be initially evaluated to determine the best way of obtaining the required gain and bandwidth. One of the results of such an exploratory program is the more precise determination of achievable repeater performance parameters which can then be used in the system equations for a more precise system layout. As the development proceeds, the system plan must be refined to take into account, more and more precisely, the knowledge gained during the development process. Finally, computed and achieved system performance are compared to pinpoint any oversights and provide insight for future developments.

13.8 SIGNAL SHAPING

In the preceding development of techniques for the design and analysis of analog cable systems, which assumed a flat level point at the repeater output, it was found that the highest frequency channel had the most thermal noise, primarily due to the higher cable loss at high frequencies. If instead, signal amplitudes in the noisy channels are increased while those in the quiet channels are reduced, the net effect is to more nearly equalize the signal-to-noise ratio in the various channels. This can be accomplished without changing the total signal power and yet improve the worst channel noise performance by several dB.

The fact that the top channel is the noisiest is evident from the previously derived noise equation

$$W_{n0} = N_R + G_R + 10 \log n + C + 88 \qquad \text{dBrnc0} \qquad (13\text{-}17)$$

The repeater gain, G_R, must compensate for cable attenuation which is known to vary approximately as the square root of frequency. This term will therefore tend to make thermal noise in top channels much greater than the channels near the bottom of the band. In the examples up to now, systems have been laid out so that the noisiest channel just meets a particular noise requirement. This suggests that some advantage could be obtained (e.g., an increase in the number of channels allowed) if the "better than necessary" performance at lower frequencies could be traded for less noise at the top of the band. Qualitatively, an increase of loss at low frequencies between zero level and the point in the system where overload occurs would allow a decrease in this loss at the higher frequencies without altering the total multichannel load. Analytically, this corresponds to making C a function of frequency with a slope opposite to that imposed by the other terms of Eq. (13-17) so than W_{n0} becomes approximately constant with frequency. Physically, the proposed modification requires the installation of a pre-emphasis network in the transmitting terminal and a complementary restoring network in the receiving terminal. The incorporation of such a feature is called signal shaping.

An outline of the computations necessary to deal with signal shaping is best developed in connection with an example. Figure 13-10(a) shows the resulting noise in an experimental system designed for a specified top channel noise performance.

By inspection of Eq. (13-17), it can be seen that the thermal noise in all channels can be made equal to that in the noisiest channel if

$$C'(f) = C_0 + W_{n0} - W_n(f) \qquad (13\text{-}18)$$

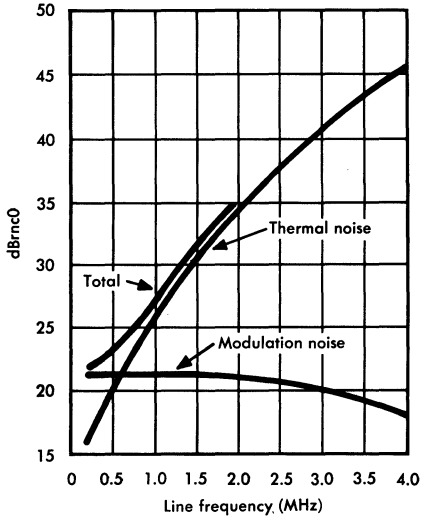
where

$C'(f)$ = new value of C at a frequency, f

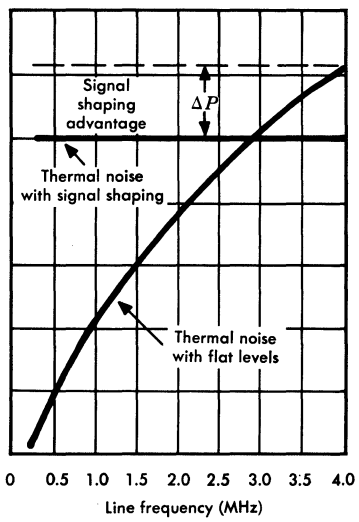
C_0 = constant C as computed for no signal shaping

W_{n0} = noise in the noisiest channel for no signal shaping

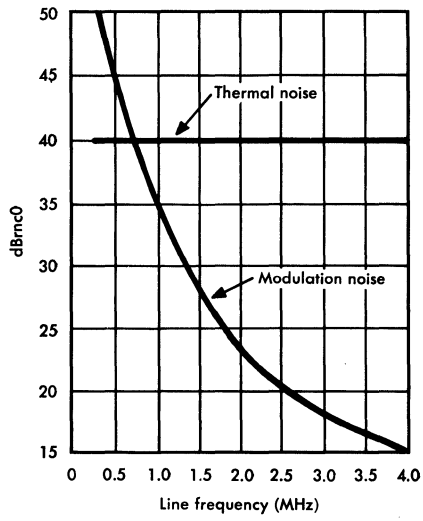
$W_n(f)$ = noise in channel centered at frequency f, with no signal shaping.



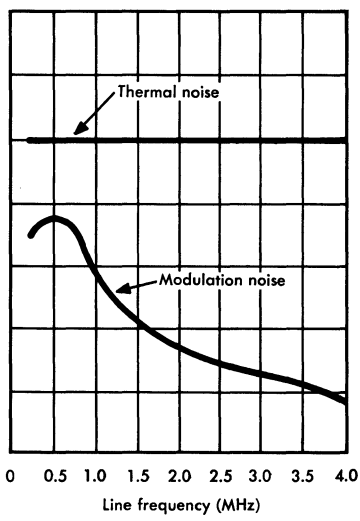
(a) Noise with flat signal levels



(b) Effect of signal level shaping on thermal noise



(c) Noise with signal level shaping



(d) Noise with shaped signal level and feedback

FIG. 13-10. Voice-frequency channel noise versus line frequency.

As the dotted line of Fig. 13-10(b) shows, the thermal noise in all channels has been increased to W_{n0}. However, all channels are now operating with the same or a larger C than initially, and thus the multichannel load at the repeaters is thereby reduced. The reduced load is the basis for the signal shaping advantage.

When the methods of Chap. 9 are used, the equivalent load for shaped levels can be calculated, and the result is a new multichannel load at the repeater output, which is smaller than the old load by an amount ΔP. It is possible, therefore, to increase the signal level by ΔP without overloading the system. As shown in Fig. 13-10(b), this reduces the noise in all channels by the same amount. The net effect compared to the original system has been to reduce the thermal noise in the noisiest channel by this difference. Hence, this reduction is known as the signal shaping advantage and is typically 3 or 4 dB in existing systems.

Modulation noise must be evaluated before final values for $C(f)$ can be established. Even for the case where $C(f) = C_0$, the W_x's resulting from second and third order products have some frequency characteristics, such as that shown in Fig. 13-10(a). The next step must be to compute the modulation noise after a constant C_0 is changed to a new $C(f)$. This requires computation of the power sum of all the individual products falling into a particular channel. The procedure for this computation is described in Chap. 10.

Figure 13-10(c) shows modulation and thermal noise corresponding to the new $C(f)$. As might have been expected, the high-level products from the high-frequency channels falling on the low-level low-frequency channels have introduced excessive noise in the latter. One method for dealing with this problem is to modify the signal shape somewhat to make total noise (rather than thermal noise) flat with frequency. The desired result can be obtained by judicious trial and error modifications of C.

Improvement of the intermodulation performance can be achieved by shaping the amplifier feedback. It can be shown that for most practical cases, the amplitude of any intermodulation product is reduced approximately by the amount of feedback at the product frequency. Thus, the high intermodulation noise at the low end of the band can be suppressed by applying additional feedback at these frequencies. A 6 dB per octave feedback slope (feedback decreasing with increasing frequency) has been used to obtain the results of Fig. 13-10(d), which shows the reduction of modulation

noise to a value below that of thermal noise. In this example, essentially the same overall performance can be obtained either by the modification of the signal shaping or by the introduction of shaped feedback. In other cases where modulation noise is greater, one or the other or a combination of these methods will result in optimized performance.

Chapter 14

Misalignment Penalties in Analog Cable Systems

In Chap. 13 the role of thermal noise and repeater nonlinearities in analog cable systems design was considered for the case of the unmisaligned system. The repeaters were ideally placed and identical in characteristics. The transmission response included no deviations from nominal, and signal levels were everywhere ideal. In reality it is not possible to achieve such conditions, and this chapter will consider the effects (on the noise performance of the system) of deviations from ideal in the transmission response. It is shown that, in general, penalties are incurred which must be included in the margins allowed in the initial design. Transmission deviations can be compensated in analog cable systems by equalization, i.e., adjustments which will provide a satisfactory overall system transmission response even though the sections making up the system may be individually misaligned. The general problem of equalization is discussed in Chap. 15. In this chapter, it is the effect of the misalignment on system S/N performance that is discussed.

Each repeater section is composed of a repeater and its associated length of cable, and any difference between average repeater gain and average cable loss will result in a net gain or loss characteristic which will be identical in each repeater section. Manufacturing variations can result in deviations that are different in each repeater section, but they usually are very small. Changes in the transmission characteristic of repeaters and cable can contribute to system misalignment as the system ages. Such aging effects will usually be similar in all repeaters and all cable sections. Temperature changes in the surrounding environment are often the dominant source of misalignment because of their effect on cable attenuation and, to a

lesser extent, on repeater gain. Temperature changes will tend to be uniform over large geographic areas so that the misalignment due to this mechanism can be expected to be the same over a whole system or at least over sizeable fractions of a system. These considerations make it reasonable that the S/N penalties incurred due to system misalignment can usually be calculated with acceptable accuracy by assuming a uniform accumulation of misalignment along the repeatered line. For each repeater section, then, the voltage gain is defined such that

$$E_{\text{OUT}} = \delta E_{\text{IN}}$$

where, in general, δ will be a function of frequency. For the portion of system to be evaluated, which includes n repeater sections, the total voltage gain, m_S, is thus

$$m_S = \delta^n$$

In the perfectly aligned systems considered previously, δ was equal to one. When δ is greater than one, a net gain per section (positive misalignment) is implied, and when δ is less than one, a net loss per section (negative misalignment) is implied. The total misalignment of the n repeater section is defined to be M_S (dB), where

$$M_S = 20 \log \delta^n$$

The system equations previously derived show the dependence of system load-carrying capacity, thermal noise, and modulation noise on repeater transmission level. All repeater levels were defined to be at the same transmission level, $-C$; however, when misalignment is present, this is no longer possible and the transmission level of each repeater is a function of its position in the system. In particular, it may be initially assumed that the level of the nth repeater is $-C + 20 \log \delta^n$. The difference in the thermal noise at zero level of this system as compared to that of the ideal system is the first topic discussed.

14.1 THE PENALTY FUNCTION

The effect of a uniformly distributed misalignment on the thermal noise accumulation in a system will be treated with the aid of Fig. 14-1. After having passed through n repeaters, the signal has experienced a net gain of δ^n. This means that the last repeater is $20 \log \delta^n$ dB closer to zero level than it would be in an ideal system.

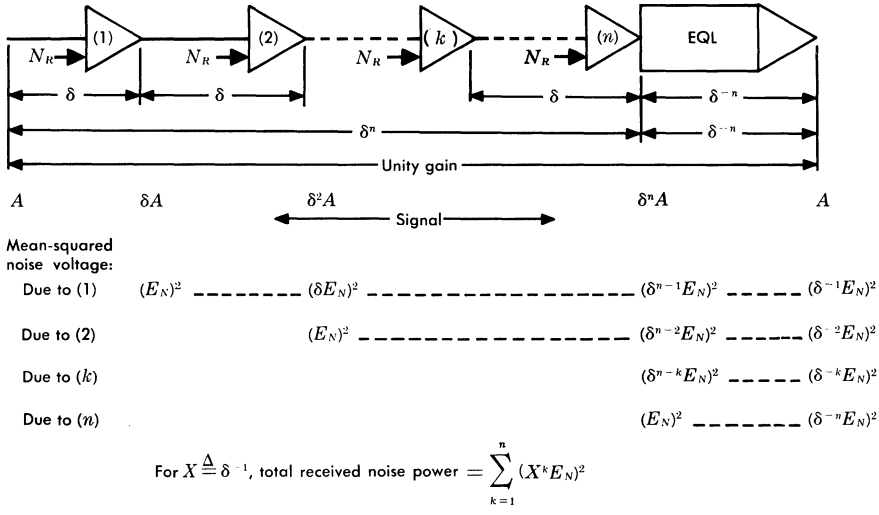


FIG. 14-1. Effect of misalignment on thermal noise.

The noise at zero level is therefore obtained by adding the number $C - 20 \log \delta^n$, rather than C, to the noise at the last repeater. This is equivalent to the insertion of an equalizer with a gain of $20 \log \delta^{-n}$ after the nth repeater, whether or not such an equalizer is physically present. Where the equalization is done at the receiver only and the noise contribution of the equalizer itself is assumed negligible, the received thermal noise power at the output of the equalizer is

$$p_N = \sum_{k=1}^{n} (X^k E_N)^2$$

$$= E_N^2 \sum_{k=1}^{n} X^{2k} \tag{14-1}$$

where X is defined as δ^{-1}.

For the aligned or ideal case, $X=1$ and

$$p_N = nE_N^2 \tag{14-2}$$

For convenience, Eq. (14-1) can be rearranged to closed form by noting that

$$\frac{1}{1-X^2} = 1 + X^2 + X^4 + \cdots + X^{2(n-1)} + X^{2n} + \cdots$$

Thus

$$\frac{X^2}{1-X^2} = X^2 + X^4 + X^6 + \cdots + X^{2n} + X^{2(n+1)} + \cdots$$

and

$$\frac{X^{2(n+1)}}{1-X^2} = X^{2(n+1)} + X^{2(n+2)} + \cdots$$

Equation (14-1) can therefore be written:

$$p_N = E_N^2 \sum_{k=1}^{n} X^{2k} = E_N^2 \frac{X^2 - X^{2(n+1)}}{1-X^2} \tag{14-3}$$

What is being sought here is the difference between the thermal noise accumulated in the misaligned condition ($X \neq 1$) and that accumulated in the aligned condition. The ratio of Eq. (14-3) defined for $X \neq 1$ to Eq. (14-3) defined for $X=1$ [i.e., Eq. (14-2)] provides this relationship:

$$\frac{(p_N) \text{ misaligned}}{(p_N) \text{ ideal}} = \left(\frac{1}{n}\right)\left(\frac{X^2 - X^{2(n+1)}}{1 - X^2}\right) \tag{14-4}$$

The thermal noise penalty due to misalignment is

$$A'_N \triangleq 10 \log \left[\frac{1}{n}\left(\frac{X^2 - X^{2(n+1)}}{1 - X^2}\right)\right] \quad \text{dB} \tag{14-5}$$

In this equation, A'_N is a function of the voltage gain of the single repeater section (X^{-1}) as well as the number of repeaters (n). It is possible to derive an approximation for Eq. (14-5) which is a function only of M_S, the total misalignment. The result is then in a more readily used form, and also brings out more clearly the quantitative relationship between cause and effect. Furthermore, it is usually quite easy to measure the transmission response of

the whole system or a section of the system, corresponding to the quantity M_S, and the per-section misalignment can be deduced from

$$\frac{M_S}{n} = 20 \log \delta \qquad \text{dB/repeater section} \tag{14-6}$$

Accordingly, the penalty A'_N will be expressed in terms of M_S. Rearranging Eq. (14-6),

$$M_S = 10 \log \delta^{2n}$$

Let Δ equal δ^n. Since X equals δ^{-1}, then

$$X^2 = \delta^{-2} = \Delta^{-2/n} \tag{14-7}$$

Rewriting Eq. (14-5) using the indicated substitution yields:

$$A'_N = 10 \log \left[\frac{1}{n} \left(\frac{\Delta^{-2/n} - \Delta^{-\frac{2}{n}(n+1)}}{1 - \Delta^{-2/n}} \right) \right]$$

$$= 10 \log \left\{ \frac{1}{n} \left[\frac{\Delta^{-2/n}\,(1 - \Delta^{-2})}{1 - \Delta^{-2/n}} \right] \right\} \tag{14-8}$$

Noting that

$$\Delta^{-2/n} = e^{\frac{\ln \Delta^{-2}}{n}}$$

$\Delta^{-2/n}$ may be expanded to the series

$$\Delta^{-2/n} = 1 + \frac{1}{n} \ln \Delta^{-2} + \frac{1}{2} \left(\frac{\ln \Delta^{-2}}{n} \right)^2 + \cdots \tag{14-9}$$

By definition,

$$-M_S = 10 \log \Delta^{-2} = 10 \log e^{\ln \Delta^{-2}}$$

from which

$$\ln \Delta^{-2} = -\frac{M_S}{4.34} \tag{14-10}$$

Substituting Eq. (14-10) into Eq. (14-9) yields

$$\Delta^{-2/n} = 1 + \frac{1}{n} \left(-\frac{M_S}{4.34} \right) + \frac{1}{2} \left(\frac{-M_S/4.34}{n} \right)^2 + \cdots \tag{14-11}$$

The higher order terms of Eq. (14-11) can be dropped if the misalignment per section, M_S/n, is small. If this condition is satisfied, then

$$\left|\frac{M_S}{4.34n}\right| << 1$$

In a practical system this relationship will almost always be true, and the approximation of Eq. (14-11) by the first two terms only is consequently a good one:

$$\Delta^{-2/n} \approx 1 - \frac{M_S}{4.34n} \qquad (14\text{-}12)$$

By use of this approximation and the equality $\Delta^{-2} = \log^{-1}(-M_S/10)$, Eq. (14-8) becomes

$$A'_N = 10 \log \left\{ \frac{1}{n} \frac{\left(1 - \frac{M_S}{4.34n}\right)\left[1 - \log^{-1}\left(-\frac{M_S}{10}\right)\right]}{1 - (1 - M_S/4.34n)} \right\}$$

Neglecting the $M_S/4.34n$ term in the numerator gives

$$A'_N = 10 \log \left[\frac{1 - \log^{-1}\left(-\frac{M_S}{10}\right)}{M_S} \right] + 6.4 \quad \text{dB} \qquad (14\text{-}13)$$

In summary, consider a string of repeater and cable sections with the first repeater at some specified transmission level. The penalty A'_N is the difference in thermal noise at zero transmission level for the case where there is a net gain or loss uniformly distributed along the length of the string, as compared to the case where all repeaters are at the same transmission level as the transmitting repeater. For purposes that will become clear, the value of A'_N determined from Eq. (14-13) is defined as the penalty function A'_0. It is plotted as a function of M_S, the total net gain, in Fig. 14-2. The approximation used to obtain this equation is valid as long as the misalignment per repeater section is much less than 4 dB.

Figure 14-2 shows that for $M_S > 0$ (i.e., net gain), the penalty is negative; in other words, the misaligned system has less thermal

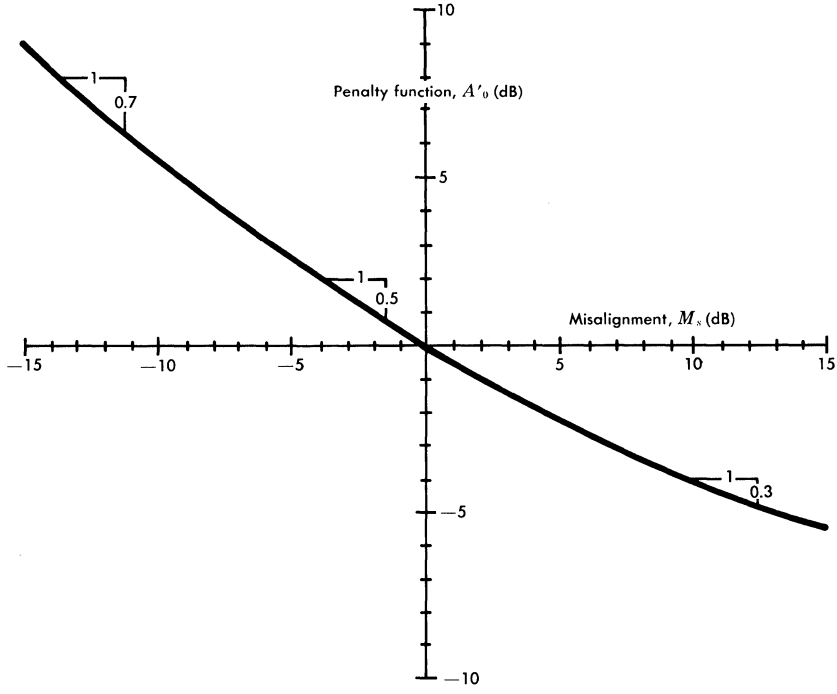
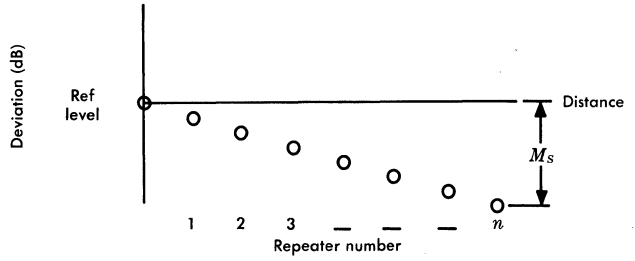


FIG. 14-2. Penalty function.

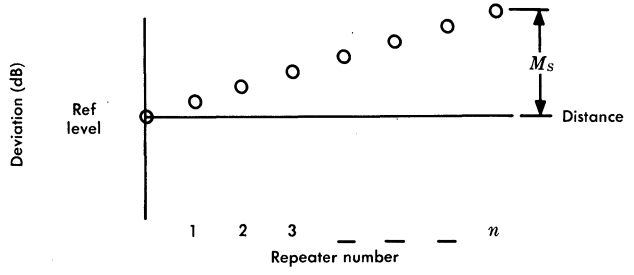
noise than the unmisaligned system. This should not come as a surprise since the transmission level of all repeaters except the first is higher in the misaligned case than in the unmisaligned case. From Chap. 13, it should be clear that the contribution of a repeater to thermal noise at zero level decreases dB for dB as the transmission level of that repeater is raised. Since misalignment is considered a source of degradation which must be held under tight control at considerable cost, it may appear startling to see the misaligned system with a lower thermal noise than that of the ideal system. The explanation of this apparent paradox is that thermal noise is not the only consideration which must be taken into account. It must be assumed that the transmission level in the ideal reference case was selected to be as high as possible consistent with load-carrying capacity in an overload-limited system, and to be optimum with regard to *total* noise on a modulation-limited system. Under these circumstances, any repeater at a level different from this optimum can be expected to degrade system performance in some respect. In the following discussion this will be seen to be the case.

14.2 PENALTIES IN OVERLOAD-LIMITED SYSTEMS

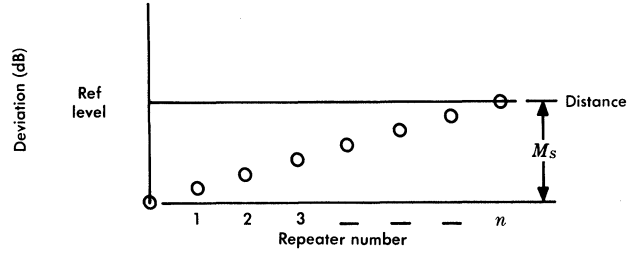
Figure 14-3(a) shows the repeater levels of a string of repeaters and cable sections with some amount of uniform negative misalignment (net loss). Equalization capability is assumed available only at the transmitting and receiving ends. The string of repeaters can thus represent a whole system which has equalizers only in the terminals or that portion of a system between any two successive line equalizers. The transmitting repeater is shown to be at the maximum



(a) Negative misalignment equalized at receiving terminal



(b) Positive misalignment equalized at receiving terminal



(c) Positive misalignment equalized at transmitting terminal

FIG. 14-3. Repeater levels on a misaligned system.

level consistent with the required load capacity as discussed in Chap. 13. Any increase in repeater levels due to a decrease in loss at the transmitting terminal (between zero transmission level and the transmitting repeater) would result in overloading the transmitting repeater. Any increase in the loss between zero level and the first repeater would further degrade thermal noise performance. Therefore the optimum strategy for this case is to leave the transmitting equalization as it would have been had there been no misalignment and to carry out the equalization of the misalignment entirely at the receiving end. This is exactly the case considered in the derivation of the penalty function, A'_0, so that for this case, the actual misalignment penalty A'_N equals A'_0.* As shown in Fig. 14-2, if M_S is negative, then A'_0 is positive, and misalignment results in increased noise.

Figure 14-3(b) shows the levels which would result for positive misalignment if the level of the transmitting repeater were held at the value that would have been selected on the basis of the required load capacity for an ideal system. The noise for such a case would be less than for the ideal case by $| A'_0 |$ dB; however, all repeaters but the transmitting one would be overloaded. In order to handle the required load, the highest level repeater can be no higher than the repeater level for an ideal system. It is thus necessary to achieve the configuration of Fig. 14-3(c) by inserting M_S dB of loss between the zero transmission level point and the transmitting repeater. Since the result of this is to lower the transmission levels of all repeaters by M_S dB, the noise at zero level for (c) is M_S dB greater than in the case for (b). Thus the net difference in thermal noise between the (c) case and the ideal case is $M_S - | A'_0 |$ which will always be a positive number. In fact, an examination of (c) shows that there is a one-for-one equivalence between the cases of (a) and (c). The transmitting repeater in (a) is at the same level as the last repeater in (c); the next repeater in (a) is at the same level as the second from the last repeater in (c) and so on. The noise contribution of a given repeater design depends only on its transmission level and not on its geographical position in the systems. The penalty for (c) must therefore equal the penalty for (a) for a given value of M_S. In other words, $M_S - | A'_0(M_S) | = A'_0(-M_S)$. Therefore, for an

*A'_N is used for misalignment penalty since it is one, but not necessarily the only, contributor to the thermal noise margin, A_N, used in the equations of the previous chapter.

overload-limited system which requires all receiving-end equalization for negative misalignment and transmitting-end equalization for positive misalignment,

$$A'_0\,(-M_S) = M_S + A'_0\,(M_S) \qquad (14\text{-}14)$$

Figure 14-4 shows the negative half of Fig. 14-2 to give A'_N as a function of $|\,M_S\,|$ for the overload-limited case.

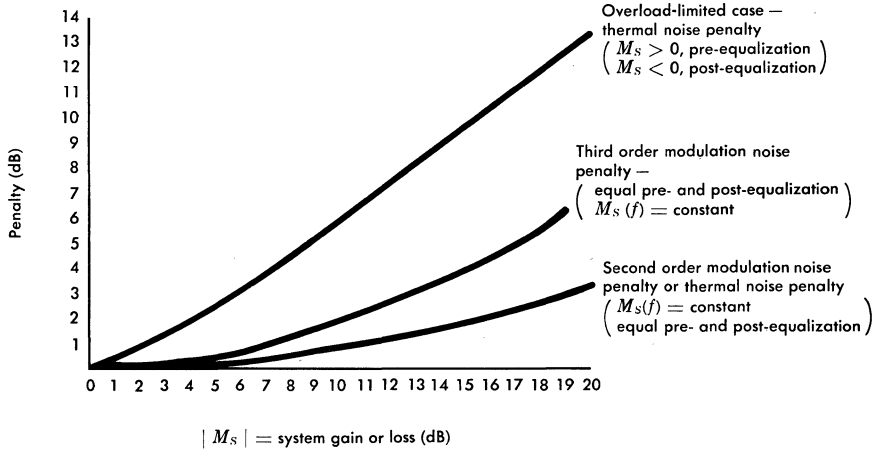
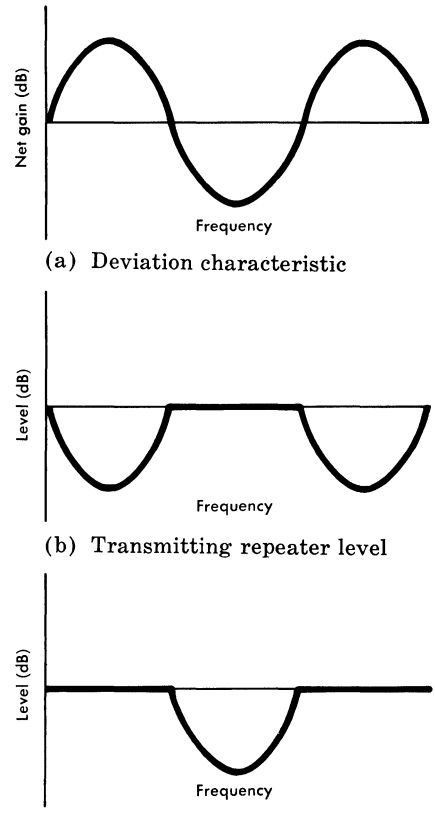


FIG. 14-4. Misalignment penalty.

In one regard, the previous discussion has been oversimplified, since M_S will usually be a function of frequency. As a consequence, the highest level repeater which determines the load-carrying capacity of the system may no longer be a flat level point. For example, if these rules for pre- and post-equalization were applied at each frequency, the system gain characteristic of Fig. 14-5(a) would result in the transmission levels of Fig. 14-5(b) and (c) at the end repeaters. It is evident that the resulting multichannel load at the end repeaters is less than for the reference case. It would thus be possible to increase all repeater levels by some flat amount (determined by methods discussed in Chap. 9) and consequently to reduce the penalty given in Fig. 14-4 by that same amount. Practically, this is not as serious a complication as it may seem, since the largest source of misalignment is usually due to the effect of temperature on the cable. Though the resulting misalignment does

have a frequency characteristic, it will be all positive or all negative across the transmission band. Under these circumstances, the use of pre-equalization for positive misalignment and post-equalization for negative misalignment will result in the highest level repeater being at a flat level point (assuming that repeaters were at a flat level in the unmisaligned reference case). Using this equalization strategy will therefore maintain the system load-carrying capacity constant, and the thermal noise penalty at each frequency can be obtained from Fig. 14-4 by using the value of $|M_S|$ at that frequency.



(a) Deviation characteristic

(b) Transmitting repeater level

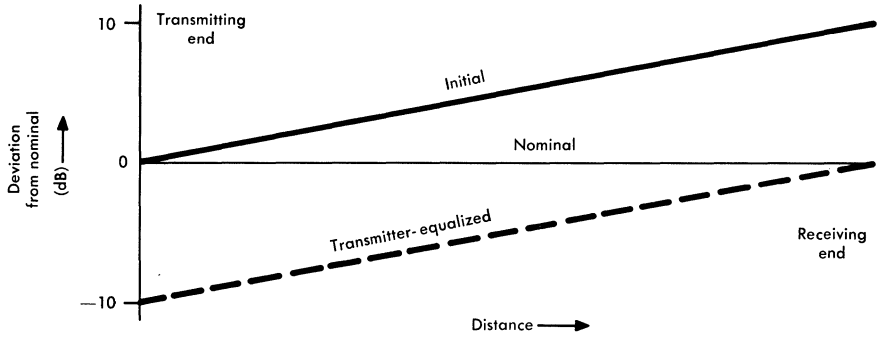
(c) Last repeater level

FIG. 14-5. Frequency characteristic of deviations in an overload-limited system.

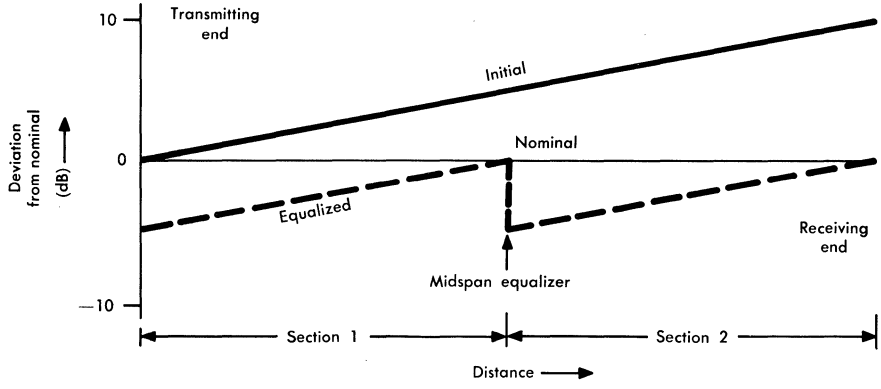
Calculation of Penalties

Consider the transmission level deviation from nominal, shown in Fig. 14-6(a) as a function of distance along the repeatered line. If the system is overload limited under ideal conditions with little or no margin, equalization must be done at the transmitter since the misalignment is positive. It is therefore necessary to pre-equalize at the transmitter by adjusting the equalizer there to a setting of −10 dB relative to nominal. The penalty incurred in thermal noise as a result, $|M_S| = 10$ dB, can be seen from Fig. 14-4 to be 6 dB. The thermal noise at zero level due to this portion of the system is consequently 6 dB greater than is the case with no misalignment.

Still assuming an overload-limited system, it is interesting to observe the effect on the thermal noise when an equalizer is added midway between the transmitter and receiver. This case is shown in Fig. 14-6(b) where it is still necessary to pre-equalize the misalignment in order to avoid overloading the line repeaters. For



(a) Equalized at transmitter



(b) Equalized at transmitter and midspan

FIG. 14-6. Example of signal level deviation and equalization.

this situation, Fig. 14-4 is entered at $|M_S| = 5$ dB, for which $A'_N = 2.7$ dB. This is the penalty associated with either half of the system between the transmitter and receiver. The total penalty, A'_{NT}, is the weighted average of the two separate penalties, i.e.,

$$A'_{NT} = [(A'_N)_1 - 3] \text{ "+" } [(A'_N)_2 - 3]$$

For this example, it is apparent that $A'_{NT} = (A'_N)_1 = (A'_N)_2 =$

2.7 dB. By halving the equalizing interval, for the same overall transmitter-to-receiver misalignment, the thermal noise penalty is reduced by somewhat more than 3 dB. Consequently, one of the choices or tradeoffs the system designer must make is between noise performance (penalties) and equalizing interval. Because the equalizing stations will usually be fairly complex, it is inherently undesirable to place them any more frequently than is absolutely necessary. On the other hand, the longer the equalizing interval, the greater the noise penalties. The approach to this question involves striking as nearly optimum a balance as possible between conflicting considerations.

Before proceeding to the effect of misalignment on modulation noise, one more example involving only thermal noise will be considered. Figure 14-7 shows four sections, with equalization assumed possible at the junction between sections. The dotted line of Fig. 14-7 shows the transmission level deviations from nominal after equalization, again assuming an overload-limited system. It can be seen that the negative misalignment of section 1 is thus post-equalized, whereas pre-equalization of the positive misalignment of section 3 is necessary, and so on.

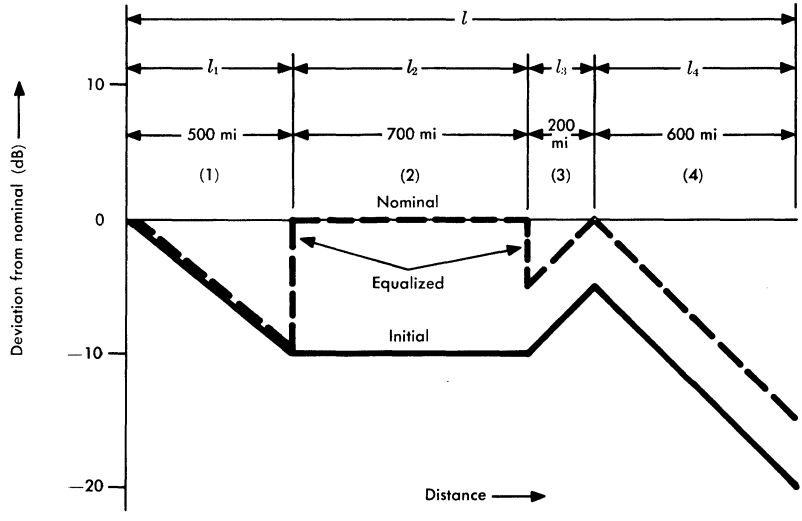


FIG. 14-7. Equalization in four blocks of an overload-limited system.

With the use of the methods employed previously, the penalties incurred in each section can be determined from Fig. 14-4; they are listed below:

Section (k)	M_k(dB)	Equalized at	$(A'_N)_k$
1	−10	Receiver	6
2	0	—	0
3	+ 5	Transmitter	2.7
4	−15	Receiver	9.7

As before, the total penalty incurred, A'_{NT}, is the weighted sum of the penalties incurred in the constitutent sections, where the weighting reflects the proportion of the total system represented in each of the sections. Defining the weighted penalty for the kth section to be $(A'_{N'})_k$,

$$(A'_{N'})_k = (A'_N)_k + 10 \log \frac{l_k}{l}$$

and

$$A'_{NT} = \text{"}\sum_k\text{"} (A'_{N'})_k \quad \text{(where "}\Sigma\text{" indicates power sum)}$$

For the example at hand, the total thermal noise penalty, A'_{NT}, becomes

$$A'_{NT} = \left(6 + 10 \log \frac{500}{2000}\right) \text{"+"} \left(0 + 10 \log \frac{700}{2000}\right)$$
$$\text{"+"} \left(2.7 + 10 \log \frac{200}{2000}\right) \text{"+"} \left(9.7 + 10 \log \frac{600}{2000}\right)$$
$$= 6.6 \text{ dB}$$

14.3 PENALTIES IN INTERMODULATION-LIMITED SYSTEMS

The effect of system misalignment on modulation noise is derived as shown in Fig. 14-8 where the approach is entirely analogous to that used previously for calculating the thermal noise penalties. As before, the results will be derived for the receiver-equalized condition from which the penalties for other equalizer placements can

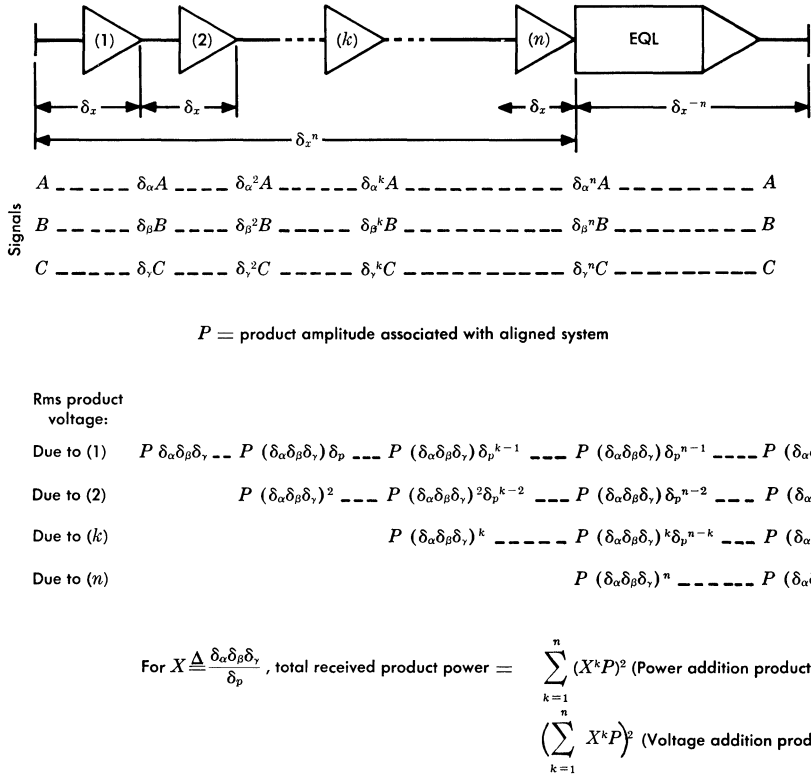


FIG. 14-8. Effect of misalignment on modulation noise.

be readily derived. There are some minor differences in the treatment of Fig. 14-8 from that of Fig. 14-1 (corresponding figure for thermal noise penalty calculation) which should be noted. The input signal is defined at three frequencies as follows: $A \cos \alpha t$, $B \cos \beta t$, and $C \cos \gamma t$. The misalignment per repeater section is assumed uniform, but may be frequency dependent. Thus δ_α, δ_β, δ_γ, and δ_p are the voltage gains of the repeater sections at the frequencies α, β, γ, and at the product frequency, respectively. Just as an identical noise figure was assumed for the repeaters in Fig. 14-1, identical M_2 and M_3 are assumed for the repeaters of Fig. 14-8. From Fig. 14-8, where

$$X \triangleq \frac{\delta_\alpha\delta_\beta\delta_\gamma}{\delta_p}$$

the penalties on modulation noise can be determined and depend on

whether the product is the type that adds randomly or in phase. For modulation products adding randomly, the penalty is

$$A'_M = 10 \log \left[\frac{1}{n} \left(\frac{X^2 - X^{2(n+1)}}{1 - X^2} \right) \right] \tag{14-15}$$

For modulation products adding in phase the penalty is

$$A'_M = 20 \log \left[\frac{1}{n} \left(\frac{X - X^{n+1}}{1 - X} \right) \right] \tag{14-16}$$

Since the equation for power addition products is identical in form to the one derived for thermal noise, the approximate form of Eq. (14-13) and the plot of Fig. 14-2 are directly applicable. It is only necessary to remember that the independent variable is now $M_M = 20 \log (\delta_p/\delta_\alpha\delta_\beta\delta_\gamma)^n$ rather than M_S = net gain.

For products adding in phase, a new approximation must be derived. In a manner analogous to that for thermal noise, it can be shown that

$$A'_M = 20 \log \left[\frac{1 - \log^{-1}(-M_M/20)}{M_M} \right] + 18.8 \tag{14-17}$$

where

$$\frac{M_M}{8.68n} << 1$$

It is useful to observe that Fig. 14-2 can be used to evaluate Eq. (14-17) if M'_S is defined to equal $M_M/2$, for then

$$\begin{aligned} A'_M &= 20 \log \left[\frac{1 - \log^{-1}(-2M'_S/20)}{2M'_S} \right] + 18.8 \\ &= 20 \log \left[\frac{1 - \log^{-1}(-M'_S/10)}{M'_S} \right] - 6 + 18.8 \\ &= 2 \left\{ 10 \log \left[\frac{1 - \log^{-1}(-M'_S/10)}{M'_S} \right] + 6.4 \right\} \end{aligned}$$

Thus, for in-phase addition

$$A'_M = 2A'_0 (M'_S) \tag{14-18}$$

Equation (14-18) means that Eq. (14-17) can be evaluated by computing $M_M = 20 \log (\delta_p/\delta_\alpha\delta_\beta\delta_\gamma)^n$, dividing the result by two, reading from Fig. 14-2 the penalty corresponding to this value,

and doubling it. The result is applicable to a post-equalized system. Before investigating the effect of different equalization strategies, it must be pointed out that, in the case of modulation misalignment penalties, the effect of having a misalignment with a frequency characteristic is considerably more complicated than was the case for thermal misalignment noise penalties. In the latter case, the penalty at a particular frequency was a function only of the misalignment at that frequency. In the case of modulation noise, misalignment penalties are a function of misalignment not only at the product frequency, but also of misalignment at each of the fundamental frequencies. In this case, the only accurate way of getting the penalty for a channel is to compute each product (with its particular penalty) and then sum the products. This is analogous to the procedure required with signal shaping.

Optimum Equalization Strategy

In order to illustrate how equalization strategy interacts with misalignment penalty, a misalignment that is constant with frequency will be assumed for the remainder of this chapter. The results from such an approach can be used in the early stages of a practical system layout to obtain an estimate of the penalty at the critical top frequency. Actual final performance optimization across the band is done by trial and error—either on a computer or by measurements on the system itself.

If it is assumed that the net gain of a string of repeater sections is constant with frequency and equal to M_S dB, then for a second order product

$$M_{M_2} = 20 \log \left(\frac{\delta_p}{\delta_\alpha \delta_\beta} \right)^n = 20 \log \delta_p{}^n - 20 \log \delta_\alpha{}^n - 20 \log \delta_\beta{}^n$$

$$= M_S - M_S - M_S = -M_S$$

Similarly, for a third order product

$$M_{M_3} = -2M_S$$

Substituting this into previously derived equations yields

$$A'_{M_2} = A'_0(-M_S)$$

$$A'_{M_3} = 2A'_0(-M_S)$$

where $A'_0(M_S)$ is the penalty function plotted in Fig. 14-2. It is

assumed that the repeater level for the reference case was selected to minimize total noise as discussed in Chap. 13. In a second order limited system this optimum was shown to occur when the thermal and modulation noise were equal. Therefore, for the misaligned case with all equalization at the receiving end, the thermal noise is $W_N + A'_0(M_S)$. By use of Eq. (14-14) and the fact that $W_2 = W_N$, the expression for modulation noise is $W_N + M_S + A'_0(M_S)$.

Instead of doing all the equalization at the receiving end, part of it could be done at the transmitting end. Assume that the fraction, r, of the total equalization required is applied at the transmitting end. Then, if M_S is the misalignment gain, all repeater levels will be lowered by rM_S dB. By the reasoning from Chap. 13,

$$\text{Thermal noise} = W_N + A'_0(M_S) + rM_S$$

$$\text{Second order modulation noise} = W_N + A'_0(M_S) + M_S - rM_S$$

As with the ideal case, the total noise will again be minimum when thermal and modulation noise are equal. This leads to a choice of $r = 1/2$. Under this condition, thermal noise, modulation noise, and total noise will each be higher by $A'_0(M_S) + M_S/2$ dB. Using Eq. (14-14), this can be written

$$A'_N = A'_{M_2} = A'_0 (\mid M_S \mid) + \frac{1}{2} \mid M_S \mid$$

and is plotted in Fig. 14-4. This plot can be used to evaluate thermal noise penalty with equal pre- and post-equalization. It can also be used for second order modulation noise penalty, for the same equalization condition and for misalignment which is flat with frequency. In both cases, $\mid M_S \mid$ is the total dB loss or gain due to the string of repeaters and cable sections.

The optimum ideal levels for a third order limited case were shown to occur when $W_3 = (W_N - 3)$ dB. When post-equalized misalignment is present, this becomes

$$\text{Thermal noise} = W_N + A'_0(M_S)$$

$$\begin{aligned}\text{Modulation noise} &= (W_N - 3) + 2A'_0(-M_S) \\ &= (W_N - 3) + 2[M_S + A'_0(M_S)]\end{aligned}$$

Introducing rM_S dB of pre-equalization as before results in:

$$\text{Thermal noise} = W_N + A'_0(M_S) + rM_S$$

$$\text{Modulation noise} = (W_N - 3) + 2[M_S + A'_0(M_S) - rM_S]$$

To find the optimum r which minimizes total noise, it is necessary to write the total noise in milliwatts:

$$\text{Total noise} = e^{0.23(W_N + A'_0 + rM_S)} + \frac{1}{2} e^{0.23(W_N + 2M_S + 2A'_0 - 2rM_S)}$$

Differentiating with respect to r and putting the result equal to zero gives:

$$0.23\, M_S\, e^{0.23(W_N + A'_0 + rM_S)} - 0.23\, M_S\, e^{0.23(W_N - 2M_S + 2A'_0 - 2rM_S)} = 0$$

$$1 - e^{0.23(2M_S - 3rM_S + A'_0)} = 0$$

which will be true if

$$2M_S - 3rM_S + A'_0 = 0$$

$$r = \frac{2}{3} + \frac{1}{3}\frac{A'_0}{M_S}$$

For $|M_S| \leq 5$ dB, Fig. 14-2 gives $A'_0/M_S \approx -1/2$ which gives $r = 1/2$ when substituted in the preceding equation. For total misalignment less than about 5 dB, equal pre- and post-equalization gives the minimum total noise for third order limited systems as it did for the second order limited case. In most practical cases, the misalignment will not be permitted to accumulate to much larger values than $|M_S| = 5$ dB without equalization, so this result applies quite accurately. If the asymptotic values shown in Fig. 14-2 are used, optimum r is shown to fall between 0.43 and 0.57 for $|M_S| \leq 20$ dB. Thus, even when misalignments larger than 5 dB are present, equal pre- and post-equalization will not be far from optimum. This will therefore be the strategy used for any modulation-limited system. The thermal noise penalty corresponding to this will be $A'_0(|M_S|) + 1/2\,|M_S|$ as shown previously. The penalty on third order voltage addition modulation noise will be $2A'_0(|M_S|) + |M_S|$ which is exactly twice the penalty for thermal noise. This is also plotted in Fig. 14-4. It is interesting to note that over the region for which $[A'_0(M_S)]/M_S = -1/2$, the thermal noise, second order modulation, and third order modulation penalties equal zero. Thus, for net gains or losses less than 5 dB, misalignment penalties can be held to insignificant values if equal pre- and post-equalization can be used. This is the case if the repeater levels set on the basis of load-carrying capacity in the ideal case would be more than $1/2\,|M_S|$ dB higher than the repeater levels based on minimizing total noise for the ideal case.

Calculation of Penalties

An example of some practical interest is based on the system configuration shown in Fig. 14-9(a). It is assumed that the section shown, which includes a transmitting equalizer, a receiving equalizer, and an equalizer midway between, is the basic building block of a long-haul system. That is, l as shown may be on the order of 100 miles, and a circuit of a thousand miles or more is achieved by successively connecting similar sections of this length. A possible variation of signal level with respect to nominal is shown in Fig. 14-9(b) prior to equalizing the deviation (solid line) and after equalizing (dashed line). In the configuration shown, the transmitting equalizer is adjusted to compensate for one-half the deviation

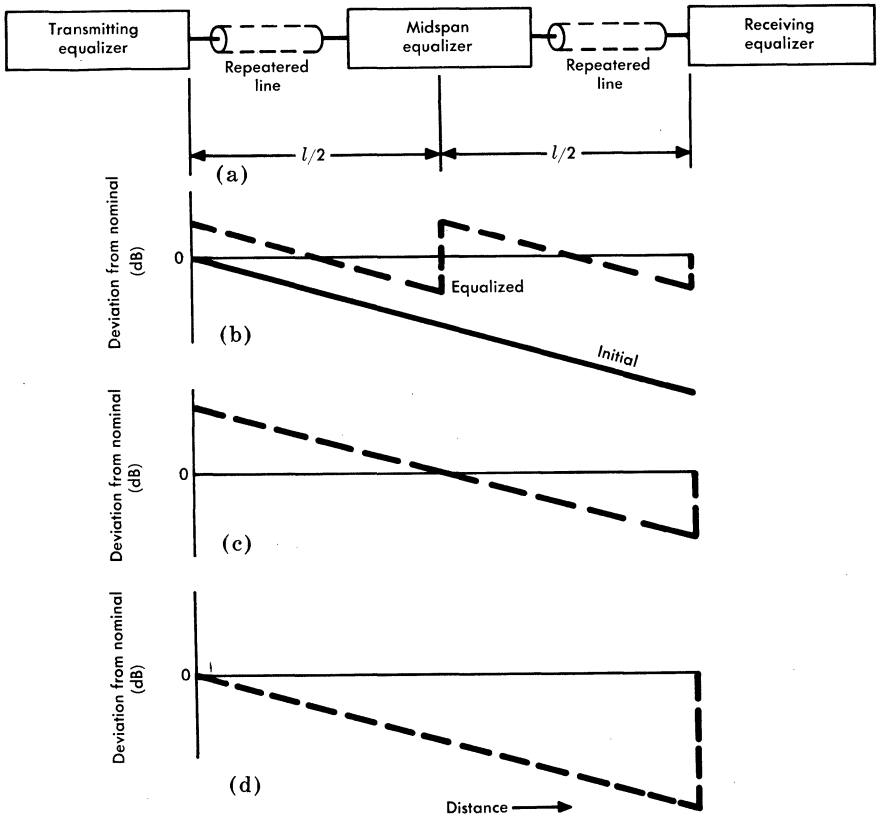


FIG. 14-9. Effect of midspan equalization on misalignment penalties.

between the transmitter and the midspan station. Thus, it compensates for $l/4$ miles of repeatered line. Similarly, the receiving equalizer is adjusted to compensate for one-half the deviation between the receiver and the midspan equalizer, again corresponding to $l/4$ miles of repeatered line. The midspan equalizer consequently serves as a post-equalizer for $l/4$ miles and a pre-equalizer for $l/4$ miles, and is adjusted to compensate for the deviation accruing in the middle $l/2$ miles of the repeatered line section.

Assuming that the system is thermal and third order limited, what are the noise penalties which result with the indicated equalizer placement? Let the total misalignment between transmitter and receiver be —8 dB. With the equalizer placement of Fig. 14-9(a), this corresponds to $M_S = -4$ dB for the section ($l/2$ in length) between equalizers. The equal pre- and post-equalized curve of Fig. 14-4 gives $A'_N = 0.1$ and $A'_{M_3} = 0.2$, for $M_S = 4$. This illustrates the fact that pre- and post-equalization hold the penalties associated with misalignments of less than 5 dB to negligible values. If the midspan equalizer is omitted, the case of Fig. 14-9(c) results. For this case, Fig. 14-4 gives $A'_N(8) = 0.6$ and $A'_{M_3}(8) = 1.2$. Since minimum total noise results when thermal noise is 1.8 dB below the total and third order modulation is 4.8 dB below the total, the penalty on total noise can be computed as follows:

$$\text{Penalty} = [(W_{TS}-1.8+0.6) \text{ "+" } (W_{TS}-4.8+1.2)] - W_{TS}$$
$$= 0.8 \text{ dB}$$

Finally, assume that pre-equalization is not possible, either because the system is not only third order limited but also overload limited or because of the physical layout of the equalization system. In this case, the situation is represented by Fig. 14-9(d). From the post-equalized curve of Fig. 14-4, $A'_N(8) = 4.6$ dB. The post-equalized case for modulation noise is not plotted on Fig. 14-4, so it will be necessary to use Eq. (14-18) and Fig. 14-2, which leads to

$$A'_{M_3}(-8) = 2A'_0(8) = -6.8 \text{ dB}$$

Using the same method as above to obtain the penalty on total noise, gives

$$\text{Penalty} = [(W_{TS}-1.8+4.6) \text{ "+" } (W_{TS}-4.8-6.8)] - W_{TS}$$
$$= 3 \text{ dB}.$$

It can be seen that misalignment has upset the optimum 3-dB

difference between thermal and third order modulation noise. If levels are adjusted to restore this relationship, the result would be approximately equal pre- and post-equalization and the associated 0.8-dB penalty computed previously would apply. Such level adjustment would be equivalent to pre-equalization, and thus the constraints assumed at the beginning of this paragraph could not apply.

This example shows the large reduction in penalties that are possible with double-ended equalization. It also shows that the misalignment penalty can be held below any specified value if the equalizers are spaced closely enough. The factors which form the basis for selecting an adequate and reasonable equalization plan to accomplish this are discussed in the next chapter.

Chapter 15

Equalization in Analog Cable Systems

The effect of transmission deviations (misalignment) on the S/N performance of analog systems was considered in the preceding chapter. These deviations were assumed to be compensated by equalizers placed at various points in the system. In this chapter, various types of equalizers are examined along with the factors determining their spacing along the system length.

In a general way, the design of an analog cable system involves the design of a collection of amplifiers and equalizers of various kinds which together with the cable form the basic transmission medium of the system. Since some of the sources of transmission deviations are time dependent while others are not, the specification of either fixed or adjustable equalizers at a particular point should depend largely on the kind of deviations they are intended to correct. A given system, in other words, will generally include both fixed *and* adjustable equalizers with each type of equalizer intended to compensate for deviations caused by different mechanisms within the system. It was apparent in Chap. 14 that, while the desired end-to-end transmission response can be achieved equally well with just about any physical placement of a given set of equalizers, the S/N penalties and the impact on repeater overload are very dependent upon the location of the equalizers. This chapter discusses the interaction of these factors as they affect the system design.

15.1 FIXED EQUALIZERS

Analog cable systems generally use fixed equalizers to provide the necessary compensation for effects which can be predicted accurately

in advance and which are essentially unchanging with time. The three major effects of this type to be discussed are the nominal cable loss, repeater-spacing deviation from nominal, and the design deviation which characterizes the repeaters as they leave the manufacturing shop.

Basic Line Repeater

In the sense that the function of the basic line repeater is the equalization of the associated cable section loss, the repeater itself performs what can be considered the first level of fixed equalization in the system.* The nominal gain required of the repeater is that determined by the analytical processes discussed in Chap. 13. For example, Fig. 15-1 shows what the required gain-frequency characteristic would be if the analysis resulted in a one-mile repeater spacing with 3/8-inch coaxial cable. It is interesting to observe that the

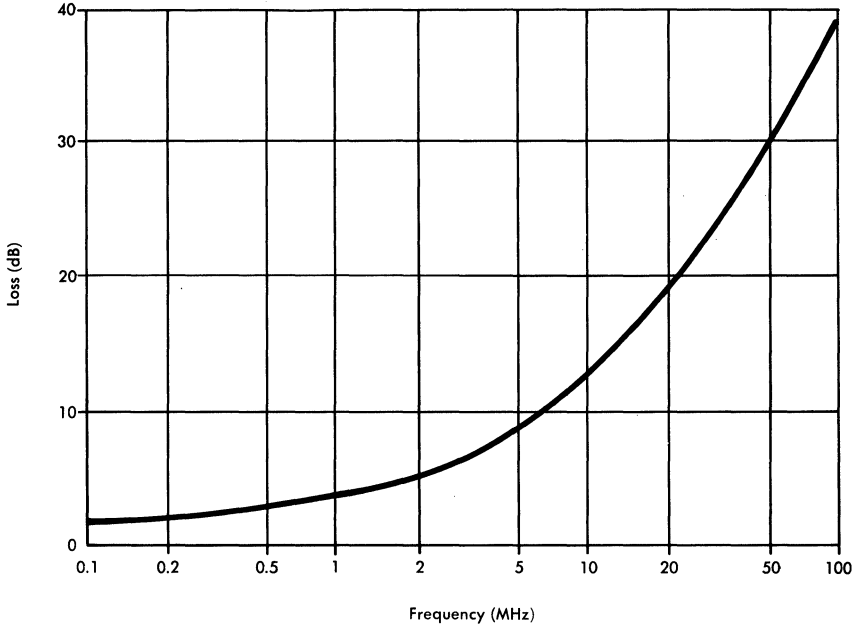


FIG. 15-1. Loss of one mile of 3/8-inch coaxial cable.

* It is not necessarily an inherent system requirement that the basic line repeater be a fixed gain device, but this is most often the case.

function of the repeaters in a 4000-mile length of this cable is the equalization of about 70,000 dB of cable loss if the top transmitted frequency is 20 MHz. For a top frequency of 60 MHz, the comparable number is 121,000 dB. The results of the previous chapter imply that misalignments must not exceed the order of 10 dB if performance penalties are to be acceptable. For the above examples this means the average net deviation must be held to the order of 0.01 per cent. It is apparent that the match between the line repeater gain and associated cable loss is of principal importance in the design of an analog cable system, and that the extent to which the match is imperfect significantly affects the requirement for additional equalization.

The first level of fixed equalization, then, is performed by the basic line repeater, which is designed to equalize the loss corresponding to the nominal repeater spacing. The physical installation of a cable system will not always permit the placement of the sequence of repeater stations at precisely the nominal interval, giving rise to a second level of fixed equalization—line building out.

Line Build-Out Networks

In cable systems provision is usually made for shorter-than-nominal repeater spacings by the use of passive line build-out networks which closely simulate the cable loss in increments of 5 to 10 per cent of the nominal spacing. A family of ten line build-out networks might be available to simulate the loss of cable lengths between 5 and 50 per cent of the nominal spacing, the connection of the appropriate network being made at the time the repeater is installed.* This permits the placement of repeaters as close as one-half the nominal spacing if this becomes a physical necessity due to geographical constraints. To achieve the desired overall system performance, it is necessary to limit the number of such short sections. In general, then, the basic building block of the analog cable system may be considered to be an amplifier (or amplifiers) and a line build-out network.

A requirement for a third level of fixed equalization results from the unavoidable mismatch between the average repeater gain and the nominal cable loss. The effect of this mismatch cannot be per-

*The selection process described here is distinguished from "adjustment" to be discussed later. The line build-out networks are therefore classed as fixed equalizers.

mitted to accumulate excessively and is usually corrected by the periodic inclusion of the third level of fixed equalization, sometimes called design deviation equalization.

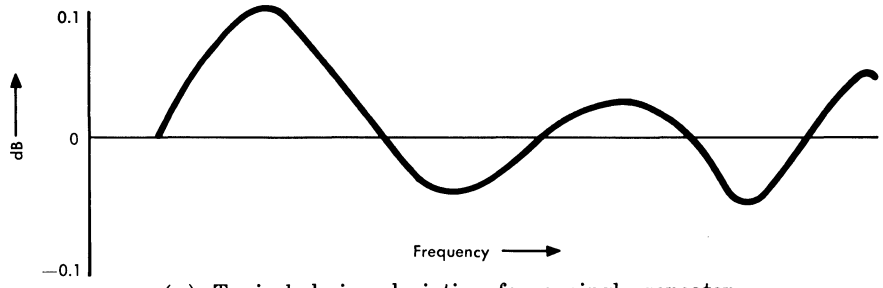
Design Deviation Equalizers

After the design of the basic line repeater is complete and its gain has been tailored to match as closely as practicable the cable loss corresponding to the nominal repeater spacing, it is possible to establish the average difference between the actual repeater gain and the objective. This difference is usually called the design deviation and might look in a particular design like the characteristic of Fig. 15-2(a). The design deviation by the nature of its definition accumulates systematically from repeater section to repeater section, and consequently the gross design deviation for a string of k repeater sections is the characteristic of Fig. 15-2(a) multiplied by k.

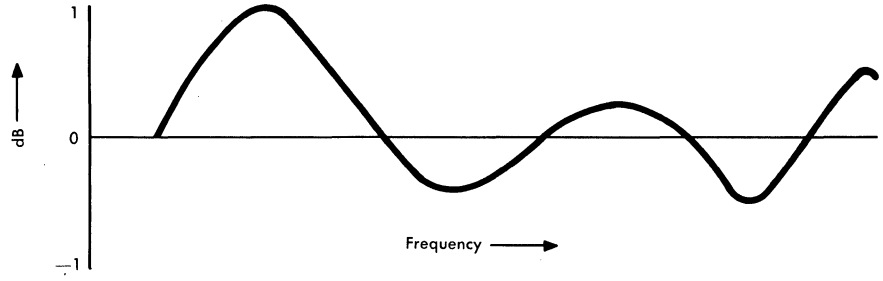
If the design deviation of Fig. 15-2(a) is to be equalized by the placement of a fixed equalizer at every tenth repeater along the line, then the accumulated design deviation will be that shown in Fig. 15-2(b) and the required equalizer response will be that shown in Fig. 15-2(c). To the extent that the characteristic of Fig. 15-2(c) precisely complements the characteristic of Fig. 15-2(b), there will be no residual design deviation for the ten-repeater section. Practical designs will be able to achieve this only approximately. The residual design deviation for the section will be greatly reduced from that of Fig. 15-2(b); however it will combine with the corresponding residuals in tandem sections to produce an overall residual which requires further attention. Since this particular residual tends to be systematic and time invariant, it can usually be compensated by the inclusion of another but different fixed equalizer. In the example being discussed, these might be placed at intervals of perhaps fifty to one hundred repeaters. If this were done, it would constitute in effect a fourth level of fixed equalization.

15.2 ADJUSTABLE EQUALIZERS

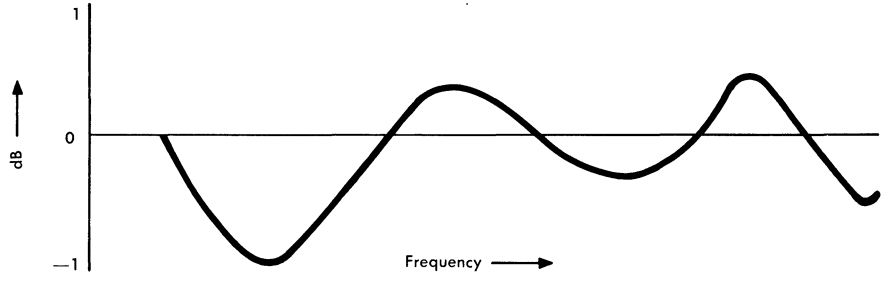
Those phenomena which cannot be specified accurately in advance of a system installation or which tend to change with time can only be corrected with equalizers having adjustable characteristics. One of the factors to be considered in selecting the type of equalizer required is the rate at which the adjustments must be made. Equal-



(a) Typical design deviation for a single repeater



(b) Cumulative design deviation for ten repeaters



(c) Fixed equalizer response required to correct for the design deviation in a ten-repeater section

FIG. 15-2. Correction of design deviation.

izers requiring adjustments only at the time of installation, for example, can usually be manually operated. Manual operation may also be satisfactory for equalizers which require only occasional adjustment. However, equalizers intended to compensate for relatively rapid variations in the transmission response will probably involve

automatic regulators.* In some cases it will be possible to associate particular equalizer characteristics with specific and well-defined causes of transmission deviations. In other cases equalizers capable of correcting a more general class of transmission deviations will be required.

Of the deviations which require adjustable equalizers, the ones to be discussed result from the cable-temperature effect, time-and/or temperature-dependent repeater effects, and time-invariant repeater effects the shapes of which are not predictable.

Cable-Temperature Effect

It will be recalled that the main objective of the basic line repeater in an analog cable system is the compensation of the nominal cable loss corresponding to the nominal repeater spacing. The nominal loss is usually defined to be the cable loss occurring at mean cable temperature. The one-mile loss of the 3/8-inch coaxial cable plotted in Fig. 15-1 applies only at 55°F, the approximate mean earth temperature within the United States. The loss of this particular cable varies with temperature at a rate of approximately 0.11 per cent per Fahrenheit degree. The variation from the nominal (55°F) loss is shown for one mile of coaxial cable in Fig. 15-3 for a ±20°F change in temperature. (This is about the maximum earth temperature variation from the local mean temperature at a depth of 4 feet—the depth at which long-haul cables are currently buried. The local mean temperature itself will vary with location over a range of about ±20°F.) It can be seen that the cable-temperature effect amounts to about ±0.38 dB/mile at 20 MHz and about ±0.67 dB/mile at 60 MHz. For 4000 miles the corresponding quantities are 1520 dB and 2680 dB, respectively.

The effect of the changing cable loss will generally be by far the largest factor affecting system performance and equalization needs. It is also an effect whose shape can be accurately specified in advance of the system installation. It is conventional, therefore, to assign a specific adjustable equalizer to the task of correcting this effect. Because of the magnitude of the effect, these equalizers will always be the most numerous of the adjustable equalizers in any system and will occur most frequently along the line. For the same

*In the jargon of analog cable systems, the automatic correction of transmission variations is often called regulation. This distinction between *regulation* and *equalization* is not made in this text. However, *regulator* will be used to mean exclusively an equalizing *device* which responds *automatically* to some stimulus. Regulator features are discussed later in this chapter.

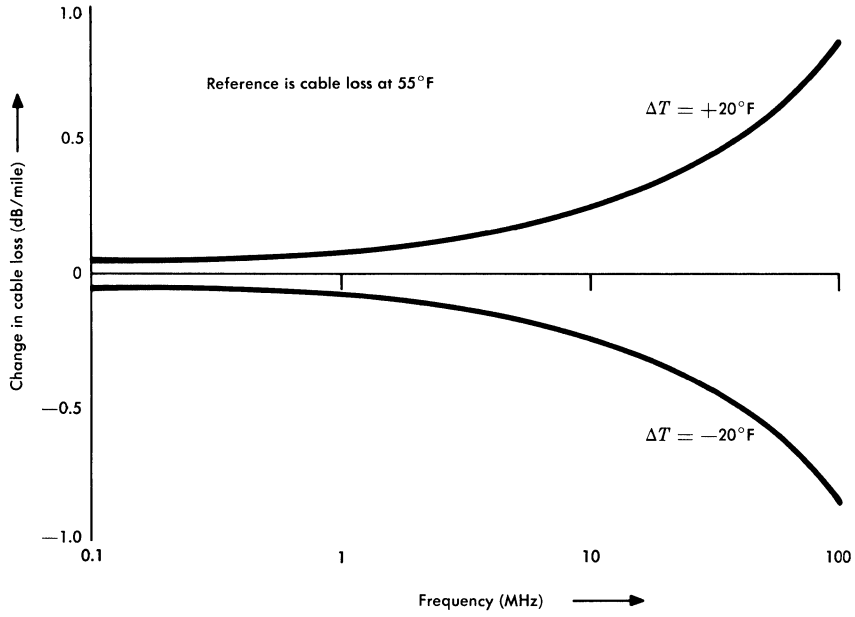


FIG. 15-3. Change in loss of one mile of 3/8-inch coaxial cable for ±20°F change in temperature.

reason they most often will be continuously adjustable equalizers and will be realized with regulator circuits of one kind or another. As a result, the repeaters incorporating these circuits have often been called regulating repeaters or temperature-regulating repeaters.

While analysis of the type covered in Chap. 13 will establish the nominal spacing of the basic line repeaters, the considerations of Chap. 14 become important in establishing the spacing of the regulating repeaters. As discussed in the latter chapter, the chief factors influencing this decision are the repeater load capacity and the system S/N performance. The choice of spacing also depends on whether double-ended or single-ended equalization is to be used.

The regulators which are likely to be used to correct for the cable-temperature effect can be either open or closed loop. The closed-loop or feedback regulator will usually be designed to respond to a pilot or test signal continuously present on the coaxial line. An open-loop regulator is designed so that it responds to some other stimulus—such as the nearby earth temperature, for example. The use of one or the other of these types of regulators is closely related

to the question of single-ended versus double-ended equalization, as well as to the accuracy with which signal magnitudes must be maintained as the signal traverses the coaxial line. The decision for a particular problem involves achieving the most appropriate balance among the following points:

1. The open-loop regulator tends to be relatively simple to realize, while the closed-loop regulator is usually more complex.
2. The open-loop regulator is usually less accurate than the closed-loop regulator.
3. The errors associated with the open-loop regulator, larger to start with, tend to accumulate in proportion to the number of tandem regulators. The errors associated with the closed-loop regulator, smaller to start with, do not tend to accumulate [1].
4. An open-loop regulator lends itself fairly readily to use as a pre-equalizer. The use of a closed-loop regulator pre-equalization plan presupposes the transmission of an error signal from down the line back to the equalizer, requiring a reverse channel for the continuous feedback of information over what might be many miles of system.

A combination of open- and closed-loop regulators will often provide the best overall compromise from the viewpoint of system performance, reliability, and maintainability. The S/N and overload advantages which result from pre- and post-equalization will, in modulation-limited systems, often justify the difficulty and increased complexity involved. In overload-limited systems, on the other hand, single-ended equalization, in which positive misalignment is wholly pre-equalized while negative misalignment is wholly post-equalized, may be unavoidable.

Consider the cable-temperature effect of Fig. 15-3 and assume that the line repeaters can tolerate a maximum signal level increase of 4 dB in the top channel on the basis of overload considerations. The regulating repeater spacing required to satisfy this requirement will presently be established for a 20-MHz top channel frequency, first using only post-equalizers and then using both pre- and post-equalizers.

From Fig. 15-3, the top channel cable loss change for a $-20°F$ change in temperature is -0.38 dB/mile. The resultant signal level deviation at 20 MHz is shown in Fig. 15-4 as a function of distance along the repeatered line. It can be seen that the signals will be 4 dB in excess of the nominal after passing through 10.5 miles of

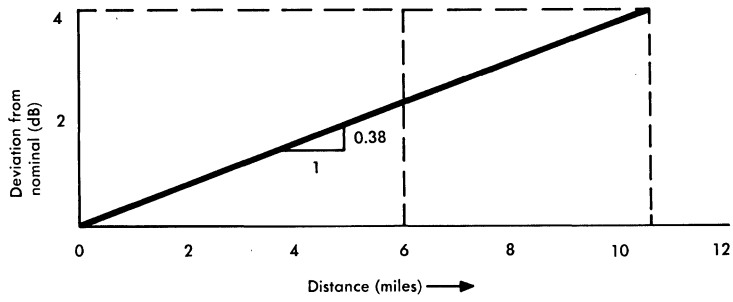
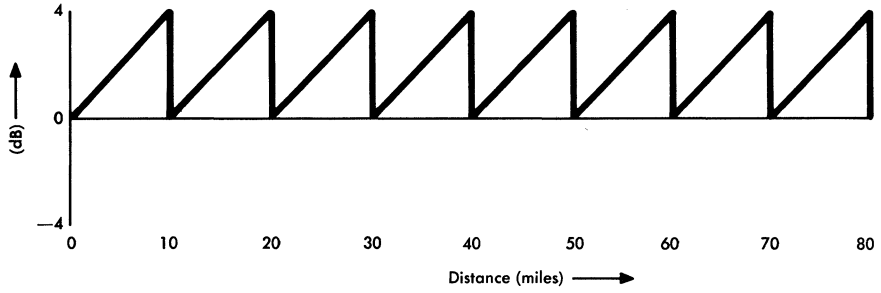


FIG. 15-4. Signal level deviation at 20 MHz for −20°F change in cable temperature.

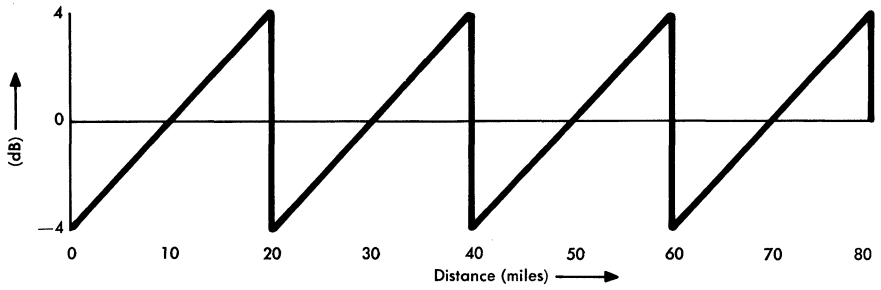
repeatered line. Consequently, the last repeater station prior to the 10.5-mile point would have to be made a regulating repeater if only post-equalization were to be used. If both pre- and post-equalization were to be employed, the signal levels would not exceed the nominal by 4 dB until 2 × 10.5 or 21.0 miles of repeatered line had been traversed.

In a relatively short length of system, the cable-temperature effects will tend to be similar, and the signal levels over that interval will look much like those of Fig. 15-5 for the 20-MHz case just considered. Figure 15-5(a) assumes post-equalization only, while both pre- and post-equalization apply in Fig. 15-5(b). It was shown in Chap. 14 that the equalizing plan of Fig. 15-5(b) results in lower S/N penalties than that of Fig. 15-5(a) and in addition requires fewer regulating repeaters. Specifically, the penalty incurred in the plan of Fig. 15-5(b) is less than 1 dB, whereas the plan of Fig. 15-5(a) results in a thermal noise penalty of −2 dB but a third order modulation penalty of 4.5 dB. This is one example of lower S/N penalties and increased equalizer spacing which may justify the increased complexity resulting from pre- and post-equalization. The need for a continuing interaction between the system analysis and the design of the system constituents is once again illustrated by this example.

The intervals at which the regulating repeaters are required will often also be appropriate for the third level of *fixed* equalization with the so-called deviation equalizers. Consequently, these are often located there with the result that the intervening basic line repeaters, with line building-out as required, remain the simplest possible build-



(a) Post-equalization only—at 10-mile intervals



(b) Pre- and post-equalization at 20-mile intervals

FIG. 15-5. Signal level deviation in 80-mile section for −20°F change in cable temperature.

ing blocks. For system reliability and ease of operation and maintenance, it is desirable to limit increased complexity to as few locations as possible.

Time-Dependent Repeater Effects

The largest of the repeater-caused changes in system transmission response with time usually results from the effect of operating temperature on the transmission response of the repeaters. If the repeaters are either buried or mounted in underground manhole-like structures (i.e., not air-conditioned), the operating temperatures will tend to track the seasonal variations in temperature. A typical variation in operating temperature for repeaters installed about four feet below the earth's surface is shown in Fig. 15-6. The shapes of the deviations due to the repeater-temperature effect are likely to be different from the time-invariant deviations discussed next, and it may or may not be attractive to compensate for them with the

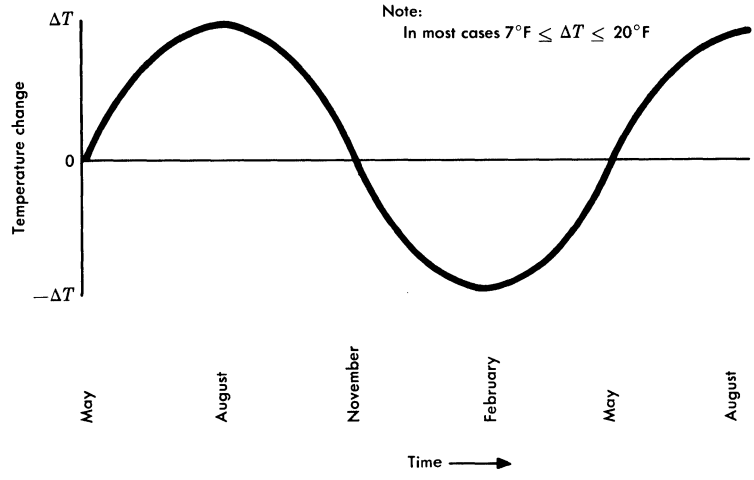


FIG. 15-6. Approximate earth temperature variation with time at 4-foot depth.

same set of adjustable equalizers. Experience in the L4 system, for example, indicates that these effects may be relatively simple functions of frequency and may consequently require a relatively simple adjustable equalizer for their correction, as compared with the relatively complex equalizers required by some of the time-invariant deviations.

Smaller and slower changes affecting system transmission response are those resulting from the aging of the elements making up the system. These can result from changes of any element in the system, but such changes in the line repeater are likely to be controlling. They will usually be small in a system composed of repeaters designed to minimize such effects.

Generally, the smallest time-dependent source of change in system transmission response is that assigned to *maintenance.* Over any significant period of time, there will be inescapable repeater failures and subsequent replacement. Since the replacement repeater will not generally have *exactly* the same response as the failed repeater which it replaces, this introduces a deviation of unpredictable shape. This class of deviations as well as the aging effects will usually be handled by manual readjustment of the general equalizers discussed later.

Time-Invariant Repeater Effects

The previous paragraphs have discussed the correction of deviations whose shape with respect to frequency is well defined and whose magnitude varies with time. Another purpose for adjustable equalizers is the correction of time invariant deviations whose shape is not known prior to installation. For this purpose, powerful families of equalizer shapes are required which are capable of applying a relatively good correction to any deviation shape which might be encountered.

The major source of deviations falling into this category is the statistical variation incurred during manufacture. This includes the impact of unavoidable variations in the components making up the line repeaters. While the deviation equalizers are used to correct for the *average* design deviation, the variations from the average, which will characterize any particular group of repeaters, must be corrected by adjustable equalizers. Since by definition this deviation is unchanging with time, it can be corrected by an adjustable equalizer which is set upon system installation and not thereafter adjusted. The nature of this deviation is such that it is likely to be a complex function of frequency and may require a correspondingly sophisticated adjustable equalizer for its correction. Various types of equalizers can be used to correct the deviations discussed in these sections. Some of the types commonly used and their main advantages and disadvantages are discussed in the following.

Bumps. One way of correcting a transmission deviation is by means of a series of equalizers of the form shown in Fig. 15-7.

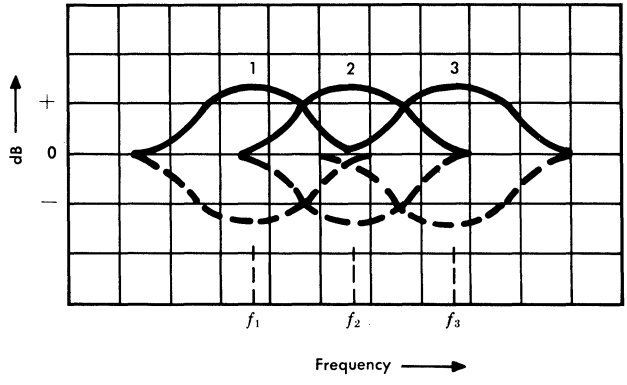


FIG. 15-7. Bump equalizer shapes.

These equalizers consist of a number of separate units each of which can be varied to give adjustment over a small range of frequencies. Thus unit 1 can be varied between the limits shown by the first solid and dotted pair of lines; unit 2 supplies the next overlapping bump, etc. Such equalizers are characterized by ease of design and manufacture. They are sometimes difficult to adjust accurately, however, because the slopes of the bumps must overlap to give continuous coverage of the frequency range. As a result there is a tendency for appreciable interactions between bumps; that is, the adjustment of one equalizer can spoil the characteristic in the region overlapped by an adjoining bump. Because of this the adjustment of this type of equalizer may be more difficult utilizing sweep measurements. If adjustments are based only on measurements at discrete frequencies (one for each degree of freedom, for example, f_1, f_2, and f_3 in Fig. 15-7), the interaction will not cause much difficulty in optimizing the result.

If the discrete frequencies can be placed in guard spaces between message channels, it may be possible to carry out the adjustment with the line in-service. The sweep technique often results in a better across-the-band equalization but can be carried out only out-of-service since the sweep would interfere with any message signals being carried at the time.

Power Series. This family of equalizer characteristics is made up of related terms which form a power series in dB versus frequency. Mathematically, the equalizers may be represented by

$$F(f) = B_1 f + B_2 f^2 + \ldots B_n f^n \qquad (15\text{-}1)$$

Each term on the right side of Eq. (15-1) represents an equalizer characteristic. The first term, for example, is linear in dB versus frequency and has a coefficient B_1 which is adjustable between predetermined plus and minus limits. This family of equalizers may be expected to be a powerful tool in the adjustment of sharp band edge characteristics at the top of the band. Negative powers of frequency (Laurent's series) would be needed at the low-frequency edge of the band. Although they have often been considered, such equalizers have not been worked out. One possible method would be to combine the terms in an orthogonal series such as Legendre's, which allows optimization of each term independently (in the least squares sense).

Cosines. The use of cosine equalizers was suggested by the observation that manual equalizers are called upon to correct a wide range

of complex deviation characteristics. Each of these characteristics is, in theory, expressible as a mathematical function in dB versus frequency, which can be analyzed in terms of its Fourier components. If equalizers can be designed to correspond to each of these components and if each can be adjusted so that its magnitude is the same as that of the corresponding Fourier component in the deviation characteristic, then, in theory, any deviation characteristic can be corrected. Figure 15-8 shows, for example, the first three terms of a Fourier series applicable to equalizing a transmitted band extend-

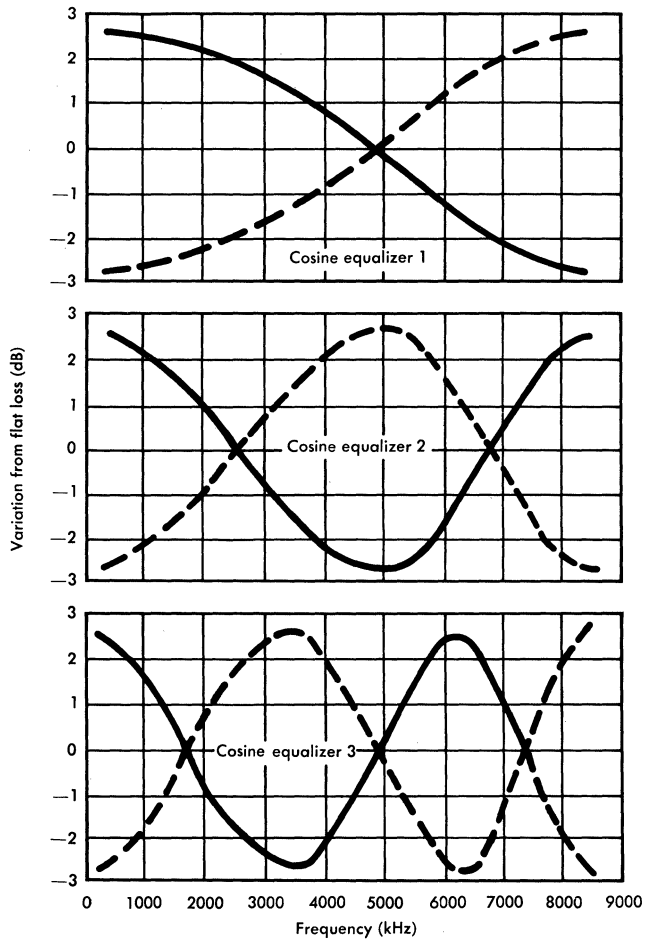


FIG. 15-8. Shapes introduced by the first three cosine harmonics.

ing from 0 to 8.5 MHz. The accuracy of the correction will depend on the precision with which each term can be set and on the number of terms for which equalizers are provided. Difficulty may be encountered at band edges, however, due to Gibbs' phenomenon.

Cosine equalizers were invented many years ago but were not immediately used in a system because a suitable method of adjustment could not be found. A solution to this problem is found in the application of sweep techniques. By these means, a dB-frequency characteristic is converted to a voltage-time characteristic. To adjust the equalizer, it is only necessary to minimize the electrical power represented by this characteristic. The simple criterion, minimum power, theoretically permits adjustment of the terms in any order, since correct compensation of any cosine term in the original wave results in a reduction of total power at the detector output. (This is true of any orthogonal series.) In practice, of course, masking effects dictate that large deviations be corrected first. If it is found during early installations of a new system that a given complement of equalizers does not achieve acceptable accuracy, equalizers corresponding to more terms in the Fourier series may be added. The use of sweep techniques makes it necessary to take the line out of service to adjust the equalizers.

Time Domain. In Chap. 29 the equivalence of echoes and transmission deviations is discussed; i.e., deviations can be considered as causing echoes or vice versa. If echoes are added which cancel those existing in a system, transmission deviations will also be corrected. A single echo, for example, corresponds to a minimum phase cosine loss ripple; the cosine equalizers discussed above are therefore a special case of the general class of time-domain equalizers. By providing taps on delay lines and means for adjusting magnitudes of the signals picked off, and by adding them to the original signal, any desired pattern of leading and lagging echoes around a signal can be provided. Used in conjunction with sweep adjustment techniques, such time domain equalizers provide adjustment of gain and delay distortion. They are an essential tool in the equalization of long broadband systems for television transmission.

15.3 EQUALIZATION DESIGN

Equalization design involves the determination of an equalizing plan and the specification of a family of equalizer networks to pro-

vide the required overall system transmission response. The selection of an equalizing plan involves the allocation and physical placement of the several kinds of equalizers chosen and is based on such considerations as S/N penalties, repeater overload, and equalizer adjustable range requirements. The choices of an equalizing plan and a family of equalizer networks are not entirely separate questions. The major considerations affecting the decisions in these areas include the following:

1. Simplicity of realization.
2. Ease of manufacture.
3. Ease of adjustment in system.
4. Association of equalizer shapes with specific causes of deviations where possible.
5. Maximum flexibility for a given level of complexity in those equalizers used for equalizing deviations whose shape is not known *a priori*.
6. Minimum residual errors in system transmission response for a given complexity of equalizer plan.
7. Whether the adjustable equalizers must be set while the system is carrying message service, or whether this can be done out-of-service.
8. Whether the adjustable equalizers associated with time-varying phenomena require continuous adjustment (as with a pilot-controlled regulator) or whether a periodic manual adjustment is sufficient.

In different system designs different problems and constraints will predominate, depending on certain basic goals which the final system is intended to realize. Consequently, it is impossible to specify a universally optimum equalizing plan or class of equalizers. A survey of existing analog cable systems would show a considerable variety of techniques and networks. Some of the equalizer types usually considered for application have been discussed in previous paragraphs.

Equalizer Selection

Once some idea of the transmission deviations which are to be equalized is available, the ability of the various kinds of equalizers

to achieve the necessary performance can be readily determined. Complex deviations and large numbers of adjustable equalizer shapes will usually require the development of appropriate computer programs for convenient analysis. Once this is done, the questions of optimization and inherent capability of a particular family of equalizers can be resolved in a fairly straightforward fashion.

It is apparent that the demands on each of the types of equalizers making up an analog cable system depend largely on the success with which the lower-order equalizers are realized. The deviation equalizer depends on the basic line repeater response. The requirements on the equalizers which correct for the response deviations present in the system initially depend on both of these and the way they interact with the cable. The specification of the equalizers for time-dependent effects should be based on the known repeater and cable dynamic behavior. The equalization plan as a whole depends on all of these. However, an equalization plan which is to be available on a timely basis often must be specified at least initially without complete information on one or more of these important ingredients, due primarily to the timing of the various phases of system design, manufacture, and evaluation.

The Equalizing Plan

It has been pointed out that the selection of the equalizing plan is not an entirely separate question from the selection of the equalizers themselves. An important point of interaction, for example, is the adjustable range which reasonably can be achieved with a particular type of equalizer. The selection of the equalizing plan for purposes of this discussion assumes that the basic layout of the system has been established as described in Chap. 13. Consequently, the nominal line repeater spacing is known and certain allowances or margins have been incorporated into the design. Selecting the equalizing plan will involve the specification of the intervals at which the different kinds of equalizers must be inserted. This selection must be consistent with the allowances made in the basic analysis for misalignment and the like, and also allow for certain circuit design details. The principles involved in selecting the equalizer layout of an analog cable system will be treated by example.

Assume that the basic analysis of a coaxial system had led to the following conclusions:

1. Nominal repeater spacing: 1 mile.

2. The added complexity of pre- and post-equalization of all significant deviations can be justified on the basis of experience with previous similar systems.
3. Repeater load capacity will permit satisfactory signal transmission up to 5 dB above nominal transmission level.
4. Remote powering and reliability factors require automatic protection of the coaxial line at 75-mile intervals. This means that parallel 75-mile sections of a multi-line cable must be sufficiently alike to be interchangeable. The 75-mile section is therefore a logical choice as the basic building block of the system.

Since pre- and post-equalization are specified, the chief remaining concerns which interact with the equalizing plan are the prevention of system (repeater) overload during the misaligned condition, and the requirements on adjustable equalizer range. The basic objective will be to space the equalizers as far apart as possible consistent with the overload limitations. It is assumed that the S/N penalty corresponding to levels which deviate ± 5 dB from nominal is consistent with the margins allowed in the system analysis leading to this layout.

Assume that the following characteristics of the cable, amplifiers, and networks have either been measured or are reasonable estimates:

1. Maximum design deviation: ± 0.10 dB/repeater.
2. Maximum repeater gain change over operating temperature range: ± 0.10 dB/repeater.
3. Top frequency cable loss change over operating temperature range: ± 0.5 dB/mile.

First consider the placement of the regulating repeaters assigned the task of compensating for the variation in cable loss with temperature. These repeaters will be able to insert corrections of the kind shown in Fig. 15-3. Making an initial allowance of 3 dB of the maximum 5-dB level deviation for the cable-temperature effect, assume (1) that the transmitting station of the 75-mile section will have only a pre-equalizer of range ± 3 dB, (2) that the receiving station will have only a post-equalizer of range ± 3 dB, and (3) that the intervening regulating repeater stations will have both a pre-equalizer and a post-equalizer, each having an adjustable range of ± 3 dB. (If possible, the two equalizers might be combined in a single equalizer network having ± 6-dB range.)

The section of line between the transmitter and the first regulating repeater, or between regulating repeaters, or between the last

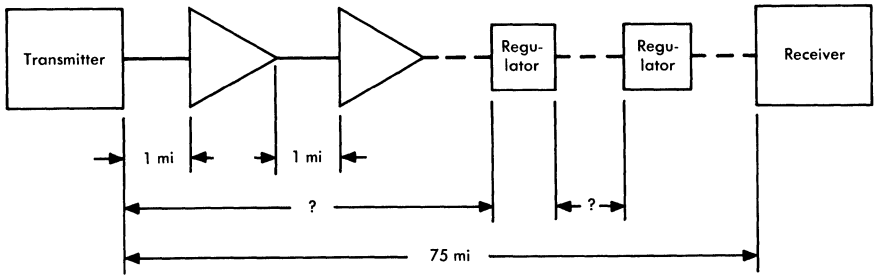
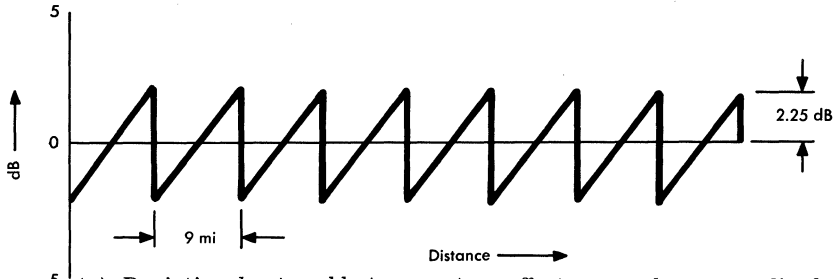


FIG. 15-9. System layout for determination of regulating repeater spacing.

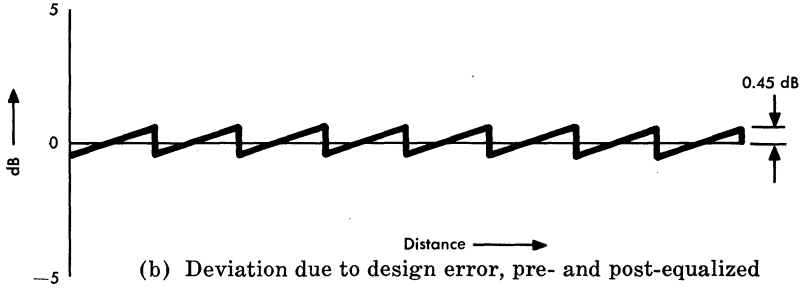
regulating repeater and the receiver (Fig. 15-9) will consequently be compensated by one pre- and one post-equalizer having a combined range of ±6 dB. Since the cable-temperature effect has been stated to be ±0.5 dB/mile, a 12-mile regulating repeater spacing may appear to be in order. Several factors tend to make it undesirable to specify this absolute maximum spacing as the actual system spacing for regulating repeaters. For example:

1. It will be difficult or impossible to establish an exact "centering" of the regulator operation, in the case of closed-loop regulators, due to deviations in the line repeaters at the pilot frequency; or, in the case of an open-loop regulator, due to the inherent inaccuracies. In the case of a mixed closed-open-loop plan, part of the closed-loop regulator range must be assigned to eliminating the open-loop regulator errors.
2. The basic line repeater gain is, in all likelihood, temperature-dependent at the pilot frequency of any closed-loop regulators which are used. Unless the regulating repeater can distinguish between changes in pilot level due to cable loss change and changes due to line repeater gain changes, some adjustable range will be used up here as well.

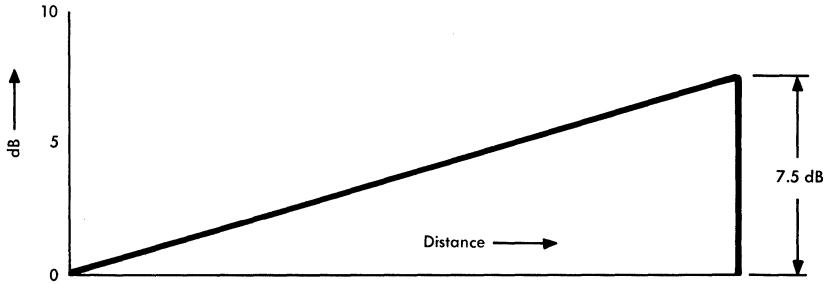
The combined result may be to make the effective change in response, for the purpose at hand, more nearly 0.6 to 0.7 dB/mile than the 0.5 dB/mile due to the cable-temperature effect alone. Consequently, a spacing of regulating repeaters at 8- to 10-mile intervals might be more reasonable. For this example, set the regulating repeater spacing at 9 miles. Figure 15-10(a) shows the signal level deviations resulting from the cable-temperature effect as equalized by the indicated plan at 9-repeater intervals.



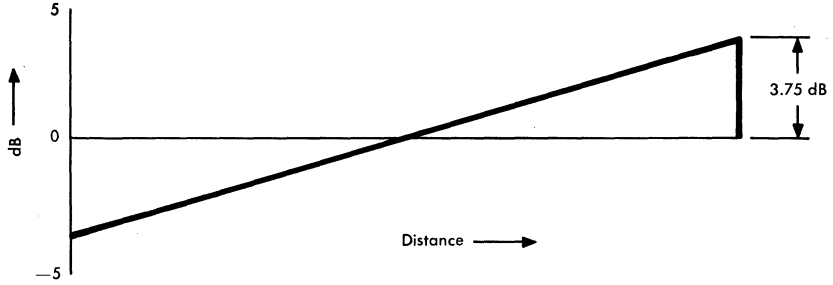
(a) Deviation due to cable-temperature effect, pre-and post-equalized



(b) Deviation due to design error, pre- and post-equalized



(c) Deviation due to repeater gain-temperature effect, prior to equalization



(d) Deviation due to repeater gain-temperature effect, pre- and post-equalized

FIG. 15-10. Signal level deviations resulting from major misalignment sources in 75-mile section.

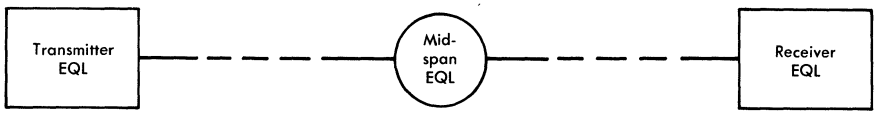
The regulating repeater station is a convenient place to locate the deviation equalizers required to correct for the line repeater design deviation. The interval is about right since the accumulated deviation could, in a worst case, approach 1 dB for 9 repeaters, and their placement there also avoids any complication of the basic line repeater stations. Since the design deviation from repeater-to-repeater within the regulating section accumulates systematically at up to 0.1 dB per repeater, the potential misalignment from this source is 0.9 dB. If this deviation is also pre- and post-equalized, the corresponding signal level deviation will be ± 0.45 dB. Figure 15-10(b) shows a possible signal level deviation associated with the repeater design error, given equal parts of pre- and post-equalization at the 9-repeater interval.

The third source of misalignment to be considered in the present discussion is the temperature-dependence of the line repeater gain. The effects of a particular temperature change on the gain of the many repeaters will tend to be alike; further, the temperature changes experienced over a 75-mile section will tend to be the same. An assumption of uniformly accumulating misalignment due to the effect of temperature on line repeater gain is usually a good one. Figure 15-10(c) shows the unequalized signal level deviations which result from 0.1 dB-per-repeater temperature effect over a 75-mile section. The equalizers required in this case may be one of the families of adjustable shapes discussed previously. Figure 15-10(d) shows the signal level deviations corresponding to Fig. 15-10(c), given equal parts of pre- and post-equalization. Two points of interest can be noted:

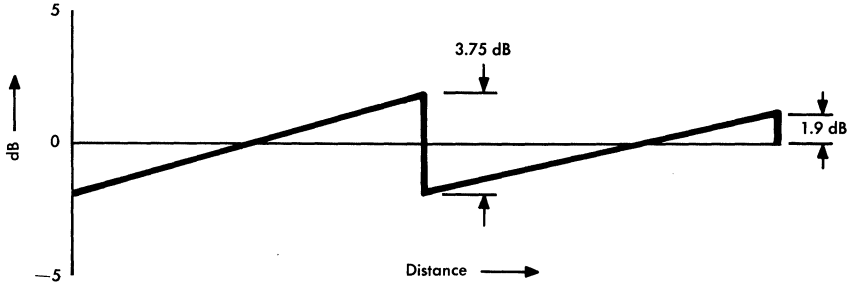
1. The addition of the deviations from all sources [Figs. 15-10(a), 15-10(b), 15-10(d)] results in exceeding the load capacity of the repeaters.*
2. The adjustment range required for the correction of the repeater-temperature effect is nearly 4 dB.

*It might be argued that the assumption that all effects will add systematically is overly pessimistic, but a closer examination will show this not to be so. Both cable and repeaters will be subjected to the same temperature variation at any particular time. Therefore if cable loss increases and repeater gain decreases for a temperature change in a particular direction, then the deviation must indeed be added. There is no similar argument for a correlation between temperature effects and design error. However, by nature, design deviations tend to be ripply with numerous zero crossings. Thus, if the two tend to cancel in one portion of the band, there will be a nearby region of the spectrum where the two effects will add; as a consequence, adding the effects is not unduly conservative.

Item 1 alone is sufficient to require one or more line equalizers (i.e., equalizers between the end stations). In addition, the difficulty in achieving a carefully controlled adjustable equalizer characteristic increases rapidly with the dynamic range required. Although a dynamic range of ±4 dB with the required linear tracking can probably be achieved in most cases, ranges of this order become increasingly elusive as the top frequencies of new systems increase. In any event, there are other factors not explicitly considered in the example which might warrant a line equalizer anyway. These include such items as the manufacturing variations in the product, deviation equalizer imperfections, and repeater aging. In the case at hand, a single line equalizer would result in a system layout for the 75-mile building block like that of Fig. 15-11(a). Although the layout of Fig. 15-11 is satisfactory from the repeater load capacity view-



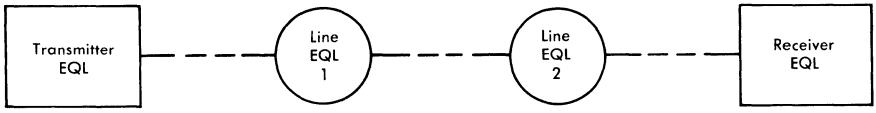
(a) Equalizer layout including a single line equalizer midway between the transmitting and receiving stations



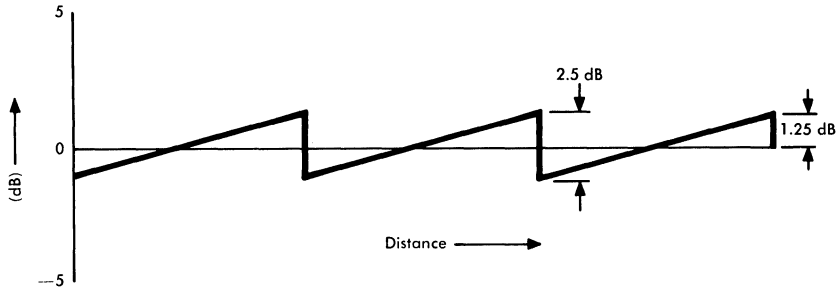
(b) Signal level deviation due to repeater gain-temperature effect for equalizer layout of Fig. 15-11(a)

FIG. 15-11. Addition of midspan equalizer.

point, it is apparent in Fig. 15-11(b) that the range required of the midspan equalizer is still nearly 4 dB. In view of the above factors, a second line equalizer might be required, producing a layout like that of Fig. 15-12(a). Alternatively, two sets of equalizer



(a) Equalizer layout including two line equalizers between the transmitting and receiving stations



(b) Signal level deviation due to repeater gain-temperature effect for equalizer layout of 15-12(a)

FIG. 15-12. Addition of two line equalizers.

networks might be included in the midspan station of Fig. 15-11, with one set assigned the task of post-equalization only and the other pre-equalization. If the layout of Fig. 15-12(a) were implemented, the signal level deviations associated with the repeater-temperature effect would be those shown in Fig. 15-12(b) after equalization.

REFERENCES

1. Elmendorf, C. H. et al. "The L3 Coaxial System," *Bell System Tech. J.*, vol. 32 (July 1953), pp. 781-1005.
2. Bode, H. W. "Variable Equalizer," *Bell System Tech. J.*, vol. 17 (Apr. 1938), pp. 229-244.
3. Lundry, W. R. "Attenuation and Delay Equalizers for Coaxial Lines," *Trans. AIEE*, vol. 68 (1949), pp. 1174-1179.
4. Blecher, F. H. et. al. "The L4 Coaxial System," *Bell System Tech. J.*, vol. 48 (Apr. 1969), pp. 819-1099.

Chapter 16

Considerations in Repeater Design for Analog Cable Systems

The preceding four chapters have developed an approach to the design and analysis of analog cable systems. In the relationships that have been derived, repeater performance plays an important part. The interaction of practical constraints (e.g., permissible cost, power, reliability) with the state of the device and circuit art will determine the performance that it is possible to achieve for a particular application at a particular time. It is not appropriate here to examine in any great depth the large body of theory and experience which has been built up in the field of amplifier design. Some general discussion of the mechanisms that determine performance and the interactions among them is provided in this chapter as background for the interdependence between repeater and system design.

16.1 BASIC DESIGN CONSIDERATIONS

Repeater performance has been characterized by the parameters of gain, noise figure, overload point, and nonlinear distortion coefficients. The way these parameters are determined by active devices, bias, and circuit configuration and the effect of feedback in controlling gain deviations and nonlinearity are treated in the following paragraphs.

Repeater Gain

The basic requirement on the repeater gain is that it match as closely as possible the nominal loss of the associated cable section. The

better the initial match, the better will be the ultimate transmission response of the system, or the simpler will be the required equalization plan, or both. The closeness of the match will often be determined in the last analysis by the degree of complexity which is judged tolerable from the viewpoints of manufacture, cost, and reliability. In some cases, size considerations are also important.

The circuit configuration used depends largely on the magnitudes of the gain and bandwidth required and on the kind of active devices available for the application. The repeater may consist of a single amplifier or two in tandem. In any event, major loop feedback will almost always be used in order to achieve the necessary amplifier linearity and reduce the effects of transistor aging, parameter variation (from unit to unit), and temperature.

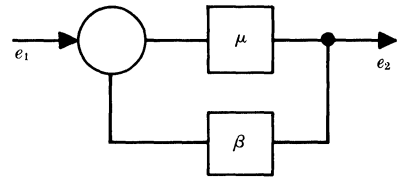


FIG. 16-1. Simplified feedback amplifier.

Equation (16-1), associated with Fig. 16-1, shows how a large amount of feedback ($|\mu\beta| >> 1$) results in an insertion gain that is approximately independent of the μ circuit:

$$\frac{e_2}{e_1} = \frac{\mu}{1-\mu\beta} \approx \frac{-1}{\beta} \qquad |\mu\beta| >> 1 \qquad (16\text{-}1)$$

As a result, variations in the active device parameters in particular have a greatly reduced impact on repeater gain. However, it is also evident from Eq. (16-1) that the reduced sensitivity to μ circuit variations does not apply to the elements making up the β circuit. Neither does it apply to any components which may be connected at the input or output of the amplifier which are, so to speak, outside the feedback loop (e.g., input and output transformers). Consequently, these parts of a repeater circuit must be treated with special care if they are not to degrade the amplifier performance excessively. Techniques used to minimize the temperature and aging effects—such as the use of low temperature coefficient components, the use of components having complementary temperature coefficients, or the use of temperature compensating elements—may be required; however, the components used in input, output, and β networks usually have a variability one or two orders of magnitude smaller than the active devices. Consequently, feedback usually results in an improvement by making the more stable passive components the controlling ones.

Repeater Noise Figure

Noise figure is discussed in detail in Chap. 8. To review briefly, the total noise figure of cascaded gain stages in a repeater is usually determined by the noise figure of the first (or lowest level) stage. The effects on noise figure of various means of coupling between the repeater input and the first stage are also discussed in Chap. 8. Generally, such coupling networks degrade the repeater noise figure by about one to four dB from that of the active devices in the amplifier string. The least degradation is usually experienced if hybrid transformer coupling is used.

It can be shown that the addition of major loop feedback has no direct effect on the noise figure of an amplifier. Noise sources near the input to such an amplifier then contribute to the noise figure independent of the amount or type of feedback. Thus, a repeater noise figure can be estimated from knowledge of the noise performance of the input transistor and the coupling network configuration.

As shown in Chap. 8, the transistor noise figure will be minimized for a particular operating condition by the selection of an optimum generator impedance. Defining the optimum generator impedance with respect to noise figure to be $R_{g(\mathrm{OPT})}$ [1].

$$R_{g(\mathrm{OPT})} = (r_b'+r_e)^2 + \left[\frac{2\alpha_0 r_e}{1-\alpha_0}(r_b' + r_e/2)\right]^{1/2} \qquad (16\text{-}2)$$

The penalty on noise figure performance due to moderate deviations from the optimum generator impedance value is fairly small in most cases. That is, the specifications of R_g within the range,

$$\frac{1}{2} R_{g(\mathrm{OPT})} \leq R_g \leq 2R_{g(\mathrm{OPT})}$$

can be expected to produce a degradation of only a few tenths of a decibel in noise figure. It is important to note, however, that the generator impedance at which the available gain of a common emitter stage is maximized will, in most cases, be very nearly the same generator impedance which will minimize the noise figure of the stage. This is a major reason that the choice of the common emitter connection for the input stage of a feedback amplifier is frequently the best choice. The relatively high gain available from the common emitter stage is then available for the overall loop gain as well as for the reduction of noise figure contributions from subsequent stages in the amplifier.

Examination of Eq. (8-45) shows that a small base resistance and large α are desirable in devices for low noise applications. It also appears that the noise figure for a given R_g can be minimized with respect to emitter current since r_e is proportional to I_e^{-1}. However, the incremental current gain of the transistor is also dependent on the magnitude of the emitter bias current, and this effect will usually cause the current at which minimum noise figure is achieved to be somewhat larger than that which could be explicitly predicted on the basis of optimizing R_g alone.

Figure 16-2 shows a typical variation in common emitter current gain as a function of emitter bias current. To achieve relative insensitivity of stage gain to the variations in I_e which result from differences in bias resistors or changes in power supply, it will usually be desirable to bias the first stage of a feedback amplifier in the "flat" region of Fig. 16-2. This will often be as important a factor in the selection of first stage bias as is the consideration of minimizing noise figure.

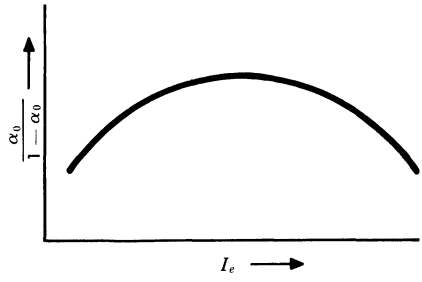


FIG. 16-2. Variation of transistor gain with emitter current.

Another factor indirectly affecting the common emitter noise figure behavior is the use of *local* feedback on the stage. This will frequently be desirable or necessary in order to achieve the desired loop gain characteristic of the feedback amplifier. It can be shown that the effective noise figure of a common emitter stage that includes a resistance, R_E, in series with the emitter lead is given by

$$n_F = 1 + \frac{r_b' + R_E}{R_g} + \frac{r_e}{2R_g} + \frac{(1-\alpha_0)\ (R_g + r_e + r_b' + R_E)^2}{2\alpha_0\, r_e\, R_g} \qquad (16\text{-}3)$$

Wherever R_E appears in Eq. (16-3), it does so in summation with r_b', and thus its effect is entirely analogous. The effect of R_E on noise figure must be considered as it assumes values which are significant compared to r_b'. It can be seen by inspecting Eq. (16-2) and by substituting $(r_b' + R_E)$ for r_b' wherever it appears that $R_{g(\mathrm{OPT})}$ will also be influenced as R_E approaches r_b'. Figure 16-3 shows the effect of R_E on minimum noise figure for a particular set of transistor parameters.

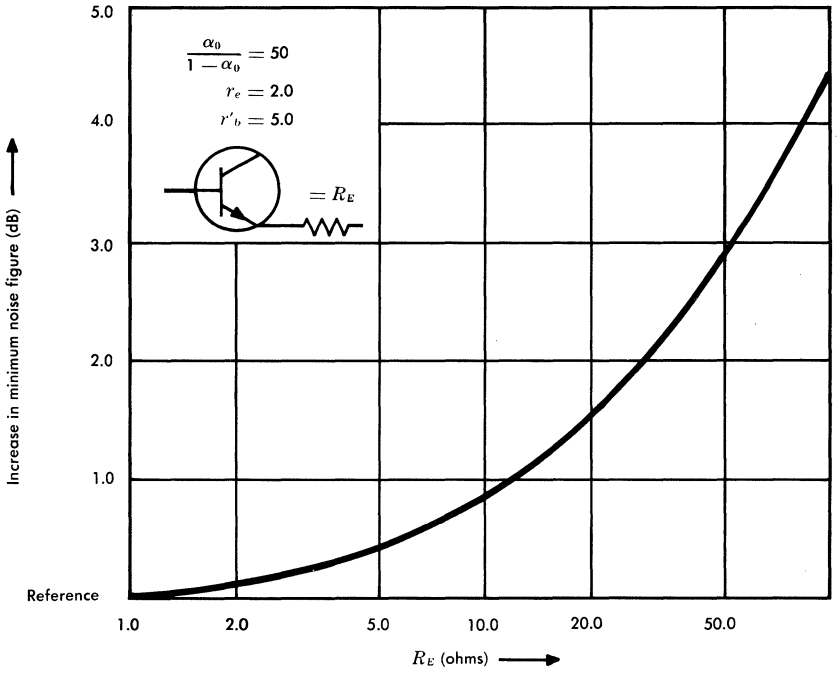


FIG. 16-3. Typical effect of emitter resistance on minimum noise figure.

Although no one solution fits all cases, the following points, in summary, will often characterize the best approach to feedback repeater design in wideband systems so far as it is determined by noise figure considerations:

1. The input stage will be common emitter.
2. There will be a minimum of local feedback on the input stage.
3. Signal coupling at the input will be via an unequal ratio hybrid transformer.

Repeater Overload

The load capacity of a repeater is described as the signal power which it can supply at its output without material effect on the linear properties of the repeater, i.e., without *overload*. The specific event which is used to define overload may vary in different applications. In all cases, overload involves excessive nonlinear behavior, and it is only the degree of excess which distinguishes the various defini-

tions. These include change in the modulation indices, M_2 and M_3, by a prescribed amount; change in repeater gain, either an increase or decrease; and damage to the repeater.

In wideband systems in which the gains of the active devices are not constant within the transmission band, overload (like the M_2 and M_3 indices) will tend to be frequency dependent. That is, the overload event will occur at different output power levels according to the signal frequency involved. Taking the case for which overload is defined as a measurable (e.g., 1/2 dB) change in the repeater gain, this can be understood by noting that both the transistor gains and overall loop feedback are going to be functions of frequency in a feedback amplifier. In particular, both will usually be less at the top of the message band than at the middle and lower frequencies. The lesser gain at high frequencies of the output transistor will require more base drive to achieve the same output power. Consequently, the output signal power at which the transistor properties change due to excessive base drive tends to be less at high frequencies than at lower frequencies. Furthermore, due to the smaller amount of negative loop feedback at high frequencies, the changes in the properties of the transistors in the μ circuit will not be suppressed as well as they are at lower frequencies. Both of these factors result in an overload point which tends to be frequency dependent in wideband systems.

This may seem to add significant complication to the overload analysis of analog cable systems since now P_R as well as C is in general a function of frequency. However, this is not usually a significant factor for two reasons. First, the frequency dependence in some cases is slight enough to warrant an approximation as a constant. Secondly, the signal shaping used to equalize the signal-to-noise ratio in all message channels always results in concentration of signal power over a fraction of the transmission band near the upper band edge. Thus, even if the overload point varies significantly over the whole message band, it is likely that the variation within that portion at which the signal power is concentrated will not be great.

The principal factors under the designer's control, which affect the load capacity of the feedback amplifier, are the same as those later discussed in detail for nonlinear distortion, namely bias conditions, load impedance, and manner of coupling to the output load. As is the case for nonlinear distortion, there is an interaction between bias and load impedance in the sense that the best d-c operating point for the power stage will be different for different load impedances.

In wideband modulation limited systems it usually turns out that the steps taken to achieve the stringent linearity requirements also result in satisfactory load capacity for the repeater. For the L4 system these steps resulted in $M_3 = -105$ dB (top channel) and $P_R = 21$ dBm, while for L5 the corresponding numbers are -115 dB and 23 dBm. (L4 is a 3600 channel, 17.5 MHz system; L5, at this writing, is expected to be a 9000 channel, 50 MHz system.)

Selected from the variety of quantitative definitions of overload, the most common ones are listed below and are discussed using Fig. 16-4 as an example. That figure shows the variation in M_3 as a function of signal power, and the change in the linear gain as a function of signal power for a wideband transistor feedback amplifier.

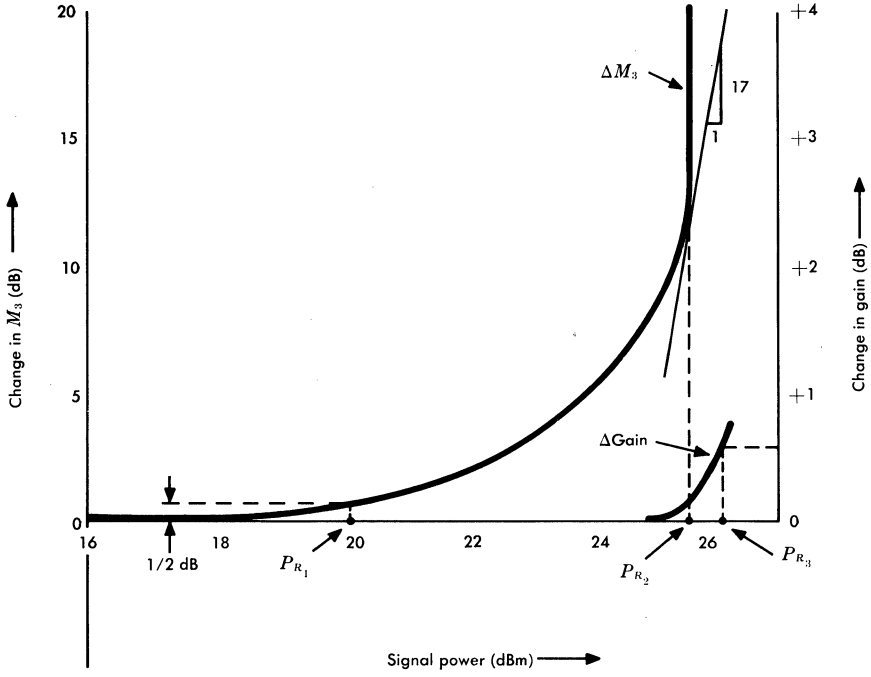


FIG. 16-4. Change in M_3 and gain versus signal power.

Case 1:

The overload point is that value of power at the output at which the amplifier M_3 has increased by 1/2 dB compared to the

low power value. (This may be approximately the power level at which a 1-dB increase in a fundamental signal at the input results in a 3.5-dB increase in third harmonic distortion product at the output.) This definition is appropriate for use in a modulation limited system. It is conservative in the sense that exceeding this power by a small amount would cause only slight performance impairment. A relatively small overload margin would be appropriate for use in conjunction with an overload point defined in this way. Relatively large variation in overload from repeater to repeater might be expected when this definition is used. This definition of overload point has been used in the L4 development and will apparently be used in the L5 development. Referring to Fig. 16-4,

$$P_{R1} = 20 \text{ dBm}$$

Case 2:

The overload point of the amplifier is that value of absolute power level at the output at which the absolute power level of the third harmonic increases by 20 dB when the input signal to the amplifier is increased by 1 dB. This in effect defines the overload point as that at which M_3 changes by 17 dB for a 1-dB increment in input signal power. The resulting overload point corresponds to gross overload and a very serious signal impairment. Presumably, the increased impairment for this power is deemed acceptable because, due to misalignment, only a few repeaters will be subjected to the maximum load, and because of the statistics of the broadband signal, even these only rarely. A more generous overload margin should be used in this case. This kind of definition has the advantage that there will be relatively small variations in the load capacity of one repeater as compared to another, and it is the definition recommended by the CCITT.* Referring to Fig. 16-4,

$$P_{R2} = 25.5 \text{ dBm}$$

Case 3:

The overload point is that value of the power at which the linear gain of the amplifier changes by 1/2 dB. As with Case 2, this corresponds to a point where gross overload is occurring. It is most appropriate when modulation is not a major consideration.

*Recommendation G.222, II[d] Plenary Assembly, New Delhi, Dec. 8-16, 1960.

Consequently, it is sometimes used for short-haul system repeaters and multiplex terminal amplifiers. From Fig. 16-4,

$$P_{R_3} = 26 \text{ dBm}$$

Thus, depending on the specific definition of the overload event, the overload point of this amplifier lies in the 6-dB range from 20 dBm to 26 dBm. The relatively large discrepancy between overload as specified by definition 1 and the latter two definitions may not have as large an effect on system design as might at first appear because of the differences in the associated margins. In other words, $P_{R_1} - A_{P_1}$ might differ much less from $P_{R_2} - A_{P_2}$ and $P_{R_3} - A_{P_3}$ than P_{R_1} differs from P_{R_2} and P_{R_3}.

Nonlinear Distortion

The linearity of a solid state feedback amplifier depends primarily on the inherent linearity of the transistors used and the nature (local or loop) and quantity of negative feedback applied. The linearity of the transistor when imbedded in a circuit is in turn largely dependent on the generator and load impedances and the selected d-c bias conditions. The overall linearity of several cascaded stages will depend on the relative phasing and amplitude of the distortion products originating in each stage and the allocation of gain among the stages. Where the tolerable level of overall amplifier distortion products is in the range required by wideband coaxial systems, the distortion originating in even the lowest level stages must be controlled. Furthermore, attention must be paid to components such as surge protection diodes and ferrite core inductors to insure that they do not contribute to repeater nonlinearity.

Among the significant nonlinearities in a transistor are those associated with the emitter-base diode characteristic, the current gain, avalanche multiplication effects, and the collector capacitance. Some of these tend to be more or less current dependent while others depend essentially on the voltages at which the device is operated.

The emitter-base diode gives rise to what is usually called the exponential nonlinearity of the transistor which results from the relationship [Eq. (16-4)] between the current in a semiconductor junction and the voltage at the junction.

$$I_e = A \exp \frac{\lambda q v}{kT} + B \tag{16-4}$$

where

k = Boltzmann's constant

q = electron change in coulombs

T = absolute temperature

λ, A, B = constants which depend on transistor parameters.

Equation (16-4) can be expressed by a Taylor series expansion of the form [2]

$$I_e = k_1 V + k_2 V^2 + k_3 V^3 \tag{16-5}$$

where

$$k_1 = \frac{1}{r_e}$$

$$k_2 = \frac{1}{2r_e^2} \cdot \frac{1}{I_e}$$

$$k_3 = \frac{1}{6r_e^3} \cdot \frac{1}{I_e^2}$$

and

$$r_e = \frac{\lambda kT}{qI_e}$$

Coefficients in a power series expansion such as this can readily be related to device M coefficients as discussed in Chap. 10. A particular measured characteristic showing the relationship between emitter current and base-emitter voltage is shown in Fig. 16-5. The data were taken on a transistor of the type used in power stages of L4 amplifiers.

The nonlinear relationship between collector and emitter current depends both on the value of emitter current and the collector-to-base voltage. The current dependent nonlinearity (so-called h_{FE} nonlinearity) has been found experimentally to obey the relation of Eq. (16-6) [2]

$$h_{FE} = \frac{h_{FE_{(\max)}}}{1 + b \log^2 (I_c/I_{c_{\max}})} \tag{16-6}$$

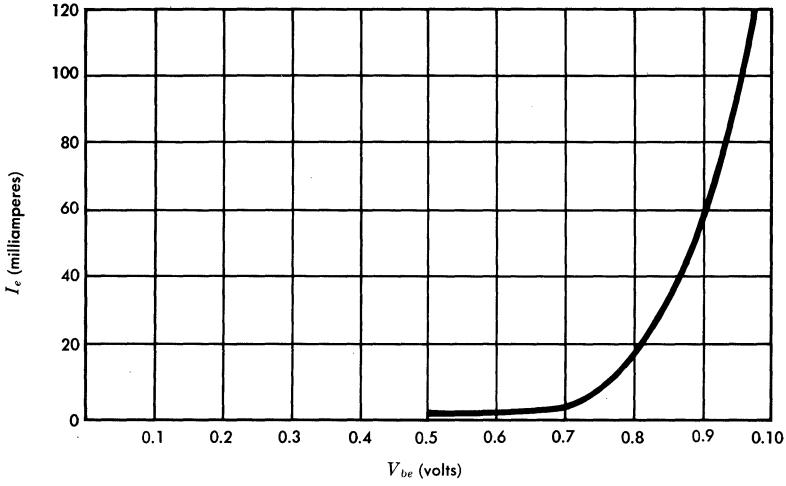


FIG. 16-5. Typical exponential nonlinearity.

where

$$h_{FE} = \frac{I_c}{I_B}$$

$h_{FE_{\max}}$ is the maximum of h_{FE}

$I_{c_{\max}}$ is the value of I_c at which $h_{FE_{\max}}$ occurs

b is a constant

The voltage dependent nonlinearity (so-called avalanche nonlinearity) results from the avalanche multiplication which occurs at high collector-base voltage. The combined effect of the h_{FE} and avalanche nonlinearities on the ratio I_c/I_e is shown in Eq. (16-7) where the second term is the avalanche multiplication factor [2].

$$\frac{I_c}{I_e} = \left[\frac{h_{FE_{\max}}}{1 + h_{FE_{\max}} + b \log^2 (I_c/I_{c_{\max}})}\right] \left[\frac{1}{(1 - V_{CB}/V_{CBO})^n}\right] \quad (16\text{-}7)$$

where V_{CBO} is the collector-base breakdown voltage (emitter open), and n is a constant dependent on the transistor material and is usually determined empirically.

The collector capacitance nonlinearity is the result of variations in the depletion layer of the reverse-biased collector-base junction as a function of collector-base voltage. The junction capacitance, C_c, is given by [2]

$$C_c = \frac{k}{(V_{CB})^{1/3}} \tag{16-8}$$

In Eqs. (16-3), (16-6), and (16-7) it is evident that the nonlinear distortion of a particular transistor is closely related to the bias conditions established for the device. These nonlinear properties can be superimposed on a linear equivalent circuit for the transistor, and the resulting performance can be analyzed for different circuit configurations and interconnections. For the present discussion, the most important results of this kind of analysis would relate to the selection of generator and load impedances for best performance. In a qualitative way, it can be seen readily that driving a common emitter stage from a high impedance (current source) will tend to reduce the effect of the nonlinearity in r_e (the exponential nonlinearity). Similarly, loading the transistor with a low impedance will tend to reduce the effect of the collector capacitance nonlinearity. In general, there will exist intermediate values of generator and load impedances which will achieve an optimum balance among the several sources of distortion so as to minimize the aggregate distortion originating in the transistor when embedded in a particular circuit. It is not uncommon in the design of such complex active circuits that the many other factors which must be considered (such as feedback shaping, realistic transformer designs, etc.) will not always permit the exactly optimum impedances to be realized.

Figure 16-2 shows a typical variation of common emitter current gain with emitter current. Operating at a quiescent emitter current such that the signal-current swings result in a minimum change in current gain suggests that operating in the flat region of the characteristic will tend to reduce the effects of the h_{FE} nonlinearity. Equation (16-5) suggests that increasing the emitter current will reduce the effects of the exponential nonlinearity. To the extent that these two nonlinearities dominate the current-dependent transistor distortion, there should exist an operating current at which the composite effect is minimized, and the optimum current will

probably fall somewhere within the flat region of Fig. 16-2. Increases beyond that current, though they presumably reduce the exponential nonlinearity, will be accompanied by a net degradation due to increased effects of the h_{FE} nonlinearity.

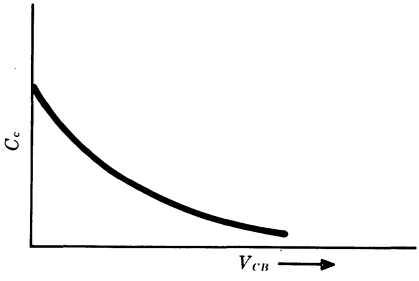


FIG. 16-6. Variation of collector capacitance with collector-base voltage.

It is possible to see qualitatively that there also exist quiescent voltages which will tend to minimize the effects of the voltage-dependent nonlinearities. Figure 16-6 shows a typical relationship between the collector capacitance, C_c, and the collector-base voltage, V_{CB}. For a particular signal swing, the effect of this nonlinearity is lessened as the quiescent V_{CB} is increased. On the other hand, it is apparent in Eq. (16-7) that the avalanche multiplication factor is minimized for small values of V_{CB}. There exists a bias condition short of the collector-base breakdown voltage which will minimize the combined effect of these nonlinearities.

Not all of the nonlinearities present in the transistor were included in the above discussion, and those that were may not be as neatly separable and independent as may have been implied. The importance of careful attention to the bias point selection cannot be overemphasized, however, and in most cases it is the most powerful single tool at the circuit designer's command for this aspect of repeater realization.

When the bias points and generator and load impedances for a transistor have been selected (presumably in such a way as to optimize transistor performance), the nonlinear distortion of a particular transistor is determined. It will be assumed for the moment that, despite the multiplicity of sources of nonlinearity, the nonlinear behavior of a transistor can be described at least approximately by an M_2 and an M_3 coefficient referred to the collector, which for clarity will be referred to as M_{2t} and M_{3t}. What then will be the M_2 and M_3 coefficients for the whole repeater? For simplicity, it is assumed that intermodulation occurring in the output stage

dominates. Consistent with Chap. 13, M_2 and M_3 will be referred to the repeater output.

In general there will be some loss, $-Q$ dB, between the collector and the repeater terminals which feed signal power into the transmission medium. From the discussion in Chap. 13, it should be clear that the effect of Q on fundamentals and harmonics leads to

$$M_2 = M_{2t} + Q - Q - Q = M_{2t} - Q$$

$$M_3 = M_{3t} + Q - Q - Q - Q = M_{3t} - 2Q$$

From these equations, it is evident that the lower the loss between collector and repeater output, the better the repeater performance. A brute force termination results in a loss, $Q = -3$, and degrades the repeater modulation coefficient, M_2, by 3 dB and M_3 by 6 dB. Just as the use of a hybrid transformer at the input to the repeater reduced the input coupling loss and therefore the repeater noise figure, so the use of the same coupling technique at the output will improve repeater nonlinearity. A coupling loss of 1 dB, i.e., $Q = -1$, is achievable with a hybrid output transformer, and this would result in a 2-dB improvement in M_2 and a 4-dB improvement in M_3 as compared to the brute force terminations. Although this discussion shows that the direct effect of using a hybrid transformer on repeater coefficients is beneficial, in a particular case it may turn out that the indirect effects are such that the net result is a degradation of performance. It has been mentioned previously that the transistor nonlinearity depends in part on its load impedance. The use of a hybrid transformer could result in a load impedance sufficiently different from optimum that the transistor linearity is degraded. Furthermore, an output hybrid transformer is in the feedback loop. As discussed later, the amount of feedback achievable in an amplifier with adequate stability margins depends on the transmission characteristic at frequencies many times that of the highest message frequency. A hybrid transformer at these frequencies can introduce large amounts of phase shift with the result that the amount of feedback possible at message frequencies is appreciably reduced. For these reasons, the more complicated output coupling network is not, in every case, to be preferred to the brute force termination.

The previous discussion has pointed out the effect of output coupling loss on the repeater modulation coefficients. The major change in going from device to repeater coefficients is due to the effect of

feedback. It can be shown that the nonlinear distortion of an amplifier with negative feedback is less than the distortion in the same amplifier without feedback by about the magnitude of the feedback factor in dB [20 log $(1-\mu\beta)$]. In the case of the effect of feedback on third order modulation, there is a further complication. The fedback second order products can be further combined with fundamentals by second order distortion in the device to result in third order products. Thus, even if a transistor had no third order distortion at all, the device used in a feedback circuit would have third order products at its output. This increased third order degradation is usually much less than the improvement introduced by the feedback in the first place. These considerations on the effect of feedback can be incorporated in the previous results relating device and repeater coefficients, yielding:

$$M_2 = M_{2t} - Q - F$$

$$M_3 = M_{3t} - 2Q - F + K_F$$

where $F = 20 \log |1-\mu\beta|$ and $K_F =$ the third order degradation in dB resulting from the feedback of second order products. A typical value of K_F could be expected to fall somewhere between 3 and 7 dB, and a value somewhere in this range can be used for early design estimates. A more precise determination can be obtained only by actual measurements on a model incorporating the device and circuit configuration under consideration.

It has been indicated in preceding chapters that the repeater modulation indices, M_2 and M_3, will not be constant with respect to frequency in a physical realization. The several nonlinear mechanisms within each transistor are clearly frequency-dependent and result in transistor behavior as a whole which is frequency dependent. In addition, the multiple source nature of the transistor nonlinearity produces at some frequencies cancellation effects which tend to be particularly sensitive to changes in the circuit environment, such as temperature or supply voltage. These cancellation effects will also tend to differ significantly from transistor to transistor within the same family and cannot usually be relied upon in the realization of amplifier objectives. Interaction among the nonlinearities in the various transistors making up a repeater also can result in cancellation and enhancement causing deviations from the simple power series behavior. The frequency shaping of the negative feedback will generally further increase the overall frequency dependence of the repeater M_2 and M_3.

These deviations from the power-series model heretofore assumed change the relationship among the several types of intermodulation products. As a specific example of what this means in practice, a 15-MHz 3α product formed from cubic distortion of a 5-MHz fundamental signal will not in general be different by 15.6 dB from a 15-MHz $\alpha + \beta - \gamma$ product formed from 14-MHz, 17-MHz, and 16-MHz fundamental signals, as is predicted by the power series representation of the amplifier transfer function. For the two conditions, there are essentially different mechanisms at work which taken together are reflected in the frequency dependence of the distortion indices. For example, the $\alpha + \beta - \gamma$ product will include an interaction component formed by the γ fundamental and the $\alpha + \beta$ product. The $\alpha + \beta$ product falls at about 30 MHz, which might be out-of-band and at which there may be relatively little feedback. Furthermore, if a power stage with a β cutoff frequency of 5 MHz is assumed, the required signal drive to the output transistor for the same output signal power is greater at the 15-MHz fundamentals, and the resultant distortion from the nonlinear mechanisms at the transistor input is correspondingly increased. These and like phenomena will tend to produce predictions of repeater distortion at 15 MHz, which are in most cases optimistic if based on 3α-type measurements. The cancellation effects mentioned above can cause exceptions in the case of any particular measurement and can be averaged out by making sufficiently numerous measurements on different repeaters and at different frequencies.

The dominant type of modulation distortion in wideband systems tends to be of the $\alpha + \beta - \gamma$ type at high frequencies, where α, β, and γ are high frequencies, and of the $\alpha - \beta$ type at low frequencies, where α and β are high frequencies. Recent practice in the characterization of repeater M_2 and M_3 reflects these tendencies. At high frequencies M_3 is calculated by adding 15.6 dB to the measured value of an $\alpha + \beta - \gamma$ product, for α, β, and γ in the high-frequency region. At low frequencies, M_2 is calculated by adding 6 dB to the measured value of a low-frequency $\alpha - \beta$ product, for α and β in the high-frequency region. In fact, M_2 and M_3 throughout the message band have come to be characterized on the basis of intermodulation measurements rather than harmonic distortion measurements.

Although it is difficult to separate in all cases the distortion due to the different kinds of intermodulation (i.e., $\alpha + \beta$, $\alpha - \beta$, $\alpha + \beta - \gamma$, etc.), noise loading procedures can be an effective way to characterize

these complex and interacting mechanisms. [If the laws of addition assumed for the different products (per Chap. 13) were not different, there would be no reason to use any other characterization method so long as the noise loading method could be instrumented properly.]

Whatever technique is used, it is ultimately possible to specify either directly or indirectly an effective M_2 and M_3 required for the systems equations of preceding chapters. The frequency dependence of M_2 and M_3, as well as C and N_F, makes the calculations cumbersome and they are usually carried out with computer assistance.

Feedback

In principle, the application of feedback to a circuit involves adding a portion of the output of the circuit to the input. The advantages to be obtained from this type of design (increased linearity and reduced sensitivity to parameter variations) have been briefly discussed. This design approach, however, introduces additional complications. Some mention of the mechanisms that limit the amount of achievable feedback, and therefore also limit the performance improvement, is in order.

From the equations previously presented, it can be seen that linearity and immunity to changes in μ improve indefinitely as the amount of feedback [$F = 20 \log (1-\mu\beta)$] is increased. This would allow, in theory, complete elimination of the modulation noise from analog systems, and all systems could be operated at the higher transmission levels permitted in the overload-limited case. This is not possible in practice; there is an upper limit to the feedback that can be achieved in an amplifier. To understand the reason for this limitation, the expression for closed-loop amplifier gain [Eq. (16-1)] is repeated.

$$\frac{e_2}{e_1} = \frac{\mu}{1-\mu\beta} = \frac{1}{\beta}\frac{\mu\beta}{1-\mu\beta}$$

As was previously stated, the insertion gain is approximately equal to $-1/\beta$. The total net gain around the feedback loops is defined to be $\mu\beta$, where μ represents the total gain provided by the active devices, and β is the loss in the portion of the loop connecting the output back to the input. From these facts it can be seen that:

$$\text{Loop gain} = 20 \log \mu\beta = 20 \log \mu + 20 \log \beta \approx 20 \log \mu - G_R$$

and consequently

$$20 \log \mu \approx 20 \log \mu\beta + G_R$$

This means that the sum of the loop gain and the insertion gain cannot exceed the total gain available from the devices (and the associated interstage networks). Thus, in no case is it possible to get loop gains in excess of the difference between the total gain, μ, and the desired insertion gain. When this case is the controlling one, an amplifier is said to be *gain limited.*

Most broadband amplifiers, however, are not gain limited, because it is the need for adequate stability margins that is controlling. For $\mu\beta = 1$ in the gain expression $\mu/1 - \mu\beta$, the denominator becomes zero. If there is any frequency, in- or out-of-band, where this condition occurs, the amplifier becomes unstable and oscillates at that frequency. If it were possible to hold $|\mu\beta| >> 1$ for all frequencies, this would not be a problem. However, any active device has some frequency above which its gain begins to decrease more or less monotonically. The rate of cutoff may be further enhanced by parasites in the amplifier circuit. There will therefore always be a frequency for which $|\mu\beta| = 1$. This is still not bad if the phase of $\mu\beta$ is not simultaneously zero. Ideally, maximum stability margins would result if that phase were 180 degrees so that the gain expression could be written $\mu/1 + |\mu\beta|$. In the transmission band the phase can often be controlled to approach this (hence the name negative feedback), but out-of-band the phase, like the gain, will change. This is due to the phase shift inherently associated with any gain-frequency characteristic, such as the gain cutoff mentioned previously. Furthermore, for very high frequencies the propagation time around the feedback loop contributes additional phase shift. Since loop gain will equal unity at some frequency and the phase shift will equal zero at some frequencies, the only way to guarantee stability is to make sure that the loop gain is below unity before the phase equals zero for the first time. To allow for device aging and variation in active device characteristics, the phase should be some reasonable number of degrees away from zero when the gain equals one, and the loop transmission should exhibit several dB of loss when the phase goes through zero. These are known as the phase and gain margins, respectively. The maintenance of adequate phase and gain margins sets an upper limit on the achievable in-band feedback. When this limit is lower than that set solely by gain considerations, the amplifier is said to be *stability limited.* The theory

relating device and circuit parameters, feedback, and top frequency is described elsewhere [3, 4, 5, 6]. For given device and circuit parameters and commonly used stability margins, the maximum achievable feedback in a stability limited case decreases about 10 dB for a doubling of the system top frequency.

It is of interest to compare the gain limited design with the stability limited design. For the gain limited case it is the value of circuit parameters at in-band frequencies that is controlling. A reduction in repeater spacing decreases the insertion gain and so permits an increase in achievable feedback. In the stability limited case it is the value of circuit parameters one or two decades above the transmission band that is controlling. A change in repeater spacing, in this case, usually does not permit an increase in feedback. In a stability limited design it is important to hold feedback loop physical size to a minimum in order to limit the additional phase shift due to propagation time. Furthermore, phase shift associated with the more complex terminating networks, desirable from the point of view of controlling noise figure and intermodulation, complicates the problem of achieving adequate amounts of feedback while maintaining sufficient stability margins. Control of the out-of-band loop gain cutoff must, of course, be done at the same time as shaping the in-band β characteristics in order to achieve the desired insertion gain shape. It is the interaction of all these requirements, in addition to others only slightly less important, that makes designing a broadband high performance analog system repeater a challenging task.

16.2 TANDEM AMPLIFIER REPEATERS

One of the first questions usually answered by the repeater designer is whether to use one or two amplifiers in the repeater realization. The single-amplifier approach will require about twice the closed loop gain (in dB) required by each amplifier of the two-amplifier approach. This, for comparable amounts of loop feedback, requires more gain and very likely more transistors in the μ path. Increasing the number of transistors will at least make it more difficult to achieve stable feedback transmission due to the higher out-of-band gain cutoff rate which results; or it may make it impossible to stabilize the feedback transmission with satisfactory margins. Meeting the noise figure and linearity objectives for the repeater in a single amplifier may be possible only with hybrid transformers at

both input and output, which would further increase the difficulty of realizing stable feedback transmission with a particular class of available transistors.

If all of the system objectives can be satisfied by a single-amplifier design while at the same time satisfactory margins are achieved in the feedback transmission, this approach would probably be used. Where the overall repeater objectives and stable feedback transmission can be achieved only by splitting the gain objective between two amplifiers and by concentrating the noise figure effort on the front end amplifier and the linearity effort on the output amplifier, the two-amplifier repeater configuration results. In a repeater configuration of two amplifiers, the interconnection between the amplifiers provides an advantageous place to insert the line build-out (LBO) networks used to permit shorter-than-nominal repeater spacing. Since the mid-repeater point is neither the highest nor the lowest TLP at the repeater station, system noise advantages can result by performing the LBO function there.

The repeaters designed for recent wideband coaxial systems have often involved two amplifiers. In a two-amplifier repeater, the division of gain between the two amplifiers is one of the freedoms open to the designer which can somewhat ease the realization of overall repeater objectives. This is normally one of the first decisions made in the design of such a repeater.

The division of the total gain between the two amplifiers making up a two-amplifier repeater interacts considerably with overall repeater noise figure and modulation distortion performance. When a two-amplifier configuration has been selected, it is advantageous in most cases to concentrate noise figure effort in the amplifier at the input section of the repeater, where the signal levels are a minimum (often called a preamplifier). By the same token, the concern for low distortion can be more or less concentrated in the amplifier forming the output section of the repeater, where the signal levels are maximum (often called a power amplifier). This is a difference in emphasis only, and some attention will have to be paid in the actual design to noise figure in the power amplifier as well as to nonlinear distortion in the preamplifier.

Assume that the repeater consists of two amplifiers (Fig. 16-7) where $G(f)$ is the gain of the preamplifier in dB; $G_2(f)$ is the gain of the power amplifier in dB; and $G_1(f) + G_2(f)$ corresponds to the loss of the associated cable section and is therefore not constant with respect to frequency. Considering the interaction between the am-

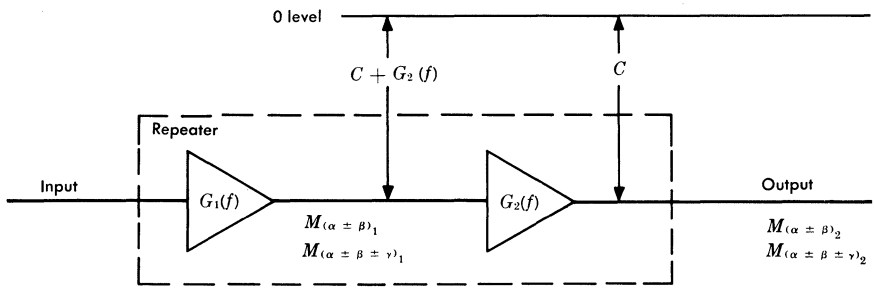


FIG. 16-7. Two-amplifier repeater configuration.

plifiers as it affects the overall modulation distortion of the repeater, define

$M_{(\alpha\pm\beta)_1}$: power of the $\alpha\pm\beta$ products at preamplifier output given 0 dBm fundamental power there at frequencies α and β;

$M_{(\alpha\pm\beta)_2}$: power of the $\alpha\pm\beta$ products at power-amplifier output given 0 dBm fundamental power there at frequencies α and β.

The power-amplifier contribution to $\alpha\pm\beta$ type modulation distortion at zero level, with *0 dBm0* fundamental signals at the amplifier output, is

$$M_{(\alpha\pm\beta)_2} - C \qquad \text{dBm} \tag{16-9}$$

Similarly, with *0 dBm0* signal power at fundamental frequencies α and β at the preamplifier output, the preamplifier contribution at zero level to $\alpha\pm\beta$ type distortion is

$$M_{(\alpha\pm\beta)_1} - \mathrm{C} + G_2(\alpha\pm\beta) - G_2(\alpha) - G_2(\beta) \tag{16-10}$$

If G_2 is flat, Eq. (16-10) becomes

$$M_{(\alpha\pm\beta)_1} - C - G_2 \tag{16-11}$$

A comparison of Eqs. (16-11) and (16-9) shows that for equal modulation indices, $M_{(\alpha\pm\beta)_1} = M_{(\alpha\pm\beta)_2}$, the contribution of the preamplifier to zero level second order distortion is less than that of the power amplifier by the gain of the power amplifier (G_2). If the gain of the power amplifier is shaped, as will generally be the

case, the corresponding difference is given in Eq. (16-10) by $G_2(\alpha) + G_2(\beta) - G_2(\alpha \pm \beta)$. Thus it is desirable that the power amplifier in such a case have relatively high gain at the fundamental frequencies and relatively low gain at the product frequencies.

A similar development for third order modulation distortion would show that the preamplifier contribution at a zero level point with 0 dBm0 fundamental power at α, β, and γ is less than the power-amplifier contribution at zero level by a factor

$$G_2(\alpha) + G_2(\beta) + G_2(\gamma) - G_2(\alpha \pm \beta \pm \gamma)$$

As with the second order distortion effects, it is desirable to keep the power-amplifier gain high at the fundamental frequencies and low at the product frequencies in order to minimize the preamplifier effect on overall repeater distortion.

To this point, C has been treated as a constant, independent of frequency. The advantages of shaping the signal level at the repeater output are discussed in Chap. 13 where it is shown that overall signal-to-noise ratio or overload advantages can be achieved by appropriately shaping the signal levels with respect to frequency at the repeater output. Because of the shape of the cable loss with frequency, the resultant signal shaping $C(f)$ will involve a higher transmission level (small C) at the frequencies near the top of the message spectrum than at the lower frequencies.

As a partial consequence of this signal shaping, the dominant source of modulation distortion in wideband coaxial systems will often be as follows:

At low message frequencies: products of the $\alpha - \beta$ type where α and β are both near the top of the band.

At high message frequencies: products of the $\alpha + \beta - \gamma$ type where α, β, and γ are all near the top of the band.

At the high message frequencies, the product frequency and the fundamental frequencies are in the same part of the spectrum. Achieving significantly different gain at the product and fundamental frequencies is thus not possible. Nevertheless, making the power-amplifier gain relatively high at the top of the message band will tend to lessen the effect of the preamplifier on the dominant third

order products. Regarding second order distortion, high power-amplifier gain at high frequencies and low power-amplifier gain at low frequencies will lessen the effect of the preamplifier on the dominant second order products.

The division of gain between the two amplifiers will also determine the extent to which each of the amplifiers will contribute to the overall noise figure behavior of the repeater. As can be seen from Eq. (8-23), the preamplifier should maintain a significant gain at all frequencies in order to control the noise figure of the repeater.

The approach which will minimize the effect of the preamplifier on overall modulation distortion and the effect of the power amplifier on overall noise figure can be summarized as follows:

1. Third order modulation noise considerations suggest high power-amplifier gain at the high message frequencies.
2. Second order modulation noise considerations suggest high power-amplifier gain at high frequencies and low power-amplifier gain at low frequencies.
3. Noise figure considerations suggest high preamplifier gain at all frequencies.

It is interesting to note that the noise figure and second order considerations produce entirely consistent requirements at the low message frequencies, namely, high preamplifier and thus low power-amplifier gain. At the higher message frequencies, where the total repeater gain required will often be fairly large (see Fig. 16-8), it will usually be possible to satisfy the requirement for high gain simultaneously in both amplifiers by something approaching an even split in the overall gain requirement.

For a 1 to 60 MHz system using 3/8-inch coaxial cable, the required gain is that of Fig. 16-8 for a repeater spacing of one mile. Using the guidelines developed above, the total gain required at 1 MHz will be allocated to the preamplifier, while the 60-MHz gain will be split evenly. Allowing for a smooth transition between these frequencies, an overall division of gain such as that of Fig. 16-9(a) results.

Since the total gain required of the repeater is so small at the lower frequencies, it is impossible to achieve negligible degradation of the repeater noise figure by the power amplifier at these frequencies. Assuming a 5-dB preamplifier noise figure and a 10-dB power-amplifier noise figure, the resultant repeater noise figure is that shown in Fig. 16-9(b), where the gain division is that shown

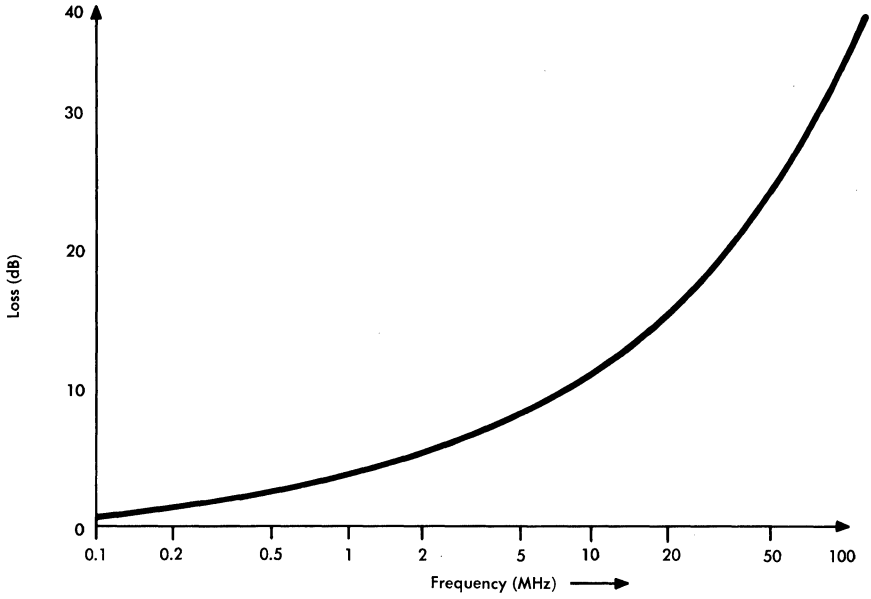
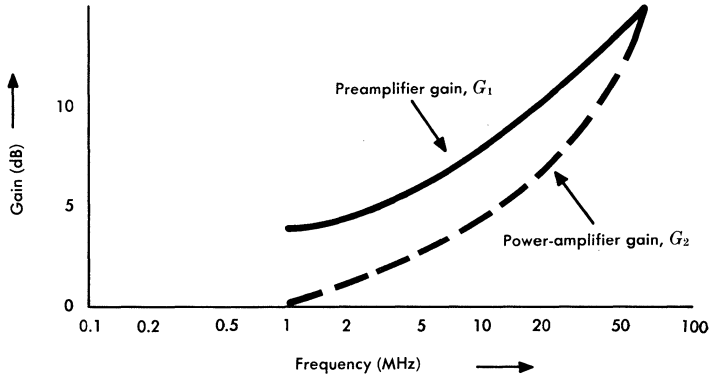


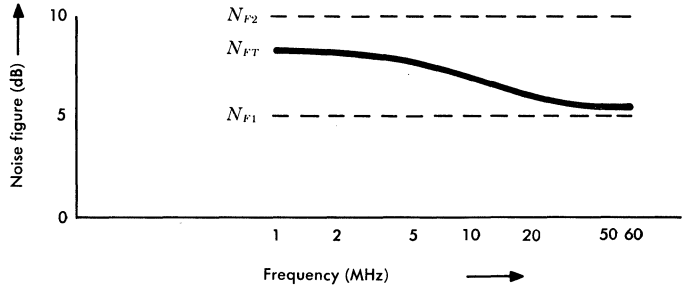
FIG. 16-8. Loss of one mile of 3/8-inch coaxial cable.

in Fig. 16-9(a). The degradation of noise figure due to the power amplifier under these conditions is less than 1 dB above 20 MHz, and about 3 dB in the 1 to 2 MHz region. The shape of the repeater noise figure as a function of frequency [Fig. 16-9(b)] is one of the factors which will determine the optimum signal shaping $C(f)$ in terms of overall system signal-to-noise ratio. The cable loss, Fig. 16-8, is another of the factors. For completeness in this respect, the final factor involved in establishing the optimum signal level shaping with frequency is the dependence of the repeater M_2 and M_3 coefficients on frequency.

Another significant source of third order modulation distortion not explicitly mentioned previously results from an interaction between the two amplifiers. It is not unlike what occurs when the feedback loop of a feedback amplifier is closed, and the resultant third order distortion includes the effect of feedback second order products modulating with fundamental signals to produce increased third order distortion. In the two-amplifier repeater, the corresponding effect is the modulation in the power amplifier of an $\alpha + \beta$ product originating in the preamplifier with a fundamental signal at γ to



(a) Division of gain



(b) Overall noise figure

FIG. 16-9. Effect of gain division on repeater noise figure in a two-amplifier configuration.

produce an $\alpha + \beta - \gamma$ interaction product. In cases where α and β are near the top of the message band, the second order product at $\alpha + \beta$ falls at near twice the top message frequency where there is usually very little loop feedback. Such products tend to be relatively high in power, and it may not be unusual for the resultant interaction product to dominate the overall repeater performance for $\alpha + \beta - \gamma$ type products. A technique that has been used at times to combat this mechanism is the insertion between the amplifiers of a low-pass filter to attenuate the out-of-band products generated within the preamplifier. Such a filter, cutting off just above the band, will naturally contribute to phase distortion within the band and the effect of such distortion on the signals to be transmitted over the system must be considered before such a technique is implemented.

16.3 REVIEW OF SYSTEM AND REPEATER DESIGN CONSIDERATIONS

This chapter is by no means exhaustive with respect to repeater or amplifier design. It may, however, serve to raise in the mind of the designer some of the important questions which greatly affect the realization of the system as a whole.

The tentative layout of an analog cable system results from an approach which is both analytical and empirical. The factors which are fixed in a particular design and those which are left variable differ from system to system. Several different attempts at a particular system design may be made during the initial phase. The results of carefully considered estimates and exploratory design efforts will usually point the way during this stage. Ultimately, sufficient information becomes available to permit a determination of the limiting repeater parameters and the required repeater spacing.

As the repeater design progresses and the basic system layout is established, equalization requirements begin to crystallize. In the case of buried cables, the effect of earth temperature on cable loss can be calculated, leading to the required spacing of the temperature regulating repeaters. In submarine cable systems this effect is quite small, and as a result regulating repeaters are not required. As the specific transmission characteristics of the repeaters become known, the design deviation equalizers can be specified as to transmission response and location in the system. Statistical studies of the anticipated manufactured product and knowledge of temperature and aging effects within the repeaters will lead to the specification of the more general class of adjustable equalizers required to achieve the overall transmission response. The questions of overload and acceptable noise penalties will further determine the equalizer layout and whether single- or double-ended equalization is to be used.

Throughout the system design, particularly during the early stage, there must be continual interaction among the system layout, system requirements, and repeater design objectives. The combination of these finally settled upon will usually represent the best compromises possible at the time which will satisfy the general system objectives such as channel capacity and signal-to-noise ratio.

REFERENCES

1. Neilsen, E. G. "Behavior of Noise Figure in Junction Transistors," *Proc. IRE*, vol. 45 (July 1957).
2. Narayanan, S. "Transistor Distortion Analysis Using Volterra Series Representation," *Bell System Tech. J.*, vol. 46 (May-June 1967).
3. Blecher, F. H. "Design Principles for Single Loop Feedback Amplifiers," *IRE Transactions on Circuit Theory*, vol. CT-4 (Sept. 1957), pp. 145-156.
4. Hakim, S. S. *Junction Transistor Circuit Analysis* (New York: John Wiley and Sons, 1962).
5. Bode, H. W. *Network Analysis and Feedback Amplifier Design* (New York: D. Van Nostrand Company, 1945).
6. Lynch, W. A. "The Stability Problem in Feedback Amplifiers," *Proc. IRE* (Sept. 1951), pp. 1000-1008.

Chapter 17

Introduction to Analog Microwave Radio Systems

This chapter and the following six chapters describe some of the important factors that influence the design, installation, and application of microwave radio systems, and demonstrate some of the methods used to optimize the design for a particular application. The ways in which these systems are similar to and different from wire transmission systems are discussed.

For the most part, the baseband load under consideration is frequency division multiplexed telephone message channels. Television, wideband data, and special service signals are also carried on these radio systems, but are not treated here.

In this chapter microwave systems are discussed in a general way to point out the nature of the problems to be considered. The next five chapters deal with the topics of radio propagation, frequency modulation theory, distortion mechanisms, and radio channel allocation. Finally, Chap. 23 describes a simplified design of a radio system.

Microwave radio relay systems currently supply about half of the Bell System toll message circuit mileage in the United States. Individual circuit lengths range from less than 20 to over 4000 miles. Route cross-sectional capacities range from less than 60 to more than 22,000 telephone message circuits. Many types of microwave radio systems carry these circuits. Short-haul radio systems are usually operated in intrastate or feeder service. Long-haul radio systems are primarily used in interstate and *backbone route* applications. Exceptions do exist as in any complex communications network, but short-haul and long-haul radio systems traditionally have been developed for overall lengths of 250 and 4000 miles, respectively.

Before proceeding, it is appropriate to make some broad comparisons between radio systems and the analog AM cable systems described previously. Although some of the concepts described in the earlier chapters are directly applicable to radio systems, others must be modified and some new ones introduced.

17.1 COMPARISON OF AM WIRE SYSTEMS AND FM RADIO SYSTEMS

Frequency Versus Amplitude Modulation

Frequency modulation (FM) is used in microwave radio systems primarily because linear amplifiers with adequate gain and power output to handle wideband AM signals at these frequencies are not available. Microwave amplifiers used in radio systems are characterized by substantially larger amplitude nonlinearity than that of analog cable system amplifiers; however, the FM signal is relatively insensitive to this type of nonlinear distortion and can thus be transmitted through amplifiers which have compression or amplitude nonlinearity with little penalty.

Thermal Noise

In both AM and FM systems, thermal noise sets the minimum allowable signal amplitude. In AM the critical point is the input to the repeater amplifier, and in FM it is the input to the radio repeater. Most of the radio repeater thermal noise originates in the first stages of the radio receiver.

Intermodulation Noise

Intermodulation noise is a major factor in the design of FM systems; however, the mechanisms which generate such noise are markedly different from those encountered in AM systems. In AM systems intermodulation noise is caused by repeater amplitude nonlinearity; in FM systems it arises primarily because of transmission gain and delay deviations. As a result, intermodulation noise in AM systems is a function of signal amplitude, but in FM systems it is a function of the amplitude of frequency deviation. The following chapters point out ways in which the frequency deviation in an FM system is analogous to the signal amplitude in an AM system.

Repeater Spacing

Radio repeater spacings (hop lengths) are determined primarily by line-of-sight path clearance and received signal strength. In relatively flat terrain, increasing path lengths will dictate increased antenna tower heights and play an economic role in repeater site selection. Transmitter power output and antenna gain will similarly enter into the economics of selection, but FM radio systems, unlike AM cable systems, are not rigidly controlled as to repeater spacing. The primary reason for this divergence lies in the transmission medium. Cable loss is measured or expressed directly in dB per mile, and doubling a length of cable multiplies its loss in dB by two. Radio path loss varies as 20 log of the path length, and therefore doubling a path length increases its loss by only 6 dB. It follows that the problem of choosing repeater spacings in a radio system is not as clear-cut as it is in the AM cable systems, where a definite solution can be found in terms of repeater performance, wire transmission variations, and system requirements. In the radio system, the problem involves tower economics, geography, fading or rain attenuation, interferences, and system requirements. Consideration of these factors results in typical microwave repeater spacings of 20 to 30 miles.

17.2 MICROWAVE RADIO SYSTEM COMPONENTS

Figures 17-1 through 17-4 show the interrelation of the major components of microwave radio systems. Circuit details are largely omitted as extraneous to the purposes of this text. From the overall point of view, it is necessary to identify only those elements having a direct bearing on the choice of system parameters.

Entrance Links

Entrance links, as typified in Fig. 17-1, form the interface between the radio equipment and the multiplex terminal equipment. They are used to compensate for cable transmission losses between the multiplex and radio equipment, which may be from 50 feet to more than 8 miles apart. In addition, they establish the transmission level points in the radio system, and provide transmission level shaping, commonly called pre- and de-emphasis. The multiplex terminals indicated are similar to, if not identical with, those used in cable systems.

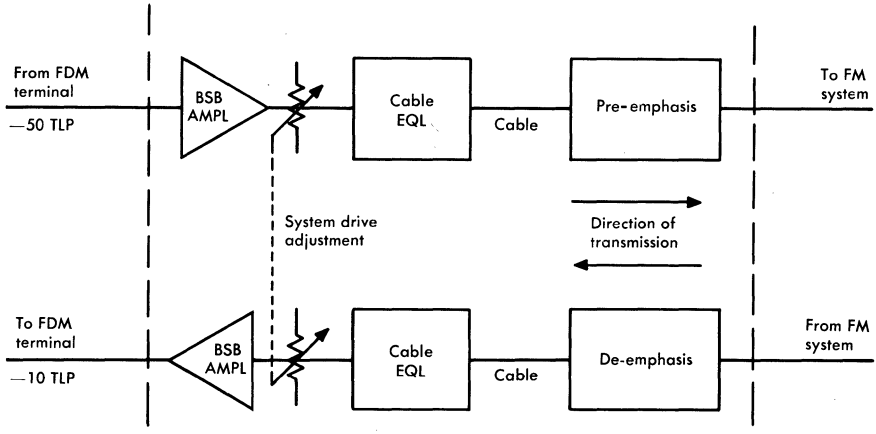


FIG. 17-1. Entrance links.

Baseband Repeaters

Figure 17-2 illustrates a baseband (BSB) radio repeater commonly used in short-haul applications. The modulating information is transferred between the receiver and the following transmitter at baseband frequencies and then retransmitted on a new frequency. At end stations, entrance links are connected at the points indicated, and the baseband connections to the multiplex terminals are thereby established. Baseband repeaters are used in short-haul service primarily to provide access to the baseband at intermediate stations for adding or dropping circuits.

Intermediate-Frequency Repeaters

Figure 17-3 illustrates an intermediate-frequency (i-f) radio repeater used in long-haul applications. In this type of repeater, sometimes called a *heterodyne* repeater, the receiver output is coupled at i-f to the following radio transmitter where it is translated upward in frequency for retransmission.

A primary advantage of this system is the avoidance of the FM-to-baseband and baseband-to-FM modulation steps used in the baseband repeater. Not only are several sources of noise avoided by eliminating unnecessary modulation steps, but also misalignment problems are reduced because an i-f repeater does not change the deviation of the signal passing through it.

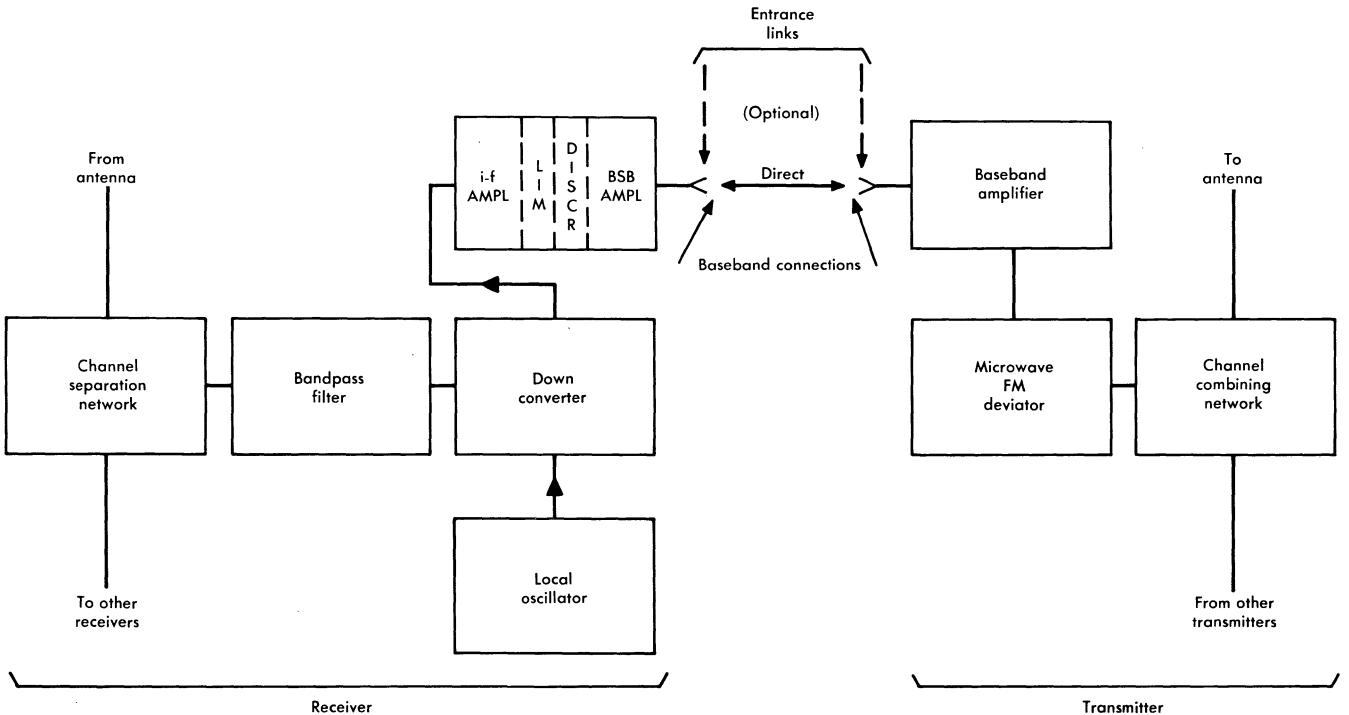


FIG. 17-2. Baseband radio repeater.

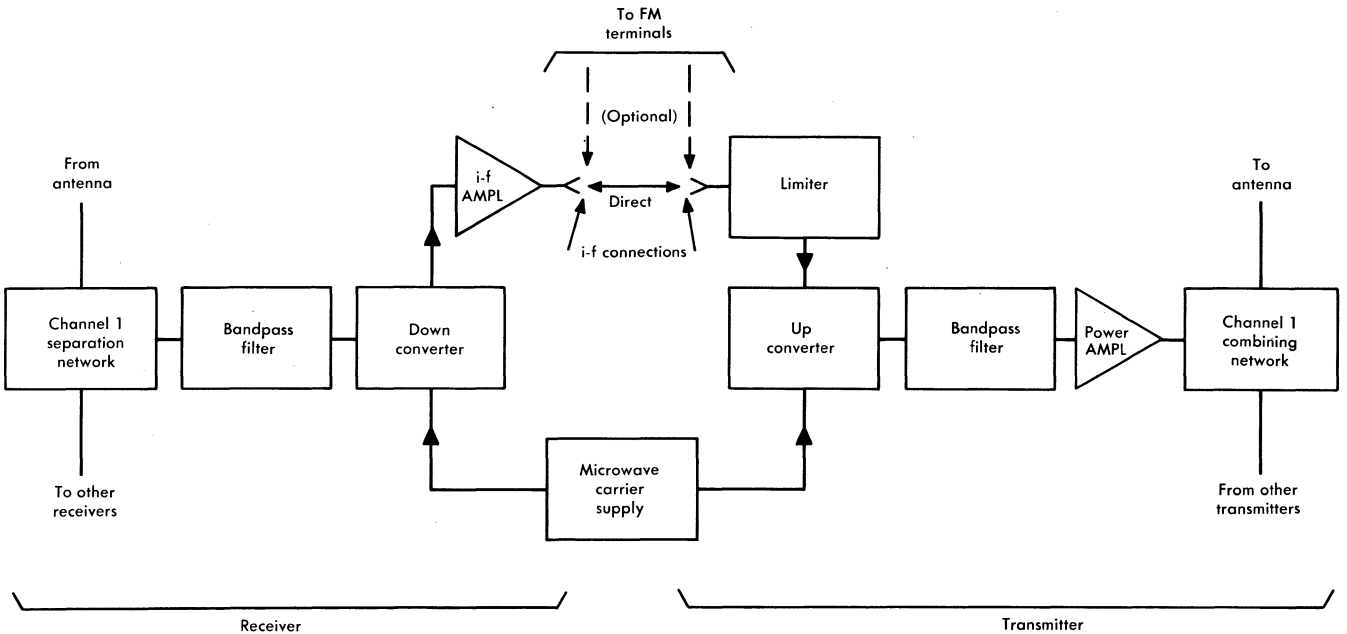


FIG. 17-3. Intermediate-frequency radio repeater.

The channel separation network at the left side of the figure extracts the Channel 1 signal and allows the other channels to continue down the waveguide to their respective separation networks. The receiver bandpass filter has high loss to all signals except those falling in Channel 1, and hence provides additional suppression of unwanted signals outside the Channel 1 frequency band. The down converter and following receiver circuits act as in a conventional radio receiver to produce a constant amplitude intermediate frequency output. The output may then be connected as required to an FM terminal receiver for demodulation or to a microwave transmitter for retransmission. In the transmitter, an amplitude limiter strips off any AM component of the signal to be transmitted. If the spurious AM components were not removed, AM/PM conversion in the microwave power amplifier and subsequent receiver circuits would degrade the system noise performance. The signal is translated to a new microwave frequency by an up-converter, and is subsequently amplified in the power amplifier and combined or multiplexed in the channel combining network with other transmitter outputs on other channels for radiation to the next station. Typical transmitter output powers for i-f repeatered radio systems of this type range from 0.5 to 10 watts. Received signals are around one-millionth of this power or —39 to —20 dBm.

FM Terminals

In an i-f repeatered radio system, the signal alternates between microwave and intermediate frequencies as it traverses first the radio paths and then the radio repeaters. At some stations, access to the baseband signal is necessary for the purposes of adding to, dropping, or otherwise modifying the message load. At these points and at the ends of the system, FM terminal transmitters are used to generate the frequency-modulated intermediate-frequency signals to be processed by the microwave radio equipment, and FM terminal receivers are used to demodulate the intermediate-frequency signals to recover baseband information.

The FM terminal transmitter shown in Fig. 17-4 generates a 70-MHz frequency-modulated output by mixing the outputs of two deviated oscillators which differ in nominal frequency by 70 MHz. The two oscillators are deviated in opposite phase or sense to relax the deviation linearity requirement on each oscillator and permit the partial cancellation of unwanted modulation products.

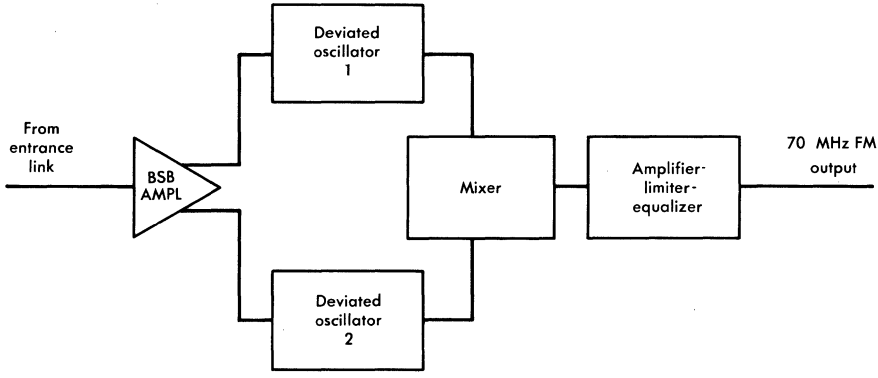


FIG. 17-4. FM terminal transmitter.

The FM receiver shown in Fig. 17-5 is used to recover the baseband signal. A limiter is used to remove amplitude variations in the signal prior to demodulation. The output of this terminal receiver is connected to an entrance link for de-emphasis and transmission to the multiplex equipment.

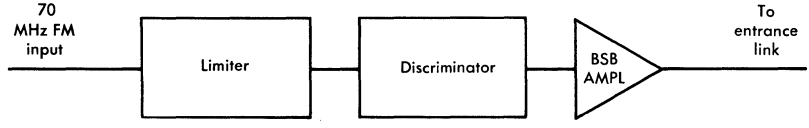


FIG. 17-5. FM terminal receiver.

17.3 PROTECTION OF SYSTEM CONTINUITY

Both wire and radio media are subject to variations with time. Cable loss varies with temperature, and broadband AM systems normally have automatic means of regulating the repeater gain to match the changing cable loss. Radio path loss varies with atmospheric conditions. For many hours of the day, the radio path is a relatively stable transmission medium; however, during certain periods, particularly during the night hours, atmospheric conditions may develop which cause fading. During fading periods, the

received signal strength on a radio channel can vary significantly, occasionally becoming somewhat higher than normal but more often dropping 20, 30, 40, or more dB below normal. When fading conditions exist, all of the radio channels will generally experience some reduction in received signal strength, but the very deep fade will affect perhaps only a single radio channel at a time. Although the deep fade (or fades) will normally move around in the band, affecting different channels at different times, deep fading may affect two or more channels simultaneously.

The automatic gain control circuit in the radio receiver compensates for fades of 25 to 40 dB, depending on the system. However, in order to insure service continuity during deep fading, an automatic protection switching system of some type is normally provided. Switching to protection channels may be done on a single hop or multiple hop basis, the primary requirement being to restore circuit continuity within about 30 milliseconds after an interruption.

An important problem in the design of a microwave system is the statistical study of outage time (i.e., service interruptions) caused by fading, equipment failures, and maintenance time as a function of the number of protection channels provided. From this study, a good engineering decision can be made on the number of protection channels required. Historically, two approaches to protection switching have evolved based on the particular frequency band of operation and the economics of the system to be protected.

Long-haul systems are composed of switching sections of one to ten hops. These systems on fully loaded 4- and 6-GHz routes use two protection channels to back up 10 or 6 working channels, respectively. Figure 17-6 lists these protection arrangements.

Short-haul systems in the 6- and 11-GHz bands are usually switched every hop on a one protection for one working channel basis. One-for-one switching is best suited to routes with one or two working channels but is less conservative of spectrum space on heavy routes than the long-haul switching systems. Crossband one-for-one diversity switching between 6- and 11-GHz systems is often used to compromise between susceptibility to rain attenuation at 11 GHz and spectrum crowding at 6 GHz.

17.4 MICROWAVE SYSTEM CHARACTERISTICS

The various Bell Laboratories designed radio systems and their allocated frequency bands are tabulated in Fig. 17-6. Other features listed are telephone message circuit capacity per radio channel, repeater type (BSB or i-f), and numbers of working and protection channels. Specific details of radio channel frequency assignments may be found in Chap. 22.

System designation	Band occupied (GHz)	Message circuits per radio channel	Repeater type	Radio channels in fully loaded route	
				Working	Protection
TD-2	3.7-4.2	600-1200	i-f	10	2
TD-3		1200	i-f	10	2
TH-1	5.925-6.425	1800	i-f	6	2
TH-3		1800	i-f	6	2
TM-1		600-900	BSB	4*	4*
TJ	10.7-11.7	600	BSB	3*	3*
TL-1		240	BSB	3	3
TL-2		600-900	BSB	3*	3*

FIG. 17-6. Selected features of radio systems.†

*TJ/TM-1 and TL-2/TM-1 are commonly used in crossband diversity.

†Figure 17-6 is intended in part as a quick directory for the articles listed below for each specific system.

REFERENCES

1. Roetkin, A. A., K. D. Smith, and R. W. Friis. "The TD-2 Microwave Radio Relay System," *Bell System Tech. J.*, vol. 30 (Oct. 1951), pp. 1041-1077.
2. Curtis, H. E., T. R. D. Collins, and B. C. Jamison. "Interstitial Channels for Doubling TD-2 Radio System Capacity," *Bell System Tech. J.*, vol. 39 (Nov. 1960), pp. 1505-1527.
3. "The TH Microwave Radio Relay System," *Bell System Tech. J.*, vol. 40 (Nov. 1961), pp. 1459-1743.
4. Gammie, J. and S. D. Hathaway. "The TJ Radio Relay System," *Bell System Tech. J.*, vol. 39 (July 1960), pp. 821-877.
5. Friis, R. W., J. J. Jansen, R. M. Jensen, and H. T. King. "The TM-1/TL-2 Short-Haul Microwave Systems," *Bell System Tech. J.*, vol. 45 (Jan. 1966), pp. 1-95.
6. "TD-3 Microwave Radio Relay System," *Bell System Tech. J.*, vol. 47 (Sept. 1968), pp. 1143-1537.

Chapter 18

Radio Propagation at Microwave Frequencies

Some knowledge of radio propagation and antennas is essential to an understanding of transmission in microwave radio systems. This chapter provides certain basic concepts concerning propagation paths, path losses, and microwave antennas and their characteristics.

18.1 PATH CHARACTERISTICS

Propagation Paths

The normal propagation paths between two radio antennas are illustrated in Fig. 18-1. The direct or free-space wave is shown in path 1, and the wave reflected from the ground is path 2. Path 3 indicates a surface wave which consists of the electric and magnetic fields associated with the currents induced in the ground. Its magnitude depends on the constants of the ground and the electromagnetic wave polarization. The sum of these three paths, taking account of both magnitude and phase, is called the ground wave. There are induction fields and secondary effects of the ground which are also a part of this wave, but these effects are negligible beyond a few wavelengths from the transmitting antenna. Path 4, called the sky wave, depends on the presence of the ionized layers above the earth that reflect back some of the energy that otherwise would be lost in outer space.

All of the paths shown in Fig. 18-1 exist in any radio propagation problem, but some are negligible in certain frequency ranges. At frequencies less than about 1500 kHz, the surface wave provides the

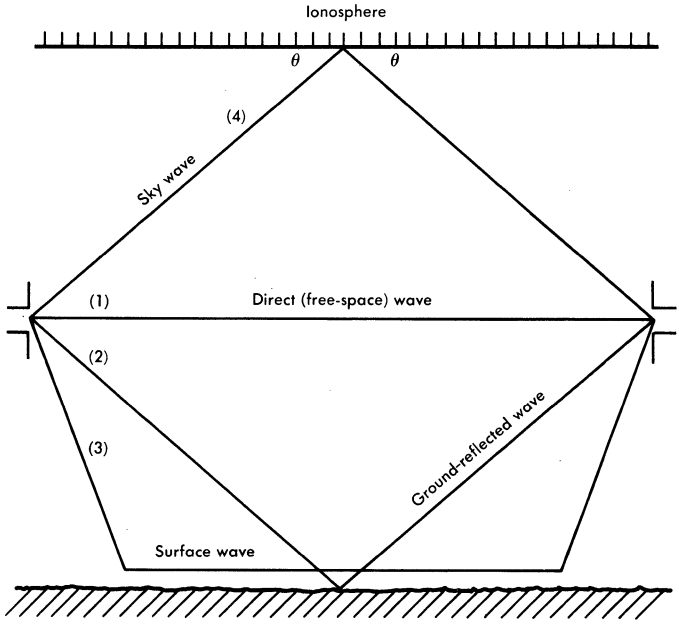


FIG. 18-1. Transmission paths between two antennas.

primary coverage, and the sky wave helps to extend this coverage at night when the absorption of the ionosphere is at a minimum. At frequencies above about 30 to 50 MHz, the free-space and ground-reflected waves are frequently the only paths of importance. At these frequencies the surface wave can usually be neglected as long as the antenna heights are not too low, and the sky wave is only a source of occasional long distance interference rather than a reliable signal for communication purposes.

At frequencies in the order of thousands of megahertz, where the microwave systems under discussion operate, the free-space wave is usually controlling on good optical paths, although in many cases attention must also be given to the reflected wave. Thus, in this chapter, the surface and sky wave propagation are neglected, and attention is focused only on those phenomena that affect the direct and reflected waves. Free-space transmission is considered first, then deviations from free-space transmission, and finally antenna properties and types.

Free-Space Path Loss

Consider first a given amount of power, p_T, radiated from an *isotropic* transmitting antenna, a point source radiating power equally in all directions. Imagine a sphere of radius d, centered upon the point source. If free-space transmission is assumed (i.e., straight line transmission through a vacuum or ideal atmosphere, with no absorption or reflection of energy by nearby objects), the radiated power density will be equal at all points on the surface of the sphere, and the total radiated power, p_T, will pass outward through the surface of the sphere. The radially directed power density at any point on the surface of the sphere, therefore, will be

$$\text{Power density} = \frac{p_T}{4\pi d^2} \tag{18-1}$$

If a receiving antenna with an effective* area, A_R, is located on the surface of the sphere, the received power, p_R, will equal the power density times the area of the antenna; that is,

$$p_R = \frac{p_T}{4\pi d^2} A_R \tag{18-2}$$

It can be shown [1] that a transmitting antenna which concentrates its radiation within a small solid angle or beam has an on-axis transmitting antenna gain with respect to an isotropic radiator of

$$g_T = \frac{4\pi A_T}{\lambda^2} \tag{18-3}$$

This is equivalent to concentrating the radiation in a solid angle of

$$\Omega = \frac{\lambda^2}{A_T} \quad \text{rad}^2 \tag{18-4}$$

and comparing this angle to 4π (radian)^2 which is a whole solid angle.

*In actual antennas, imperfect reflector illumination and re-radiated energy result in a power loss. The discrepancy between theoretical and actual gain leads to the concept of an effective area which is smaller than the physical area.

Grouping the results of the previous two paragraphs,

$$p_R = p_T \left(\frac{4\pi A_T}{\lambda^2}\right)\left(\frac{A_R}{4\pi d^2}\right) \tag{18-5}$$

It is now convenient to rearrange the terms, primarily to get the transmitting and receiving antenna gains into identical forms. In this way, both transmitting and receiving antennas are specified as to their gain relative to isotropic radiators.

$$p_R = p_T \underbrace{\left(\frac{4\pi A_T}{\lambda^2}\right)}_{\substack{\text{Trans}\\ \text{ant}\\ \text{gain}}} \underbrace{\left(\frac{4\pi A_R}{\lambda^2}\right)}_{\substack{\text{Rec}\\ \text{ant}\\ \text{gain}}} \left(\frac{\lambda}{4\pi d}\right)^2 \tag{18-6}$$

Stated in decibels, the ratio of p_T to p_R equals

$$10 \log \frac{p_T}{p_R} = -10 \log \frac{4\pi A_T}{\lambda^2} - 10 \log \frac{4\pi A_R}{\lambda^2} + 20 \log \frac{4\pi d}{\lambda} \tag{18-7}$$

The manipulation of antenna gain terms in the previous paragraph results in a distance and frequency-dependent term which is called the *free-space path loss* between isotropic radiators.

$$\text{Free-space path loss in decibels} = 20 \log \frac{4\pi d}{\lambda} \tag{18-8}$$

This term is plotted in Fig. 18-2 for representative path lengths and frequencies.

Section Loss

Section loss is defined as the loss in decibels between a radio transmitter output and the following radio receiver input. It includes the loss as determined by Eq. (18-7), plus all the waveguide and network losses at both ends of the hop. Addition of these factors to Eq. (18-7) defines the section loss in dB as:

$$\text{Section loss} = 20 \log \frac{4\pi d}{\lambda} - 10 \log \frac{4\pi A_T}{\lambda^2} - 10 \log \frac{4\pi A_R}{\lambda^2} + \text{waveguide losses} + \text{network losses} \tag{18-9}$$

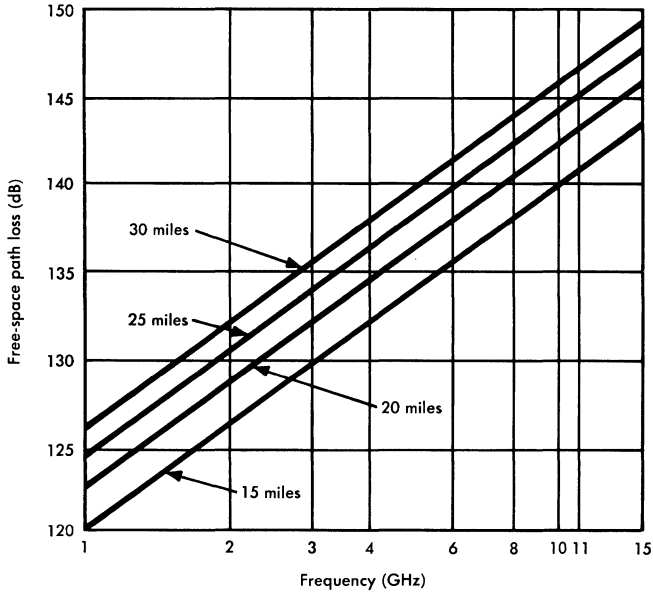


FIG. 18-2. Free-space path loss versus frequency and path length.

The following example illustrates these relations.

Example 18.1

Problem

Determine the receiver input power for a 4-GHz radio hop given the following conditions:

Transmitter output power	= 37.0 dBm (5 watts)
Gain of each antenna	= 39.6 dB
Loss of networks in receiving waveguide	= 1.9 dB
Loss of networks in transmitting waveguide	= 1.9 dB
Transmitting waveguide loss	= 2.1 dB
Receiving waveguide loss	= 2.1 dB
Frequency	= 4.0 GHz
Path length	= 28.5 miles

Solution

From Fig. 18.2, the free-space path loss at 4 GHz for a path length of 28.5 miles is 137.5 dB. Thus, using Eq. (18-9), the section loss is

$$\text{Section loss} = 137.5 - 39.6 - 39.6 + 2(2.1 + 1.9)$$

$$= 66.3 \text{ dB}$$

The input power to the radio receiver is equal to the transmitter output power minus the section loss. Thus,

$$\text{Receiver input power} = 37.0 - 66.3$$

$$= -29.3 \text{ dBm}$$

Antenna Heights and Path Clearance

Up to this point, only free-space transmission has been considered. The presence of the earth and the nonuniformity of the atmosphere may markedly affect the actual operating conditions.

For a large percentage of the time, the path loss of a typical microwave link can be made to approximate closely the calculated free-space loss. This can be done by engineering the path between antennas to provide an optical line-of-sight transmission path which has adequate clearance with respect to surrounding objects. This clearance is necessary not only to keep the path loss under normal atmospheric conditions from deviating from the free-space value, but also to reduce severe fading problems during abnormal conditions.

The importance of adequate clearance can be seen by considering Fig. 18-3, which shows the profile of the path between two antenna sites. For the antenna heights shown, the distance H represents the clearance of the line-of-sight path, AB, and the intervening terrain. Path ACB represents a secondary transmission path via reflection from a projection. With no phase reversal at the point of reflection, the signal from the two paths would partially cancel whenever AB and ACB differed by an odd multiple of a half wavelength. When the grazing angle of the secondary wave is small, which is typically the case, a phase reversal will normally occur at the point of reflection. Therefore, whenever AB and ACB differ by an odd multiple of a half wavelength, the energies of the received signals add, rather than cancel. Conversely, if the two paths differ by a whole number of wavelengths, the signals from the two paths will tend to cancel.

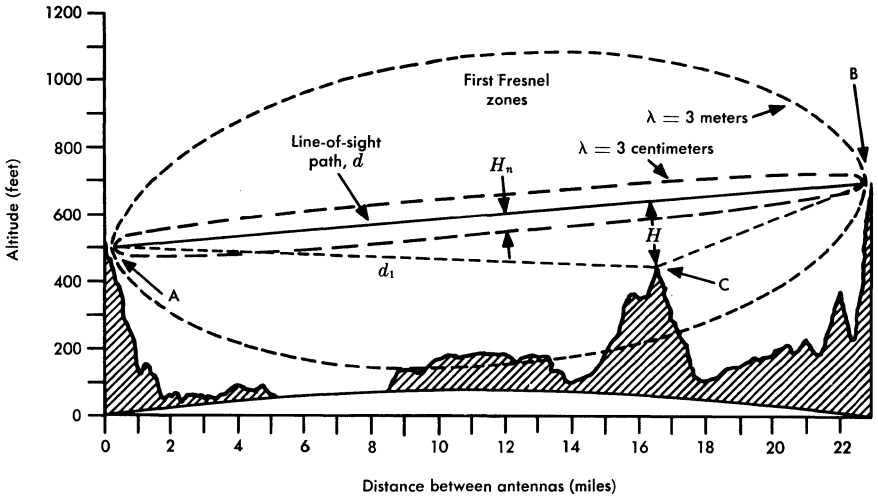


FIG. 18-3. Typical profile plot showing first Fresnel zones for 100 MHz and 10 GHz.

The amount of clearance is generally described in terms of Fresnel* zones. All points from which a wave could be reflected with an additional path length of one-half wavelength form an ellipse which defines the first Fresnel zone. Similarly, the boundary of the nth Fresnel zone consists of all points from which the delay is $n/2$ wavelengths. For any distance, d_1, from antenna A, the distance H_n from the line-of-sight path to the boundary of the nth Fresnel zone is approximated by the parabola:

$$H_n = \sqrt{\frac{n\lambda d_1 (d-d_1)}{d}} \qquad (18\text{-}10)$$

where λ, the wavelength, and d and d_1, the path lengths, are measured in identical units. The boundaries of the first Fresnel zones for $\lambda = 3$ meters (100 MHz) and $\lambda = 3$ centimeters (10 GHz) in the vertical plane through AB are shown in Fig. 18-3. In any plane normal to AB, the Fresnel zones are concentric circles.

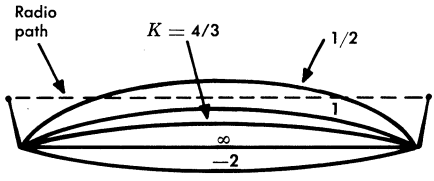
Measurements have shown that to achieve a normal transmission loss approximately equal to the free-space loss, the transmission path should pass over all obstacles with a clearance of at least 0.6 times the first Fresnel zone distance, and preferably by an amount equal to the first Fresnel zone distance. However, because of refraction

*Fresnel (Frå-nell').

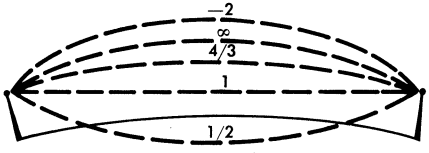
effects, greater clearance is usually provided in order to reduce deep fading under adverse atmospheric conditions.

The effective path clearance varies with time because radio waves seldom follow truly straight lines. Atmospheric refraction, resulting from variations in the dielectric permittivity with height, causes the radio signal to bend slightly from its ideal straight line path. This effect can be visualized by assuming that the radio wave does travel *exactly* in a straight line but over an earth with a fictitious radius which is either greater or less than the true earth radius. On the average, the radio wave is bent downward *as if* the earth radius were 4/3 of its actual value. For this reason, it is frequently convenient to plot the elevations of a path on special profile paper on which the earth's radius is assumed to be 4/3 of its actual value. The radio path is then plotted as a straight line and the earth's curvature appears to flatten.

The effective earth radius factor K (ratio of effective earth radius to true earth radius) is a function of atmospheric conditions and may be as low as 1/2 for a small percentage of the time. This corresponds to a so-called earth bulge [2] condition and may result in a considerable increase in path loss over a wide range of frequencies unless adequate path clearance is provided. On the other hand, when the effective earth radius factor is infinite, it is as if the earth were completely flat, and long-range interference from same-channel stations may result because the shielding by the earth curvature has temporarily been removed.



(a) Effective earth profiles versus K



(b) Actual radio paths versus K

FIG. 18-4. Effect of K on radio paths.

The deviations in the curvature of the radio waves and the corresponding values for the effective earth radius factor are contrasted in Fig. 18-4 for several values of K. An infinite effective radius does not imply a limit; negative values of K can be pictured as a "depressed" earth, but more importantly, negative values represent conditions for which the atmosphere acts like a duct or waveguide for propagation over relatively long distances.

In determining suitable tower heights, a profile plot of the terrain between the proposed antenna sites is obtained, and the worst obstacle in the path, such as the ridge shown in Fig. 18-3, is located. This obstacle is then used as a leverage point from which the most suitable antenna height at each location can be chosen to provide the proper clearance. Path testing using portable antennas is frequently done to verify the appropriateness of paths and determine optimum antenna heights.

Fading

Substandard atmospheric refraction ($K < 1$) may transform a line-of-sight path into an obstructed one, because the effective path clearance becomes zero or negative. This situation can happen under conditions of heavy ground fog or extremely cold air over warm earth. The result is a substantial increase in path loss over a wide frequency band. The magnitude and frequency of occurrence of this type of slow, flat fading can be reduced only by the use of greater antenna heights.

The more common form of microwave fading on paths with adequate clearance is a relatively fast, frequency selective type of fading caused by interference between two or more rays in the atmosphere. The separate paths between transmitter and receiver are caused by the irregularities (second and higher order derivatives) in the variations in dielectric permittivity with height. The refraction effect mentioned earlier depends on the average slope (first derivative) of the same variation in dielectric permittivity. The transmission margins that must be provided against both types of fading are important in determining the overall system parameters.

An interference type fade can have any depth, but fortunately the deeper the fade the less frequently it occurs and the shorter its duration when it does occur. Figure 18-5 shows the median duration of fades of various depths on a 4-GHz system with typical repeater spacings of about 30 to 35 miles. It will be noted that the median duration of a 20-dB fade is about 30 seconds, and the median duration of a 40-dB fade is about 3 seconds. At any given depth of fade, the duration of 1 per cent of the fades may be as much as ten times or as little as 1/10 of the median duration.

Multipath fading occurs primarily at night on typical 4-GHz line-of-sight paths. During the day or whenever the lower atmosphere is thoroughly mixed by rising convection currents and winds, the signals

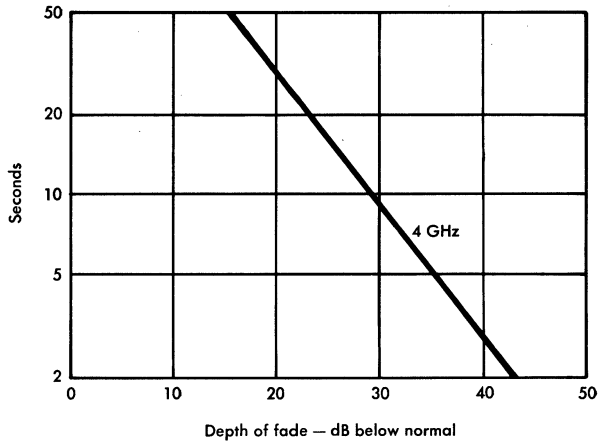


FIG. 18-5. Median duration of fast fading.

on line-of-sight paths are normally steady and at or near the predicted free-space values. On clear nights with little or no wind, however, sizable irregularities or layers can collect at random elevations, and these irregularities in refraction result in multipath transmission on path lengths of the order of a million wavelengths or longer. Multipath fading tends to build up during the night with a peak in the early morning hours and then to disappear as the layers are broken up by the convection caused by the heat of the early morning sun.

Both the number of fades and the percentage of time below a given level tend to increase as either the repeater spacing or the frequency increases. Multiple paths are usually overhead, although ground reflections can be a factor in some cases. The effects of multipath fading can be minimized by the use of either frequency or space diversity.

Absorption

Rainfall and water vapor also produce pronounced attenuation effects at the higher microwave frequencies. It is well known that certain absorption bands occur in the spectrum of visible light, and the theory of these absorption bands indicates that they should be found throughout the electromagnetic spectrum. The first absorption

band due to water vapor peaks at about 22 GHz, and the first absorption band due to the oxygen in the atmosphere peaks at about 60 GHz [3].

The effect of rain on microwave radio propagation in the region of 4 to 6 GHz is small relative to the losses introduced by other causes of fading. At higher frequencies, however, rain attenuates radio transmission to a much greater degree. The radio energy is absorbed and scattered by the rain drops, and this effect becomes more pronounced as the wavelength approaches the size of the raindrops. Figure 18-6 indicates the estimated atmospheric absorption for vari-

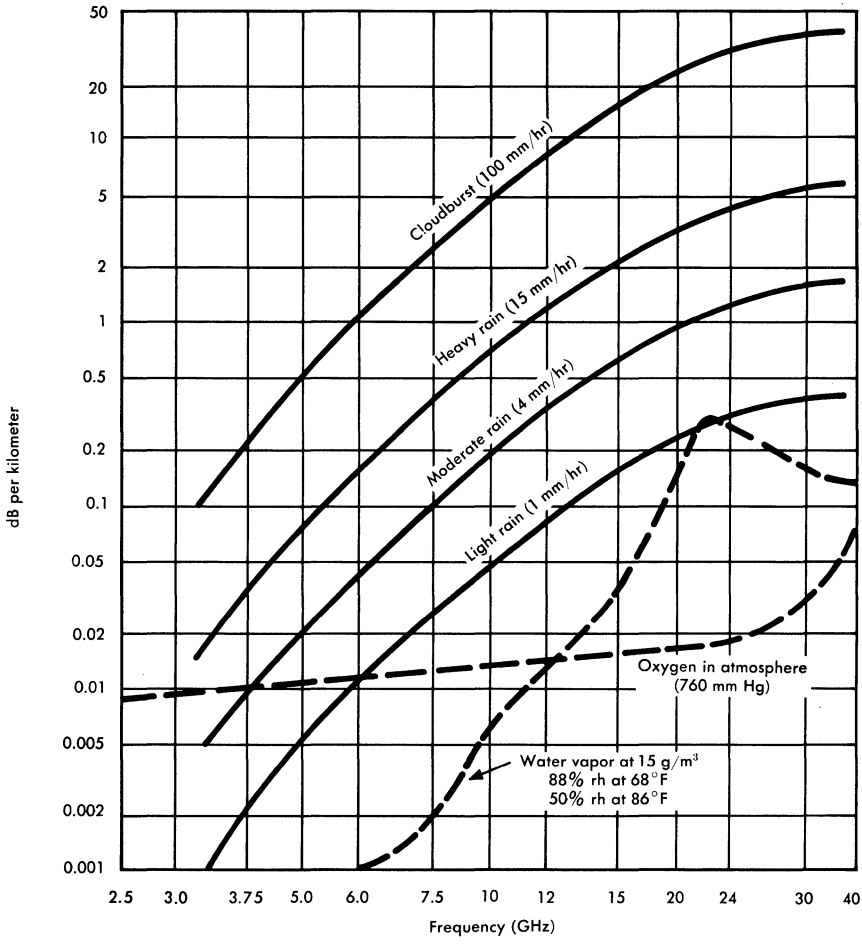


FIG. 18-6. Estimated atmospheric absorption.

ous conditions of rainfall. From this figure it is evident that rain attenuation must be considered in any system operating at frequencies of 10 GHz or above and perhaps at lower frequency in areas where heavy rains occur frequently. It is also evident that rain attenuation over any individual microwave band is almost independent of frequency, and therefore no protection is offered by the use of inband frequency diversity.

18.2 MICROWAVE ANTENNAS

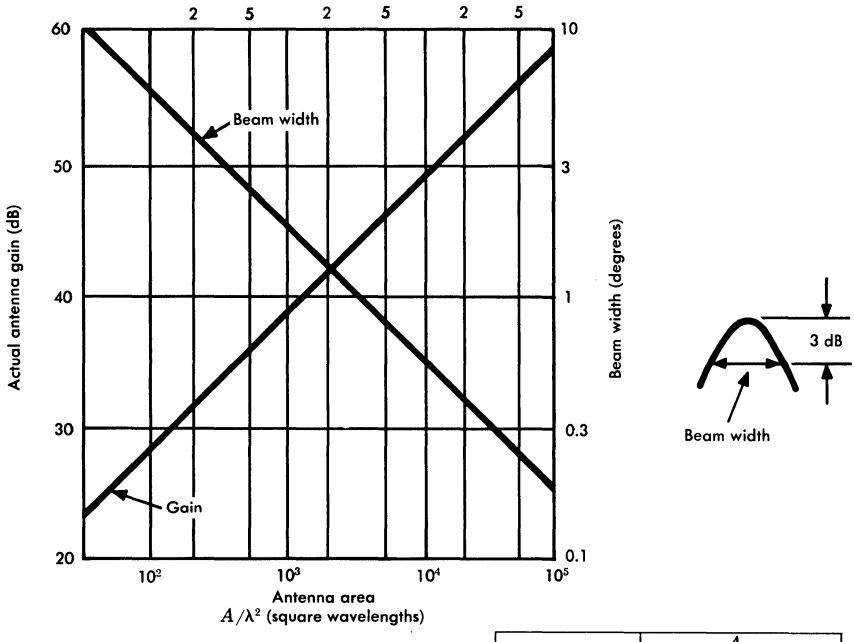
Antenna Characteristics

Many antenna characteristics are important in microwave systems. The first of these, antenna gain, has already been defined. An antenna has gain because it concentrates the radiated power in a narrow beam rather than sending it uniformly in all directions. Since it reduces section loss, high antenna gain is obviously desirable.

Closely associated with antenna gain is beam width. Since an antenna achieves gain by concentrating power in a narrow beam, the width of the beam will decrease as the antenna gain is increased. Antennas used in microwave systems ordinarily have half-power beam widths of the order of one degree (see Fig. 18-7). A narrow beam minimizes interference from outside sources and adjacent antennas. A very narrow beam, however, imposes severe mechanical stability requirements and leads to problems in antenna lineup and fading.

All of the energy from an antenna does not lie in the direction of the main beam; some of it is concentrated in minor beams called sidelobes, which are potential sources of interference into or from other microwave paths. Figure 18-8 illustrates the relationship between the main beam and sidelobes for a horn-reflector antenna, which is discussed later.

Several antenna characteristics are important in evaluating the coupling of interference between adjacent antennas or radio paths. The *front-to-back ratio* of an antenna may be defined as the ratio of its maximum gain in the forward or intended direction to its maximum gain in the region of its backward direction. The front-to-back ratio of an antenna in an actual installation may be 20 to 30 dB less than its isolated or free-space value because of foreground reflections from objects in or near the main transmission lobe or beam. The front-to-back ratio of the antenna is critical in repeater



Note:
Abscissa is actual antenna area, and "actual antenna gain" is taken to be 3 dB below theoretical.

Antenna	$\frac{A}{\lambda^2}$ 4 GHz	6 GHz	11 GHz
8-foot paraboloid	$8(10)^2$	$2(10)^3$	$6(10)^3$
4-foot paraboloid	$5(10)^2$	$5(10)^2$	$1.5(10)^3$
8-foot square	$1(10)^3$	$2.5(10)^3$	$8(10)^3$

FIG. 18-7. Approximate antenna gain and beam width.

systems, especially when the same signal frequencies are to be used in both directions from one station. Additional characteristics involving two or more antennas at the same station are *side-to-side coupling* and *back-to-back coupling*. These factors express in dB the coupling losses between antennas carrying transmitter output signals and antennas carrying receiver input signals. Typical transmitter outputs are some 60 dB higher in level than receiver input levels, and accordingly the coupling losses must be high to avoid unwanted interferences, particularly when the desired signal is fading.

Use of Polarization

To improve adjacent channel discrimination and to facilitate the design of channel combining and dropping networks, it is common

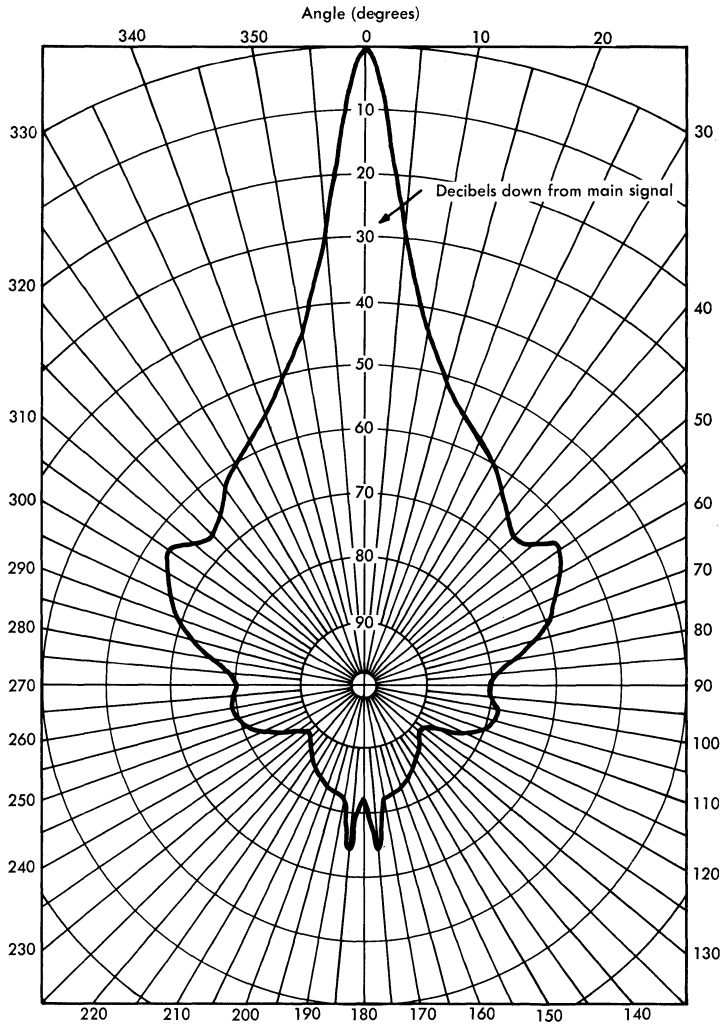


FIG. 18-8. Smoothed azimuthal directivity of horn-reflector antenna with vertically polarized signal at 3740 MHz.

practice in microwave relay systems to interleave alternate radio channel frequencies on horizontal and vertical polarizations of the transmitted signal. Polarization refers to the alignment of the electric field in the radiated wave.

Another orthogonal system with left- and right-hand circular polarization could be used, but the practical problems of maintaining

polarization discrimination over relatively wide bandwidths and in the presence of reflections do not make it attractive.

When energy is radiated in one polarization, a small portion may be converted to the other polarization by imperfections in the antenna system and path. The ratio of the power received in the desired polarization to the power received in the opposite polarization is called the cross-polarization discrimination. Cross-polarization discriminations of 25 to 30 dB for an entire hop are routinely obtained with ordinary antenna systems.

Typical Microwave Antennas

Parabolic Antenna. The parabolic (or dish) antenna consists of a paraboloid reflector illuminated with microwave energy by a feed system located at the focus. Depending on the design, one to four waveguide runs in one or two radio bands may be fed to the antenna simultaneously. These antennas in the 5- to 10-foot diameter range are widely used in short-haul systems and occasionally on lightly loaded long-haul routes where economic considerations control the choice of antenna systems.

Horn-Reflector Antenna. In the horn-reflector antenna, Fig. 18-9, a vertically mounted horn tapering outward from the focal point is used to illuminate a section of a parabolic surface which then reflects the energy outward. Because of the design and size of the horn, the impedance match of this antenna to its waveguide feed is very good, the return loss being between 40 and 50 dB. It is a broadband antenna and can be used with both vertical and horizontal polarization in the 4-, 6-, and 11-GHz bands. Its nominal characteristics are tabulated in Fig. 18-10, and Fig. 18-8 shows its horizontal directivity. The horn-reflector antenna has small sidelobes and radiates very little power to the rear, resulting in a nominal 70-dB front-to-back ratio. Measurements made at 6 GHz on a large number of antenna installations have shown that in horn-reflector antenna systems, side-to-side coupling and back-to-back coupling, as well as cross-polarization discrimination, follow approximately normal distributions. The mean and standard deviations of these distributions are listed in Fig. 18-11. The side-to-side and back-to-back coupling of the antenna system will vary considerably from location to location as a result of foreground reflections and leakage of energy at the joints of the waveguide run feeding the antennas.

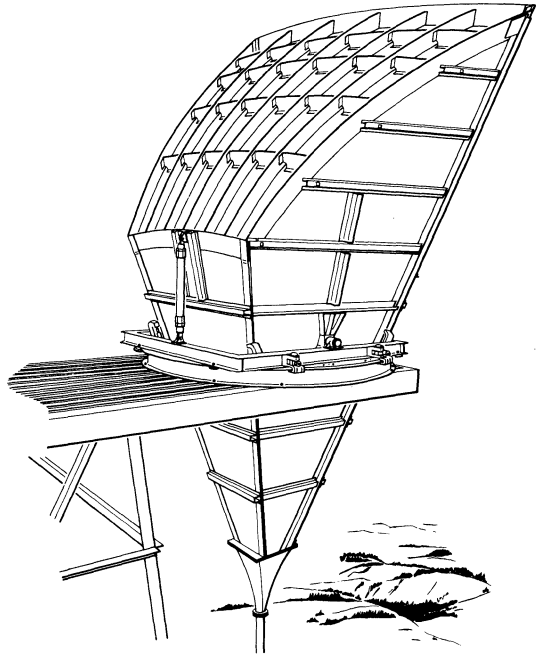


FIG. 18-9. Horn-reflector antenna.

Frequency	4 GHz		6 GHz		11 GHz	
Polarization	Vert	Hor	Vert	Hor	Vert	Hor
Midband gain (dB)	39.6	39.4	43.2	43.0	48.0	47.4
Front-to-back ratio (dB)	71	77	71	71	78	71
Beam width (azimuth) (degrees)	2.5	1.6	1.5	1.25	1.0	0.8
Beam width (elevation) (degrees)	2.0	2.13	1.25	1.38	0.75	0.88
Sidelobes (dB below main beam)	49	54	49	57	54	61
Side-to-side coupling (dB)	81	89	120	122	94	112
Back-to-back coupling (dB)	140	122	140	127	139	140

FIG. 18-10. Horn-reflector antenna characteristics for a particular pair of antennas without any waveguide system attached.

	Mean	Standard deviation
Side-to-side coupling (same polarization)	102	8.1
Side-to-side coupling (opposite polarization)	109	9.0
Back-to-back coupling (same polarization)	125	10.3
Back-to-back coupling (opposite polarization)	127	10.3
Cross-polarization discrimination	28	5.0

FIG. 18-11. Horn-reflector antenna and its waveguide system characteristics in dB at 6 GHz.

REFERENCES

1. Kraus, J. D. *Antennas* (New York: McGraw-Hill Book Company, Inc., 1950).
2. Schelleng, J. C., C. R. Burrows, and E. B. Ferrell. "Ultra Short-Wave Propagation," *Proc. IRE*, vol. 21, no. 3 (March 1933), pp. 427-463.
3. Medhurst, R. G. "Rainfall Attenuation of Centimeter Waves: Comparison of Theory and Experiment," *Trans. IEEE*, AP-13, no. 4 (1965), pp. 550-564.

Chapter 19

Properties of FM and PM Signals

The discussion of angle-modulated signals in Chap. 5 is extended in this chapter as background for the chapters on FM system analysis which follow. Expressions are derived for the FM spectrum when the modulating signal consists of one or more sinusoids; the analysis is then extended to more complex modulating signals. High-index FM and phasor representation of angle-modulated signals are discussed, and finally, the problem of unavoidable amplitude modulation on the FM signal and the effects of limiters on this problem are examined. For more detailed discussions of particular points, the reader is referred to standard texts on modulation theory [1, 2].

19.1 FREQUENCY ANALYSIS OF FM AND PM SIGNALS

For amplitude modulation, the frequency components of the modulated signal consist of a carrier, an upper sideband, and a lower sideband. The frequency components of the upper sideband have the same form as the components of the modulating signal except that they have been translated upward in frequency by an amount equal to the carrier frequency. The lower sideband is a mirror image of the upper sideband about the carrier frequency. This is illustrated in Fig. 19-1. For every component at frequency f_m in the modulating signal, there are components in the modulated signal at frequencies $f_c + f_m$ and $f_c - f_m$, where f_c is the carrier frequency. In a sense then, superposition holds since the effect produced by any particular modulating component does not depend on the other modulating components which are present. If the highest frequency

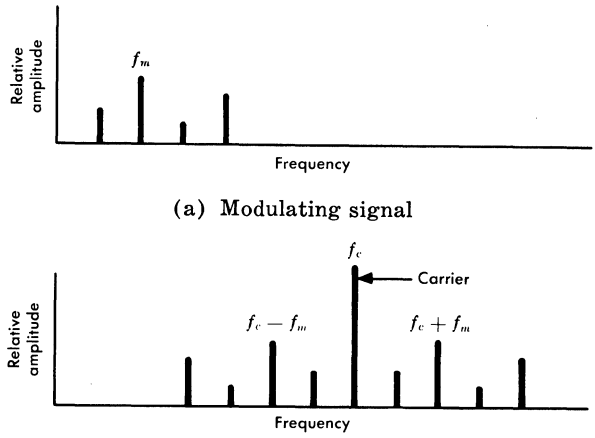


FIG. 19-1. Frequency spectrum of an amplitude-modulated signal.

component in the modulating signal is f_T, the modulated signal is restricted to the frequency range $f_c - f_T$ to $f_c + f_T$, and the required bandwidth is $2f_T$ centered at f_c.

In the case of frequency modulation, the frequency components of the modulated signal are more complexly related to the components in the modulating signal. In a strict mathematical sense, a single modulating sinusoid produces an infinite number of sideband components, although most are negligibly small. The multiplicity of sideband components complicates the frequency spectrum of an FM signal. In addition, the sideband components produced by any single-frequency component in the modulating signal depend on all the frequency components in the modulating signal. Hence, superposition does not apply.

Is it really advantageous to deal with the frequency components of an FM signal in view of this difficulty? At the present time, the answer seems to be that this is the best way known. The transmission characteristics of networks, interstages, and other transmission paths are specified as a function of frequency. Imperfect transmission at any particular frequency will affect only those frequency components of the signal which are at that frequency, but this in turn may cause serious impairment to the signal being trans-

mitted if the imperfection is not properly equalized before the FM discriminator. Furthermore, the problem of determining the required bandwidth depends on the location of all of the important frequency components in the signal. So in spite of the difficulty, some knowledge of the frequency components of an FM signal is essential.

Modulation by a Single Sinusoid

The analysis of angle-modulated signals starts with a single sinusoidal modulation which produces a peak phase deviation of X_1 radians. From Eq. (5-19),

$$M(t) = A_c \cos (\omega_c t + X_1 \cos \omega_1 t) \tag{19-1}$$

Functions of this type may be resolved into summations of sinusoids by application of the following Bessel function identities:

$$\sin (\alpha + X \sin \beta) = \sum_{n=-\infty}^{\infty} J_n(X) \sin (\alpha + n\beta) \tag{19-2}$$

$$\cos (\alpha + X \sin \beta) = \sum_{n=-\infty}^{\infty} J_n(X) \cos (\alpha + n\beta) \tag{19-3}$$

$$\sin (\alpha + X \cos \beta) = \sum_{n=-\infty}^{\infty} J_n(X) \sin\left(\alpha + n\beta + \frac{n\pi}{2}\right) \tag{19-4}$$

$$\cos (\alpha + X \cos \beta) = \sum_{n=-\infty}^{\infty} J_n(X) \cos\left(\alpha + n\beta + \frac{n\pi}{2}\right) \tag{19-5}$$

Here $J_n(X)$ is the Bessel function of the first kind of nth order and of argument X. Values of $J_n(X)$ for several values of X are listed in Fig. 19-2. A more complete tabulation of values may be obtained in References 2 and 3. Note that the argument, X, is the index of modulation.

The identity of Eq. (19-5) applied to the signal of Eq. (19-1) yields

$$M(t) = A_c \sum_{n=-\infty}^{\infty} J_n(X_1) \cos\left(\omega_c t + n\omega_1 t + \frac{n\pi}{2}\right) \tag{19-6}$$

	$X = 1/2$	$X = 1$	$X = 2$	$X = 3$	$X = 10$
$J_0(X)$	0.938	0.765	0.224	−0.260	−0.246
$J_1(X)$	0.242	0.440	0.577	0.339	0.043
$J_2(X)$	0.031	0.115	0.353	0.486	0.255
$J_3(X)$	0.003	0.020	0.129	0.309	0.058
$J_4(X)$	0.000	0.002	0.034	0.132	−0.220

FIG. 19-2. Values of $J_n(X)$ for several values of X.

The first few terms may be written as

$$M(t) = A_c\Big\{J_0(X_1)\cos\omega_c t + J_1(X_1)\cos\left[(\omega_c+\omega_1)t+\frac{\pi}{2}\right] + J_{-1}(X_1)\cos\left[(\omega_c-\omega_1)t-\frac{\pi}{2}\right] + J_2(X_1)\cos\left[(\omega_c+2\omega_1)t+\frac{2\pi}{2}\right] + J_{-2}(X_1)\cos\left[(\omega_c-2\omega_1)t-\frac{2\pi}{2}\right]+\cdots\Big\} \tag{19-7}$$

Because of the identity

$$J_{-n}(X) = (-1)^n J_n(X) \tag{19-8}$$

it follows that $M(t)$ can be written as

$$M(t) = A_c\Big\{J_0(X_1)\cos\omega_c t + J_1(X_1)\cos\left[(\omega_c+\omega_1)t+\frac{\pi}{2}\right] + J_1(X_1)\cos\left[(\omega_c-\omega_1)t+\frac{\pi}{2}\right] - J_2(X_1)\cos(\omega_c+2\omega_1)t - J_2(X_1)\cos(\omega_c-2\omega_1)t+\cdots\Big\} \tag{19-9}$$

Equation (19-9) shows that a single sinusoidal modulating signal produces sets of sidebands displaced from the carrier by multiples of the modulating frequency. These successive sets of sidebands are often referred to as first order sidebands, second order sidebands, etc., the magnitudes of which, relative to the carrier, are determined by the coefficients $J_1(X_1)$, $J_2(X_1)$, etc., respectively. As Fig. 19-2 shows, the higher order sidebands rapidly become un-

important as the index of modulation, X, becomes less than unity. Thus if the index is sufficiently small, the spectrum of a PM signal resembles that of an AM signal. For larger values of X, the value of $J_n(X)$ starts to decrease rapidly as soon as $n = X$.

Modulation by Two Sinusoids

When two or more sinusoids are applied simultaneously to an angle modulator, a superposition of sideband spectra will not suffice to describe the resulting output. A derivation of the output spectrum for two-sinusoid modulation begins with the equation

$$M(t) = A_c \cos (\omega_c t + X_1 \cos \omega_1 t + X_2 \cos \omega_2 t) \tag{19-10}$$

It is convenient to start the analysis by writing Eq. (19-10) in the form

$$M(t) = A_c \cos \left(\frac{\omega_c t}{2} + X_1 \cos \omega_1 t + \frac{\omega_c t}{2} + X_2 \cos \omega_2 t \right) \tag{19-11}$$

Using the trigonometric identity

$$\cos (A + B) = \cos A \cos B - \sin A \sin B \tag{19-12}$$

Eq. (19-11) can be written as

$$M(t) = A_c \left\{ \cos \left(\frac{\omega_c t}{2} + X_1 \cos \omega_1 t \right) \cos \left(\frac{\omega_c t}{2} + X_2 \cos \omega_2 t \right) - \sin \left(\frac{\omega_c t}{2} + X_1 \cos \omega_1 t \right) \sin \left(\frac{\omega_c t}{2} + X_2 \cos \omega_2 t \right) \right\} \tag{19-13}$$

The identities of Eqs. (19-2) through (19-5) can now be applied; thus,

$$M(t) = A_c \left\{ \left[\sum_{n=-\infty}^{\infty} J_n(X_1) \cos \left(\frac{\omega_c t}{2} + n\omega_1 t + \frac{n\pi}{2} \right) \sum_{m=-\infty}^{\infty} J_m(X_2) \cos \left(\frac{\omega_c t}{2} + m\omega_2 t + \frac{m\pi}{2} \right) \right] - \left[\sum_{n=-\infty}^{\infty} J_n(X_1) \sin \left(\frac{\omega_c t}{2} + n\omega_1 t + \frac{n\pi}{2} \right) \sum_{m=-\infty}^{\infty} J_m(X_2) \sin \left(\frac{\omega_c t}{2} + m\omega_2 t + \frac{m\pi}{2} \right) \right] \right\} \tag{19-14}$$

This equation may be written as

$$M(t) = A_c \Bigg\{ \sum_{n=-\infty}^{\infty} \sum_{m=-\infty}^{\infty} J_n(X_1) J_m(X_2) \left[\cos\left(\frac{\omega_c t}{2} + n\omega_1 t + \frac{n\pi}{2}\right) \cos\left(\frac{\omega_c t}{2} + m\omega_2 t + \frac{m\pi}{2}\right) - \sin\left(\frac{\omega_c t}{2} + n\omega_1 t + \frac{n\pi}{2}\right) \sin\left(\frac{\omega_c t}{2} + m\omega_2 t + \frac{m\pi}{2}\right) \right] \Bigg\} \quad (19\text{-}15)$$

By means of Eq. (19-12) further reduction can be obtained to give

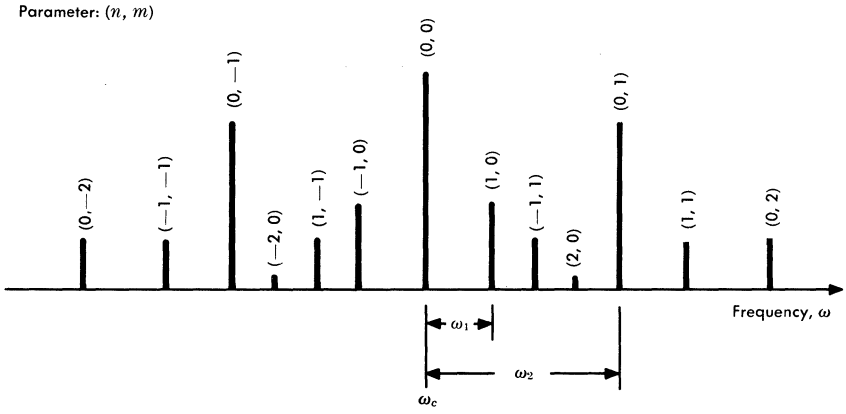
$$M(t) = A_c \sum_{n=-\infty}^{\infty} \sum_{m=-\infty}^{\infty} J_n(X_1) J_m(X_2) \cos\left[(\omega_c + n\omega_1 + m\omega_2) t + \frac{(n+m)\,\pi}{2} \right] \quad (19\text{-}16)$$

This equation is the desired result. It indicates that there will be sideband components displaced from the carrier by all possible multiples of the individual modulating frequencies. However, there will also be components displaced by all possible sums and differences of multiples of the modulating frequencies, and therefore superposition does not apply.

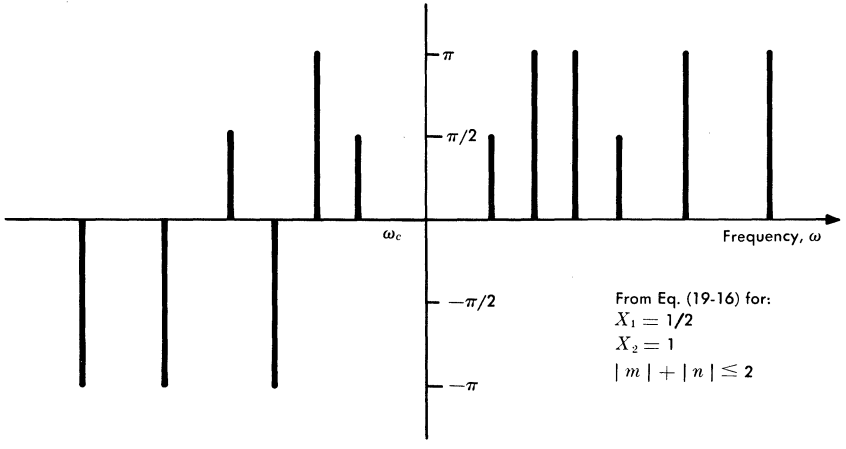
Figure 19-3 shows the amplitude and relative phase spectra of the zero, first, and second order components obtained from Eq. (19-16) for the condition $X_1 = 1/2$ and $X_2 = 1$. In the general case, the order of the component is equal to the sum of the magnitudes of the orders of the Bessel functions used to compute the amplitude of that component. For example, a second order component in Eq. (19-16) is any component for which $|m| + |n| = 2$.

Modulation by Three or More Sinusoids

The preceding discussion of two-sinusoid modulation illustrates that superposition is inapplicable to the angle-modulation problem. The complexity of solution increases rapidly as more modulating



(a) Amplitude spectrum



(b) Phase spectrum

FIG. 19-3. Amplitude and relative phase spectra for two-sinusoid modulation.

signals are added. It can be shown that for three-sinusoid modulation, the resultant formula similar to Eq. (19-16) will be

$$M(t) = A_c \sum_{n_1=-\infty}^{\infty} \sum_{n_2=-\infty}^{\infty} \sum_{n_3=-\infty}^{\infty} J_{n_1}(X_1) J_{n_2}(X_2) J_{n_3}(X_3)$$

$$\cos\left[(\omega_c + n_1\omega_1 + n_2\omega_2 + n_3\omega_3)t + (n_1 + n_2 + n_3)\frac{\pi}{2}\right] \quad (19\text{-}17)$$

The pattern established by Eq. (19-16) and Eq. (19-17), when extended to N sinusoids, leads to the general result

$$M(t) = A_c \sum_{n_1=-\infty}^{\infty} \cdots \sum_{n_N=-\infty}^{\infty} \prod_{r=1}^{N} J_{n_r}(X_r) \cdot \cos\left(\omega_c t + \sum_{r=1}^{N} n_r \omega_r t + \sum_{r=1}^{N} n_r \frac{\pi}{2}\right) \tag{19-18}$$

In this equation, the symbol Π denotes that all N of the J_{n_r} coefficients are multiplied together.

It is apparent from Eq. (19-18) that the determination of the spectral components of an FM or PM signal is a very formidable task even for relatively small values of N. Fortunately, in many practical cases of interest, the index of modulation is sufficiently low that the amplitudes of the various components can be obtained from approximate expressions derived from Eq. (19-18). These expressions are obtained by applying the series expansions of the Bessel functions to the amplitude coefficient for each basic spectral component. For example, the carrier frequency component is obtained from Eq. (19-18) by setting all $n_r = 0$. Then if A_0 denotes the carrier amplitude,

$$A_0 = A_c \prod_{r=1}^{N} J_0(X_r) \tag{19-19}$$

Equation (19-19) can be approximated by using the power series expansion given in Reference 3 as $J_0(X) = 1 - X^2/4 + X^4/64 \ldots$. This approximation, after multiplying, collecting terms, and neglecting all powers greater than fourth yields

$$A_0 \approx A_c\left(1 - \frac{1}{4}\sum_{r=1}^{N} X_r^2 + \frac{1}{16}\sum_{r=1}^{N-1}\sum_{s=r+1}^{N} X_r^2 X_s^2 + \frac{1}{64}\sum_{r=1}^{N} X_r^4\right) \tag{19-20}$$

Equation (19-20) can be written more compactly if an additional small approximation is made. This approximation is made by doubling the coefficient of the last term and thereby introducing an error equal to

$$1/64 \sum_{r=1}^{N} X_r^4$$

However, in so doing, the last two terms of the equation may be combined to reduce Eq. (19-20) to

$$A_0 \approx A_c\left[1 - \frac{1}{4}\sum_{r=1}^{N} X_r^2 + \frac{1}{32}\left(\sum_{r=1}^{N} X_r^2\right)^2\right] \tag{19-21}$$

The series expansion for e^{-a} is

$$e^{-a} = 1 - a + \frac{a^2}{2} - \cdots \tag{19-22}$$

It should now be observed that Eq. (19-21) has the same form as this series expansion. If the parameter D_ϕ is defined as

$$D_\phi = \frac{1}{2}\sum_{r=1}^{N} X_r^2 \tag{19-23}$$

it follows that Eq. (19-21) can be expressed as

$$\begin{aligned} A_0 &\approx A_c\left(1 - \frac{D_\phi}{2} + \frac{D_\phi^2}{8}\right) \\ &\approx A_c\, e^{-D_\phi/2} \end{aligned} \tag{19-24}$$

Equation (19-24) is the approximation desired. It is valid provided that the peak phase deviation, X_r, of each of the modulating signal components is small, thereby permitting the approximations used to obtain Eqs. (19-20) and (19-21). More specifically, it is required that $\sum_{r=1}^{N} X_r^2$ be less than unity. The parameter D_ϕ has an important significance. Since it is assumed that each modulating signal component is sinusoidal, the rms phase deviation produced by each component is equal to $X_r/\sqrt{2}$, or the mean square deviation is $X_r^2/2$. Hence, D_ϕ is the mean-square phase deviation of the total modulating signal, and $\sqrt{D_\phi}$ equals the rms phase deviation resulting from the total signal.

The techniques used to obtain Eq. (19-24) can be applied to find approximate expressions for the amplitudes of the sideband components. Figure 19-4 tabulates the results that can be obtained.

The utility of the approximations tabulated in Fig. 19-4 arises from the fact that numerical methods can be used to calculate the

power spectrum of the FM signal for any baseband signal, provided that (1) the baseband signal can be expressed as a finite summation of N sinusoids and (2) the mean-square phase deviation, D_ϕ, is no greater than 0.5. A baseband load consisting of multiplexed telephone channels may be simulated by representing the talkers by a number of uniformly spaced sinusoids with power level and baseband frequency corresponding to the power spectral density of the composite baseband load. Therefore, the power spectrum of an FM signal in a microwave system carrying multiplexed telephone channels and with mean-square phase deviation of less than 0.5 can be approximated by using the expressions listed in Fig. 19-4. In making the calculations, a power summation is made of all the products falling at each sideband frequency. The product-count results given in Chap. 10 are directly applicable.

Component	Frequency	Amplitude and relative phase
Zero order (carrier)	ω_c	$e^{-D_\phi/2}$
First order	$\omega_c \pm \omega_r$	$\pm \frac{1}{2} X_r e^{-D_\phi/2}$
Second order	$\omega_c \pm 2\omega_r$	$\frac{1}{8} X_r^2 e^{-D_\phi/2}$
Second order	$\omega_c \pm \omega_r \pm \omega_s$	$\pm \frac{1}{4} X_r X_s e^{-D_\phi/2}$
Third order	$\omega_c \pm 3\omega_r$	$\pm \frac{1}{48} X_r^3 e^{-D_\phi/2}$
Third order	$\omega_c \pm 2\omega_r \pm \omega_s$	$\pm \frac{1}{16} X_r^2 X_s e^{-D_\phi/2}$
Third order	$\omega_c \pm \omega_r \pm \omega_s \pm \omega_t$	$\pm \frac{1}{8} X_r X_s X_t e^{-D_\phi/2}$

$$D_\phi = \frac{1}{2} \sum_{r=1}^{N} X_r^2 \leq \frac{1}{2}$$

Sign of amplitude term is determined by giving X_r, X_s, and X_t the same signs as ω_r, ω_s, and ω_t.

FIG. 19-4. Approximate expressions for amplitudes of spectral components (relative to carrier) for N-sinusoid modulation.

Phase Modulation by a Band of Random Noise

When the baseband signal consists of many single-sideband, frequency-multiplexed telephone channels, it is often convenient to simulate the baseband signal by an equivalent band of random noise. The determination of the sideband spectrum when the modulating signal consists of random noise involves considerable analysis [4]. A particular case of interest, often assumed in the analysis of a radio system, is that of pure phase modulation by a band of random noise extending uniformly across the baseband from 0 to f_T hertz. For this case, the power density of the sideband spectrum has been shown to be

$$s(f) = \frac{e^{-D_\phi}}{2f_T}\left\{D_\phi \binom{1-x}{2}^0 + \frac{D_\phi^2}{2!}\binom{2-x}{2}^1 + \frac{D_\phi^3}{3!2!}\left[\binom{3-x}{2}^2 - 3\binom{1-x}{2}^2\right] + \frac{D_\phi^4}{4!3!}\left[\binom{4-x}{2}^3 - 4\binom{2-x}{2}^3\right] + \cdots\right\} \tag{19-25}$$

where

$s(f)$ = power spectral density (watts per hertz of bandwidth).

D_ϕ = mean-square phase deviation of the total noise signal (radians squared).

f_T = top baseband frequency (Hz).

$x = \dfrac{|f_c - f|}{f_T}$, where f_c is the carrier frequency.

$\binom{\ }{\ }$ indicates that the enclosed term goes to zero when the quantity $n - x$ is negative.

In this equation, the factor e^{-D_ϕ} represents the power in the modulated carrier relative to an unmodulated carrier of unity amplitude. Note that this is the same result as tabulated in Fig. 19-4 for the zero order (carrier) component, since the power is proportional to the square of the amplitude, and $(e^{-D_\phi/2})^2 = e^{-D_\phi}$.

Figure 19-5 shows the quantity $s(f)2f_T$ expressed in dB with respect to the unmodulated carrier, for several values of the root-mean-square phase deviation, $\sqrt{D_\phi}$. Although a curve is shown

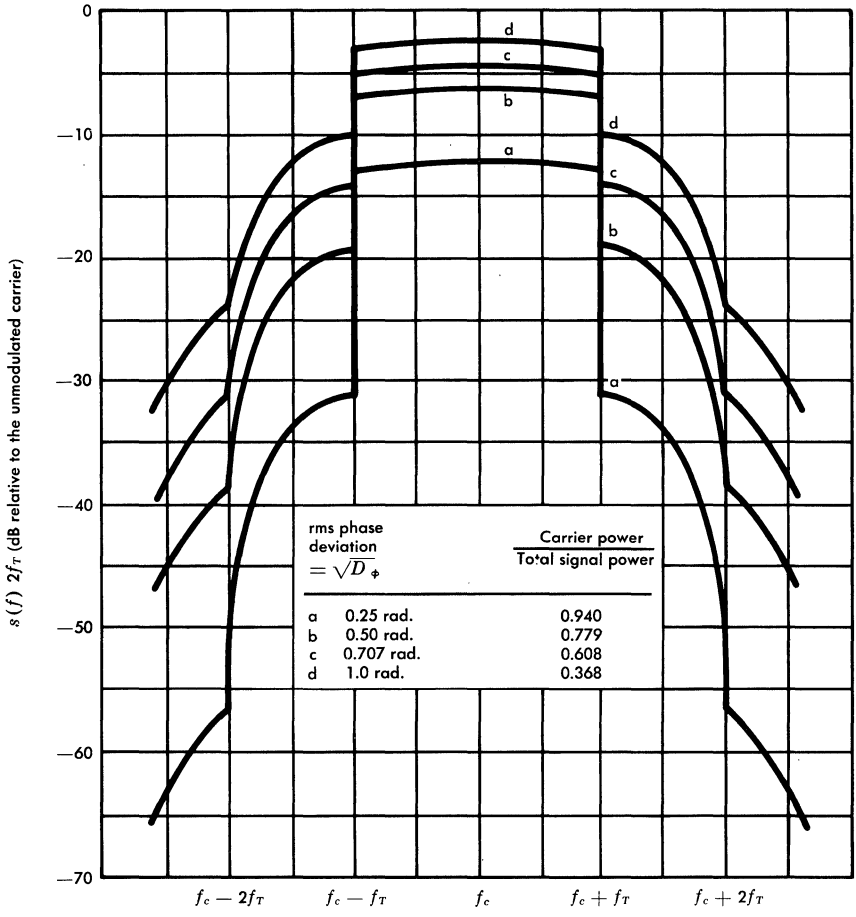


FIG. 19-5. Sideband spectra of a carrier phase-modulated by a baseband signal consisting of a flat band of random noise which extends from 0 to f_T Hz.

for $\sqrt{D_\phi} = 1$, approximations made in the analysis necessitate that $\sqrt{D_\phi} \leq 0.707$ (i.e., $D_\phi \leq 1/2$) for greatest accuracy. This is the same restriction imposed on the approximations made in deriving the coefficients for Fig. 19-4.

A qualitative understanding of the shape of the spectra in Fig. 19-5 can be obtained from the following considerations. First order sideband components are formed by the modulation of the carrier and the individual components of the baseband or modulating signal.

These sideband components will fall within the band bounded by $f_c \pm f_T$. Since the spectrum of the modulating signal has been assumed flat, the spectrum of the first order sideband components will also be flat versus frequency. This is true even though the amplitude of each sideband component will, of course, be a function of the amplitude of all the other components, as previously discussed in connection with Eq. (19-16).

Second order sideband components, which fall within the band bounded by $f_c \pm 2f_T$, arise from combinations involving the carrier frequency and any second order combination of baseband frequencies such as $\alpha + \beta$, $\alpha - \beta$, or 2α. The number of products formed is greatest in the vicinity of the carrier, with the result that the power in the second order sidebands is maximum around f_c and drops off to zero at frequencies greater than $f_c \pm 2f_T$.

In a similar manner, third order sideband components, which fall in the region bounded by $f_c \pm 3f_T$, arise from combinations of the carrier with third order combinations of baseband frequencies. Again, more products are formed near the carrier frequency, so that the power in the third order sidebands has a broad maximum in the $f_c \pm f_T$ portion of the spectrum and drops to zero at $f_c \pm 3f_T$.

The result of power addition of the higher order components to the first order sidebands accounts for the curvature in the spectrum between $f_c - f_T$ and $f_c + f_T$ in Fig. 19-5. Notice that this curvature increases in going from a low-phase deviation (curve a) to a high deviation (curve d). This is because the power in the second and third order sidebands builds up relatively rapidly as the phase deviation increases. This is analogous to the way second and third order modulation products increase relative to the fundamental as the input power to a nonlinear device is increased. The same effect accounts for the relatively slow falloff of higher order sidebands shown by curve d, as against the rapid falloff of curve a.

Quantitatively, the spectrum for an rms phase deviation in Fig. 19-5 is interpreted as follows. The modulated carrier power is understood to be $A_c^2 e^{-D_\phi}$. The region from $f_c - f_T$ to $f_c + f_T$ is dominated by first order sidebands of power density $A_c^2 s(f)$ per hertz of bandwidth. For example, given a phase-modulated signal of -30 dBm total power with rms phase deviation, $\sqrt{D_\phi}$, of 0.5 radian due to noise extending from 0 to 3 MHz, the carrier power would be $-30 + 10 \log 0.779 = -31.1$ dBm, and the sideband power density per 3 kHz in

the first order sideband region would be

$$\text{Density} = -30 \text{ dBm} - 7 \text{ dB} + 10 \log \frac{3 \times 10^3}{6 \times 10^6}$$

$$= -70 \text{ dBm/3 kHz} \qquad (19\text{-}26)$$

Spectra for High Modulation Index

In much of the preceding analysis, approximations have been made which have restricted the results to low-index systems, where the mean-square phase deviation is one-half or less. The spectrum of the FM signal may be approximated by an alternative technique applicable to sufficiently high-index systems and generally known as the quasi-stationary method. The basic idea may be illustrated by considering an FM signal in which the carrier is 70 MHz and the modulating signal is a 100-Hz square wave which deviates the carrier ±1 MHz. Then, half of the time the FM signal is at 71 MHz and half of the time at 69 MHz. Thus, the spectrum would have two spikes, or concentrations of power, at 69 MHz and at 71 MHz. Each would carry half the total power of the unmodulated carrier. If the modulating signal is triangular so that the frequency sweeps linearly back and forth between 69 MHz and 71 MHz, the spectrum is clearly essentially continuous and of uniform amplitude between 69 MHz and 71 MHz.

Although the quasi-stationary approach often gives useful results, it should always be viewed with suspicion and used with caution. For low-index systems, it gives wrong results and should be rejected outright. For example, suppose the 70-MHz carrier is to be deviated ±100 kHz by a 1-MHz square wave. Then the quasi-stationary method says that half the power is at 69.9 MHz and half at 70.1 MHz. This is completely wrong. There is no power at those frequencies. A correct analysis, using Eq. (19-18) or the approximations of Fig. 19-4, shows the spectrum to have components spaced at 1-MHz intervals around 70 MHz. In fact, to pass a 1-MHz square wave reasonably well, spectrum components out to ±10 MHz from the carrier would need to be transmitted.

The quasi-stationary approach may often be used when the index is greater than 10, and the low-index approach of Fig. 19-4 may be used when the mean-square deviation, D_ϕ, is less than one-half. This leaves an area of medium indices where no suitable approximation

is at present known and for which any analytical approach becomes quite difficult. One technique used is to examine a low-index case and a high-index case and argue that the medium-index case falls between.

19.2 PHASOR REPRESENTATION OF ANGLE MODULATION

Phasor diagrams of amplitude- and angle-modulated signals were introduced in Chap. 5 to give physical representations for the analysis. In the following discussions similar diagrams will be presented as a lead-up to the diagram for higher-index angle modulation.

Low Modulation Index

An amplitude-modulated signal with sinusoidal modulation of index m may be represented by the time function

$$M(t) = (1 + m \cos \omega_1 t) \cos \omega_c t$$

$$= \cos \omega_c t + \frac{m}{2} \cos (\omega_c + \omega_1) t + \frac{m}{2} \cos (\omega_c - \omega_1) t \quad (19\text{-}27)$$

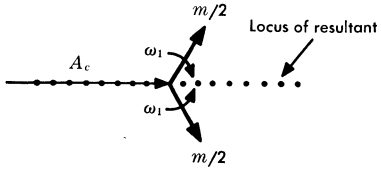


FIG. 19-6. AM phasor representation.

The phasor representation of this signal, as shown in Fig. 19-6, consists of a carrier and two contra-rotating sideband signals.

With amplitude modulation, the carrier signal is unchanged by the modulation process. Instead, the total signal power increases and the information appears as additional energy distributed in sidebands. Variations in the amplitude modulation index, m, do not change the number of sideband components, only their amplitude.

For comparison, a sinusoidally modulated angle-modulation signal is represented in Fig. 19-7. The dotted line in Fig. 19-7 is the locus of the resultant formed by the carrier and the first order sidebands. This approximation is good only for very low values of phase modulation index, X, and is used in Chap. 20 to analyze the effects of random noise.

Higher Modulation Index

A phasor diagram such as the one shown in Fig. 19-8 where $X=1$ radian may be used to illustrate the various sideband relationships derived previously for single-sinusoid modulation. This figure should be compared with Fig. 19-7 which represents the low-index case.

The locus of the resultant five-component approximation is curved and closely follows the signal locus, which by definition is a segment of a circle with radius equal to the amplitude of the unmodulated carrier. It is important to recognize that the resultant signal amplitude and consequently the total signal power do not change with phase or frequency modulation. Instead, the power originally in the unmodulated carrier is redistributed among the carrier and its sidebands. A continuous improvement in signal-to-noise ratio is obtained in angle modulation systems with increasing peak deviation, as a result of the coherent or in-phase addition of many sideband pairs in contrast with the random addition of noise components.

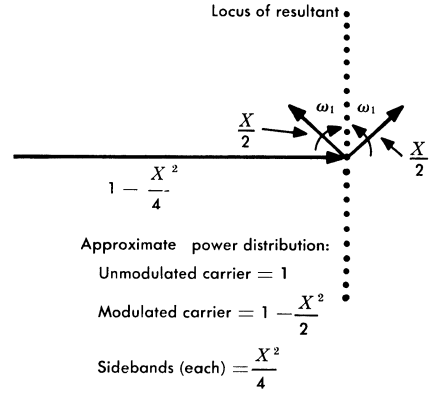


FIG. 19-7. Low-index FM/PM phasor diagram.

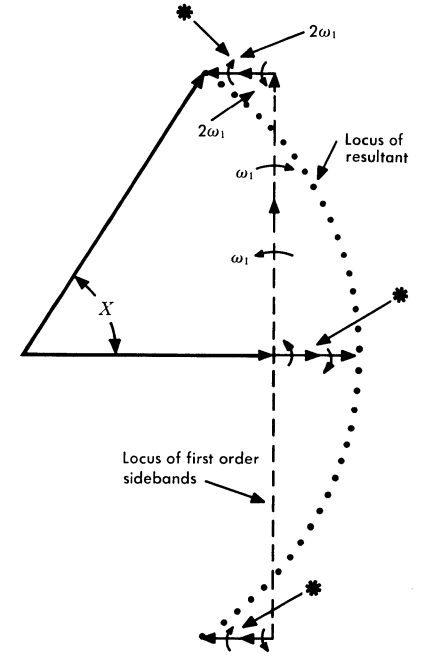


FIG. 19-8. Phasor diagram for $X = 1$.

19.3 EFFECTS OF LIMITING

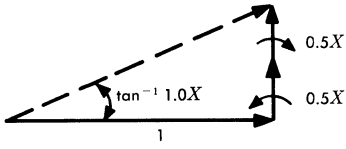
It was shown in Chap. 10 that an FM signal could pass through a device with a nonlinear amplitude transfer characteristic without significant distortion provided the carrier frequency was properly chosen with respect to the top base-

band frequency and peak frequency deviation. However, some devices used in FM systems, such as traveling wave tube amplifiers, have a phase nonlinearity called AM/PM conversion, which can produce significant distortion when the FM signal is also amplitude-modulated. Since it is impossible to prevent residual AM in radio relay systems, limiters are used to suppress the AM prior to the major AM/PM conversion devices.

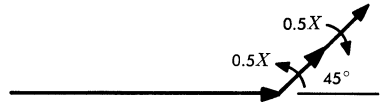
A limiter is a highly nonlinear device which suppresses any incidental amplitude modulation of a carrier with little effect on phase modulation. An *ideal* limiter would remove AM completely and have no effect at all on PM or FM. This is strictly a mathematical concept. A *real* limiter reduces AM to a fraction of its original value. In a good design, the AM index may be reduced by a factor of 100; this is frequently referred to as 40 dB of limiting (i.e., 20 log 100 = 40 dB). Furthermore, in an actual limiter there is some conversion of the AM at the input to PM at the output. In a good limiter the PM index will be only a small fraction of the AM index, perhaps 2 per cent; this is often measured in degrees. In this example, the AM/PM conversion is 0.13 degrees per dB. In contrast, a traveling wave tube may have 6 degrees per dB of AM/PM conversion.

A basic difficulty in the application of limiters to FM systems is that they are highly nonlinear. As a result, many familiar concepts based on the principles of tandem linear networks have to be abandoned. As an illustration of this, consider a transmission phase characteristic which rotates each of the sideband phasors of a low-index FM signal by 45 degrees clockwise. Figure 19-9(a) shows the undistorted signal at $t = 0$; here the carrier amplitude is assumed to be unity, and use is made of the low-index approximation to represent the first order sidebands as $0.5X$, where X is the modulation index. The phasor diagram after transmission distortion is shown in Fig. 19-9(b). Each sideband phasor can be resolved into components as shown in Fig. 19-9(c), and rearranged as in Fig. 19-9(d). This is clearly a combination of AM and PM. An *ideal* limiter removes the AM components, leaving Fig. 19-9(e). This is a pure PM signal, but the index of modulation (i.e., peak phase deviation) has been reduced from X to $0.707X$. The baseband output will thus be reduced 3 dB.

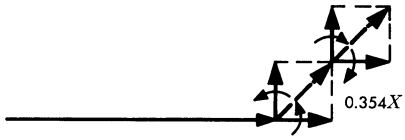
Phase distortion in combination with a limiter has thus produced amplitude compression at baseband. There is no phase equalization which can restore the sidebands in Fig. 19-9(e) to their original



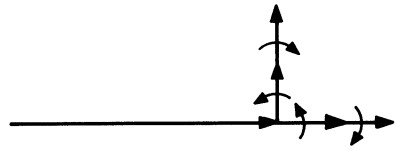
(a) Undistorted wave



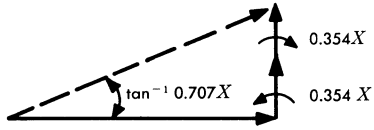
(b) Phase distortion = −45° for each sideband



(c) Resolution into components



(d) Component phasors rearranged



(e) After limiting

FIG. 19-9. Effects of phase distortion and limiting (phase modulation index = X, $t = 0$).

amplitude of $0.5X$; only a gain equalizer can do that. In fact, it is clear that a phase equalizer, placed in the system somewhere between limiters and intended to correct for the original phase distortion (by rotating each sideband phasor 45 degrees counterclockwise), will result only in a repetition of the process, so that a second 3-dB transmission loss at baseband will ensue.

In general, it can be demonstrated that there is no one-to-one correspondence between the transmission characteristic of the network ahead of the limiter and the necessary equalizer which follows. As a result, measurements for equalization purposes which are made through limiters by ordinary sweep-frequency techniques are not always useful. One partial solution to this problem is the use of

a frequency-modulated carrier which is swept slowly across the band of interest. The visual delay and the differential gain and phase measuring sets use this approach.

REFERENCES

1. Black, H. S. *Modulation Theory* (Princeton, N. J.: D. Van Nostrand Company, Inc., 1953).
2. Goldman, S. *Frequency Analysis, Modulation, and Noise* (New York: McGraw-Hill Book Company, Inc., 1948).
3. Dwight, H. B. *Mathematical Tables* (New York: Dover Publications, 1958).
4. Abramson, N. "Bandwidth and Spectra of Phase- and Frequency-Modulated Waves," *Trans. IEEE, Communications Systems,* vol. CS-11 (Dec. 1963), pp. 407-414.

Chapter 20

Random Noise in FM and PM Systems

From earlier discussions of random noise it will be recalled that thermal noise determines a lower limit to the random noise level in any electrical circuit and that additional noise may be expected from other sources such as electron tubes and transistors. In this chapter, the effect of random noise in phase- and frequency-modulated systems is considered.

The chapter begins with the development of a very useful system equation for predicting the noise at baseband which occurs when an FM signal is perturbed by additive random noise that has a constant spectral density. This equation is commonly used for thermal noise calculations in FM systems. Other sources of random noise with nonconstant spectral densities, such as shot noise and $1/f$ noise, are considered conceptually in the discussion of pre-emphasis. Also discussed in this chapter is the phenomenon of breaking.

20.1 DEVELOPMENT OF BASIC FM SYSTEM NOISE EQUATION

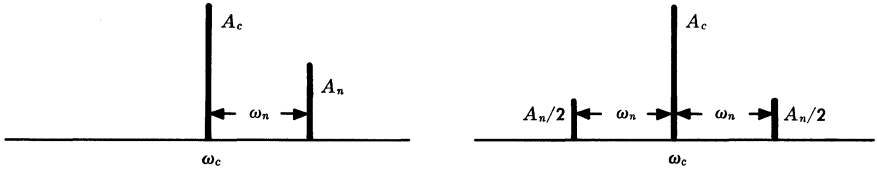
When random noise with constant spectral density is added to an FM signal, an unwanted modulation of the carrier occurs. The amount of unwanted carrier deviation depends on the relative magnitudes of the carrier and the noise. Upon demodulation, this unwanted modulation becomes a random noise at baseband whose spectral shape depends on whether an FM or PM demodulator is used. At the output of a PM system, the noise voltage is flat with frequency, whereas the noise voltage at the output of an FM system increases linearly with frequency. This is commonly referred to as the triangular noise spectrum of an FM system. These facts are developed

in the following analysis which leads to an equation for the random noise at 0 TL in an FM system.

Unwanted Modulation of a Carrier

The analysis begins by examining the unwanted phase modulation produced on an otherwise unmodulated carrier by an interfering sinusoid. The principles of the analysis are then extended to determine the modulation caused by a flat band of random noise.

PM Due to an Interfering Sinusoid. Consider the case where a sinusoid of peak amplitude A_n is separated by a frequency ω_n from a carrier of peak amplitude A_c. This case is shown in Fig. 20-1(a).



(a) Before limiting (b) After limiting

FIG. 20-1. Spectrum—carrier plus interfering sinusoid.

The analysis proceeds under the assumption that $A_c >> A_n$, and therefore the phase deviation, θ, shown in Fig. 20-2 can be assumed equal to tan θ. Thus, the peak phase deviation due to a single interfering sinusoid is

$$\text{Peak phase deviation} = \frac{A_n}{A_c} \quad \text{rad} \tag{20-1}$$

Also shown in Figs. 20-1 and 20-2 is the result of limiting the amplitude of the composite signal. The effect is to turn the single interfering sinusoid into a pair of contrarotating sidebands of amplitude $A_n/2$. These sidebands are coherent; thus, the peak deviation of the carrier is still A_n/A_c. Limiting, therefore, removes amplitude components from the signal and reduces its total power but does not reduce the interference on the recovered signal.

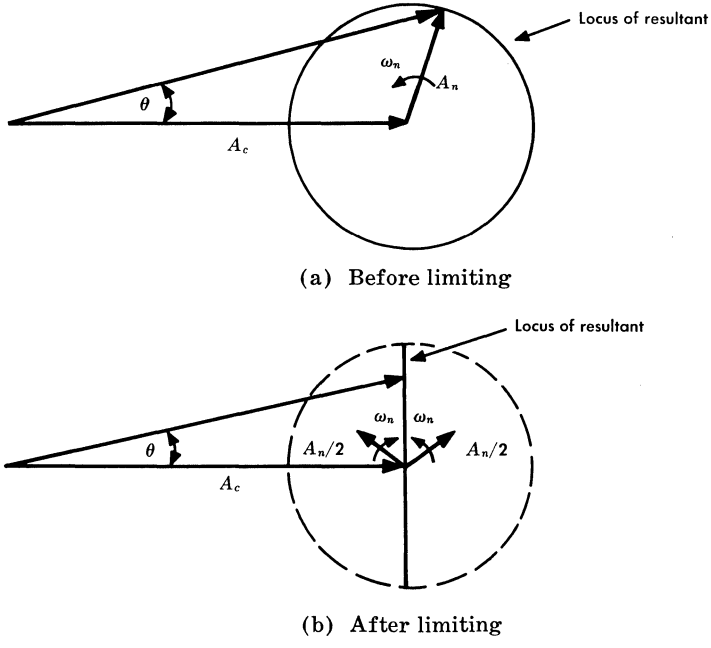


FIG. 20-2. Phasor diagram—carrier plus interfering sinusoid.

Since the phase modulation is sinusoidal, the rms phase deviation is equal to the peak phase deviation divided by $\sqrt{2}$. Hence,

$$\text{Rms phase deviation} = \frac{A_n}{A_c\sqrt{2}} \quad \text{rad}$$

$$= \frac{a_n}{A_c} \quad \text{rad} \tag{20-2}$$

where $a_n = A_n/\sqrt{2}$ is the rms amplitude of the interference. The rms phase deviation will be useful later since, for random noise, the rms voltage is more easily defined than the peak voltage.

Later, use will be made of the fact that a noise component at a frequency of either $\omega_c + \omega_n$ or $\omega_c - \omega_n$ produces a baseband output at the same frequency, ω_n. When two such noise components are simultaneously present (as they usually are), they add on a power basis since, in general, they arise from uncorrelated voltages.

FM Due to an Interfering Sinusoid. The frequency modulation produced by a sinusoidal interference is easily obtained from a knowledge of the phase modulation since the instantaneous frequency deviation is defined as the time derivative of the instantaneous phase deviation. When the carrier is much larger than the interference,

$$\text{Instantaneous phase deviation} = \frac{A_n}{A_c} \sin(\omega_n t + \theta_n) \text{ rad} \tag{20-3}$$

Taking the derivative with respect to time,

$$\text{Instantaneous frequency deviation} = \frac{A_n}{A_c} \omega_n \cos(\omega_n t + \theta_n) \text{ rad/sec} \tag{20-4}$$

Therefore,

$$\text{Peak frequency deviation} = \frac{A_n}{A_c} \omega_n \quad \text{rad/sec}$$

$$= \frac{A_n}{A_c} f_n \quad \text{Hz} \tag{20-5}$$

The rms frequency deviation resulting from the interfering sinusoid is

$$\text{Rms frequency deviation} = \frac{A_n}{A_c \sqrt{2}} \omega_n \quad \text{rad/sec}$$

$$= \frac{A_n}{A_c \sqrt{2}} f_n \quad \text{Hz}$$

$$= \frac{a_n}{A_c} f_n \quad \text{Hz} \tag{20-6}$$

The peak frequency deviation is a function of the difference frequency, f_n. Consequently, sinusoidal components which are well displaced from the carrier frequency produce larger frequency deviations than sinusoidal components close to the carrier frequency. For the rms phase deviation, it is sufficient to know the ratio of the rms interference voltage, a_n, to the peak carrier voltage, A_c. For the rms frequency deviation, it is necessary to take into account the frequency of the interference.

In Chap. 5 it was pointed out that the index of modulation (or peak phase deviation) of a carrier modulated by a single sinusoid can be expressed as the peak frequency deviation divided by the modulating signal frequency. Comparison of Eqs. (20-1) and (20-5) shows that the same relation applies here for the case of modulation resulting from the presence of an interfering sinusoid. Thus, for sinusoidal modulation a useful relation to remember is:

$$\begin{matrix}\text{Peak phase deviation} \\ \text{(or index of modulation)}\end{matrix} = \frac{\text{peak frequency deviation}}{\text{modulating frequency}} \tag{20-7}$$

PM Due to Random Noise. The effects of interference due to a band of random noise about the carrier are considered next. Of interest are (1) the total noise which appears in the baseband (important for television transmission, for example) and (2) the noise in a particular baseband slot (for example, the noisiest channel in a telephone multiplex group).

The method of analysis developed thus far in this chapter is directly applicable [1]. The random noise can be assumed to consist of a sufficiently large number of sinusoidal components of incommensurable frequency, of equal amplitude, and of arbitrary phase. It is convenient to analyze the system noise on a per-hertz basis; thus, a band of noise N hertz wide will be thought of as equivalent to N approximately uniformly-spaced sinusoids. Let A_n equal the peak amplitude of a sinusoid having the same power as a band of noise one hertz wide. Then, if it is again assumed that the total noise power or, more specifically, the peak total noise amplitude is small relative to the carrier, the resulting phase modulation is approximately

$$\sum_{n=1}^{N} \frac{A_n}{A_c} \sin\,(\omega_n t + \theta_n) \qquad \text{rad} \tag{20-8}$$

For the conditions assumed, superposition holds. That is, in the random noise case the phase modulation of the carrier is equal to the summation of the phase modulation components which would have been produced by the input noise components individually. It is shown later that if the noise power is not small relative to the carrier power, nonlinear noise effects will occur.

When a single interfering sinusoid was being considered, it was possible to write directly an expression for the peak phase deviation,

Eq. (20-1). The peak value of N interfering sinusoids can be defined only if a known phase relationship exists between the sinusoids, which is not the case here. Thus, for random noise it is not possible to write an expression for the peak phase deviation which is analogous to Eq. (20-1). The rms phase deviation, however, can be defined. Let the total rms voltage produced by the N sinusoids be α_N. For the case assumed, where $A_1 = A_2 = A_n$, it follows that this rms voltage is

$$\alpha_N = \sqrt{\sum_{n=1}^{N} \left(\frac{A_n}{\sqrt{2}}\right)^2}$$

$$= \sqrt{N\left(\frac{A_n}{\sqrt{2}}\right)^2}$$

$$= \frac{A_n}{\sqrt{2}}\sqrt{N}$$

$$= a_n\sqrt{N} \qquad (20\text{-}9)$$

where $a_n = A_n/\sqrt{2}$ is the rms amplitude of each sinusoidal component. The rms phase deviation of Eq. (20-8) can then be written as

$$\text{Total rms phase deviation due to band of random noise} = \frac{\alpha_N}{A_c} \quad \text{rad}$$

$$= \frac{A_n}{A_c\sqrt{2}}\sqrt{N} \quad \text{rad}$$

$$= \frac{a_n}{A_c}\sqrt{N} \quad \text{rad} \qquad (20\text{-}10)$$

When $N = 1$, the above expressions reduce to those previously defined for the rms phase deviation due to a single sinusoid; that is,

$$\text{Rms phase deviation due to a 1-hertz band of noise} = \frac{A_n}{A_c\sqrt{2}} \quad \text{rad}$$

$$= \frac{a_n}{A_c} \quad \text{rad} \qquad (20\text{-}11)$$

Noise at Baseband Frequencies

The preceding analysis can now be applied to the specific problem of deriving equations for the random noise in the baseband of both PM and FM systems.

PM System Noise. In the following discussion, it is assumed that the random noise is flat versus frequency over the band from $f_c - f_1$ to $f_c + f_1$, as shown in Fig. 20-3(a), and that the carrier power is much greater than the noise power. The rms phase deviation due to the noise in a 1-hertz band at $f_c + f_n$, where $f_n \leq f_1$, is a_n/A_c radians, as shown in Eq. (20-11). This phase modulation will cause an rms noise voltage to appear at the output of a PM detector in a 1-hertz band centered at baseband frequency f_n. A second noise voltage of equal magnitude will also appear at f_n due to the noise in a 1-hertz band centered at $f_c - f_n$. Since the two noise voltages are uncorrelated, they will add on a power basis. Thus, the total rms noise voltage in a 1-hertz band centered at frequency f_n will be proportional to $a_n \sqrt{2}/A_c$. The ratio $a_n \sqrt{2}/A_c$ is, of course, the rms phase deviation produced by two 1-hertz bands of noise [Eq. (20-10) for the case $N = 2$], where the specific interpretation has been made that one band is f_n hertz above the carrier and the other is f_n hertz below the carrier. Thus, for a PM system, at the output of a radio receiver with a phase modulation detector,

$$\text{Rms noise voltage in a 1-hertz band centered at baseband frequency } f_n = \frac{1}{k} \frac{a_n \sqrt{2}}{A_c} \tag{20-12}$$

where

$\frac{1}{k}$ = the transfer constant of a PM detector expressed in volts per radian of phase deviation.

and

$\frac{a_n \sqrt{2}}{A_c}$ = the rms phase deviation, in radians, due to two bands of noise, each 1 hertz wide and located at $f_c + f_n$ and $f_c - f_n$. The carrier frequency is f_c.

It is evident from this analysis that in a PM system the baseband noise voltage per hertz due to flat random noise is independent of f_n. Hence, the noise spectrum at the output of a PM detector is flat versus frequency. The total baseband noise voltage is equal to the equivalent voltage resulting from the power summation of all the

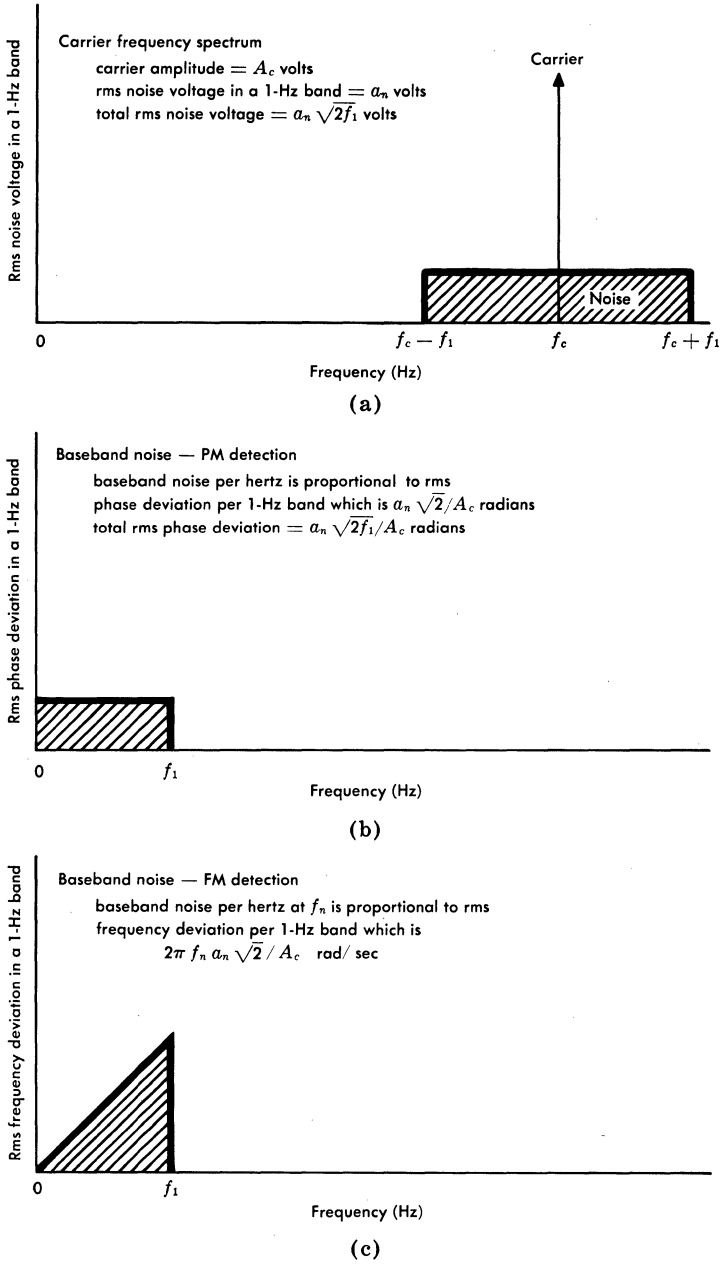


FIG. 20-3. Addition of a flat band of noise to a carrier and resultant baseband noise in PM and FM systems.

1-hertz band noise voltages. If the baseband extends from 0 to f_1 hertz, the total rms noise voltage at the output of the PM detector will be the power summation of f_1 voltages, each of which has an rms amplitude given by Eq. (20-12). Thus,

$$\text{Total rms noise voltage in PM system baseband} = \sqrt{f_1 \left(\frac{1}{k}\frac{a_n\sqrt{2}}{A_c}\right)^2}$$

$$= \frac{1}{k}\frac{a_n}{A_c}\sqrt{2f_1} \qquad (20\text{-}13)$$

where

$\frac{a_n}{A_c}\sqrt{2f_1}$ = the total rms phase deviation, in radians, due to a band of noise extending from $f_c - f_1$ to $f_c + f_1$.

Note that this is the same as Eq. (20-10) for the case $N = 2f_1$.

Figure 20-3(b) illustrates the baseband noise spectrum for a PM system. If multiplexed telephone channels were transmitted over such a system, all channels would have equal noise. To find the noise in dBrnc0, a relationship between the rms phase deviation in the radio system and the power at the zero transmission level point is needed. This relationship could be stated in a number of ways; a similar problem occurs in the FM case and is discussed in more detail later.

FM System Noise. As in the case of the PM system noise derivation, it will be assumed that the random noise is flat versus frequency over the band from $f_c - f_1$ to $f_c + f_1$, as shown in Fig. 20-3(a), and that the carrier power is much greater than the noise power. The baseband noise voltage in an FM system is proportional to the rms frequency deviation rather than to the rms phase deviation of the carrier. The instantaneous frequency deviation caused by the noise can be obtained by differentiating the expression for the instantaneous phase deviation, Eq. (20-8). The result would be a summation of N sinusoidal terms, each of which has a peak amplitude of $\omega_n A_n/A_c$ or an rms amplitude of $\omega_n a_n/A_c$ volts. These amplitudes are, of course, the same as those for the single-sinusoid interference case, Eqs. (20-5) and (20-6), and represent the peak and rms frequency deviations, respectively, which are caused by the noise at frequency $f_c + f_n$ or $f_c - f_n$. The total rms noise voltage in a 1-hertz band centered at f_n at the output of an FM detector will

be the power summation of the rms noise voltages due to the noise at both of these frequencies. Thus, for an FM system, at the output of an FM detector,

$$\text{Rms noise voltage in a 1-hertz band centered at baseband frequency } f_n = \frac{1}{k_1}\frac{\omega_n a_n \sqrt{2}}{A_c} \tag{20-14}$$

where

$\frac{1}{k_1}$ = the transfer constant (or deviation sensitivity) of the FM detector expressed in volts per radian per second of frequency deviation.

and

$\frac{\omega_n a_n \sqrt{2}}{A_c}$ = the rms frequency deviation, in radians per second, due to two 1-hertz bands of noise, one at $f_c + f_n$ and the other at $f_c - f_n$.

Note that the baseband noise voltage per hertz in an FM system varies directly with ω_n and therefore increases linearly with baseband frequency. This is the so-called triangular noise spectrum of an FM system and is illustrated in Fig. 20-3(c).

The total rms noise voltage at the output of the FM detector will be directly proportional to the total rms frequency deviation of the carrier. The total rms frequency deviation may be obtained by integrating the mean-square frequency deviation produced by each 1-hertz band of noise and then taking the square root. This is, of course, analogous to a power summation of the components. Thus, if the baseband extends from 0 to f_1 hertz, the total rms frequency deviation is given by

$$\text{Total rms frequency deviation of carrier} = \sqrt{\int_0^{f_1} \left(\frac{2\pi f a_n \sqrt{2}}{A_c}\right)^2 df}$$

$$= \frac{2\pi a_n \sqrt{2}}{A_c}\sqrt{\frac{f_1^3}{3}} \quad \text{rad/sec}$$

$$= \frac{a_n \sqrt{2}}{A_c}\sqrt{\frac{f_1^3}{3}} \quad \text{Hz} \tag{20-15}$$

The total rms frequency deviation for a portion of the spectrum is found by changing the limits of integration. A case of particular

interest is that in which the bandwidth under consideration is small compared to its separation from the carrier. For example, the top channel in a telephone multiplex signal might occupy a 3-kHz band at a baseband frequency of around 4 MHz in a typical radio system. Let the bandwidth at some baseband frequency, f_1, be equal to δf, where $\delta f << f_1$. For this case, the rms frequency deviation will be approximately equal to the power summation of δf components, each having an amplitude given by Eq. (20-14). Thus,

Rms frequency deviation due to two δf bands of noise, one at $f_c + \omega_1/2\pi$ and the other at $f_c - \omega_1/2\pi$

$$\approx \sqrt{\delta f \left(\frac{\omega_1 a_n \sqrt{2}}{A_c}\right)^2}$$

$$\approx \frac{\omega_1 a_n \sqrt{2\delta f}}{A_c} \quad \text{rad/sec}$$

$$\approx \frac{a_n f_1 \sqrt{2\delta f}}{A_c} \quad \text{Hz} \qquad (20\text{-}16)$$

In practice, the baseband for radio systems extends from f_B to f_T, and the relationship $\delta f << f_1$ is valid over this region for any given message channel. Therefore, Eq. (20-16) is not restricted to the top message channel.

It frequently happens that the carrier and noise levels are known in terms of power rather than voltage. For this reason, Eq. (20-16) will be written in an alternative form using the following relations:

$$p_n = \frac{a_n^2}{R} \quad \text{watts/hertz} \qquad (20\text{-}17)$$

and

$$p_c = \frac{A_c^2}{2R} \quad \text{watts} \qquad (20\text{-}18)$$

Here, R is the circuit impedance at the point where the noise and carrier voltages are defined, while p_n and p_c are the corresponding noise and carrier average powers. Substituting in Eq. (20-16) for a_n and A_c gives the following alternative expression for the

rms frequency deviation due to the two noise bands, each of width δf.

$$\text{Rms frequency deviation due to two } \delta f \text{ bands of noise, one at } f_c + \omega_1/2\pi \text{ and the other at } f_c - \omega_1/2\pi \approx f_1 \sqrt{\frac{p_n\,\delta f}{p_c}} \text{ Hz} \qquad (20\text{-}19)$$

Noise at 0 TL in an FM System

To determine the noise in dBrnc0 from Eq. (20-16) or Eq. (20-19), a relationship between the rms frequency deviation of a signal in the radio system and the power of that signal at 0 TL is needed. If the deviation sensitivity, $1/k_1$, of the FM detector and the transmission level at the output of the detector were both known, the noise power at 0 TL could be determined. However, it is unlikely that either of these values will be known during the early design stages of a radio system. As an alternative, use can be made of the relationship between the multiplex signal which the system is to handle and the peak frequency deviation which this signal will produce. Chapter 9 discusses and shows how to determine the peak load, P_s, which a multiplexed telephone system must be designed to carry. If this load is applied to an FM system, the peak frequency deviation produced in the system is taken to correspond to the peaks of a sinusoidal baseband signal having a power of P_s dBm0. The peak frequency deviation for the system is usually established early in the design, at least tentatively, so that this factor together with the value of P_s for the load to be carried will permit the system noise to be evaluated. If the peak deviation is ΔF and the variation is sinusoidal, the rms deviation is $\Delta F/\sqrt{2}$. Let $\overline{F}$ denote this rms value, which will now be assumed to correspond to the power (or rms voltage) of P_s dBm0. Clearly, any rms frequency deviation equal to $\overline{F}$ will produce P_s dBm0. Conversely, the baseband power, P_N, in dBm0 produced by any other rms frequency deviation, $\overline{f}$, can be expressed as follows:

$$P_N = P_s + 20 \log \frac{\overline{f}}{\overline{F}} \qquad \text{dBm0} \qquad (20\text{-}20)$$

where

$\overline{F}$ = rms frequency deviation produced by the baseband signal, P_s

$\overline{f}$ = rms frequency deviation produced by any other signal having a power of P_N dBm0.

It follows, then, that if $\bar{f}$ represents the rms frequency deviation due to a band of noise, P_N will represent the resulting baseband noise power. For the case of noise in any message channel, an approximate expression for $\bar{f}$ is given by Eq. (20-19). Substituting this equation in Eq. (20-20) gives

$$P_N = P_s + 20 \log \frac{f_1}{\bar{F}} \sqrt{\frac{p_n \delta f}{p_c}} \qquad \text{dBm0} \qquad (20\text{-}21)$$

In practice, p_n and p_c are usually expressed in dBm per hertz and dBm, respectively. Furthermore, reference is usually made to the peak frequency deviation, $\Delta F = \sqrt{2}\ \bar{F}$, rather than the rms deviation, $\bar{F}$. For these reasons it is convenient to rewrite Eq. (20-21) in the form

$$P_N = P_s + 20 \log \frac{f_1}{\Delta F} + 10 \log 2\delta f + P_n \text{ dBm/Hz} - P_c \text{ dBm} \qquad \text{dBm0} \qquad (20\text{-}22)$$

Equation (20-21) or Eq. (20-22) gives the noise in dBm0 in a narrow band δf hertz wide due to flat random noise. The multiplexed telephone channel is located at baseband frequency f_1, and ΔF is the peak frequency deviation of the FM system. Normally, multiplexed telephone channels are spaced at 4-kHz intervals and have an effective noise bandwidth of 3 kHz. Thus, δf is normally assumed to be 3 kHz.

Equation (20-21) is the basis for a simple procedure which is occasionally used to determine the noise at 0 TL due to flat random noise in the FM portion of the system. Substituting $\Delta F = \sqrt{2}\ \bar{F}$ permits writing Eq. (20-21) as

$$P_N = P_s + 20 \log \frac{f_1}{\Delta F} + 10 \log \frac{2\delta f p_n}{p_c} \qquad \text{dBm0} \qquad (20\text{-}23)$$

The calculation is first made for a message channel located at a baseband frequency equal to the peak frequency deviation. This makes the term 20 log $f_1/\Delta F$ equal zero. The ratio of carrier power to noise power in a band $2\delta f$ then determines the quantity 10 log $p_c/2\delta f p_n$ dB. The noise at 0 TL is then this number of dB below P_s. Since the noise varies at 6 dB per octave as a function of channel location, it is a simple matter to determine the noise in any other

channel. In practice, when p_n and p_c are given in dBm per hertz and dBm, respectively, Eq. (20-22) rather than Eq. (20-21) is used for this calculation.

Equation (20-23) may be modified to obtain the annoyance in a single message channel for a single radio hop by substituting 3 kHz for δf and noting that 0 dBm0/3 kHz = 88 dBrnc0. This gives

$$W_N = 88 + P_s + 20 \log \frac{f_1}{\Delta F} + 37.8 + 10 \log \frac{p_n}{p_c} \quad \text{dBrnc0}$$

$$W_N = 125.8 + P_s \text{ dBm0} + 20 \log \frac{f_1}{\Delta F} + P_n \text{ dBm/Hz} - P_c \text{ dBm} \quad \text{dBrnc0} \quad (20\text{-}24)$$

20.2 SYSTEM NOISE AND PRE-EMPHASIS

In the preceding discussion an equation was derived which is frequently used for calculating the thermal noise contribution of a single repeater in a multirepeater FM system. There are, of course, other sources of noise in repeaters and terminals. Some of these other noise contributors are briefly discussed in the following, and the means of adding random noise in multirepeater systems is described. Also presented is a standard method for improving the noise performance of an FM system. This method, called pre-emphasis, is a level shaping process which approximately equalizes the noise in the various message channels.

Sources of Noise

There are numerous sources of random noise in FM systems. Thermal noise has already been mentioned. Some other sources are:

1. Shot noise, which appears in any oscillator as a result of the irregularity of electron flow and the bandwidth of the oscillator resonant circuit. It is usually characterized in FM systems by a flat baseband noise spectrum which crosses the triangular noise spectrum at a low frequency.
2. Flat and $1/f$ noise, which are introduced by baseband amplifiers in entrance links, FM terminals, and baseband repeaters.

By careful design, these contributors are held to values considerably lower than the top message circuit noise caused by thermal

noise at the receiver input. However, at mid-baseband and lower frequencies, these secondary sources dominate and therefore must be included in the overall design.

The summation of the various noise spectra resulting from front-end thermal noise, oscillator shot noise, baseband amplifier noise, and any other source *not* under the influence of the presence of modulation is called *idle noise*. This summation is performed on a power basis because of the uncorrelated nature of the various noise sources.

Addition of Random Noise in Multiple Hops

The random noise introduced in each radio hop will be superimposed on the desired modulating signal on a power basis. If the system is designed to have identical received power levels at each receiver and the repeaters perform identically, the noise will add at a rate of 10 log N dB where N is the number of hops.

Pre-emphasis and De-emphasis

The effect of pre-emphasis on idle noise is readily predictable and is illustrated in the following discussion. The effect of pre-emphasis on intermodulation noise is similar but much more complicated and is not treated in this text.

Suppose that the idle noise power density at the baseband output of an FM radio system has the form

$$d_n = d_{no}\left[1 + \left(\frac{f}{f_o}\right)^2\right] \tag{20-25}$$

where d_{no} is the power density of a flat noise spectrum resulting, for example, from shot noise, and f_o is the frequency at which the flat noise spectrum and the FM triangular noise spectrum cross as shown in Fig. 20-4. The resulting noise density at this point is 3 dB above the flat noise floor.

The unpreemphasized message signal power density spectrum at the same baseband output point may be represented by a flat spectrum of density

$$d_s = d_{so} \text{ watts/hertz} \qquad f_B \leq f \leq f_T \tag{20-26}$$

This baseband signal and the noise spectrum are represented by solid lines in Fig. 20-4.

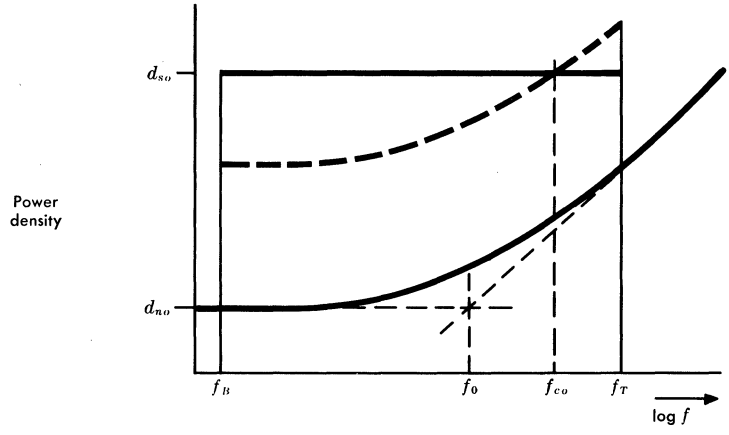


FIG. 20-4. Baseband noise spectra.

An examination of Fig. 20-4 shows that the signal-to-noise ratio at the bottom baseband frequency exceeds the signal-to-noise ratio at the top frequency. The ratio of these two signal-to-noise ratios, which can be obtained from Eqs. (20-25) and (20-26), is $1 + (f_T/f_o)^2$ when $f_B << f_o$.

The first step in pre-emphasis is to shape the signal power density spectrum to match the curvature of the noise power density spectrum, i.e.,

$$d_{sp} = d_{so}\left[1 + \left(\frac{f}{f_o}\right)^2\right] \qquad f_B \leq f \leq f_T \tag{20-27}$$

However, this shaping effectively increases the total signal power over the original flat spectrum total power*, $d_{so}\, f_T$, by the ratio

$$\text{Power ratio} = \frac{1}{f_T}\int_0^{f_T}\left[1 + \left(\frac{f}{f_o}\right)^2\right] df = 1 + \frac{1}{3}\left(\frac{f_T}{f_o}\right)^2 \tag{20-28}$$

In order to reestablish the original total modulating signal power and thereby maintain the original rms deviation of the FM signal, the signal power must be reduced by the ratio given by Eq. (20-28).

*At this point f_B is assumed to be zero. There is usually very little error introduced by this step.

The resultant signal density is shown as a dashed curve in Fig. 20-4. The pre-emphasis function in its entirety is

$$\text{Final pre-emphasis shape (dB)} = 10 \log \left[\frac{1 + \left(\frac{f}{f_o}\right)^2}{1 + \frac{1}{3}\left(\frac{f_T}{f_o}\right)^2} \right] \qquad (20\text{-}29)$$

The frequency f_{co} which appears in Fig. 20-4 is called the crossover frequency of the pre-emphasis function and is defined as the frequency at which pre-emphasis causes no change in the transmission level point of the baseband signal as it appears in the radio system. It is therefore the frequency at which pre- and de-emphasis have no effect on the system idle noise.

For this relatively simple pre-emphasis shape, it is possible to construct a curve of top circuit improvement and approximate bottom circuit degradation as a function of the amount of pre-emphasis used. This curve is shown in Fig. 20-5 illustrating that in order to improve top message channel noise, the noise performance of the lower message channels must be degraded. The theoretical limit of top message channel idle noise improvement is 10 log 3, or 4.77 dB which occurs when the pre-emphasis function is simply a 6 dB per octave slope. This shape could be used only in an FM system with

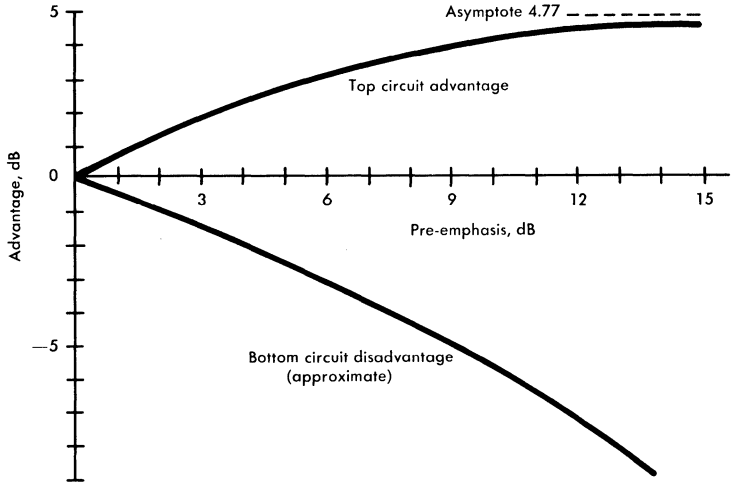


FIG. 20-5. Pre-emphasis advantage and disadvantage.

negligible flat noise components. The *amount* of pre-emphasis is usually designated by the difference in dB between the top and bottom message channel transmission levels or $10 \log\left[1 + \left(\frac{f_T}{f_o}\right)^2\right]$ in this example.

Television Predistortion

Pre-emphasis is also used for television transmission on FM microwave systems although not for the purpose of noise equalization. For TV transmission, the baseband shaping is generally referred to as predistortion rather than pre-emphasis although it is a linear process which simply lowers the signal level for components falling below about 1 MHz.

As discussed elsewhere, a TV signal is more tolerant of high-frequency noise than low-frequency, so that pre-emphasis, as it applies to improving the noise performance at the high end of the baseband, is not required. Instead, the signal is predistorted ahead of the FM terminal transmitter to improve the transmission of the color information. In practice, a given channel may at any time have to carry either a black and white or a color TV signal. Because of this, predistortion is always used even though it may not be essential for black and white transmission.

There are two ways of viewing the manner in which predistortion helps the transmission of a color signal. First, from the quasi-stationary viewpoint, the exact frequencies at which the 3.58-MHz color subcarrier and its sidebands are being transmitted through the system vary according to the amount of grayness in the picture, which is changing at a relatively slow rate. These changes in transmission frequencies result in variations of the amplitude and phase (called differential gain and phase) of the 3.58-MHz color carrier; the variations affect the hue and intensity of the colors. Predistortion, by reducing the amplitude of the low-frequency components of the TV signal, reduces the excursions of the 3.58-MHz carrier and accordingly reduces the differential gain and phase.

From a second and more precise viewpoint, the transmission distortions of the system produce intermodulation products of the TV signal. These intermodulation products fall in the color band and affect the color transmission. Reducing the amplitude of the low-frequency components reduces the magnitude of the modulation products in which they are involved and thereby improves the color performance.

20.3 BREAKING REGION

The discussion in Section 20.1 treated the phase and frequency modulation produced by random noise when the total noise power is much less than the carrier power. If this is the case, the signal-to-noise ratio in the baseband output varies linearly with the signal-to-noise ratio in the FM or PM portion of the system. When the carrier power is less than ten to forty times the noise power, this linearity no longer holds; the baseband output signal-to-noise ratio decreases faster than does the FM or PM signal-to-noise ratio. In this region, which is referred to as the breaking region, the system rapidly becomes unusable.

The phenomenon of breaking is examined by first considering the case where a carrier is subjected to a relatively large interfering sinusoid; the results are extended to the case of a carrier imbedded in a relatively high level flat band of random noise.

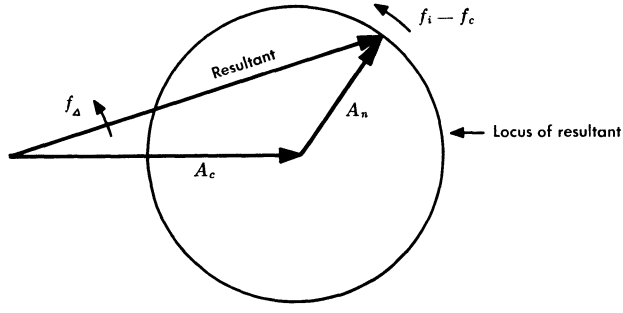


FIG. 20-6. Carrier plus sinusoidal interference, phasor diagram.

With reference to the phasor diagram of Fig. 20-6, the instantaneous frequency deviation, f_Δ, of the resultant signal formed by adding an interfering sinusoid to a carrier may be derived as

$$f_\Delta(t) = (f_i - f_c)\frac{A_n}{A_c}\left\{\frac{\cos 2\pi\,(f_i - f_c)t + \dfrac{A_n}{A_c}}{1 + 2\dfrac{A_n}{A_c}\left[\cos 2\pi\,(f_i - f_c)t\right] + \left(\dfrac{A_n}{A_c}\right)^2}\right\} \tag{20-30}$$

For small values of A_n/A_c, this expression is equivalent to Eq. (20-4). However, as A_n approaches A_c in amplitude, the resultant signal is far from sinusoidal, becoming impulsive when $A_n = A_c$. A plot of f_Δ for several representative values of A_n/A_c is shown in Fig. 20-7.

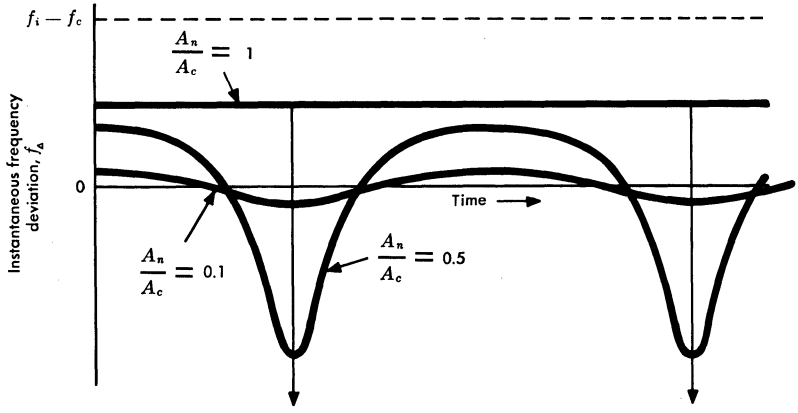


FIG. 20-7. Instantaneous frequency deviation versus time for two sinusoids.

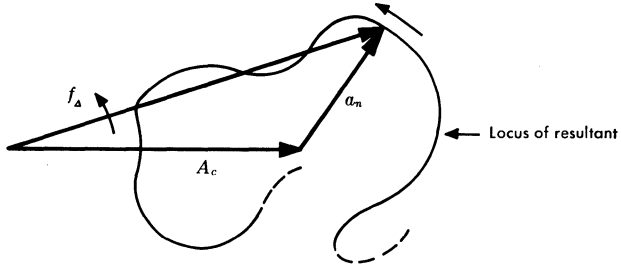


FIG. 20-8. Carrier plus noise, phasor diagram.

From Eq. (20-30) and by reference to Fig. 20-8, it may be inferred that in the case of a carrier imbedded in a flat band of random noise, as the noise peaks approach the carrier in amplitude, the resulting output from an FM discriminator will contain impulses or spikes which tend to have a flat spectrum versus frequency. Because

of the high peak factor of random noise, this situation occurs even when the total noise power is roughly 10 to 15 dB below the carrier power. The nature of the resulting noise in the telephone circuits is impulsive and is therefore especially damaging to data transmission.

Figure 20-9 illustrates the shape of the baseband noise spectrum as a function of carrier-to-noise ratio in one hop of a particular radio system. The message channel noise above 2 MHz is seen to increase dB for dB with reducing carrier level, whereas the message channel noise below 2 MHz increases at a faster rate. This baseband spectrum can be viewed as resulting from the flat spectral density due to breaking being superimposed on the normal triangular FM noise spectrum.

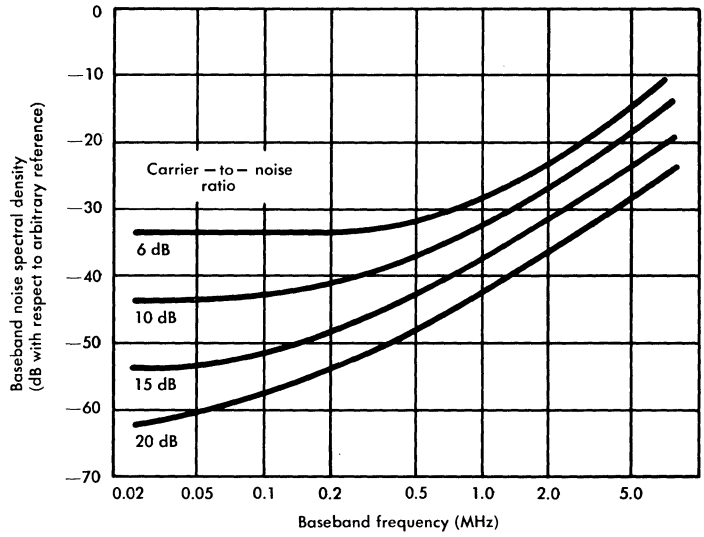


FIG. 20-9. Baseband noise versus frequency.

Example 20.1

Problem

Determine the noise in the noisiest channel at the output of an FM system resulting from the thermal noise contributed by all repeaters.

The system parameters are assumed as follows:

1. Baseband signal: This signal will consist of 1000 single-sideband multiplexed message channels. Each channel is effectively 3 kHz wide; the channel spacing is 4 kHz. The baseband signal is thus 4 MHz wide and is assumed to extend from 0 to 4 MHz.
2. Repeaters: The system consists of 100 repeaters in tandem. Each repeater has an input carrier power of —30 dBm and a noise figure of 12 dB. The repeater bandwidth is 20 MHz.
3. Peak frequency deviation: 4 MHz.

Solution

First, a check should be made to insure that the noise power at the input of any given receiver is small compared to that of the received carrier. Thermal noise at 290°K is — 174 dBm/Hz. The noise power in a 20-MHz band is $10 \log (20 \times 10^6) = 73$ dB higher, or — 101 dBm. Increased by the 12-dB noise figure, the total noise power at the input of any given repeater is —89 dBm. With an unmodulated carrier power of —30 dBm, the carrier-to-noise ratio of any unfaded hop is 59 dB, which allows for more than 40 dB of fading before breaking.

When the methods of Chap. 9 are used, P_s for this system is 25.4 dBm0 (based on $\sigma = 5$ dB).

Because of the triangular noise spectrum, the top baseband channel is the noisiest in an unpreemphasized FM system. Since thermal noise is —174 dBm/Hz and the noise figure is 12 dB, the noise power density at the input to a single repeater is —162 dBm/Hz. The carrier at this point is —30 dBm.

Substituting the above results into Eq. (20-24) yields

$$W_N = 125.8 + P_s + 20 \log \frac{f_1}{\Delta F} + P_n - P_c \qquad \text{dBrnc0}$$

$$= 125.8 + 25.4 + 20 \log \frac{4}{4} - 162 + 30$$

$$= 19.2 \text{ dBrnc0}$$

For 100 repeaters the noise power would be 20 dB higher or approximately 39 dBrnc0. Depending on the flat noise floor obtainable on this system, pre-emphasis could reduce this total by 3 to 4 dB for a resultant of 35 to 36 dBrnc0 of idle noise in each voice circuit.

REFERENCE

1. Rice, S. O. "Properties of a Sine Wave Plus Random Noise," *Bell System Tech. J.*, vol. 27 (Jan. 1948), pp. 109-157.

Chapter 21

Intermodulation Noise in FM and PM Systems

There are numerous ways in which intermodulation noise can be generated in FM transmission systems. In all cases, the noise which is produced is a nonlinear function of the applied modulating signal. However, the type of nonlinearity is not the same in all cases, nor are the principal system parameters associated with the noise mechanism the same in all cases. Because of these differences, a general treatment of intermodulation noise in FM systems is not presented; instead, the principal noise sources are analyzed separately.

The two most widely studied sources of intermodulation noise in FM systems are low-order transmission deviations and echoes. Experience has shown that these two sources are the major contributors to FM system noise, and for that reason they are considered in this chapter in some depth. Other noise contributors such as "AM/PM intermodulation noise" and "pseudo-echo intermodulation noise" can also be significant sources of noise if certain relatively straightforward design considerations are not implemented.

As previously mentioned, transmission deviations are a major source of intermodulation noise in FM systems. Transmission deviations are defined as any deviations in the transmission path from the ideal characteristics of constant gain and linear phase (or constant delay) for all frequency components of the FM signal. The transmission deviations examined in this chapter are linear time-invariant gain and phase shapes which are prescribed functions of frequency. No new frequencies are produced in the FM signal as it passes through the transmission deviations. However, the relative amplitude and phase information of the carrier and sidebands is altered, and this is interpreted by the demodulator as additional modulation and hence causes distortion in the recovered signal. The method presented in this chapter is just one of many ways of anal-

yzing this noise phenomenon. Other techniques with varying degrees of complexity are treated elsewhere in the literature [1, 2]; however, because of the difficulty of dealing with this topic, there is presently no exact analytical approach to the subject.

Another major contributor to intermodulation noise is echoes. As the name implies, this mechanism arises when the incident FM signal is combined with one or more echoes as a result of secondary transmission paths. As will be seen later in the chapter, this mechanism also produces distortion in the recovered modulating signal. Instead of transmission deviation values, the important parameters in this case are the echo levels and their time delays.

It will be seen from the analysis in this chapter how transmission deviations produce baseband gain shaping as well as intermodulation noise. Baseband gain shaping typically shows up in radio relay systems as a roll-off with increasing frequency which usually is not difficult to control in modern systems. Since the analysis presented here inherently precipitates both baseband gain shaping components and intermodulation noise components, the former will also be briefly discussed where appropriate.

21.1 INTERMODULATION NOISE DUE TO LOW-ORDER TRANSMISSION DEVIATIONS

An insight into the process by which transmission deviations cause undesired components in the demodulated output signal can be obtained from the following discussion. Assume that an FM signal with several sideband components is applied to a network which has ideal transmission for the entire FM signal except at the frequency of one of the sideband components. The amplitude of this particular sideband component is slightly altered. This is equivalent to adding to the applied FM signal a small extraneous signal at the frequency of this particular component. Hence, the output signal from the nonideal transmission network may be thought of as consisting of the applied FM signal plus a small extraneous signal. As a result, the demodulated output signal may also be considered as consisting of two components: the desired signal, which is proportional to the input modulating signal, and an undesired or interference signal. Although the source of the extraneous signal is very different from the source of random noise components, the effect is much the same. The equivalent of both amplitude and phase modulation of the applied FM signal will occur.

Consider two cases which illustrate these principles. In the first case, the modulating signal is a single sinusoid. The resulting FM signal consists of a carrier and a number of sideband components separated from the carrier by integral multiples of the modulating frequency. Transmission deviations at any of these frequencies will distort the signal being transmitted. For example, altering a first order sideband component by a small transmission deviation is equivalent to adding a small extraneous sinusoidal signal at the frequency of this component, as shown in Fig. 21-1(a). The principal effect is a small undesired phase modulation at a frequency equal to the frequency difference between the carrier and the altered sideband component. In this case, the difference frequency would be equal to the modulating frequency, and the sinusoidal baseband output would merely be altered in amplitude and phase. On the other hand, if the transmission deviation were such as to alter a second order sideband component, the baseband output would contain an unwanted second harmonic of the modulating frequency. This is shown in Fig. 21-1(b).

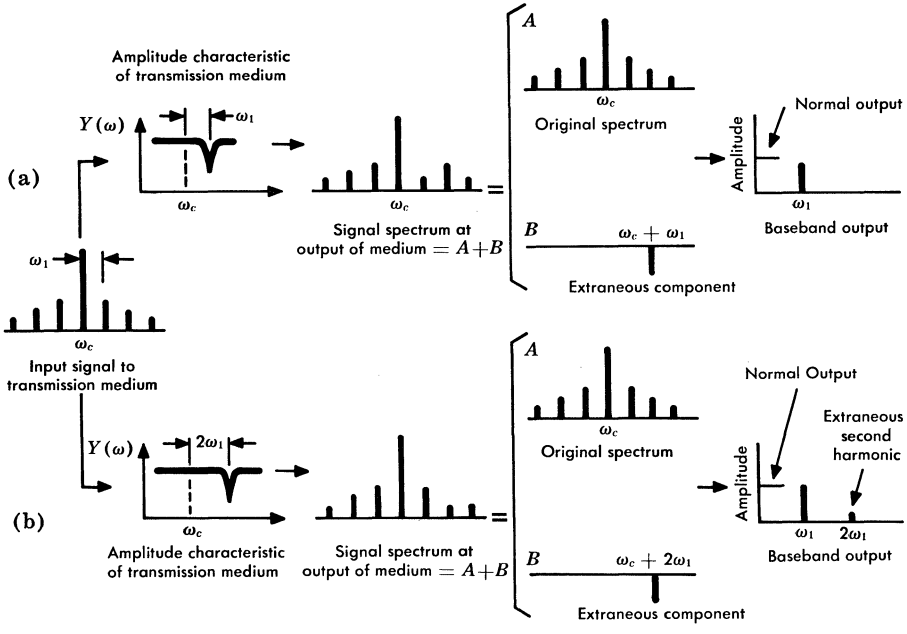


FIG. 21-1. Simplified illustration of baseband distortion caused by a transmission deviation.

In another case, the modulating signal consists of two sinusoids. When the two sinusoids have radian frequencies ω_1 and ω_2, there are sideband components at each of the frequencies $\omega_c + n\omega_1 + m\omega_2$. Consider the case of $n=1$ and $m=-1$, that is, a frequency component at $\omega_c + \omega_1 - \omega_2$. If the amplitude or phase of this component of the FM signal is altered during transmission, the result will be equivalent to unwanted phase modulation at a frequency $\omega_1 - \omega_2$. At the output of the system, therefore, the two original baseband components, as well as an unwanted component at the difference frequency of the two original sinusoids, would be present.

These examples illustrate that transmission deviations in an FM system can introduce baseband frequency components at the output of a system which did not exist at the input. In this sense, transmission deviations in an FM system have an effect similar to amplifier nonlinearity in an AM system. These intermodulation products vary with the modulation index much as intermodulation products in AM systems vary with signal amplitude. With a given transmission deviation, the ratio of the signal to the unwanted products decreases as the index of modulation increases. After demodulation, equalizers cannot eliminate these intermodulation products; therefore it is necessary to equalize the system ahead of the limiter and demodulator if the products are to be reduced.

Derivation of Distortion Terms

As stated previously, there is no exact solution to the problem of determining the intermodulation noise which is produced by transmission deviations. One approximate approach which has proven successful in practice for well equalized systems is to represent the transmission characteristic as a power series gain and phase function up to fourth order. The normalized transmission characteristic is as follows:

$$Y_N(\omega) = [1 + g_1(\omega - \omega_c) + g_2(\omega - \omega_c)^2 + g_3(\omega - \omega_c)^3 + g_4(\omega - \omega_c)^4]\, e^{j[b_2(\omega-\omega_c)^2 + b_3(\omega-\omega_c)^3 + b_4(\omega-\omega_c)^4]} \quad (21\text{-}1)$$

where

ω_c = carrier frequency in rad/sec

g_1, g_2, g_3, g_4 = linear, parabolic, cubic, and quartic gain coefficients, respectively

b_2, b_3, b_4 = parabolic, cubic, and quartic phase coefficients, respectively.

The linear phase shape is omitted since it is equivalent to constant delay and does not introduce distortion. The transmission characteristic is normalized with respect to the carrier frequency such that the transmission at the carrier frequency is unity. Particular forms of transmission characteristics are shown in Fig. 21-2.

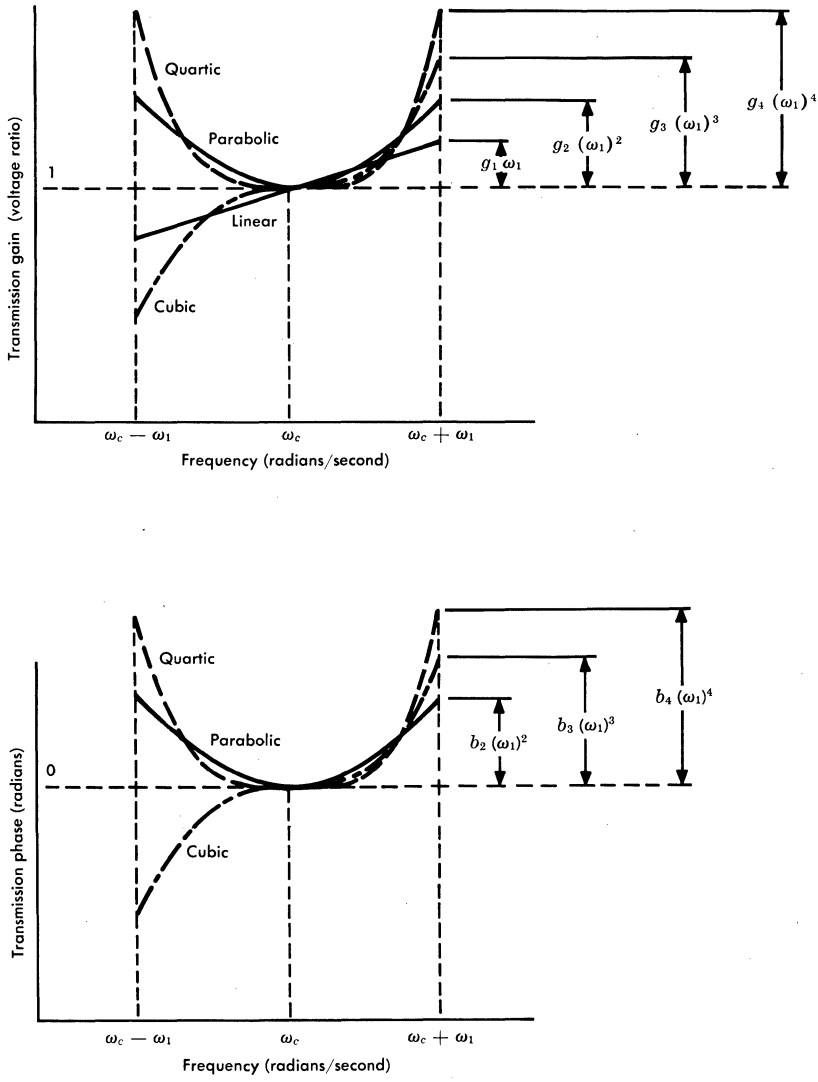


FIG. 21-2. Low-order transmission gain and phase shapes.

The input FM signal is represented as follows:

$$e_1(t) = A_c \cos [\omega_c t + \phi(t)] \tag{21-2}$$

For convenience, A_c is set equal to unity for the following analysis. Equation (21-2) can then be written in exponential notation as

$$e_1(t) = \text{Re}\left\{e^{j[\omega_c t + \phi(t)]}\right\} \tag{21-3}$$

The symbol Re indicates that only the real part of the expression represents the actual signal. The Re will be dropped in order to simplify the following expressions. This gives

$$e_1(t) = e^{j\omega_c t}\, e^{j\phi(t)} \tag{21-4}$$

The input signal has been represented as a function of time, whereas the transmission characteristic has been expressed as a function of frequency. To determine the effect of the transmission characteristic on the input signal, it is necessary to determine first the spectrum of the input signal. This can be done using Fourier transforms:

$$\text{Direct transform, } G(\omega) = \mathfrak{F}\,[f(t)]$$

$$= \int_{-\infty}^{\infty} f(t)\, e^{-j\omega t} dt \tag{21-5}$$

$$\text{Inverse transform, } f(t) = \mathfrak{F}^{-1}[G(\omega)]$$

$$= \frac{1}{2\pi} \int_{-\infty}^{\infty} G(\omega)\, e^{j\omega t} d\omega \tag{21-6}$$

Thus, the spectrum of the input signal, $G_1(\omega)$, is given by the direct Fourier transform of $e_1(t)$ as

$$G_1(\omega) = \mathfrak{F}\,[e_1(t)] \tag{21-7}$$

$$= \int_{-\infty}^{\infty} e^{j\omega_c t}\, e^{j\phi(t)}\, e^{-j\omega t} dt \tag{21-8}$$

When the transmission characteristic of the transmission path to which the input signal is applied is $Y(\omega)$, the spectrum of the output signal, $G_2(\omega)$, is equal to

$$G_2(\omega) = Y(\omega) G_1(\omega) \tag{21-9}$$

Thus every frequency component of the input signal is multiplied by the transmission characteristic at that frequency to obtain the output component. Substitution of Eq. (21-7) into Eq. (21-9) gives

$$G_2(\omega) = Y(\omega)\,\mathfrak{F}[e_1(t)] \tag{21-10}$$

The output signal, $e_2(t)$, is given by the inverse Fourier transform of the output spectrum, $G_2(\omega)$, as

$$e_2(t) = \mathfrak{F}^{-1}[G_2(\omega)] \tag{21-11}$$

Substitution of Eqs. (21-9) and (21-6) into Eq. (21-11) gives

$$\begin{aligned} e_2(t) &= \mathfrak{F}^{-1}[Y(\omega)G_1(\omega)] \\ &= \frac{1}{2\pi}\int_{-\infty}^{\infty} Y(\omega)G_1(\omega)e^{j\omega t}d\omega \end{aligned} \tag{21-12}$$

In this equation, ω is a variable of integration which disappears when the limits are evaluated. It is therefore possible to replace ω by $\omega + \omega_c$ without changing the value of the integral. Equation (21-12) may then be written as

$$\begin{aligned} e_2(t) &= \frac{1}{2\pi}\int_{-\infty}^{\infty} Y(\omega+\omega_c)G_1(\omega+\omega_c)e^{j(\omega+\omega_c)t}d\omega \\ &= \frac{e^{j\omega_c t}}{2\pi}\int_{-\infty}^{\infty} Y(\omega+\omega_c)G_1(\omega+\omega_c)e^{j\omega t}d\omega \end{aligned} \tag{21-13}$$

In shorter form this becomes

$$e_2(t) = e^{j\omega_c t}\,\mathfrak{F}^{-1}[Y(\omega+\omega_c)G_1(\omega+\omega_c)] \tag{21-14}$$

From Eq. (21-8) the expression for $G_1(\omega + \omega_c)$ can be obtained by replacing ω by $\omega + \omega_c$:

$$\begin{aligned} G_1(\omega+\omega_c) &= \int_{-\infty}^{\infty} e^{j\omega_c t}e^{j\phi(t)}e^{-j(\omega+\omega_c)t}dt \\ &= \int_{-\infty}^{\infty} e^{j\phi(t)}e^{-j\omega t}dt \\ &= \mathfrak{F}[e^{j\phi(t)}] \end{aligned} \tag{21-15}$$

Substitution of Eq. (21-15) into Eq. (21-14) gives

$$e_2(t) = e^{j\omega_c t}\mathfrak{F}^{-1}\left\{Y(\omega + \omega_c)\,\mathfrak{F}[e^{j\phi(t)}]\right\} \tag{21-16}$$

This is the desired result. It shows that the effect of a transmission characteristic $Y(\omega)$ on an FM signal can be expressed in terms of the effect of a transmission characteristic $Y(\omega + \omega_c)$ on the modulation term, $e^{j\phi(t)}$, of the FM signal. The modulation term is the same as the actual FM signal except that the carrier has been shifted from ω_c to zero frequency. The transmission characteristic $Y(\omega + \omega_c)$ is the same as the original characteristic $Y(\omega)$ except that it is shifted downward in frequency by an amount ω_c. Thus, the transmission shape that is centered at ω_c in $Y(\omega)$ is moved downward and is centered at zero frequency in $Y(\omega + \omega_c)$. Likewise, since $Y(\omega)$ is double-sided, from the definition of the Fourier transform, the transmission shape that is centered at $-\omega_c$ in $Y(\omega)$ is centered at $-2\omega_c$ in $Y(\omega + \omega_c)$.

To make use of the completely general equation, Eq. (21-16), both the positive and the negative frequency shapes must be included in $Y(\omega + \omega_c)$. However, if the assumption is made that the FM signal is of sufficiently low index and that the transmission medium passes only frequencies in the vicinity of the carrier frequency, $\pm f_c \pm b$, with $b/f_c << 1$, then the product of $Y(\omega + \omega_c)$ and $\mathfrak{F}[e^{j\phi(t)}]$ in Eq. (21-16) can be approximated to a high degree by just the product of the transmission characteristic which is centered at *zero* frequency in $Y(\omega + \omega_c)$ and $\mathfrak{F}[e^{j\phi(t)}]$. As mentioned earlier, one such representation for this transmission characteristic which has proven useful in practice is the normalized characteristic of Eq. (21-1), or in a more appropriate form

$$Y_N(\omega + \omega_c) = (1 + g_1\omega + g_2\omega^2 + g_3\omega^3 + g_4\omega^4)\, e^{j(b_2\omega^2 + b_3\omega^3 + b_4\omega^4)} \tag{21-17}$$

If the exponential is expanded using

$$e^x = 1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} + \cdots$$

and only the terms up to the fourth power of ω are retained,

$$Y_N(\omega + \omega_c) = [1 + g_1\omega + g_2\omega^2 + g_3\omega^3 + g_4\omega^4]\,[(1 - \frac{b_2^2\omega^4}{2}) + j(b_2\omega^2 + b_3\omega^3 + b_4\omega^4)] \tag{21-18}$$

Multiplying and rearranging terms yields

$$Y_N(\omega + \omega_c) = 1 + g_1\omega + (g_2 + jb_2)\omega^2 + [g_3 + j(b_3 + g_1b_2)]\omega^3 + [g_4 - \frac{b_2^2}{2} + j(b_4 + b_3g_1 + b_2g_2)]\omega^4 \quad (21\text{-}19)$$

The substitution of Eq. (21-19) into Eq. (21-16) results in, after considerable work,

$$e_2(t) = \sqrt{[1 + P(t)]^2 + [Q(t)]^2}\cos[\omega_c t + \phi(t) + \theta_i(t)] \quad (21\text{-}20)$$

where

$$\theta_i(t) = \arctan\frac{Q(t)}{1 + P(t)} \quad (21\text{-}21)$$

and

$$P(t) = g_1\phi'(t) + g_2\phi'^2(t) + b_2\phi''(t) + 3(b_3 + g_1b_2)\phi''(t)\phi'(t) + g_3[\phi'^3(t) - \phi'''(t)] + \left(g_4 - \frac{b_2^2}{2}\right)(-4\phi'''\phi' - 3\phi''^2) + (b_4 + b_3g_1 + b_2g_2)(6\phi''\phi'^2 - \phi'''') \quad (21\text{-}22)$$

and

$$Q(t) = -g_2\phi''(t) + b_2\phi'^2(t) + (b_3 + g_1b_2)[\phi'^3(t) - \phi'''(t)] - 3g_3\phi'(t)\phi''(t) + (b_4 + b_3g_1 + b_2g_2)(-4\phi'''\phi' - 3\phi''^2) + \left(g_4 - \frac{b_2^2}{2}\right)(\phi'''' - 6\phi''\phi'^2) \quad (21\text{-}23)$$

where $\phi'(t)$ denotes $d\phi(t)/dt$, etc.

When $P(t) << 1$ and $Q(t) << 1$, as is usually the case, Eq. (21-20) can be approximated as

$$e_2(t) = [1 + P(t)]\cos[\omega_c t + \phi(t) + Q(t)] \quad (21\text{-}24)$$

Therefore, when the transmission deviations are small, the amplitude modulation is given approximately by $P(t)$, and the unwanted phase modulation is given approximately by $Q(t)$. Figure 21-3 lists the various phase and amplitude terms found by substituting Eqs. (21-22) and (21-23) into Eq. (21-24).

It should be noted from Eq. (21-21) that a closer approximation to the phase modulation $\theta_i(t)$ would be $Q(t)[1 - P(t)]$. This improves the approximation since $-P(t)\,Q(t)$ yields interaction terms. These additional terms were considered in the development of Fig. 21-3.

Type of transmission deviation	PM		AM plus AM/PM conversion*	
	Equalizable after detection	Not equalizable after detection	Equalizable after detection	Not equalizable after detection
Linear gain			$g_1\phi'$	$-\frac{1}{2}g_1^2\phi'^2+\frac{1}{3}g_1^3\phi'^3$
Parabolic gain	$-g_2\phi''$	$g_2^2\phi'^2\phi''$		$g_2\phi'^2+\frac{1}{2}g_2^2\phi''^2$
Cubic gain		$-3g_3\phi'\phi''$	$-g_3\phi'''$	$g_3\phi'^3$
Quartic gain	$g_4\phi''''$	$-6g_4\phi'^2\phi''$		$-4g_4\phi'\phi'''-3g_4\phi''^2$
Parabolic phase (linear delay)	$-\frac{1}{2}b_2^2\phi''''$	$b_2\phi'^2$	$b_2\phi''$	$2b_2^2\phi'\phi'''+b_2^2\phi''^2$
Cubic phase (parabolic delay)	$-b_3\phi'''$	$b_3\phi'^3$		$3b_3\phi'\phi''$
Quartic phase (cubic delay)		$-4b_4\phi'\phi'''-3b_4\phi''^2$	$-b_4\phi''''$	$6b_4\phi'^2\phi''$
Interaction:				
g_1g_2		$g_1g_2\phi'\phi''-g_1^2g_2\phi'^2\phi''$		$-g_1g_2\phi'^3$
g_1g_3		$3g_1g_3\phi'^2\phi''$		$g_1g_3\phi'\phi'''$
g_1b_2	$-g_1b_2\phi'''$			$2g_1b_2\phi'\phi''-2g_1^2b_2\phi'^2\phi''$
g_1b_3		$-4g_1b_3\phi'\phi'''-3g_1b_3\phi''^2$	$-g_1b_3\phi''''$	$3g_1b_3\phi'^2\phi''$
g_2b_2		$-4g_2b_2\phi'\phi'''-3g_2b_2\phi''^2$	$-g_2b_2\phi''''$	$4g_2b_2\phi'^2\phi''$

*Multiply all terms by AM/PM conversion coefficient in radians.

FIG. 21-3. Phase modulation caused by small, low-order transmission deviations and AM/PM conversion.

Distortion in the Recovered Baseband Signal. If a limiter can successfully remove the amplitude variations before the signal reaches a device which converts amplitude variations to phase variations, the amplitude modulation of the signal may be dismissed. If, however, the amplitude variations reach an AM/PM converter, they become a second source of distortion. The unwanted phase components in the signal, whether introduced directly by transmission deviations or by transmission deviations followed by AM/PM conversion, will eventually be demodulated along with the desired modulation.

Some of the baseband distortion components fall without frequency translation directly on the modulating signal causing them. These components are directly equalizable at baseband frequencies since their effect is only to change baseband gain and not to cause interference at new frequencies. Terms which are translated in frequency or which otherwise give rise to new baseband frequency components are nonequalizable at baseband frequencies. The equalizable and nonequalizable components are listed separately in Fig. 21-3 for direct PM and also for indirect PM which results from AM/PM conversion.

The order of the various products in Fig. 21-3 may be demonstrated by assuming a modulating input signal, $\phi(t) = kV(t)$. The PM terms from parabolic gain, for example, assuming no AM/PM conversion, would be

$$\text{Output distortion} = \phi_D(t) = -kg_2V''(t) + k^3g_2^2V'^2(t)\,V''(t)$$

$$= -kg_2V''(t) + \frac{k^3g_2^2}{3}\frac{d}{dt}V'^3(t) \qquad (21\text{-}25)$$

Since the desired modulation is $\phi(t) = kV(t)$, the ratio of undesired to desired terms is

$$\frac{\phi_D(t)}{\phi(t)} = -g_2\frac{V''(t)}{V(t)} + \frac{k^2g_2^2}{3}\frac{dV'^3(t)/dt}{V(t)} \qquad (21\text{-}26)$$

The first term does not change with deviation sensitivity, k, and as noted before, it is equalizable since it has an output spectrum identical in components to $V(t)$. The second term varies in amplitude directly as the square of k. This variation conforms with third order intermodulation noise in that the product-to-input level ratio increases 2 dB with a 1-dB change in system deviation. The order of the nonequalizable terms of Fig. 21-3 is listed in Fig. 21-4 except for the interaction terms which are omitted.

Transmission shape	PM	AM plus AM/PM
Linear gain		Second and third
Parabolic gain	Third	Second
Cubic gain	Second	Third
Quartic gain	Third	Second
Parabolic phase (linear delay)	Second	Second
Cubic phase (parabolic delay)	Third	Second
Quartic phase (cubic delay)	Second	Third

FIG. 21-4. Order of nonequalizable modulation products.

From Fig. 21-4 two general rules can be stated [3]. For intermodulation noise due to transmission deviations the rule is:

Even-order gain and delay transmission deviations cause odd-order noise.

Odd-order gain and delay transmission deviations cause even-order noise.

For "AM/PM intermodulation noise" the rule is (for those transmission deviations that cause significant relative noise):

Even-order gain and delay transmission deviations cause even-order noise.

Odd-order gain and delay transmission deviations cause odd-order noise.

There are other useful properties of intermodulation noise [3], the knowledge of which can make the evaluation of system performance more effective. For example, doubling a transmission deviation coefficient will cause the intermodulation noise to increase 6 dB for most transmission deviations. Also, there are some transmission deviations which produce intermodulation noise that increases as much as 10 dB in typical systems when the number of multiplexed telephone channels is increased by a factor of 1.5.

PM Distortion — Sinusoidal Baseband Signals

The results summarized in Fig. 21-3 have numerous applications in practice. To demonstrate one relatively simple application, consider the case of a baseband signal consisting of two sinusoids being applied to a PM system which has only a parabolic phase transmission distortion. (Later in the chapter the much more complicated case of multiplexed telephone channels is considered.) When the baseband signal given by

$$V(t) = A_1 \cos \omega_1 t + A_2 \cos \omega_2 t$$

is applied to a phase modulator, the resulting PM signal is

$$e_1(t) = A_c \cos (\omega_c t + kA_1 \cos \omega_1 t + kA_2 \cos \omega_2 t)$$

This signal is then applied to the transmission system which has a parabolic phase transmission deviation. The phase modulation at the output of the transmission system will consist of the desired signal, $\phi(t)$, and a distortion term, $\phi_D(t)$. The distortion can be obtained directly from Fig. 21-3. Assuming no AM/PM conversion occurs,

$$\phi_D(t) = -\frac{1}{2} b_2^2 \phi''''(t) + b_2 \phi'^2(t)$$

The total baseband output after demodulation is obtained by letting $\phi(t) = kV(t)$ and dividing $\phi(t) + \phi_D(t)$ by k, or

$$\text{Baseband output} = V(t) - \frac{1}{2} b_2^2 V''''(t) + b_2 k V'^2(t)$$

Referring to the input signal and taking the appropriate derivatives,

$$V'(t) = -A_1\omega_1 \sin \omega_1 t - A_2\omega_2 \sin \omega_2 t$$

and

$$V''''(t) = A_1\omega_1^4 \cos \omega_1 t + A_2\omega_2^4 \cos \omega_2 t$$

Then by substitution,

$$\text{Baseband output} = A_1 \cos \omega_1 t + A_2 \cos \omega_2 t - \frac{1}{2} b_2^2 A_1 \omega_1^4 \cos \omega_1 t$$

$$- \frac{1}{2} b_2^2 A_1 \omega_2^4 \cos \omega_2 t + b_2 k A_1^2 \omega_1^2 \sin^2 \omega_1 t$$

$$+ 2 b_2 k A_1 A_2 \omega_1 \omega_2 \sin \omega_1 t \sin \omega_2 t$$

$$+ b_2 k A_2^2 \omega_2^2 \sin^2 \omega_2 t$$

Using the appropriate trigonometric substitutions and omitting the d-c terms gives the desired result:

$$\text{Baseband output} = A_1\left(1 - \frac{b_2{}^2\omega_1{}^2}{2}\right)\cos\omega_1 t + A_2\left(1 - \frac{b_2{}^2\omega_2{}^2}{2}\right)\cos\omega_2 t$$

$$- \frac{A_1{}^2 b_2 k\omega_1{}^2}{2}\cos 2\omega_1 t - \frac{A_2{}^2 b_2 k\omega_2{}^2}{2}\cos 2\omega_2 t$$

$$+ A_1 A_2 b_2\omega_1\omega_2\cos(\omega_2 - \omega_1)t$$

$$- A_1 A_2 b_2\omega_1\omega_2\cos(\omega_2 + \omega_1)t \qquad (21\text{-}27)$$

To demonstrate the effect of AM/PM conversion, assume that instead of parabolic phase distortion the PM signal encounters a parabolic *gain* distortion followed by a device having AM/PM conversion. It will be seen that some of the resulting distortion components will be similar for the two cases. As the first step, the amplitude modulation on the signal may be obtained from Eq. (21-24) where the amplitude is seen to be

$$\text{Amplitude} = 1 + P(t) \qquad \text{volts}$$

$$= 20\log[1 + P(t)] \qquad \text{dB}$$

For very small amplitude variations, which will typically be the case, the following approximation can be used

$$20\log[1 + P(t)] \approx 8.686\,P(t)$$

To obtain the additional phase modulation in radians resulting from an AM/PM conversion constant of Φ degrees/dB, multiply $P(t)$ and the conversion constant as follows

$$\phi_D(t)_{\text{AM/PM}} = \Phi\frac{\pi}{180} \times 8.686\,P(t) = 0.1516\,\Phi\,P(t) \qquad \text{radians}$$

Letting $\phi(t) = kV(t)$ in Fig. 21-3 and assuming only parabolic *gain* distortion yields

$$P(t) = g_2 k^2 V'^2(t) + \frac{g_2{}^2 k^2}{2} V''^2(t)$$

Thus,

$$\phi_D(t)_{\text{AM/PM}} = 0.1516\,\Phi\left[g_2k^2V'^2(t) + \frac{g_2^2k^2}{2}V''^2(t)\right]$$

or

$$V_D(t)_{\text{AM/PM}} = 0.1516\,\Phi\left[g_2kV'^2(t) + \frac{g_2^2k}{2}V''^2(t)\right]$$

The constant 0.1516 which appears in this expression has wide application in this type of problem. It permits a ready translation of an AM/PM conversion constant in degrees per dB to a factor linking a small amplitude modulation index to a small phase modulation index

[PM index: X (radians)] = 0.1516 [Φ degrees/dB] [AM index: m]

Continuing the analysis of modulation by two sinusoids, the resultant distortion terms (other than d-c terms) are

$$\begin{aligned}V_D(t)_{\text{AM/PM}} = 0.1516\,\Phi\Bigg\{&-\frac{A_1^2k}{2}\left[g_2\omega_1^2 - \frac{(g_2\omega_1^2)^2}{2}\right]\cos 2\omega_1 t\\ &-\frac{A_2^2k}{2}\left[g_2\omega_2^2 - \frac{(g_2\omega_2^2)^2}{2}\right]\cos 2\omega_2 t\\ &+A_1A_2k\left[g_2\omega_1\omega_2 + \frac{(g_2\omega_1\omega_2)^2}{2}\right]\cos(\omega_2-\omega_1)t\\ &-A_1A_2k\left[g_2\omega_1\omega_2 - \frac{(g_2\omega_1\omega_2)^2}{2}\right]\cos(\omega_2+\omega_1)t\Bigg\}\end{aligned} \qquad (21\text{-}28)$$

Example 21.1

Problem

Find the baseband distortion, using Eqs. (21-27) and (21-28), for a practical PM system with the following parameters:

$$k = \frac{1}{2}\ \text{rad/volt}$$

$$A_1 = A_2 = 1\ \text{volt}$$

$$\omega_1 = 2\pi \times 10^6\ \text{rad/sec}$$

$$\omega_2 = 8\pi \times 10^6\ \text{rad/sec}$$

Assume the parabolic phase to be 0.1 radian at a frequency 10 MHz from the carrier.

Solution

Solving for b_2,

$$b_2 (2\pi \times 10^7)^2 = 0.1 \text{ radian}$$

$$b_2 = \frac{0.1}{(2\pi \times 10^7)^2} \text{ sec}^2/\text{rad}$$

Substituting the above into Eq. (21-27) gives:

$$\text{Baseband output} = 1^{-}\cos \omega_1 t + 0.999 \cos \omega_2 t$$

$$- 0.00025 \cos 2\omega_1 t - 0.004 \cos 2\omega_2 t$$

$$+ 0.002 \cos (\omega_2 - \omega_1) t - 0.002 \cos (\omega_2 + \omega_1) t$$

Here it is seen that the transmission deviation has changed the amplitudes of the two original sinusoids and that second order terms have appeared. It is also seen that the amplitude changes of the original sinusoids increase rapidly as a function of baseband frequency. Although a baseband equalizer could be used to correct these amplitudes, the intermodulation terms would remain.

To solve for the AM/PM distortion terms, assume the AM/PM conversion constant to be 6 degrees per dB and the parabolic gain to be 1 dB at a frequency 10 MHz from the carrier

$$g_2 (2\pi \times 10^7)^2 = 0.122$$

$$g_2 = \frac{0.122}{(2\pi \times 10^7)^2} \text{ sec}^2/\text{rad}^2$$

Substituting the above into Eq. (21-28) gives

$$V_D(t)_{\text{AM/PM}} = - 0.00056 \cos 2\omega_1 t - 0.0089 \cos 2\omega_2 t$$

$$+ 0.0022 \cos (\omega_2 - \omega_1) t - 0.0022 \cos (\omega_2 + \omega_1) t$$

Here again the distortion terms are seen to be second order in conformity with Fig. 21-4. Note that the sign of these terms depends on the polarity of the AM/PM conversion factor.

FM System Noise with Multiplexed Telephone Channels

The distortion produced by transmission deviations is increasingly difficult to deal with as the modulating signal becomes more complex.

There are, however, other procedures which can be used in such cases. Two methods which are used in practice for estimating the amount of intermodulation noise are given in the following.

Product Count Method. For the purposes of discussion, assume an FM system with parabolic phase distortion and a load consisting of multiplexed telephone channels. For an FM system, $\phi'(t) = k_1 V(t)$. In Fig. 21-3 the PM unequalizable distortion is given by $b_2\phi'^2(t)$. The corresponding FM distortion is $b_2 d\phi'^2(t)/dt$. Hence, the total output signal, assuming no AM/PM conversion and neglecting equalizable distortion, can be written as

$$V_{\text{out}}(t) = V_{\text{in}}(t) + b_2 k_1 \frac{d}{dt}[V_{\text{in}}(t)]^2 \qquad (21\text{-}29)$$

In this form, the expression closely resembles the expression for second order intermodulation in amplitude modulation systems. The difference between this expression and the power series used in AM systems is the taking of a derivative, which is equivalent to multiplying the voltage spectrum by $j\omega$. Therefore, the noise voltage amplitude in a single telephone channel due to the distortion term $b_2 k_1 V_{\text{in}}^2(t)$ can be multiplied by the center frequency (in the baseband signal) of the telephone channel, ω_1 radians per second, to obtain the effect of the derivative. Because of this multiplier, the intermodulation noise is worst in the top telephone channel, where $\alpha + \beta$ products are dominant. The calculations therefore will be made for that channel and type of product.

First, let $V_{\text{in}}(t)$ be represented by two sinusoids, such that

$$V_{\text{in}}(t) = V_\alpha \cos \omega_1 t + V_\beta \cos \omega_2 t$$

When this equation is substituted into Eq. (21-29) and terms falling at $\omega_1 + \omega_2$ are collected, there results an output $\alpha + \beta$ product, $V_{\alpha+\beta}(t)$, given by

$$V_{\alpha+\beta}(t) = -b_2 k_1 V_\alpha V_\beta (\omega_1 + \omega_2) \sin (\omega_1 + \omega_2) t$$

To evaluate interference into any particular telephone channel, it is convenient to substitute $\omega_{\alpha+\beta} = \omega_1 + \omega_2$. This gives

$$V_{\alpha+\beta}(t) = -b_2 k_1 V_\alpha V_\beta \omega_{\alpha+\beta} \sin \omega_{\alpha+\beta} t$$

In this form it is apparent that, for a given $\omega_{\alpha+\beta}$, the magnitude of the interference is not dependent on the frequencies of the fundamentals. This is not the case for all forms of intermodulation. If

the squares of the amplitudes of both sides of this equation are halved, the resulting expression may be used to relate signal and product levels at a zero transmission level point:

$$\frac{V^2_{\alpha+\beta}}{2} = 2b_2^2k_1^2\,\frac{V_\alpha^2}{2}\,\frac{V_\beta^2}{2}\,\omega^2_{\alpha+\beta}$$

or in terms of average power across a 1-ohm resistor,

$$p_{\alpha+\beta} = 2b_2^2k_1^2\,p_\alpha\,p_\beta\,\omega^2_{\alpha+\beta}$$

The constant $H_{\alpha+\beta}$ was defined in Chap. 10 as the power in dBm0 of the $\alpha + \beta$ product resulting from two 0 dBm0 sinusoids (0.001 watt each at 0 TLP), or

$$H_{\alpha+\beta} = 10 \log\left[\frac{2b_2^2k_1^2\omega^2_{\alpha+\beta}\,(0.001)^2}{0.001}\right] \text{dBm0} \qquad (21\text{-}30)$$

where k_1 is specified in units referred to 0 TLP. The $\alpha+\beta$ product annoyance to a telephone customer has been given in Eq. (10-27) which can be written as

$$W_{\alpha+\beta} = H_{\alpha+\beta} + K_{\alpha+\beta} \qquad \text{dBrnc0} \qquad (21\text{-}31)$$

where the constant $K_{\alpha+\beta}$ depends on talker statistics and the number of channels to be transmitted.

Example 21.2

Problem

Using the product count method, find the top telephone channel noise in a system having the following parameters: (1) the baseband consists of 1000 frequency-multiplexed telephone channels in a band which extends from 0 to 4 MHz; (2) the signal is transmitted over an unpreemphasized FM system; (3) the peak frequency deviation is 4 MHz; and (4) in the FM transmission path there is an unequalized linear envelope delay (parabolic phase) distortion of one nanosecond per megahertz (ns/MHz).

Solution

A 1 ns/MHz delay slope is equivalent to $b_2 = 10^{-15}/4\pi$ sec^2/rad as is developed later. For this problem, $\omega_{\alpha+\beta} = \omega_T = 8\pi \times 10^6$ rad/sec, and for 1000 talkers P_s is 25.4 dBm0 so that a 25.4 dBm or 0.346 watt sinusoid at 0 TLP results in $2\sqrt{2}$ MHz rms deviation.

Therefore,

$$k_1^2 = \frac{(2\sqrt{2} \times 2\pi \times 10^6)^2}{0.346} = 92\pi^2 \times 10^{12} \qquad \frac{\text{rad}^2}{\text{watt-sec}^2}$$

Substituting the above parameters into Eq. (21-30) yields a value for $H_{\alpha+\beta}$ of -51.4 dBm0. Using the relations developed in Chap. 10, $K_{\alpha+\beta}$ is found to be about 79.6 dB for this system. Therefore, the top telephone channel noise is estimated by this method to be $-51.4 + 79.6 = 28.2$ dBrnc0.

Noise Loading Method. In the laboratory and field testing of message systems, noise loading is often used to evaluate actual performance by simulating a "live" load with bandlimited gaussian noise. The average power density currently used in the Bell System to simulate a busy-hour load for radio systems carrying 600 or more telephone message channels is -16 dBm0 per 4 kHz. This level is derived in Chap. 9.

In practice, this load is presented to the system under test with narrow frequency slots suppressed, and a selective detector is used at the system output to measure the noise appearing in the narrow slots. The noise loading test is outlined in Fig. 21-5.

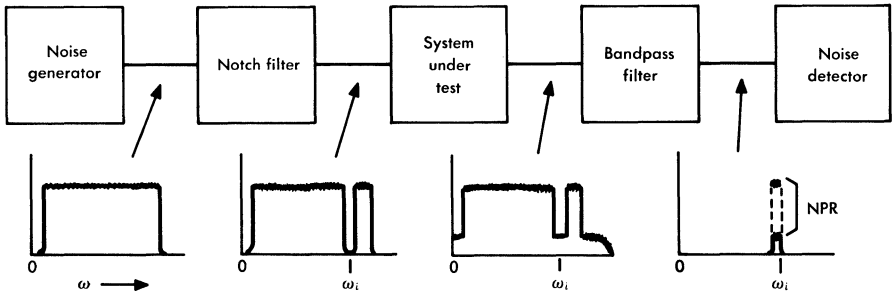


FIG. 21-5. Noise loading test.

The noise loading test may be simulated analytically as shown in the following. This method of analysis reduces product counting to a convolution process and also allows the evaluation of preemphasis, or level shaping advantage, which is highly tedious by the product count method. The illustration again assumes a linear

envelope delay distortion and begins with Eq. (21-29),

$$V_{out}(t) = V_{in}(t) + b_2 k_1 \, d/dt \, [V_{in}(t)]^2$$

From this point, the methods introduced in Chap. 10 (page 267 *et seq.*) may be used to evaluate the distortion noise spectrum. The following is intended to indicate the method rather than to stand as a rigorous proof.

The solution of Eq. (21-29) is analogous to the solution of the system shown in Fig. 21-6.

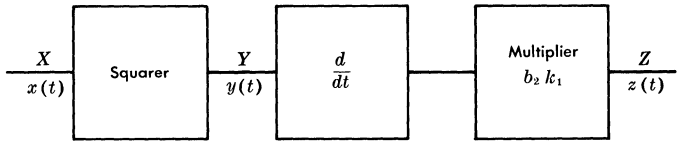


FIG. 21-6. System analog for second order modulation.

In this system, $x(t)$, which is equivalent to $V_{in}(t)$, is assumed to be a gaussian process with zero mean. The autocorrelation function, $\Re_x(\tau)$, of $x(t)$ is defined as

$$\Re_x(\tau) \equiv \lim_{T \to \infty} \frac{1}{2T} \int_{-T}^{T} x(t) \, x(t+\tau) \, dt = E[x(t) \, x(t+\tau)] \qquad (21\text{-}32)$$

where E[ ] stands for the expected or mean value of the quantity inside the brackets. The autocorrelation function of the input and the input power density spectrum form a Fourier transform pair as follows:

$$S_x(\omega) = \mathfrak{F}[\Re_x(\tau)] = \int_{-\infty}^{\infty} \Re_x(\tau) e^{-j\omega\tau} d\tau \qquad (21\text{-}33)$$

and

$$\Re_x(\tau) = \frac{1}{2\pi} \int_{-\infty}^{\infty} S_x(\omega) e^{j\tau\omega} d\omega \qquad (21\text{-}34)$$

Finally, the mean of the squared value of the input signal is

$$E[x^2(t)] = \Re_x(0) = \frac{1}{2\pi} \int_{-\infty}^{\infty} S_x(\omega) \, d\omega \qquad (21\text{-}35)$$

This statement simply says that $\Re_x(0)$ equals the average power of the input signal $x(t)$ since $S_x(\omega)$ is the input power density in

watts per hertz assuming a 1 ohm load impedance for convenience. From Fig. (21-6), $\mathcal{R}_y(\tau)$, the autocorrelation of the square of $x(t)$, may be written [4]

$$\mathcal{R}_y(\tau) = E[x^2(t)\, x^2(t+\tau)] = \mathcal{R}_x^{\,2}(0) + 2\,\mathcal{R}_x^{\,2}(\tau) \qquad (21\text{-}36)$$

The term $\mathcal{R}_x^{\,2}(0)$ is the square of the average power of the signal, and as a constant it contributes no distortion spectrum. As a result, the distortion spectrum at Y is found by using Fourier transform theory to be

$$S_y(\omega) = 2\int_{-\infty}^{\infty} \mathcal{R}_x^{\,2}(\tau)\, e^{-j\omega\tau} d\tau = \frac{2}{2\pi}[S_x(\omega) * S_x(\omega)] \qquad (21\text{-}37)$$

where * indicates a convolution. Finally, since taking the derivative of $V(t)$ is equivalent to multiplying $S(\omega)$ by ω^2, the power spectral density at point Z is evaluated as

$$S_z(\omega) = \frac{k_1^2 b_2^2 \omega^2}{\pi}[S_x(\omega) * S_x(\omega)] \qquad (21\text{-}38)$$

It follows that the signal-to-noise or noise-power ratio at any baseband frequency in an FM system with parabolic phase or linear delay slope will be

$$\mathrm{NPR}(\omega) = 10 \log \left\{ \frac{\pi\, S_{V_{\text{in}}}(\omega)}{b_2^2 k_1^2 \omega^2\, [S_{V_{\text{in}}}(\omega) * S_{V_{\text{in}}}(\omega)]} \right\} \qquad (21\text{-}39)$$

Figure 21-7 is a compilation of the primary baseband noise contributors for selected transmission deviations; it was obtained by means similar to the preceding. One equation is to be used for direct PM distortion and the other for AM distortion followed by AM/PM conversion. These equations list only the dominating terms for practical systems and do not include interaction terms (e.g., $g_1 b_2$) which could be appreciable in some systems [3].

Example 21.3

Problem

Using the noise loading method, find the top telephone channel noise in a system having the following parameters (identical with

Amplitude and phase transmission deviations

$$\mathrm{NPR}(\omega)_{\mathrm{dB}} = 10 \log \left[\frac{S(\omega)}{\frac{k_1^2}{\pi}\left\{ b_2^2\omega^2 + \left(\frac{3g_3}{2}\right)^2\omega^4 + 4b_4^2\omega^6 \right\} [S(\omega) * S(\omega)] + \frac{3k_1^4}{2\pi^2}\left\{ b_3^2\omega^2 + 4g_4^2\omega^4 \right\} [S(\omega) * S(\omega) * S(\omega)]} \right]$$

Amplitude and phase transmission deviations plus AM/PM conversion

$$\mathrm{NPR}(\omega)_{\mathrm{dB}} = 10 \log \frac{1}{(0.1516\Phi)^2} \left[\frac{S(\omega)}{\frac{k_1^2}{\pi}\left\{ g_2^2\omega^2 + \left(\frac{3b_3}{2}\right)^2\omega^4 + 4g_4^2\omega^6 \right\} [S(\omega) * S(\omega)] + \frac{3k_1^4}{2\pi^2}\left\{ g_3^2\omega^2 + 4b_4^2\omega^4 \right\} [S(\omega) * S(\omega) * S(\omega)]} \right]$$

Notes: 1. For unpreemphasized FM, $S(\omega) = \frac{\pi p_o}{\omega_T}$ for $-\omega_T < \omega < \omega_T$; p_o is total reference drive in watts referred to 0 TLP.

2. Φ = AM/PM conversion constant in degrees per dB.

3. $k_1 = \sqrt{2000}\,\pi \times$ [peak deviation due to 0 dBm0 sinusoid in Hz].

FIG. 21-7. FM system noise power ratios for selected transmission deviations.

those of Example 21.2): (1) the baseband signal consists of 1000 frequency-multiplexed telephone channels in a band which extends from 0 to 4 MHz; (2) the signal is transmitted over an unpre-emphasized FM system; (3) the peak frequency deviation is 4 MHz; and (4) in the FM transmission path there is an unequalized linear delay slope of 1 ns/MHz. Also, assume that a flat noise spectrum of −16 dBm0 per 4 kHz is used to represent the telephone load. This is equal to +14 dBm0 or 0.025 watts distributed from 0 to 4 MHz for 1000 channels.

Solution

The various components of Eq. (21-39) are evaluated as follows:

1. $0.025 \text{ watt} = \dfrac{1}{2\pi}\displaystyle\int_{-\omega_T}^{\omega_T} S_{V_{\text{in}}}(\omega)\, d\omega = \dfrac{\omega_T}{\pi} S_{V_{\text{in}}} \qquad -\omega_T < \omega < \omega_T$

   $\text{therefore, } S_{V_{\text{in}}} = \dfrac{0.025\,\pi}{\omega_T} \quad \text{watt/Hz} \qquad -\omega_T < \omega < \omega_T$

2. $S_{V_{\text{in}}}(\omega) * S_{V_{\text{in}}}(\omega) = \left(\dfrac{0.025\,\pi}{\omega_T}\right)^2 (2\omega_T - |\,\omega\,|) \qquad \text{watt}^2/\text{Hz}$

   where $-2\omega_T < \omega < 2\omega_T$

3. $b_2 = \dfrac{10^{-15}}{4\pi} \qquad \text{sec}^2/\text{rad}$

4. $k_1^2 = 92\,\pi^2 \times 10^{12} \qquad \text{rad}^2/\text{watt-sec}^2$

Using these components in Eq. (21-39) yields

$$\text{NPR}(\omega) = 10 \log \left[\pi \frac{0.025\pi}{\omega_T} \frac{\text{watt}}{\text{Hz}} \div \frac{10^{-30}}{16\pi^2} \frac{\text{sec}^4}{\text{rad}^2} \times 92\pi^2 \right.$$

$$\left. \times\, 10^{12} \frac{\text{rad}^2}{\text{watt-sec}^2} \times 64\pi^2 \times 10^{12} \frac{\text{rad}^2}{\text{sec}^2} \times \frac{(0.025\pi)^2}{\omega_T} \frac{\text{watt}^2}{\text{Hz}} \right]$$

Finally, evaluating NPR at the top telephone channel frequency which is $\omega_T = 8\pi \times 10^6$ rad/sec,

$$\text{NPR} = 40.4 \text{ dB}$$

Since the original signal density was assumed to be −16 dBm0 per 4 kHz ≈ −17 dBm0 per 3 kHz = 71 dBrnc0, the top telephone

channel noise is

$$71 - 40.4 = 30.6 \text{ dBrnc0}$$

This result is in reasonable agreement with the 28.2 dBrnc0 value obtained by the product count method. A discussion of the differences between the results of these methods is presented in Chap. 10. For modern high-capacity radio systems, the noise loading method is considered to be a better estimator of system performance and is used extensively in present-day system evaluation and design.

Addition of Noise Contributors

In the preceding pages two noise contributors in FM systems have been analyzed: (1) intermodulation noise due to transmission deviations, and (2) AM/PM intermodulation noise. Even though different, these two contributors have the same property of being functions of the baseband signal. Hence, it would be expected that they would be correlated to some degree. This would mean that combining the two noise power density spectra together assuming random addition (power addition) might not be sufficient in general.

The addition of these two noise contributors was considered analytically in Reference 5. It was found that certain conditions exist under which the correlation can be significant. To illustrate this in a practical problem, a representative FM radio relay system was examined and it was found that the power density spectrum for the correlated sum of the two noise contributors was substantially different from the power addition of the individual noise spectra. In the top channel, the correlated noise power was about 4.5 dB lower than the noise resulting from a power sum.

Further Application of Figure 21-7

Figure 21-7 may be useful in analyzing intermodulation noise in an existing FM system. Two or three sinusoidal baseband signals may be substituted for the noise loading signal, and measurements made of the sum and difference product levels as functions of the input frequencies. A partial separation of the contributors may then be made on the basis of their frequency dependence. For example, an $\alpha + \beta$ product which increases at a 6 dB per octave rate would indicate intermodulation noise due to delay slope (b_2) and/or AM/PM intermodulation noise due to parabolic gain (g_2).

Envelope Delay Distortion

In practice, envelope delay distortion (EDD) is measured rather than phase distortion. Envelope delay is defined as the derivative of the phase characteristic $\theta(\omega)$ of the transfer function $Y_N(\omega)$ with respect to radian frequency, i.e.,

$$\text{Envelope delay} = \frac{d}{d\omega}\theta(\omega) \qquad \text{sec} \tag{21-40}$$

where $\theta(\omega)$ represents the phase characteristic $\theta(\omega + \omega_c)$ normalized with respect to ω_c. Envelope delay measures the time required to propagate a change in the envelope of a signal (the actual information-bearing part of the signal) through the system. If $\theta(\omega)$ is proportional to frequency around ω_c, the envelope delay will be a constant for all frequencies and there will be no distortion of the transmitted signal. For the more general case, $\theta(\omega)$ will not be linear but will instead include higher order components. This has been indicated in Eq. (21-1) where, by ignoring the quartic phase component, $\theta(\omega)$ is written as

$$\theta(\omega) = b_2(\omega - \omega_c)^2 + b_3(\omega - \omega_c)^3 \tag{21-41}$$

For the general case, the derivative of the phase characteristic and hence the envelope delay will not be constant but will instead contain terms that are functions of frequency. The envelope delay distortion is equal to the envelope delay minus the constant delay term and thus includes all of the nonconstant terms. As an example, the envelope delay distortion of the normalized phase characteristic given by Eq. (21-41) is

$$\text{EDD} = 2b_2(\omega - \omega_c) + 3b_3(\omega - \omega_c)^2 \tag{21-42}$$

Linear EDD is usually expressed in terms of nanoseconds per megahertz. For 1 ns/MHz, evaluating the coefficient of the linear term of Eq. (21-42) yields $b_2 = (1 \times 10^{-9})/(4\pi \times 10^6) = 1/4\pi \times 10^{-15}$ sec^2/rad. In Example 21.1, $b_2 = 0.1/(2\pi \times 10^7)^2 = (1/\pi)(1/4\pi \times 10^{-15})$ sec^2/rad, indicating a linear EDD of about 0.32 ns/MHz.

In FM systems, many circuits in the i-f and r-f parts of the system have bandpass characteristics. Inherent in these are considerable

amounts of parabolic EDD. If the parabolic EDD is not centered exactly on ω_c, that is, if the bandpass characteristic is not centered exactly on the carrier, linear EDD results. To show this, let the EDD be

$$\text{EDD} = 3b_3(\omega - \omega_0)^2$$

Expanding around $(\omega - \omega_0) = (\omega_c - \omega_0)$ in a Taylor series yields

$$\text{EDD} = 3b_3\,[(\omega_c - \omega_0)^2 + 2(\omega_c - \omega_0)(\omega - \omega_c) + (\omega - \omega_c)^2]$$

Defining $(\omega_c - \omega_0) = \Delta\omega$ and discarding the constant term leaves a component of parabolic delay equal in magnitude to the original shape plus a new linear term:

$$\text{EDD} = 3b_3(\omega - \omega_c)^2 + 6b_3\Delta\omega\,(\omega - \omega_c)$$

The preceding analysis illustrates that in an FM system the distortion at baseband due to a given transmission characteristic depends on the location of the carrier frequency. This is analogous to an AM system in which the modulation distortion depends on the bias points selected for the active elements.

21.2 INTERMODULATION NOISE DUE TO ECHOES

Echoes, which are a significant source of intermodulation noise in FM systems, may be generated in a number of ways. They may result from impedance mismatches, waveguide mode conversion, "sneak" transmission paths at harmonic or product frequencies, and other related effects. The conditions for the generation of an echo are a desired primary transmission path and an unwanted secondary path over which a fraction of the original signal arrives at the receiver, displaced in time from the primary signal.

Figure 21-8 illustrates a situation commonly encountered in waveguide runs. Here an echo of amplitude ratio r relative to the desired signal, is displaced by an absolute time delay, T, which does not vary significantly over the frequency range of one radio channel.

Derivation of the Distortion Term

If an angle-modulated signal $e_1(t) = A_c \sin [\omega_c t + \phi(t)]$ is transmitted over the system shown in Fig. 21-8, the resultant transmitted

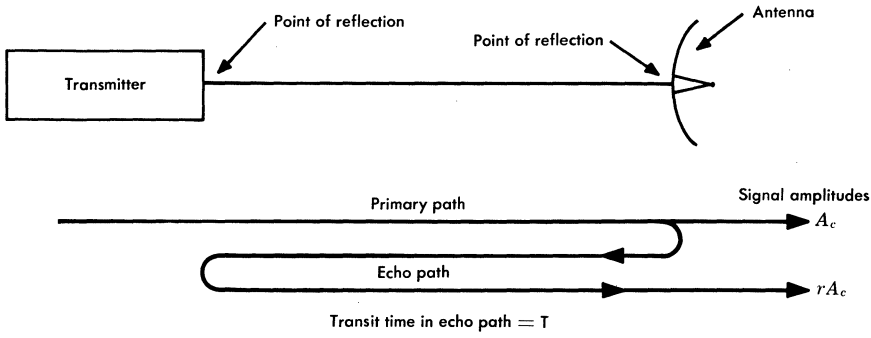


FIG. 21-8. Echo in a transmitter-antenna system.

signal will be

$$e_2(t) = A_c \{\sin [\omega_c t + \phi(t)] + r \sin [\omega_c t - \omega_c \mathsf{T} + \phi(t - \mathsf{T})]\} \quad (21\text{-}43)$$

This expression may be rewritten in exponential form as

$$e_2(t) = A_c \operatorname{Re} [-je^{jx} - jre^{j(x+y)}] \quad (21\text{-}44)$$

where $x = \omega_c t + \phi(t)$ and $y = -\omega_c \mathsf{T} + \phi(t - \mathsf{T}) - \phi(t)$.

Rearranging terms,

$$e_2(t) = A_c \operatorname{Re} [-je^{jx} (1 + re^{jy})]$$

For $r << 1$ this may be written

$$e_2(t) = A_c \operatorname{Re} \left(-je^{jx} e^{re^{jy}}\right)$$

$$= A_c \operatorname{Re} \left\{e^{r \cos y} [-je^{j(x + r \sin y)}]\right\}$$

The last expression may be approximated as

$$e_2(t) = A_c (1 + r \cos y) \sin (x + r \sin y) \quad (21\text{-}45)$$

Substituting for the variables in Eq. (21-45) yields

$$e_2(t) = A_c \{1 + r \cos [-\omega_c \mathsf{T} + \phi(t - \mathsf{T}) - \phi(t)]\}$$
$$\cdot \sin \{\omega_c t + \phi(t) + r \sin [-\omega_c \mathsf{T} + \phi(t - \mathsf{T}) - \phi(t)]\} \quad (21\text{-}46)$$

After limiting, the phase modulation will be

$$\phi(t) + \phi_D(t)$$

where

$$\phi_D(t) = r \sin [-\omega_c \mathsf{T} + \phi(t - \mathsf{T}) - \phi(t)] \quad (21\text{-}47)$$

Noise Contour Chart

The distortion term $\phi_D(t)$ given above produces intermodulation noise since it is a nonlinear function of the modulating signal, $\phi(t)$. This distortion is treated in Reference 6 where a flat gaussian noise spectrum representation of the baseband signal is used to derive the intermodulation noise falling in the top multiplexed telephone channel as a function of the relative amplitude and absolute time delay of the echo. The results of this analysis are depicted as a contour chart in Fig. 21-9.

Example 21.4

Problem

To illustrate the use of Fig. 21-9, find the top telephone channel noise produced by an echo in an FM system having the same parameters as that of Example 21.2 with 1000 message channels in the band of 0 to 4 MHz and a peak frequency deviation of 4 MHz. Assume a 50-dB echo with an absolute time delay of 250 nanoseconds.

Solution

For this system, P_s was found to be 25.4 dBm0, and if white noise of −16 dBm0 per 4 kHz is used to simulate the busy-hour load, the rms frequency deviation will be

$$25.4 + 3 - (-16 + 10 \log 1000) = 14.4 \text{ dB below 4 MHz}$$

or

$$\sigma = 0.76 \text{ MHz} \qquad \text{rms}$$

In this sequence 25.4 is P_s, 3 is the adjustment from peak to rms voltage for a sinusoid, and −16 + 10 log 1000 is the total drive

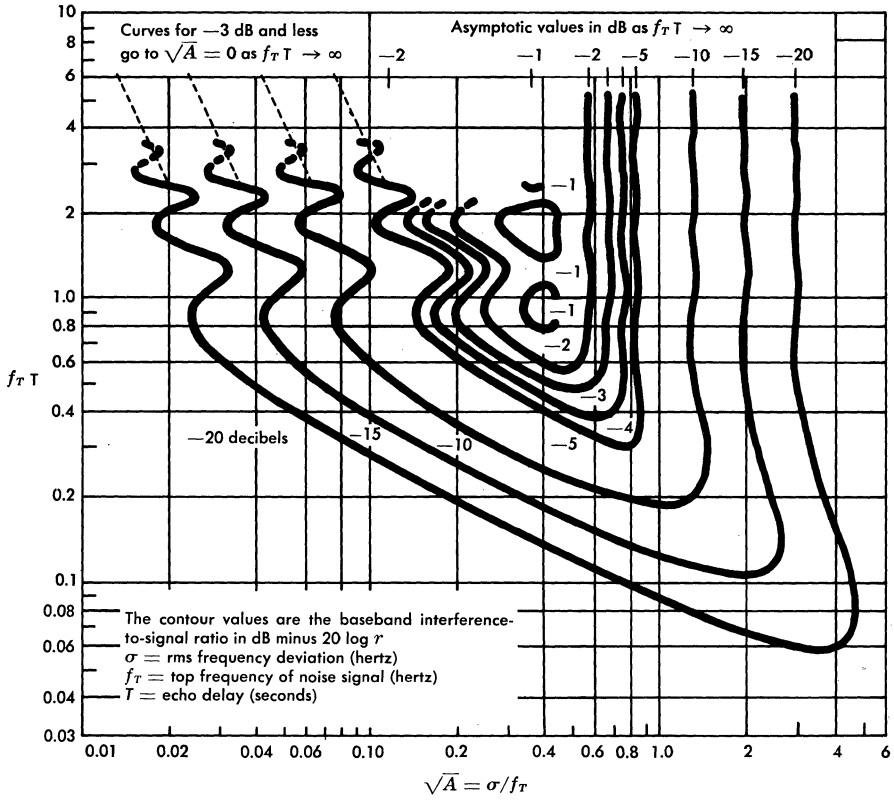


FIG. 21-9. Contours of constant interference in the top channel of an unpre-emphasized multichannel FM system.

simulating 1000 talkers, 25 per cent active. The term $\sqrt{A} = \sigma/f_T$ which is used to enter the abscissa of Fig. 21-9 will be 0.76/4, or 0.19. For a path with a 50-dB echo and an absolute time delay, T, of 250 ns, the ordinate value $f_T T$ would be $4 \times 10^6 \times 250 \times 10^{-9} = 1.0$, and the chart would yield a contour value of −4 dB. When added to the relative echo level of 20 log $r = -50$ dB, the resultant is a noise power ratio of 54 dB. As in Example 21.3, the resultant top telephone channel noise is found by subtraction from 71 dBrnc0; that is, 71 − 54 = 17 dBrnc0.

Figure 21-9 as given here applies only to an unpreemphasized FM signal. The use of pre-emphasis will reduce the actual top channel noise produced by an echo to a value below that calculated (typically by 3 to 4 dB).

Amplitude and Delay Distortion Resulting from Echoes

If the input signal to the system of Fig. 21-8 is simply $e_1(t) = A_c \sin \omega t$, where ω may take on differing values, and the substitutions $x = \omega t$ and $y = -\tau\omega$ are made in Eq. (21-45), the resulting signal for $r << 1$ is of the form

$$e_2(t) = A_c\,(1 + r\cos \tau\omega)\sin\,(\omega t - r\sin \tau\omega) \qquad (21\text{-}48)$$

An inspection of this equation will show that the effect on the signal is equivalent to transmission through a device with gain characteristic $1 + r \cos \tau\omega$ which has ripples spaced in frequency with an interval of $1/\tau$ Hz and a peak-to-peak amplitude of 17.37 r dB. Similarly, the equivalent phase characteristic is found to be

$$\phi(\omega) = \frac{-r\sin \tau\omega}{1 + r\cos \tau\omega} \approx -r\sin \tau\omega$$

Taking the derivative of this with respect to ω to find envelope delay yields

$$\tau(\omega) = \frac{d\phi}{d\omega} = -r\,\tau\cos \tau\omega$$

This component also has ripples spaced with an interval of $1/\tau$ hertz, and the peak-to-peak envelope delay distortion is $\tau_{pp} = 2r\tau$.

In Example 21.4, a 50-dB echo with 250-ns delay was assumed. The basic ripple spacing for this echo is 1/250 ns, or 4 MHz. For a 50-dB echo, r will be 0.00316, and the peak-to-peak amplitude ripple will be 0.055 dB. Finally, the peak-to-peak delay ripple may be calculated as 1.58 ns.

It should be noted that the equivalence between echoes and ripples in gain and phase holds rigorously only for minimum-phase networks such as passive unequalized filters. Microwave repeaters do not have minimum phase properties because gain and phase equalizers are used to reduce their transmission deviations.

Discussion of Antenna Echo Objectives

To illustrate how echoes set a requirement on antenna system design, suppose the noise allocation to waveguide echoes in a 4000-mile system is 30 dBrnc0. In such a system there will be about 250 waveguide runs from radio transmitters and receivers to antennas at the tops of the towers. Hence, each antenna could be

allocated 30 — 10 log 250 = 6 dBrnc0, if the individual contributors add on a power basis. Since a 50-dB echo has been shown to produce +17 dBrnc0 in the previous illustration, the echo requirement on each antenna would be 61 dB instead of 50 dB. To limit echoes to this level or below places severe requirements on allowable irregularities in the waveguides. In dominant-mode waveguide systems, echoes usually arise from mismatches at the ends of the run, as in Fig. 21-8 or irregularities in the run. In other words, the requirement basically states that the sum of the return losses of the equipment and any waveguide discontinuity be 61 dB. This is usually interpreted as requiring that all impedances have 31-dB return loss minimum.

In waveguide systems capable of carrying higher order modes, as typified by the circular waveguide feeding the horn-reflector antenna, echo paths frequently exist in these higher order modes. The objective for echoes extends to these paths as well, indicating the need for precise antenna orientation, waveguide alignment, and the extension of return loss or reflection coefficient requirements to higher order modes as well as to the dominant mode.

REFERENCES

1. Liou, M. L. "Noise in an FM System Due to an Imperfect Linear Transducer," *Bell System Tech. J.*, vol. 45 (Nov. 1966), pp. 1537-61.
2. Rice, S. O. "Second and Third Order Modulation Terms in the Distortion Produced when Noise Modulated FM Waves Are Filtered," *Bell System Tech. J.*, vol. 48 (Jan. 1969), pp. 87-141.
3. Cross, T. G. "Intermodulation Noise in FM Systems Due to Transmission Deviations and AM/PM Conversion," *Bell System Tech. J.*, vol. 45 (Dec. 1966), pp. 1749-1773.
4. Laning, J. H., Jr. and R. H. Batten. *Random Processes in Automatic Control* (New York: McGraw-Hill Book Company, 1956), pp. 82-85.
5. Cross, T. G. "Power Density Spectrum of the Sum of Two Correlated Intermodulation Noise Contributors in FM Systems," *Bell System Tech. J.*, vol. 46 (Dec. 1967), pp. 2437-52.
6. Bennett, W. R., H. E. Curtis, and S. O. Rice. "Interchannel Interference in FM and PM Systems under Noise Loading Conditions," *Bell System Tech. J.*, vol. 34 (May 1955), pp. 601-636.

Chapter 22

Frequency Allocation

An important consideration in the design of multiple channel microwave systems is the choice of appropriate microwave, intermediate, and beating oscillator frequencies. A carefully chosen frequency plan can do much to enhance the capacity of a microwave radio relay system and reduce the design problems in its physical realization. A poor choice, conversely, may limit capacity, reduce reliability, and ultimately increase the costs of system manufacture, installation, and maintenance.

To illustrate the general problem, reference will be made to Fig. 22-1 which shows a typical long-haul microwave radio relay system. Two broadband radio channels, one carrying frequency-multiplexed message channels and the other a television signal, may be traced in this figure from their baseband sources at the left of the figure to their destinations at the right. The building blocks in this figure were discussed in Chap. 17 and are shown here assembled as a complete system.

The 70-MHz i-f outputs of the FM terminal transmitters at the left of Fig. 22-1 are applied to two radio transmitters for translation to microwave frequencies and amplification. The frequency translation is sufficiently different for the two channels that the two signals may be combined by microwave filters, transmitted over a radio hop, and then separated at a distant receiving station without excessive interchannel interference. The frequency translation is shown diagrammatically in Fig. 22-2.

At the receiving end of a radio hop, the FM microwave signals are separated by channel separation networks similar, if not identical, to the channel combining networks. They are then translated to the common intermediate frequency range. After amplification and

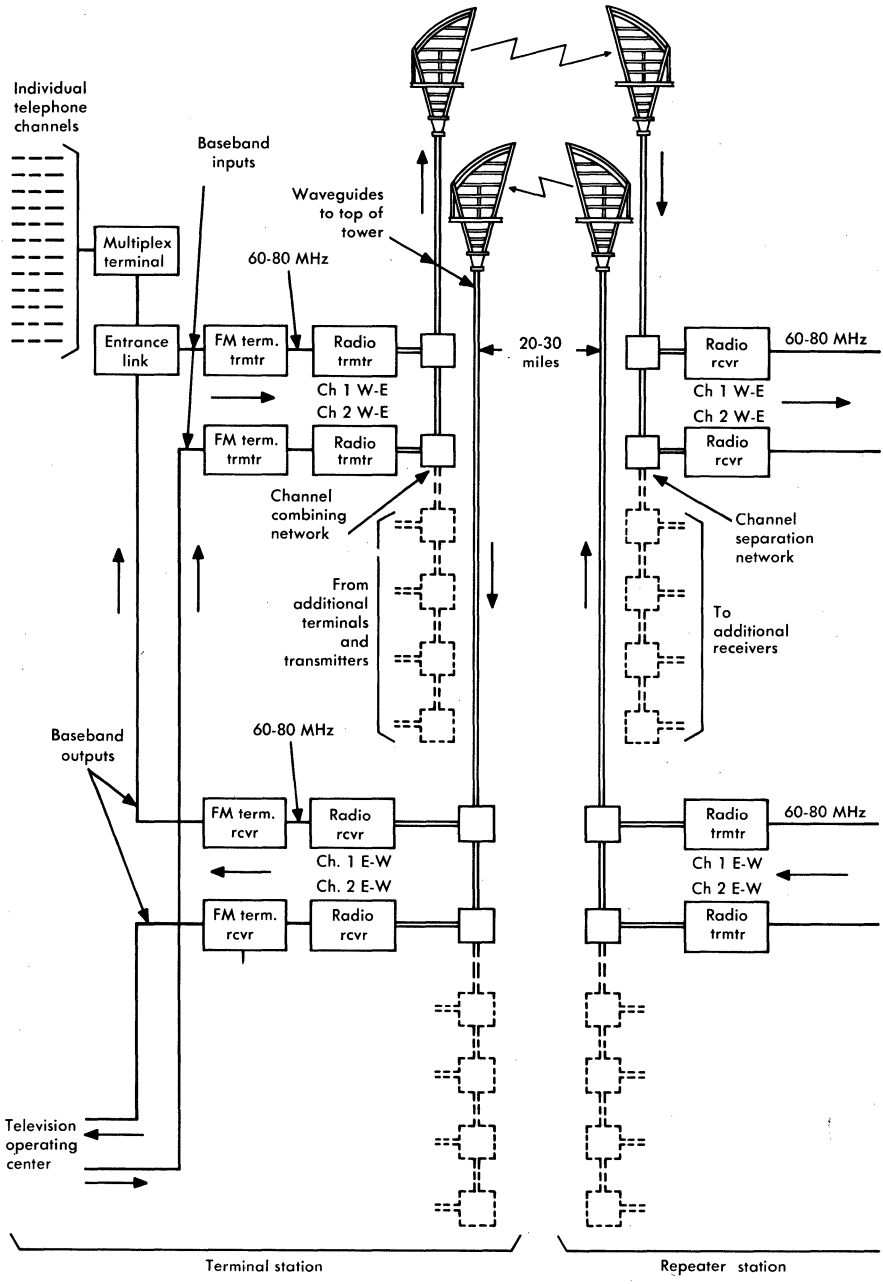


FIG. 22-1. Typical microwave relay system—block diagram.

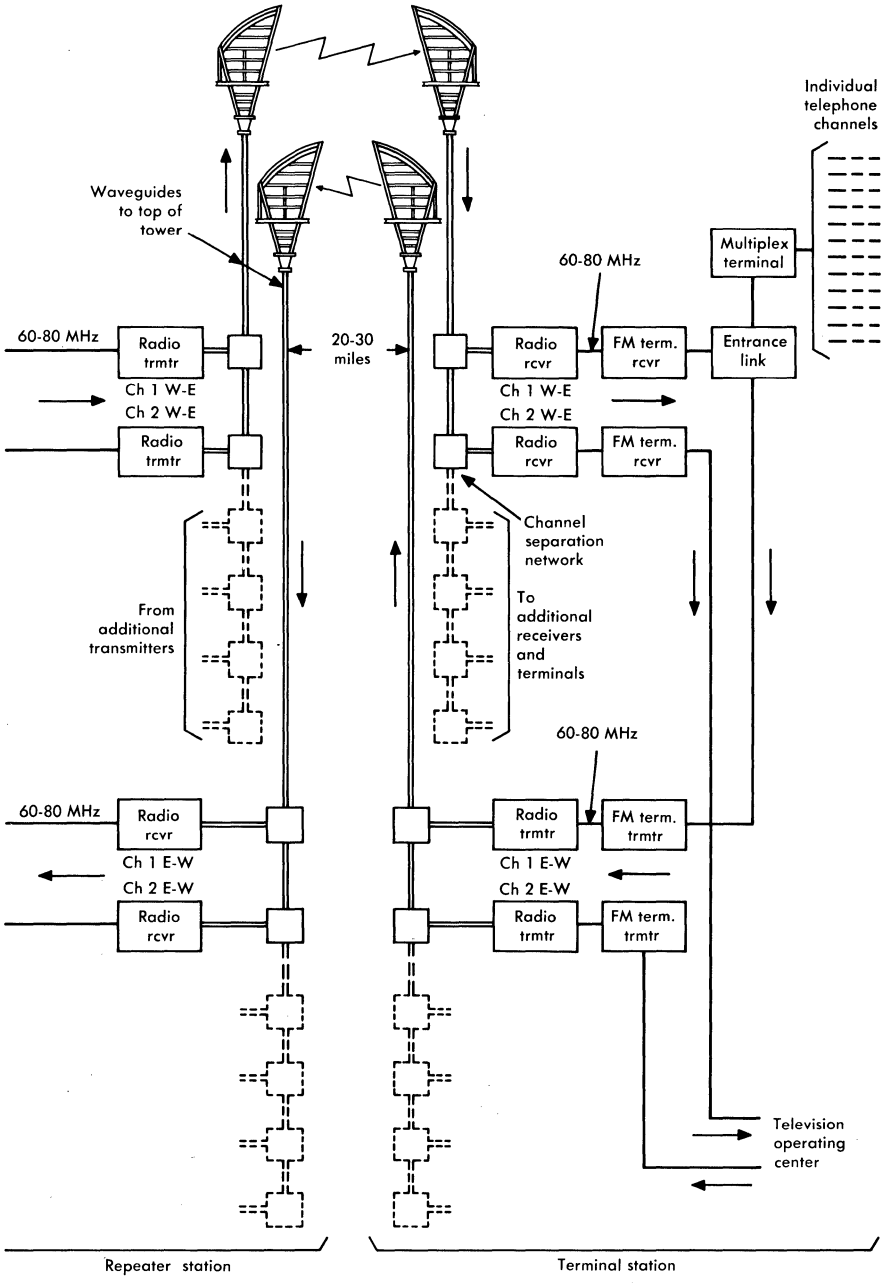


FIG. 22-1. Typical microwave relay system—block diagram (continued).

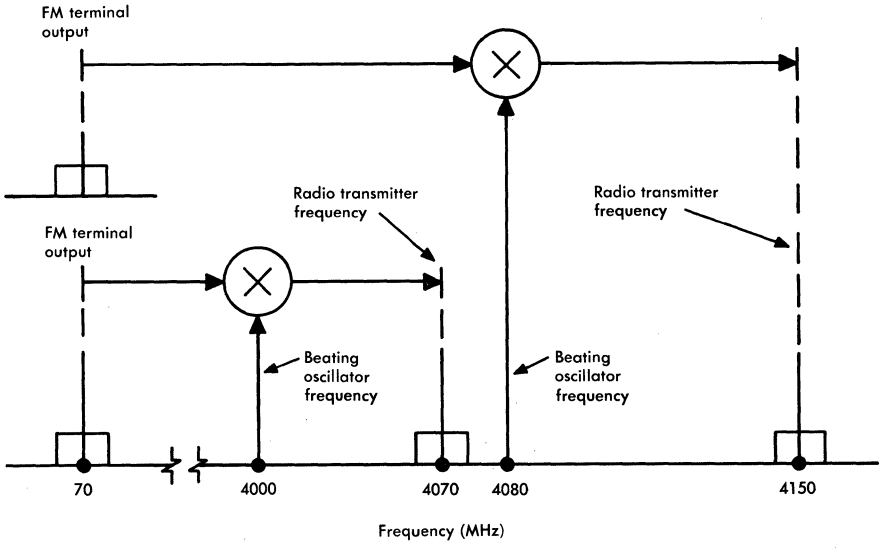


FIG. 22-2. Typical frequency translations.

other processing, the signals are retranslated to new microwave frequencies or demodulated to baseband depending on their destination.

A number of FM signals can be multiplexed at microwave frequencies and transmitted by a single antenna. This results in overall economy but at the same time gives rise to possible interchannel interference. As more radio channels are to be accommodated in the design, the sources of interference tend to increase. At the same time, the individual radio channel widths must decrease to stay within the overall allocated microwave spectrum. How then are the number and width of radio channels determined? At what frequencies should they be located? What spacing should be provided between channels to avoid interference? These questions are the crux of the frequency allocation problem. As in most engineering problems the solution is a compromise between economy and performance, and the factors which influence this compromise are considered qualitatively in this chapter.

First, consideration is given to some of the factors which determine both the channel bandwidth and the number of microwave channels a given system may have. These factors include legal

restrictions, type of service, message objectives, frequency deviation, and fade margin. For particular systems some of these factors are more important than others.

Next, some typical types of interference which may occur in multichannel microwave systems are considered. It is shown how certain frequency allocation techniques are helpful in minimizing these interferences.

Finally, the frequency allocations of Bell Laboratories designed microwave relay systems are listed to illustrate the results of these methods of allocation.

22.1 FACTORS INFLUENCING CHANNEL BANDWIDTH

Available Microwave Bands

One of the more obvious limitations on channel bandwidth is the available bandwidth allocated by the Federal Communications Commission (FCC). For fixed, common carrier service, the presently allocated bands above 890 MHz are given in Fig. 22-3.

Band (GHz)	Band edges (MHz)	Band width (MHz)
2	2,110-2,130	20
	2,160-2,180	20
4	3,700-4,200	500
6	5,925-6,425	500
11	10,700-11,700	1000

FIG. 22-3. Common carrier frequency allocations above 890 MHz.

Up-to-date information on available frequency allocations may be obtained from the *FCC Rules and Regulations,* Volume II, Part 2.

To date, the 4-GHz band has been used primarily for long-haul or "backbone" routes, while the 11-GHz band, which is more affected by rain attenuation, has been used for short-haul or "feeder" routes. The 6-GHz band is presently shared by long- and short-haul systems. The 2-GHz common carrier band has had only limited application because of the narrow bandwidths allowed. Some of the service channels for long-haul system administration and switching are carried in this band.

Desirable Baseband Width

In deciding how to utilize the available microwave bandwidth, a choice must be made between a few very broad microwave channels or a greater number of narrow ones. An important consideration in making this choice is the width of the baseband signal to be transmitted.

The choice of baseband width is often influenced by the requirements for television transmission. For example, in the TD-2 and TJ systems one of the baseband signals is a standard television signal requiring a 4-MHz baseband width. In the TH-1 system a 10-MHz baseband width was chosen so that the Bell System could accommodate 10-MHz high resolution television signals. In each of these systems, the television signals have influenced the choice of baseband width. A basic building block of telephone multiplex is the mastergroup of 600 message channels which has a top baseband frequency in the vicinity of 3 MHz. This block was found to be comparable to a television signal with respect to noise and distortion requirements for the TD-2 system.

For the TH system with a 10-MHz baseband, a single mastergroup of 600 circuits would be uneconomical of spectrum space and, accordingly, this system was designed to carry three mastergroups or 1800 message channels (see Fig. 6-9).

The newer TD-3 system is designed for two mastergroups or 1200 message channels using the same frequency plan as the older TD-2 system at 4 GHz but with higher power, lower input circuit noise, and improved distortion characteristics. Interestingly enough, the TD-2 system, originally designed to carry 360 to 480 message channels, can be modified to carry 1200 by applying new technology to certain critical areas. The resulting noise performance is not equal to the current objectives of the newer systems but meets the original objectives for TD-2.

Another consideration which imposes a limitation on baseband width is the influence of selective fading. As the width of the baseband and hence the i-f band is increased, selective fading makes the carrier-to-sideband phase relationship increasingly unstable.

Yet another consideration is rate of growth. A microwave system which operates in a band 500 MHz wide has a large ultimate

capacity. On some routes only a portion of this capacity may be required when the system is first installed. If a system has a cross-section of several microwave channels, the equipment for the individual channels may be obtained as needed rather than all at once.

Frequency Deviation

Frequency deviation has already been introduced as a factor in the determination of microwave bandwidth. A low value of peak frequency deviation leads to a system which is limited by idle noise and which requires high r-f power to meet its objectives.

On the other hand, a large frequency deviation requires much greater bandwidths, resulting in the uneconomic use of the available microwave spectrum, difficulties in obtaining linear FM modulators capable of large frequency deviations, and problems in the design of i-f circuits with adequate bandwidth. In practice, the deviation is selected to optimize the noise performance of the system by bringing the thermal noise and intermodulation noise into balance. This optimum deviation generally turns out to be 4 to 5 MHz and clearly makes a significant contribution to the required channel width.

Signal Quality

Although Carson's rule (Chap. 5) is a valuable guide in system design, it should be remembered that FM signals have sidebands which theoretically extend to infinite frequency. The rule is therefore a compromise between quality and economy, since some distortion is always introduced when the higher order sidebands are removed by filters or by other bandlimiting circuits. On the other hand, if two FM signals are placed in adjacent microwave channels and are not bandlimited, the higher order sidebands of each may extend into the other channel and cause interference. By taking into account the particular signals to be transmitted in adjacent channels with specified spacing, it is possible to choose a filter shape which will minimize signal degradation due to these two causes. In any case, however, the signal quality will decrease and the filter requirements will become more severe as microwave channels are spaced closer together.

In long-haul systems designed for "backbone" routes where efficient bandwidth utilization is very important, adjacent microwave channels are nearly contiguous, with little or no separating guard band so that channel width and channel spacing are about equal.

Short-haul, or feeder systems, on the other hand, do not normally need the full transmission capacity of the microwave band, and adjacent channels are separated by a guard band to ease channel filter requirements. In this case, channel spacing is larger than channel width.

Before leaving the subject of signal quality, a few words should be said about Carson's rule which is mentioned here and in Chap. 5. As a criterion for determining minimum channel bandwidth, it is a good approximation at high and low modulation indices. In the intermediate range, however, its reliability is less clearly defined. Nevertheless, since the rule maintains its applicability at modulation extremes and in practice has proved reasonably satisfactory in the intermediate range, it is considered a suitable starting point in the development of frequency allocation plans.

Cost

To establish a microwave radio route, suitable repeater locations are needed. New roads often have to be constructed, buildings and towers have to be built, and antenna systems must be installed. These costs tend to be fixed and independent of the number of radio or telephone channels; therefore, as the number of telephone channels is increased, the fixed cost per telephone channel is reduced. For a given loading per radio channel, this can be accomplished by moving adjacent radio channels closer together in frequency. However, as the channel separation is reduced, filter requirements become more severe, tolerances on i-f and microwave frequencies become tighter, and, as a result, the system becomes more complex and expensive. In general, then, it is found that the total cost consists of some fixed costs and some variable costs. The fixed costs per channel tend to decrease as more telephone channels are added, whereas the variable cost per channel tends to increase after a certain point. The optimum point depends largely on the present and projected demand for route capacity for either television or message signals or both. Long-haul systems designed for large capacity use less frequency space per talker than do short-haul systems which frequently use only two or four radio channel assignments on a given hop.

Protection Against Deep Fades

As was pointed out in Chap. 18, microwave channels are subject to fades of various magnitudes and durations. The increase in path loss which occurs during moderate fades can be compensated by the

radio receiver automatic gain control. The penalty in these cases is a corresponding increase in idle noise contributed by the faded hop. Deep fades, however, may seriously degrade the system noise performance. Fortunately, deep fades tend to be frequency-selective so that if the allotted band is wide enough and if there are several microwave channels in the system, one or two of them may be reserved for protection purposes. When a regular channel fails as a result of a fade or equipment failure, or when it is necessary to perform routine maintenance on the channel, its signal may be transferred to a spare channel by rapid switching techniques so that there is little or no signal degradation or interruption. Such protection greatly increases the reliability of the system and is very important on heavily loaded routes.

It must be realized, however, that such protection reduces the route capacity of the system. This makes it economically desirable on long-haul systems to let one or two radio channels act as spares for several working channels and to switch sections of up to ten hops rather than single hops only. The cost of the relatively complex switching equipment can be justified in such cases since it is spread over a large number of telephone circuits.

In short-haul systems the situation is usually different. The smaller number of radio channels carried by the system makes the cost of switching equipment per telephone circuit more significant. At the same time the need for more frequent dropping and adding of circuits at intermediate points along the route demands shorter switching sections. The objective in short-haul systems is therefore to simplify the switching equipment as much as possible. This is customarily done by providing protection on a one-for-one basis. By parallel feeding two transmitters with the same baseband signal on each hop, no transmitter-end switching is necessary, and an automatic switch selects the better of the two channels at the receiving end. This simplified switching system is less efficient in the use of available bandwidth than are the long-haul switching systems.

At 11 GHz, another significant source of outage is rain. Rain attenuation is not frequency-selective within the 11-GHz band, and frequency diversity within the band therefore offers no protection. A one-for-one diversity plan using the TM-TL systems offers an attractive solution. Here the 11-GHz and 6-GHz channels operate as crossband diversity pairs providing the reliability of 6-GHz transmission without using duplicate microwave channels in the more desirable 6-GHz band.

It is apparent from the preceding that the number of protection channels and the switching methods adopted for a group of working channels are very important design considerations. The statistical analysis of system outage time due to fading, equipment failures, and maintenance practices, as well as a detailed study of economics, is involved.

In summary, then, the factors of available bandwidth, baseband width, frequency deviation, signal quality, cost per telephone channel, and protection against fading all must be considered in determining the frequency allocation of a microwave system.

22.2 INTERFERENCE IN MICROWAVE CHANNELS

Consideration is now given to the various types of interference which may occur in a multichannel microwave system and their influence on the choice of frequency allocation. Six sources of interference are:

1. In-channel interference.
2. Image channel interference.
3. Adjacent channel interference.
4. Direct adjacent channel interference.
5. Limiter transfer action.
6. Single-frequency interference.

Each of these is considered separately. External interferences are assumed to be kept within limits by FCC regulations and are not discussed here.

In-Channel Interference

Several sources of r-f interference are illustrated in Fig. 22-4 which shows in block form four consecutive repeaters and five potential interference paths. Separate receiving and transmitting antennas are shown, although in some systems only a single antenna may be used. The two most potentially serious interference paths are those labeled 1 and 2. In these paths, high-level signals from transmitting antennas interfere with the low-level signals at the receiving antenna. In determining permissible interference levels, it is important to note that the signal level at the receiving antenna should be the signal level expected when the desired channel is experiencing the deepest allowable fade. The high-level signal from path 2 will be reduced by the back-to-back ratios of the two antennas, but only the side-to-side loss between antennas attenuates the signal from path 1.

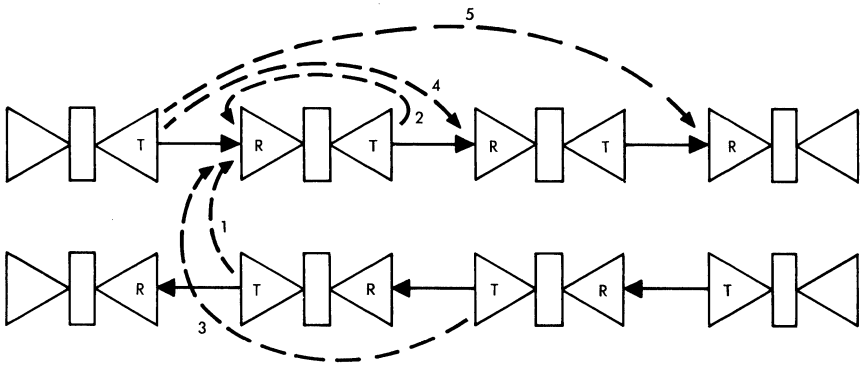
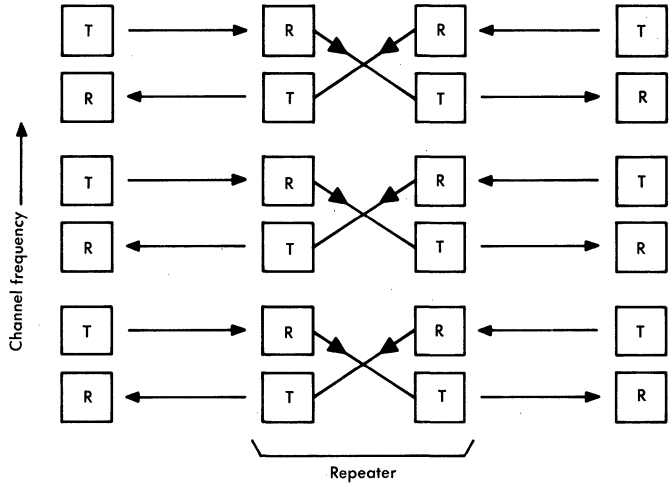


FIG. 22-4. Radio frequency interference.

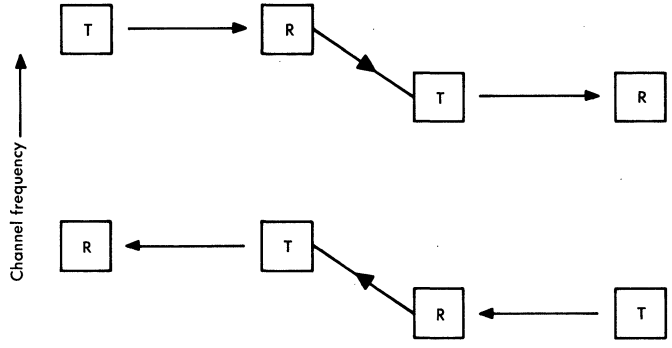
In practice, this amount of loss is not adequate. Additional loss might be introduced if the desired signal were cross-polarized with respect to the interfering signal, but help from this source should not generally be counted on. This results from the complex and unknown nature of the coupling path, particularly with respect to its influence on polarization direction. The problem of excessive coupling is avoided, however, by using different transmitting and receiving frequencies at a given repeater. One arrangement of this type, known as a two-frequency allocation, is shown in Fig. 22-5(a).

Another potential source of interference is shown as path 3 in Fig. 22-4. With a two-frequency allocation, two signals are received from opposite directions on the same frequency, and normally will be at about equal field intensity. In this case, the interference will be reduced by the front-to-back ratio of a single antenna, which may be about 70 dB for a horn-reflector antenna but only about 40 to 50 dB for a parabolic antenna. Further advantage might be obtained by the use of orthogonal polarizations, but this is of little practical value in the case of foreground reflections or scattering. Engineering studies show that a minimum front-to-back ratio of about 66 dB is required in long-haul systems and except for a few cases where severe foreground reflections occur, this requirement is met by the horn-reflector antenna.

Microwave systems using the parabolic antenna may not have adequate front-to-back ratios when using a two-frequency allocation. The four-frequency allocation shown in Fig. 22-5(b) avoids this



(a) Two-frequency plan



(b) Four-frequency plan

FIG. 22-5. Frequency allocation plans.

problem and is commonly used with short-haul systems. The split-channel plan, used at 6 GHz for short-haul TM-1 radio and listed later in this chapter, is a four-frequency plan derived from the long-haul TH-1 plan by splitting the channels and utilizing half of the resulting reduced-width channels for each direction of transmission.

Two other sources of interference are shown by the paths labeled 4 and 5 in Fig. 22-4. In the commonly used allocation plans of Fig. 22-5, the path 4 type interference presents no problem since frequency

frogging is used in adjacent hops. Over-reach interference, represented by path 5, is potentially troublesome but can be reduced to tolerable proportions by locating the transmission path in every third hop slightly out of line.

Image Channel Interference

Image channel interference is illustrated in Fig. 22-6. Two signals with carrier frequencies of 11,000 and 11,140 MHz are shown. These signals are separated with filters and are applied to modulators where they are translated to an intermediate frequency of 70 MHz. Suppose a beating oscillator frequency of 11,070 MHz is used for the 11,000-MHz signal. If the filters were ideal, there would be no problem. If, however, the rejection of filter 1 at 11,140 MHz is inadequate, this signal will beat with 11,070 MHz to form an extraneous 70-MHz i-f output. This effect is known as image channel interference, where the image channel is defined as the channel which differs in frequency from the beating frequency by the same amount as the desired channel but is on the other side of the beating frequency.

Incidentally, in this example the 11,070-MHz beating oscillator of channel 1 could cause trouble by entering the channel 2 receiver

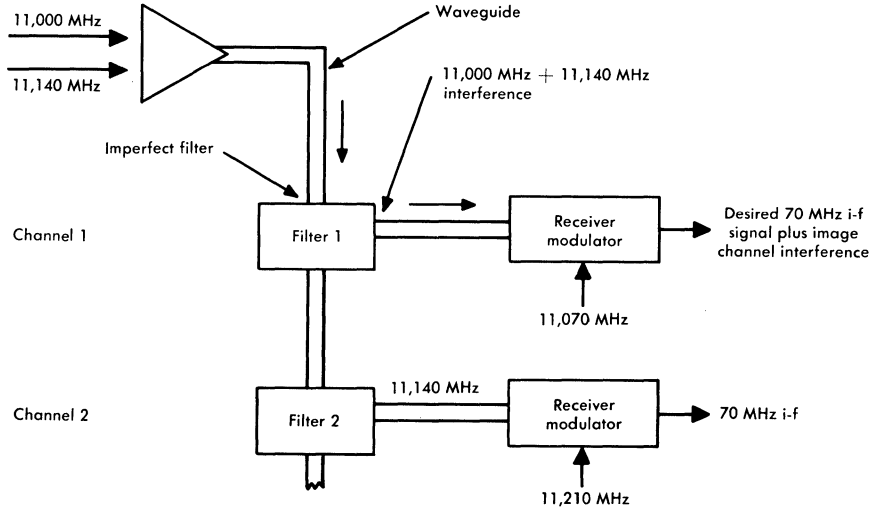


FIG. 22-6. Image channel interference.

modulator and mixing with its 11,140-MHz input signal to form a 70-MHz interference. This example illustrates two of the interactions which must be considered in modulator and filter design.

Adjacent Channel Interference

Adjacent channel interference, mentioned previously, occurs when two FM channels are spaced in frequency so that the sidebands from one extend into the other. Figure 22-7 shows how this can happen, by making use of the power spectral density of a carrier which has been phase-modulated by an equivalent baseband signal consisting of random noise. Although the interference can be prevented by removing the higher order sidebands with filters before the two signals are combined, this cannot be done without causing some distortion of the signal in each channel.

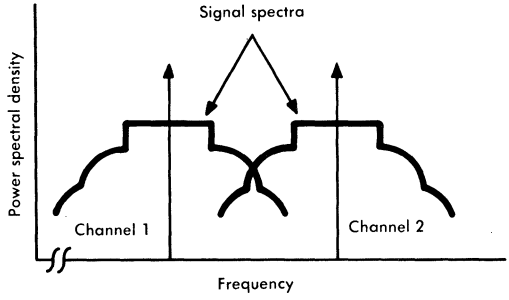


FIG. 22-7. Adjacent channel interference.

It is evident from Fig. 22-7 that channel spacing will have an important influence on the problem of filtering overlapping sidebands. This fact has given rise to a rule which is sometimes used in the design of long-haul systems. The rule states that channel spacing should be at least three times the top baseband frequency to ensure that second order sidebands from an interfering channel will not overlap first order sidebands in an adjacent channel. This rule generally leads to practically realizable channel filters.

Like many rules, this one is not rigid and any tentative allocation plan must be studied in detail from the point of view of sideband overlap. Such a study must take into account the nature of the modulating signals which determine the sideband energy distribution in the FM spectrum. The important advantages to be obtained by

using orthogonal polarizations in adjacent channels should also be taken into account.

Direct Adjacent Channel Interference

Interference from overlapping sidebands will be garbled or unintelligible since the disturbing sidebands are not correlated with the disturbed carrier. In systems with closely spaced channels, however, a more complicated form of adjacent channel interference has been noted in which the interference is intelligible. This type of interference, in which the signal on the adjacent channel appears as an identical signal in the disturbed channel, has been termed direct adjacent channel interference (DACI).

The transfer mechanism involves interference with the zero crossings of the desired signal and tends to take place in amplifier stages where compression or limiting occurs. The mechanism producing this type of interference has been studied [1] but is still not fully understood.

Limiter Transfer Action

The use of limiters in the radio repeater may result in an interference between channels which is referred to as limiter transfer action. The basic mechanism of this interference is illustrated in Fig. 22-8. Three adjacent radio channels with carriers spaced 20 MHz apart are shown. Channel 2 is assumed to be cross-polarized with respect to channels 1 and 3. Assume that channel 1 is carrying a baseband sinusoid at 7 MHz with a sufficiently low index of modulation that only the first-order sidebands need to be considered. At station A, the output of channel 1 will therefore consist of a carrier and single-frequency sidebands 7 MHz on each side of the carrier. At station B, the upper sideband of the channel 1 signal will appear as an interfering sinusoid 13 MHz from the center frequency of the channel 2 receiver. The amplitude of the interference which reaches the channel 2 limiter will depend on the cross-polarization discrimination between channels 1 and 2 on the station A to station B path, as well as the channel 2 receiver gain 13 MHz from center frequency. The channel 2 carrier and the interference represent a composite AM and PM signal at the input to the limiter, the AM component of which will be removed by the limiter. The PM component remains, however, and at the limiter output there will appear the carrier and sidebands 13 MHz on each side of the carrier. This signal is transmitted by channel 2 at station B. At station C, the upper sideband

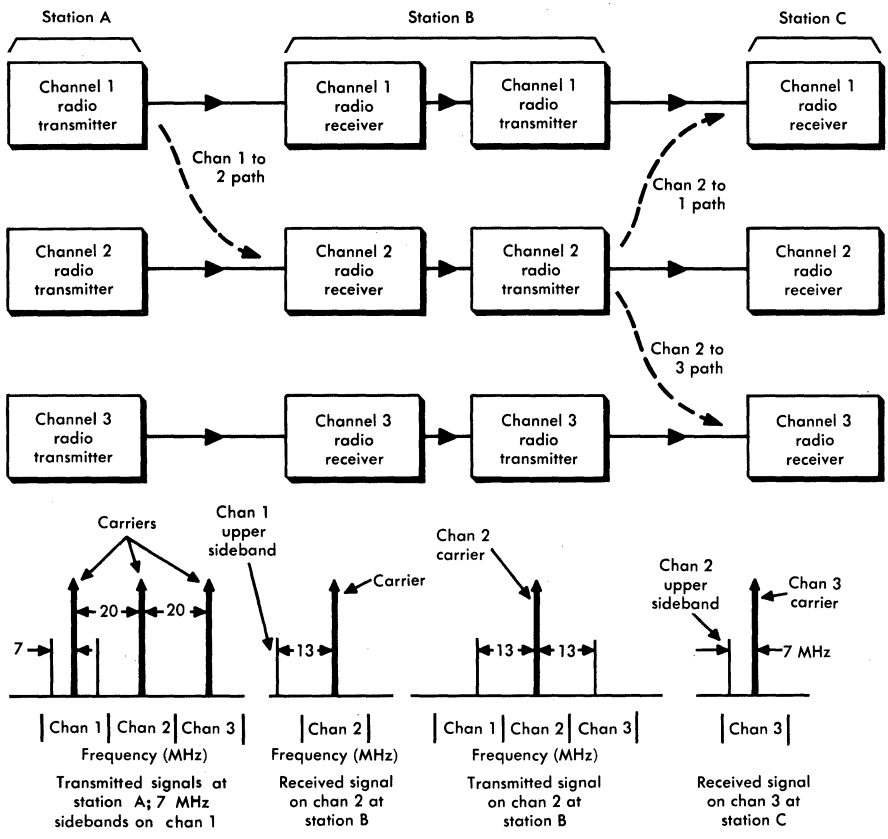


FIG. 22-8. Limiter transfer action.

of the channel 2 signal will appear as an interference 7 MHz from center frequency in the channel 3 receiver. The amplitude of this interference will depend on the cross-polarization discrimination between channels 2 and 3 on the station B to station C path, as well as the loss of the channel 2 radio transmitter 13 MHz from its center frequency. Hence, by this mechanism, a baseband signal in one channel may induce an interference at the same baseband frequency in another channel. In addition, of course, the channel 2 lower sideband can couple back to the channel 1 receiver at station C. In either case, an analysis should also include a study of the frequency tolerances of the various beating frequency oscillators in-

volved in the paths to determine the actual frequency range about the nominal within which the interference may appear at the third station.

Tests have shown that limiter transfer action is in exact accordance with elementary FM theory. The mechanism may become an important consideration when a high-level sinusoid is present on a channel or, as in the case of the auxiliary channels in TH-1, when a carrier is located near the edge of a broadband radio channel. For either condition, the choice of the frequency allocation plan may affect the seriousness of the problem in the radio system. As seen from the discussion, the losses to be considered are two cross-polarization discriminations and the loss, with respect to center frequency, of one repeater at the frequency of the interfering sideband. In addition, allowance should be made for the transfer of energy from one side of the carrier to the other by the limiter, which causes the ratio of the carrier to one sideband at the limiter output to be 6 dB lower than the carrier-to-interfering sideband ratio at the limiter input.

Single-Frequency Interference

It should be apparent that single-frequency interference may appear in the radio channel by way of any of the interference mechanisms previously described, whenever the interfering channel contains single-frequency components in its sidebands. In this case, the presence of the interference is dependent on the modulation in the interfering channel. Equally important are interferences from carriers or beating oscillator signals in the system. Some of the mechanisms already described may cause interferences due to the presence of these signals; in other cases, different mechanisms are involved.

In-channel interference will be single-frequency in character whenever the index of modulation in the disturbing channel is low or when, in the extreme, the carrier is unmodulated. In either case, the interference will consist primarily of the high-level carrier component, and may be treated as a single-frequency interference. It might seem that interference from a carrier on the "same" channel would not be very serious since ideally both carriers will have the same frequency. Neither signal, however, is likely to be exactly at its nominal frequency, and a frequency difference of several megahertz may actually exist, thereby placing the interference in the active part of the baseband of the disturbed channel.

If the frequency allocations have been judiciously selected, same-channel interference at radio frequencies is rare. It might occur, for example, when freak transmission conditions result in "over-reach" from a distant repeater at a time when the disturbed repeater is subject to a moderate fade. Same station crosstalk at i-f is, however, a much more probable occurrence since repeater stations and terminals are rich in 70 and 74 MHz signals.

A single-frequency interference objective of −68 dBm0 in a voice channel has been used in the engineering of some systems such as TH-3. This leads to a desired carrier-to-interference ratio requirement which will range from 60 to over 100 dB depending on frequency offset. In some cases, a more lenient requirement may be applied if one of the signals causing the interference is derived from a source which has residual frequency modulation. For example, the carrier from an FM terminal transmitter may contain considerable low-frequency modulation due to oscillator shot noise and power supply harmonics and will therefore jitter in frequency. As a result, the interference from this carrier will not be a clean sinusoid but will instead be smeared or spread over a narrow frequency band. To the listener on a telephone channel, the interference may sound more like a "burble" and will be less objectionable than a single-frequency tone. Hence, a 10-dB reduction in the single-frequency interference requirement has been used in the past for these cases to make the subjective effect of the "burble" approximately equal to that of a clean tone.

If an interference recurs regularly along the radio route (e.g., at each repeater station), it is important to determine whether or not the interference always appears at exactly the same baseband frequency. If it does, allowance may have to be made for systematic (in-phase) addition of the interference. In many cases, however, the permissible frequency differences in beating oscillator signals at the repeater stations will be sufficient to prevent systematic addition, and the interferences will instead be spread over some region about the nominal frequency in these multiexposure cases.

22.3 FREQUENCY ALLOCATIONS FOR EXISTING SYSTEMS

The discussion of frequency allocation has thus far been largely qualitative. In the following figures the allocation plans for the various Bell Laboratories developed radio relay systems are presented with commentary where appropriate to illuminate special details.

In the following diagrams, Figs. 22-9 through 22-12, the vertical scale represents frequency, and the height of the small boxes indicates the bandwidth of a single channel. In each case, the frequency arrangements at a single repeater have been shown. The situation at adjacent repeaters will differ only to the extent that transmitters and receivers will be interchanged, since a given transmitted signal in any of the figures will be received at the adjacent station on the same frequency.

The beating oscillator frequency associated with each channel is shown in each figure. Thus, in Fig. 22-10, TH-1 channel 18 receiver uses a beating oscillator frequency of 6078.6 MHz. The signal is then transmitted on channel 28 using a transmitter beating oscillator frequency of 6330.7 MHz.

At this point it should be evident that diagrams of this type effectively display large amounts of information about the frequency allocation of a system. In the following sections, therefore, only brief descriptions will be included. The individual diagrams should be studied and compared.

4-GHz Long Haul

The TD-2 and TD-3 systems both use the frequency plan of Fig. 22-9. This is a two-frequency plan using four antennas (two in each direction) at each repeater station.

The TD-2 system originally provided six two-way microwave channels, one of which was used for protection. Single polarization was employed, and on any one hop either channels 1 to 6 or channels 7 to 12 could be used. Both groupings could not be used simultaneously because of excessive adjacent channel coupling. By using dual polarization and improved discrimination at i-f, it has been possible to reduce adjacent channel interference to the point where all 12 channels can be used simultaneously on the same path. This is known as interstitial operation and has doubled the system capacity. Interstitial operation includes an additional protection channel providing ten working and two protection channels.

6-GHz Long Haul

In contrast to the identical TD-2 and TD-3 plans at 4 GHz, the 6 GHz TH-1 and TH-3 plans shown in Fig. 22-10 diverge significantly. Although they operate on the same radio channel assignments, the beating oscillator assignments have been changed in the

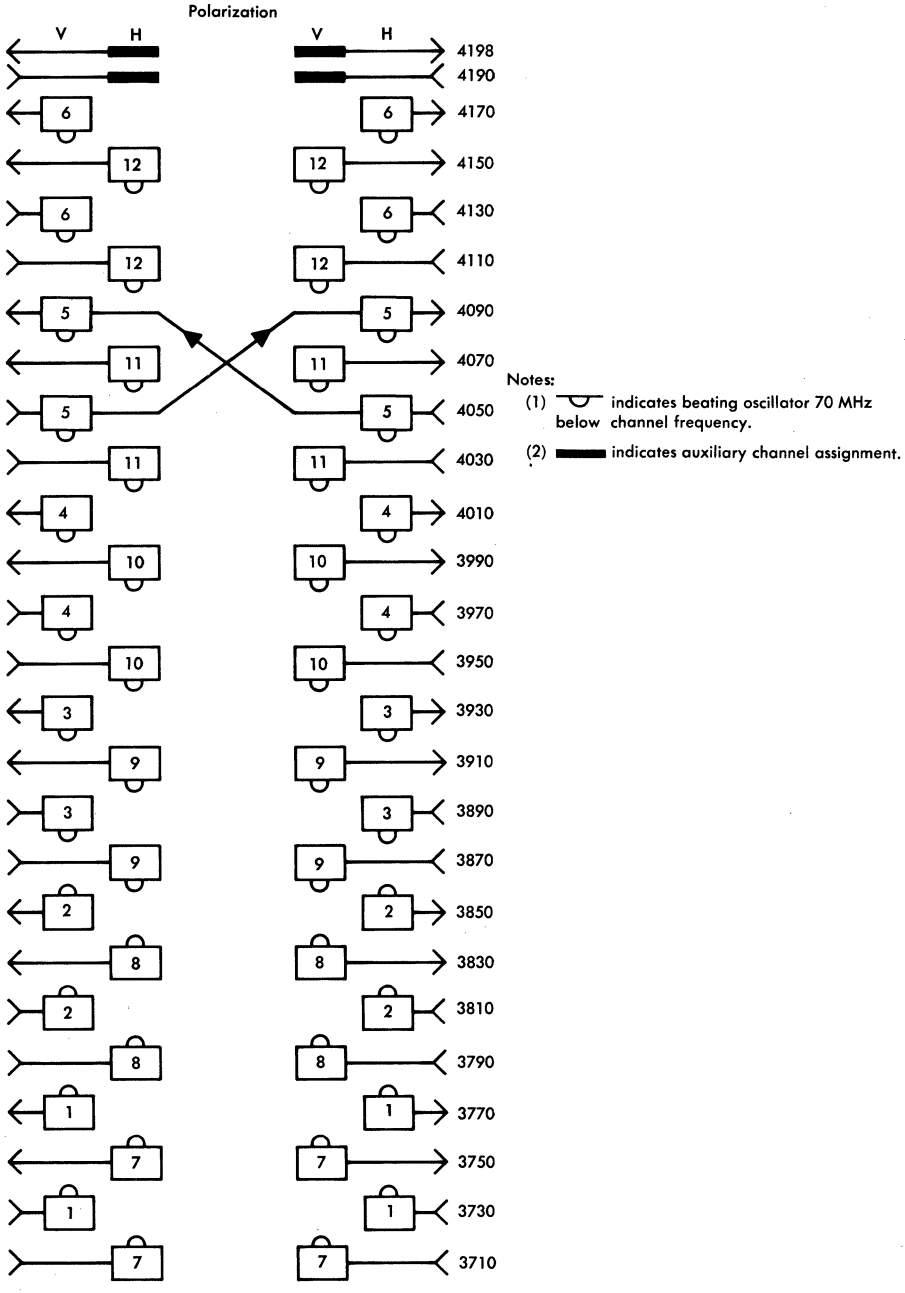
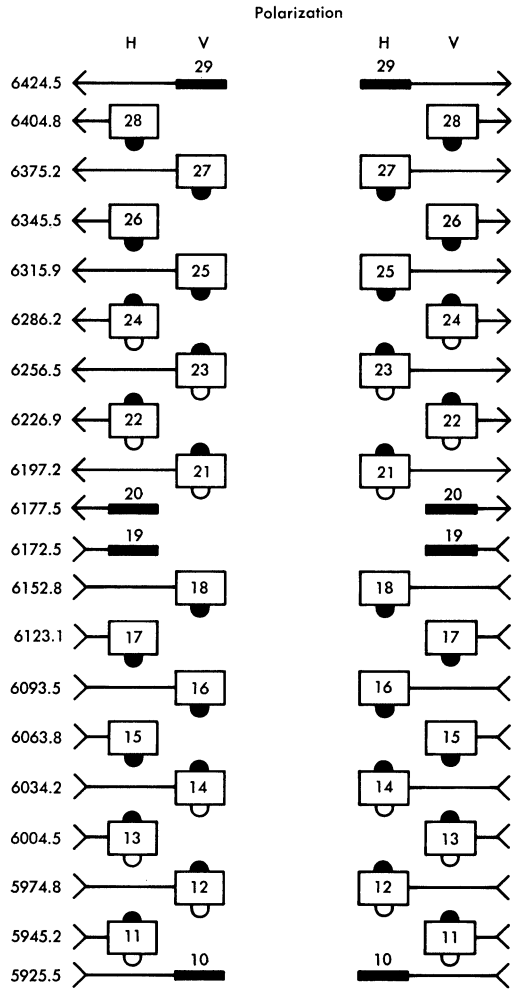


FIG. 22-9. TD frequency plan.



Notes:

(1) ⌒ indicates beating oscillator assignment above channel frequency.

(2) ∩ indicates TH-3 beating oscillator assignment if different from TH-1.

(3) Oscillator frequency offset for TH-1 is 74.1 MHz.

(4) Oscillator frequency offset for TH-3 is 70 MHz.

(5) ▬ indicates TH-1 auxiliary channel assignment.

FIG. 22-10. TH frequency plans.

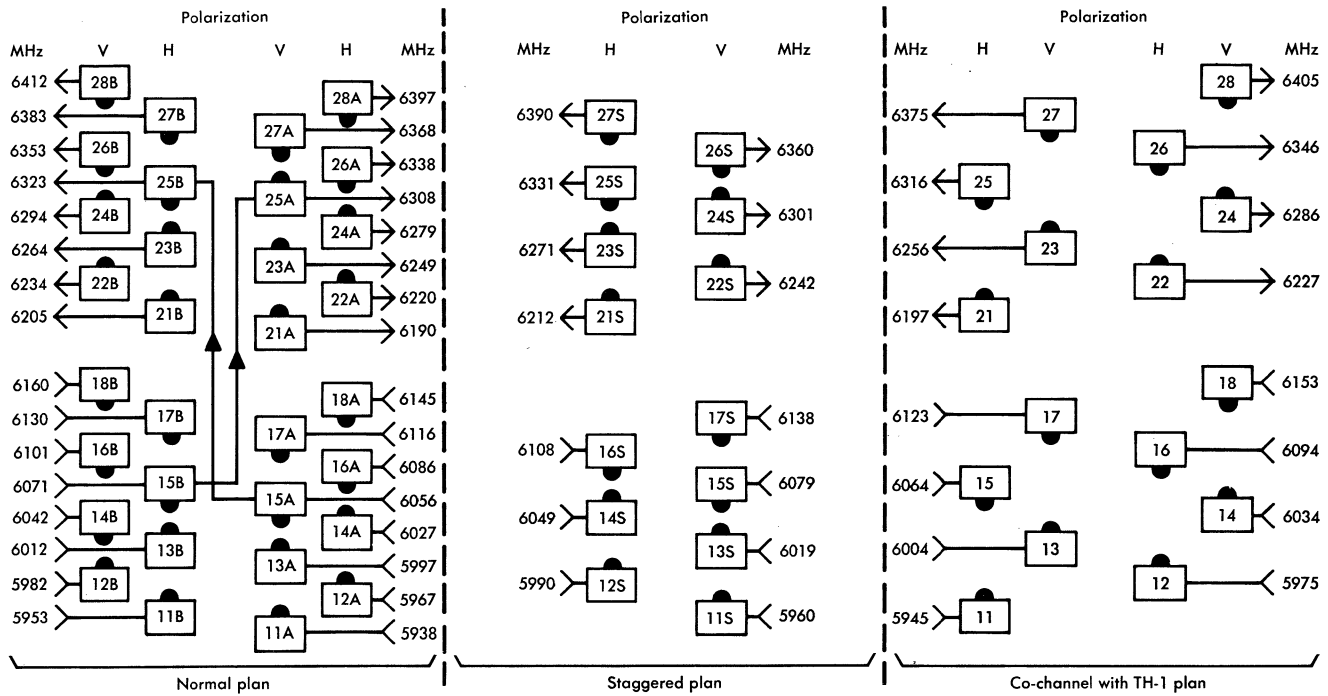
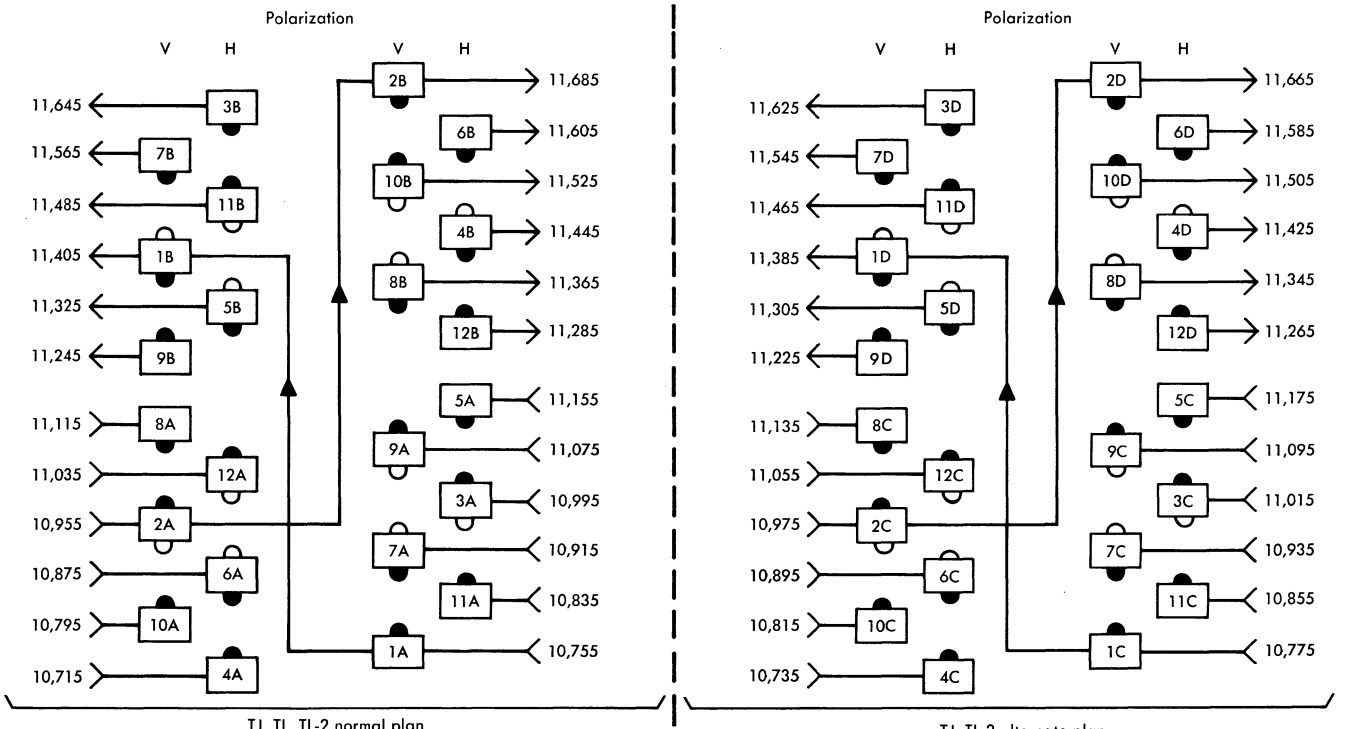


FIG. 22-11. TM-1 frequency plans at 6 GHz.



Note:

(1) ● indicates beating oscillator 70 MHz *above* channel frequency.

(2) ∩ is TJ assignment if different from TL, TL-2.

FIG. 22-12. 11-GHz frequency plans.

newer TH-3 plan to make the intermediate frequency 70 MHz, which is compatible with the other radio systems. It may be seen in Fig. 22-10 that all of the TH-3 beating frequencies are below their respective channel frequencies. This departure from the TH-1 plan results in improved transmitter modulator characteristics. This improvement and the shift to an intermediate frequency of 70 MHz impose stringent requirements on spurious couplings; e.g., the output of a channel 18 transmitter falls in the image region of a channel 24 receiver.

One outstanding feature of the TH plans and all that follow is that the transmitters at any given station are placed in one half of the allocated band and the receivers in the other half. This reduces side-to-side interference between adjacent channels (path 1 in Fig. 22-4) but increases the likelihood of simultaneous fading of channels in a given direction by reducing the frequency range by half.

Choice of the Intermediate Frequency

Several factors influence the choice of the intermediate frequency. The overriding influence in recent designs has been compatibility with earlier systems.

Using the development in Section 10.2, it can be easily shown that a lower bound for an intermediate frequency in a typical radio system would be about 30 MHz. Practical limitations of stray capacitance, cable loss, and active device parameters historically have set the upward bound at approximately 100 MHz.

The original TD-2 plan (Fig. 22-9) placed the beating oscillator frequencies half way between the channel assignments in order to avoid oscillator leakage interference. The allowable intermediate frequencies under this restriction are odd multiples of 10 MHz.

Finally, the suppression of image channel interference generally requires about 100 dB of rejection ahead of the receiver modulator at the image frequency. To achieve this suppression with realistic filters requires that the i-f be sufficiently high that the image band occurs over approximately 100 MHz from the channel center frequency.

The almost universal use of a 70-MHz intermediate frequency in Bell System microwave radio systems is based on these factors. A notable exception is the TH-1 plan which uses 74.1 MHz and was designed for minimum beating oscillator interference with the channel plan of Fig. 22-10. In the development of TH-1, no need was

seen for i-f compatibility with other systems. The more modern concepts of system restoration over mixed facilities have led to the hybrid TH-3 plan, again using 70 MHz.

6-GHz Short Haul

The short-haul plans at 6 GHz have evolved from the long-haul plans in an attempt to reduce intersystem interference. The three plans listed in Fig. 22-11 were generated as follows:

1. Split-frequency plan. Each long-haul channel assignment was split to derive a four-frequency plan for eight half-width channels in place of eight full-width TH-1 (long-haul) channels.
2. Staggered plan. This plan assigns channels half-way between TH (long-haul) assignments primarily to avoid interference with crossing routes.
3. Co-channel plan. This plan uses the regular TH (long-haul) assignments directly to obtain four half-width channels in a four-frequency plan.

11-GHz Short Haul

The 11 GHz plans in Fig. 22-12 are each designed for twelve two-way radio channels on a four-frequency basis. The two plans are not intermixed on a given route but are intended for crossing routes to minimize interference.

REFERENCES

1. Ruthroff, C. L. "A Mechanism for Direct Adjacent Channel Interference," *Proc. IRE*, vol. 49 (June 1961), p. 1091.
2. Kinzer, J. P. and J. F. Laidig. "Engineering Aspects of the TH Microwave Radio Relay System," *Bell System Tech. J.*, vol. 40 (Nov. 1961), pp. 1485-1491.

Chapter 23

Illustrative Radio System Design

This chapter applies the material of previous chapters to a simplified design of an actual system. Obviously, more considerations exist than can be discussed here, but the following treatment will be useful in tying together some of the topics already presented.

Many factors are involved in a system design, and a straightforward procedure is not always possible. The designer must provide a system which meets certain transmission and reliability objectives, and at the same time he must try to minimize the overall cost to the ultimate customer. The system should be designed for easy maintenance. The design chosen should be capable of being developed within reasonable time limits. The time allowed may be a few months in some cases, or several years in others. In general, there is no single solution to the overall problem. Not all of the considerations can be reduced to mathematical terms, and many can best be resolved by judgment and experience based on previous designs.

As might be expected, the early design work often consists of calculating the performance of several possible arrangements. Extrapolations from previous designs may be helpful. These preliminary calculations can then be used as a guide for further work.

23.1 SYSTEM OBJECTIVES

The system to be developed for illustration is a new 6-GHz long-haul facility to be used on existing 4-GHz TD-2 routes. Some preliminary objectives for this system are listed as follows:

1. The length of the system is to be 4000 miles with 150 hops.
2. The noise performance should be 41 dBrnc0 maximum in the noisiest multiplexed telephone channel for unfaded busy-hour

conditions and 55 dBrnc0 maximum with one hop faded by 40 dB. (The objective of 41 dBrnc0 is 3 dB more stringent than the earlier TH-1 objective of 44 dBrnc0.)

3. The capacity per microwave channel is to be determined.
4. Antennas are to be existing horn-reflector antenna systems with additional networks for 6 GHz.

23.2 DESIGN PROCEDURES

The first choice to make is that of a frequency allocation plan. The 6-GHz band is no longer uncongested as it was for the design of TH-1. The presence of numerous systems (both Bell System and otherwise) using frequency plans derived from TH-1 therefore forces the adoption of that channelization plan for the new system for compatibility in spectrum occupancy. The channel frequency spacing used in TH-1 is 29.7 MHz. If 4-MHz peak frequency deviation is assumed as a starting point (because it is used in other systems), the top modulating frequency according to Carson's rule would be

$$\frac{29.7 - 8}{2} \approx 11 \text{ MHz}$$

However, another rule (page 536), intended to prevent second order sidebands from overlapping with first order sidebands, is more restrictive and leads to a slightly lower top modulating frequency,

$$\frac{29.7}{3} \approx 9.9 \text{ MHz}$$

This frequency is comparable to the 8.524-MHz top frequency of a newly developed 1800 message circuit multiplex, and this load will be assumed for the remainder of the discussion.

With regard to the antenna system, the average value of the section loss [defined in Eq. (18-9)] for TD-2 is known to be 66 dB. Making corrections from 4 GHz to 6 GHz for increased path loss and antenna gains and allowing for slightly higher waveguide losses yields a section loss of 63 dB.

The next design step is to allocate the system noise objective of 41 dBrnc0 among the various noise contributors. A tentative division is made in Fig. 23-1, which assumes that the noise contributors add on a power basis. The allocations marked by a † in Fig. 23-1 are performance estimates of existing equipment. The r-f crosstalk allocation, similarly, is a carryover from previous systems. The remain-

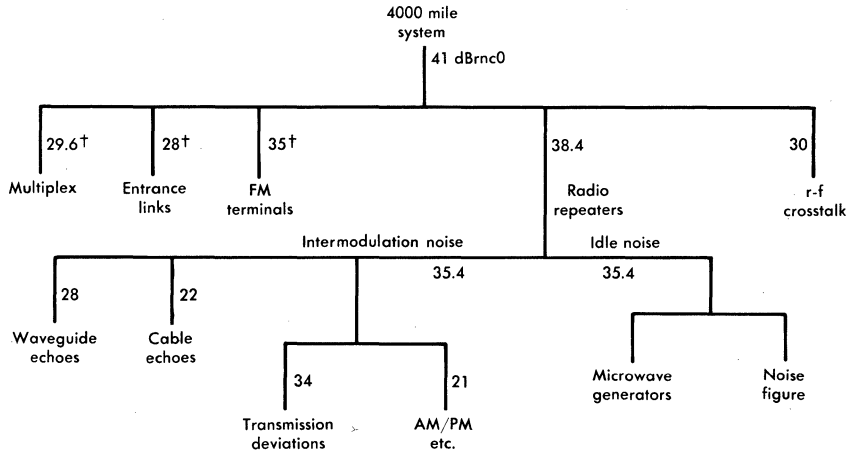


FIG. 23-1. Tentative noise allocation.

ing noise is divided equally between idle and intermodulation noise since experience has shown that radio systems are essentially idle and second order intermodulation noise limited.

Each of these is a noise objective for a complete 4000-mile, 150-hop system. Thus the objectives for the various building blocks of the system can be allocated. If any objective in the preliminary allocation cannot be attained, the division of objectives must be rearranged accordingly.

The idle noise objective of 35.4 dBrnc0 for 150 radio repeaters may be translated by use of Eq. (20-24) to a per-hop relationship of noise figure and transmitted power as follows. Substituting $\Delta F = 4$-MHz peak, $f_1 = f_T = 8.524$ MHz*, $P_n = -174 + N_F$ dBm/Hz (N_F is the repeater noise figure in dB), and $P_s = 27.6$ dBm0 for 1800 circuit loading into Eq. (20-24) yields:

$$35.4 - 10 \log 150 > 27.6 + 20 \log \frac{8.524}{4}$$

$$+10 \log (6 \times 10^3) - 174 + N_F - P_c + 88$$

or

$$27.6 \text{ dBm} \geq N_F - P_c$$

*The top channel is usually the noisiest in radio systems.

where P_c is the received carrier power in dBm at the point of noise figure measurement. Translating this level by the section loss of 63 dB, the required transmitter power and noise figure relationship is found to be

$$35.4 < P_T - N_F$$

If an estimate of 3.3 dB is made for the pre-emphasis advantage in the top message channel, the relationship becomes

$$32.1 < P_T - N_F$$

Practical noise figures for microwave repeaters tend to run between 7 and 12 dB and in this case would imply the need for a +39 to +44 dBm transmitter output power, currently obtainable only with traveling wave tube amplifiers. A traveling wave tube with +40.5 dBm output capability has been developed for this band. A higher output power could be obtained with further development but will not be needed since a repeater having a noise figure of 8.4 dB at normal signal power is within the reach of present technology.

The noise resulting from a 40-dB fade in one hop can be derived by using Eq. (20-24) and letting the section loss increase 40 dB to 103 dB so that P_c becomes $40.5 - 103 = -62.5$ dBm. The pre-emphasis advantage is again estimated as 3.3 dB. Thus, the noise is

$$W_N \text{ (one faded hop)} = 27.6 + 20 \log \frac{8.524}{4} + 10 \log (6 \times 10^3)$$

$$- 174 + 8.4 + 62.5 + 88 - 3.3$$

$$= 53.6 \text{ dBrnc0}$$

When this noise is added to the noise from the rest of the system, the sum is within the objective of 55 dBrnc0.

In retrospect, the noise figure value of 8.4 dB used in this last calculation is somewhat pessimistic. At low signal levels in practical repeaters, only the receiving-modulator preamplifier controls the overall noise figure since the other noise sources are masked by increased i-f amplifier gain. The reduction in noise figure may amount to 1 or even 2 dB.

The next step is to test for breaking on the faded hop. This is done by comparing the carrier level during a 40-dB fade with the total

thermal noise referred to the receiver input in an estimated channel noise bandwidth of 30 MHz:

$$-62.5 - [-174 + 8.4 + 10 \log (30 \times 10^6)] = 28.3 \quad \text{dB}$$

The resulting 28.3 dB signal-to-noise ratio assures that breaking will not be a problem in this system because breaking does not start until the signal-to-noise ratio drops to 10 to 15 dB.

The preliminary breakdown of the intermodulation noise objective for transmission deviations results in a per-hop allocation of 12.2 dBrnc0, assuming power addition as an initial estimate only. A computation along the lines of the noise spectrum analysis of Chap. 21 is next used to generate a set of figures showing curves of intermodulation noise in the top channel versus transmission deviation values, for each of the seven transmission deviations discussed in Chap. 21. These curves are derived assuming a representative preemphasis function and the system parameters previously used. An iterative comparison of the intermodulation noise objective with noise produced by the various transmission deviations is next used to set limits on the allowable transmission deviations in the radio repeater.

A similar procedure is used to set the objectives on antenna and waveguide echoes, AM/PM intermodulation noise, and any other noise contributors. The selectivity requirements for the various filters which are used in the system are based largely on the interference mechanisms outlined in Chap. 22. Many other factors will influence the final design in a manner similar to those discussed.

Chapter 24

Introduction to Digital Transmission

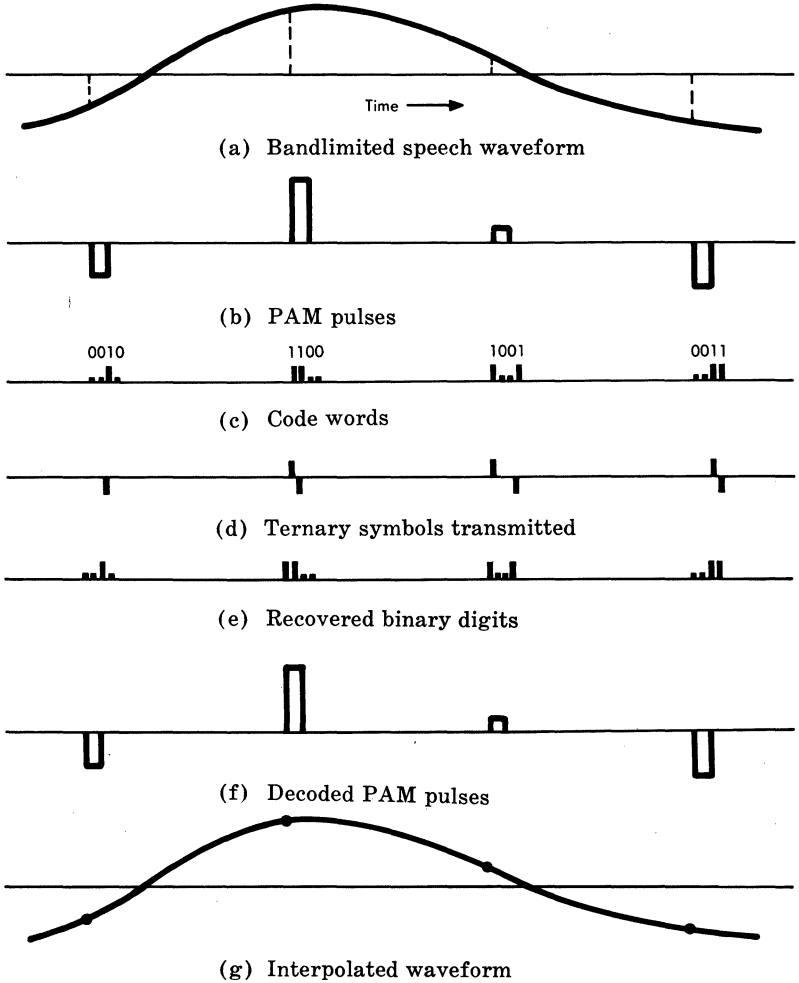
Previous chapters have discussed transmission systems in which the signals applied to the transmission media are continuous functions of the message waveform. It has been shown how a sinusoidal carrier can have either its amplitude or phase continuously varied in accordance with the message. In digital transmission systems the applied signals are discrete in both time and amplitude. Signals that are discrete only in time, such as pulse amplitude modulated (PAM) signals, or discrete only in amplitude, such as facsimile data, will not be considered here as digital signals since they cannot be carried by digital transmission lines containing regenerative repeaters. In the simplest case either a pulse or a space (no pulse) is transmitted in each unit of time. The stream of pulses and spaces can be thought of as binary numbers that represent analog signals to which sampling and appropriate coding rules have been applied. Although binary signals are easiest to generate, the actual signals usually found on cable media have more than two symbols. The most prevalent are ternary using three symbols (i.e., positive pulse, negative pulse, and no pulse) and quaternary using two different amplitude pulses for each polarity.

The advantage of converting message signals into digital form is the ruggedness of the digital signal. Quantizing noise associated with analog-to-digital conversion occurs at the terminal, where it can be controlled by assigning a sufficient number of digits for each sample. Once these signals are in digital form, additional impairments are negligible in a properly engineered system since each repeater transmits a new waveform in response to the received symbol and will almost always do so correctly in spite of noise and interference in the media. This process, called regeneration, provides the primary

advantages for digital transmission. The price paid for this ruggedness is increased bandwidth relative to that required for the original signal. It will be seen that in many transmission environments this ruggedness results in substantial economies [1].

24.1 SIGNAL PROCESSING

Analog signals, including message signals, are converted into digital form in steps depicted in Fig. 24-1. First the signal, which has



(a) Bandlimited speech waveform

(b) PAM pulses

(c) Code words

(d) Ternary symbols transmitted

(e) Recovered binary digits

(f) Decoded PAM pulses

(g) Interpolated waveform

FIG. 24-1. Signal processing steps in digital transmission.

been bandlimited by a low-pass filter, is sampled as shown in Fig. 24-1(b). These samples are called PAM pulses.

Although discrete in time, PAM pulses are not suitable for digital transmission because they are not discrete in amplitude. The next step is analog-to-digital conversion and is accomplished by a coder. The coder converts each PAM sample into a binary number called a code word, Fig. 24-1(c). (The number of binary digits used and the rules applied to represent each sample affect the faithfulness of reproduction.) Binary digits as generated by the coder are then translated into ternary or other multilevel digital signals acceptable to the transmission facility, Fig. 24-1(d).

At the receiving terminal the binary digits are recovered, and then the code words are decoded into a PAM form by a digital-to-analog converter, Fig. 24-1(f). The pulses are passed through an interpolation filter to recover the original message signal. Due to the finite number of code words available at the coder, the decoded PAM samples can have only a finite number of discrete amplitudes; therefore some error between the original and the recovered signal may be expected. This error is called quantizing distortion, which is the controlling signal impairment encountered in digital transmission.

24.2 DIGITAL HIERARCHY

In addition to being characterized by the discrete nature of the signals and the use of regenerative repeaters, digital transmission is also characterized by the use of time division multiplexing. Multiplexing of signals in digital form allows interconnection of digital transmission facilities with different signaling speeds. The digital hierarchy shown in Fig. 24-2 forms the basis of an interconnected digital network [2].

The interface between the digital system and the analog system is made by digital terminals which convert the incoming analog signals to a digital form suitable for application to a digital transmission facility. Terminals that multiplex many message channels for application to a single digital line are called digital channel banks. Digital multiplexers form the interface between digital transmission facilities of different rates. They combine digital signals from several lines in the same level of the hierarchy into a single pulse stream suitable for application to a facility of the next higher level in the hierarchy. A more detailed description of the various parts of the digital hierarchy follows.

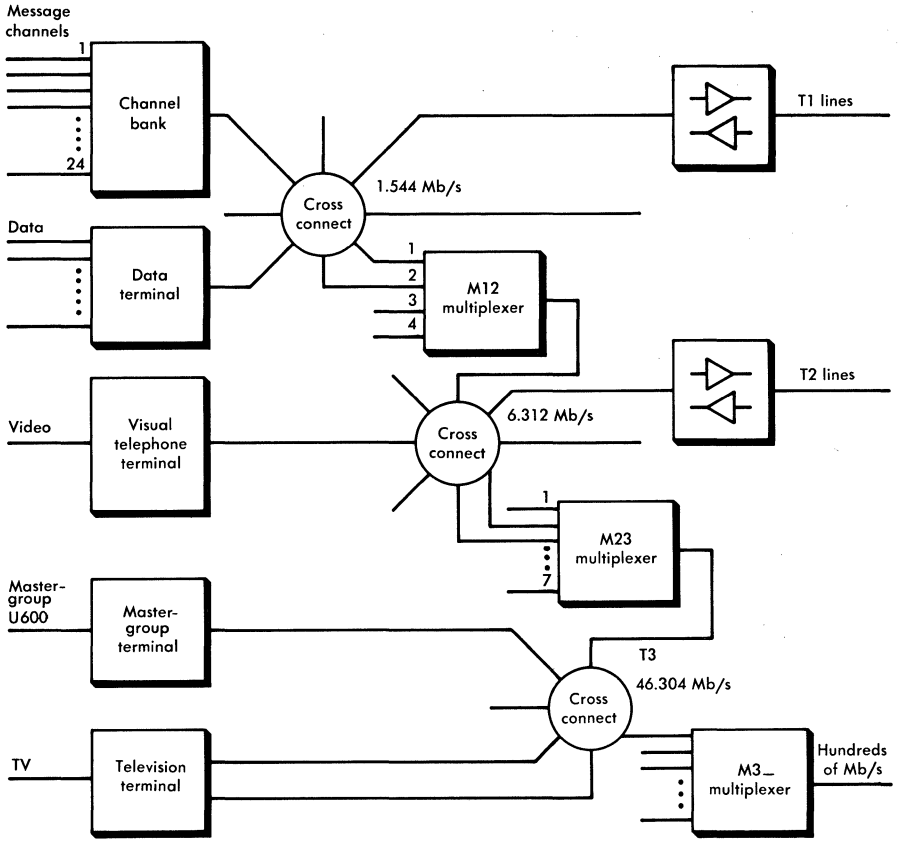
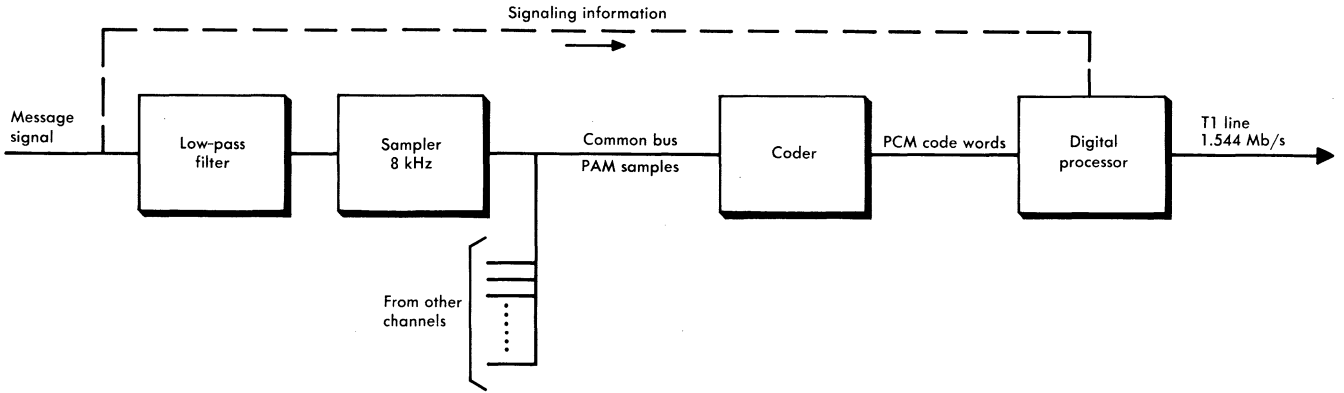


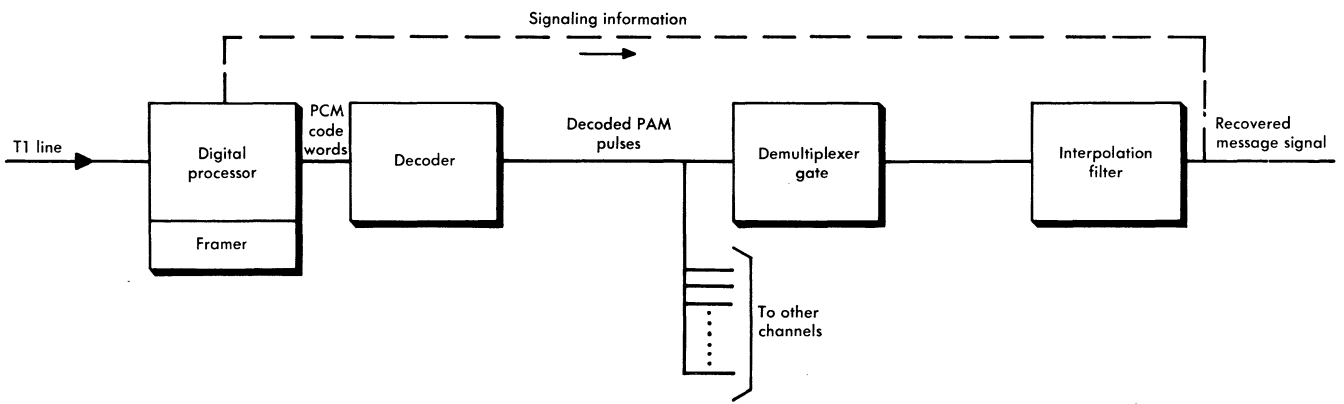
FIG. 24-2. Digital hierarchy.

Channel Banks

Digital channel banks multiplex many voice-frequency signals and code them into digital form. The incoming message signal is applied to the transmitting portion of the channel bank, Fig. 24-3. The signal is passed through a low-pass filter which limits the signal to a bandwidth less than one-half of the sampling frequency. The signal is then sampled at an 8-kHz rate. Samples from many message channels are applied to a common bus. Since the gates operate sequentially, the voltage on the common bus is seen to be a time division multiplexed version of all the PAM samples, Fig. 24-4. The voltage on the common bus is processed by the coder, which is shared by all the channels to produce PCM code words.



(a) Transmitting portion



(b) Receiving portion

FIG. 24-3. Channel bank block diagram.

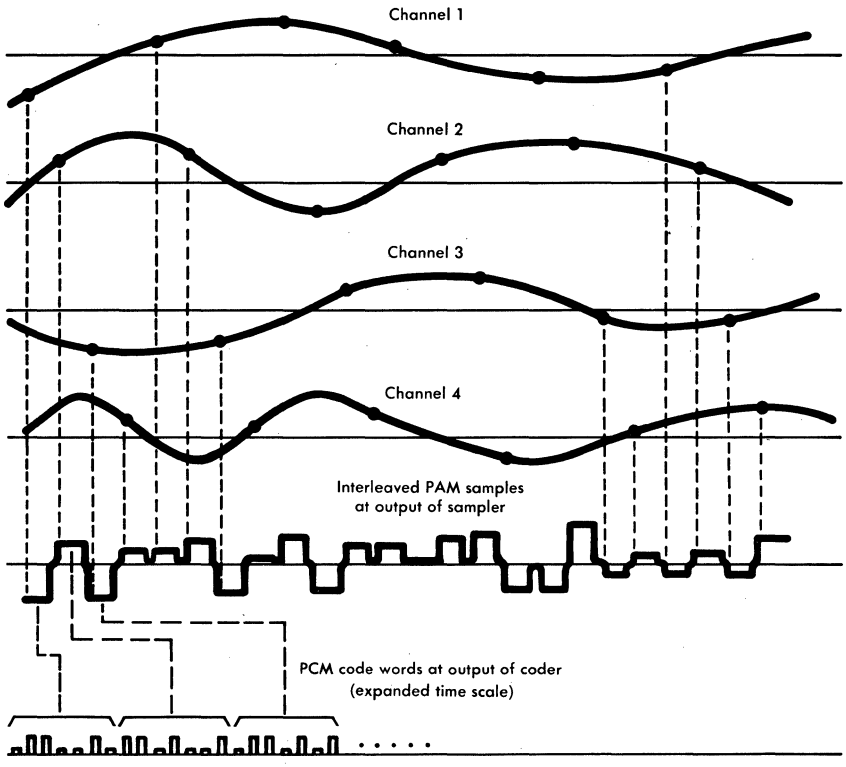


FIG. 24-4. Signal waveforms in a channel bank.

Before this digital signal can be applied to a transmission line, additional digital processing is usually necessary. First, the signaling information for each incoming message channel is multiplexed with the coder output. Next, in order to permit identification of the PCM code words at the receiver, framing information is inserted. Finally, the binary signals are converted to a form acceptable to the digital transmission line.

The D1 channel bank [3] converts 24 message channels to digital form. Each channel is coded into a 7-digit binary word. A signaling digit is then multiplexed to each word associated with the channel. Finally, a framing digit is multiplexed, resulting in the format shown in Fig. 24-5. Thus the total number of digits per sampling cycle is

$$(7 + 1) \times 24 + 1 = 193$$

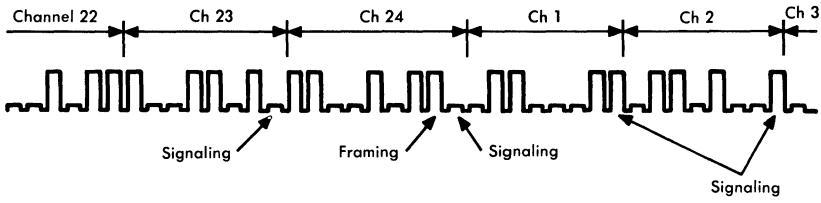


FIG. 24-5. D1 channel bank format.

The block of 193 digits is called a frame. Since there are 8000 frames per second, a digital capacity of 1.544 megabits per second (Mb/s) is required.

The receiving portion of the channel bank performs the inverse operations. The incoming signal from the digital line is first converted to binary form. A framing circuit searches for and synchronizes to the framing bit pattern, which insures that the locally generated timing pulses are in synchronism with the incoming pulse train. The signaling digits are sorted out and directed to the individual channels, and the PCM code words are delivered to the decoder. The output of the decoder is a series of quantized PAM pulses which are demultiplexed and applied to the individual interpolation filters. Interpolation produces an analog signal which is applied to the receiving voice-frequency line.

Single-Channel Terminals

When the bandwidth of the signals to be transmitted is such that after digital conversion it occupies the entire capacity of a digital transmission line, a single channel terminal is provided.

Mastergroup and Television Terminals. Two examples of such signals that are treated in a similar manner are the mastergroup signal and the commercial television signal. A mastergroup terminal allows message channels that have already been frequency division multiplexed to be applied to a digital facility without demultiplexing each channel to baseband. This is especially important because signals are expected to travel over both analog and digital facilities.

The block diagram for both the mastergroup and television terminals is shown in Fig. 24-6. The signal processor provides frequency shifting for the mastergroup signal and provides d-c restoration for the television signal. The mastergroup frequency extends from 564 to 3084 kHz. By shifting this band in frequency, it is

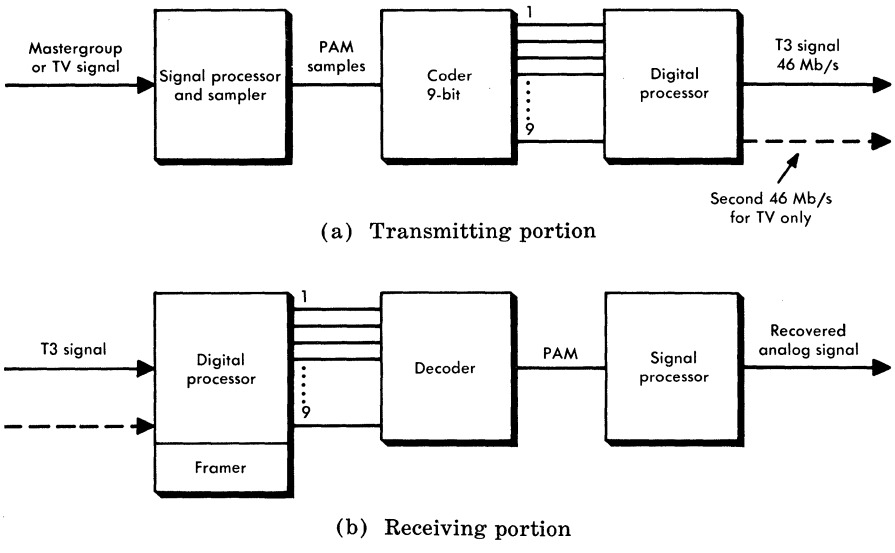


FIG. 24-6. Mastergroup and television terminal block diagram.

possible to sample at about 5.1 MHz. Sampling of television signals is at twice the mastergroup rate, or about 10.2 MHz.

To meet the transmission requirements, nine binary digits per sample are required for both mastergroup and television signals. The digital rate from the coder is therefore about 46 Mb/s for the mastergroup and 92 Mb/s for the television signal.

The digital processor shown in Fig. 24-6 has a three-fold purpose. It converts the parallel digital output from the coder to serial form, inserts framing information, and converts the serial binary signal to a form suitable for transmission. In addition, for the television terminal the 92-Mb/s signal must be split into two 46-Mb/s signals since only the 46-Mb/s speed belongs to the digital hierarchy.

The receiving portion of the terminal performs the inverse of the transmitting portion functions.

Visual Telephone Terminal. For economic reasons, it becomes desirable to code the video signal of PICTUREPHONE service into the T2 capacity of 6.312 Mb/s. To permit both adequate detail and constrast resolution, some specific characteristics of the signal and some subjective effects of motion pictures can be exploited.

In regions of low detail in a picture, such as simple backgrounds, the eye is sensitive to abrupt brightness error but insensitive to gradually increasing errors. In regions of dense detail, the eye is insensitive to brightness errors but is rather sensitive to position error or jitter of the edges of the detail. These signal characteristics and subjective effects suggest the use of differential pulse code modulation as means of bandwidth conservation in the digitizing of visual telephone signals. In this form of modulation the difference between the present sample and the previous sample is coded. Since these differences are expected to be smaller than the samples themselves, a smaller number of digits is needed for coding. The difference between the present sample and the corresponding sample in the last frame is expected to be even smaller. To exploit frame-to-frame correlation, however, requires storage of a full picture.

Data Terminals

An increasing portion of communications traffic involves signals other than voice. Most of these are data signals which are binary in form with transitions between levels occurring randomly.

The function of a data terminal is to accept a serial periodic or nonperiodic data signal and convert it to a synchronous stream acceptable to the digital transmission facility. Presently, the data rates generated by customer equipment are a fraction of the data rate capabilities of digital lines; thus, present data terminals are designed to transmit several data signals over the digital line. In some cases, data and message signals are multiplexed onto the same bit stream.

Data signals could be sampled directly; however, to preserve transition time accuracy the sample rate would have to be quite high, resulting in an excessive bit rate, especially for a signal with infrequent transitions. A more efficient method would be one which codes the transition times [4]. A typical format of data coding is shown in Fig. 24-7. When there is no transition, a signal of all ones is transmitted. A first zero, called the *address* bit, indicates the presence of a transition. Following the address bit is a *code* bit which indicates in which half of the address timing interval the actual transition took place. This reduces the time quantizing error by a factor of two compared to the direct sampling method. The last bit is a *sign* bit that indicates whether the direction of transition is zero to one or vice-versa. It is clear that the address bit, which in-

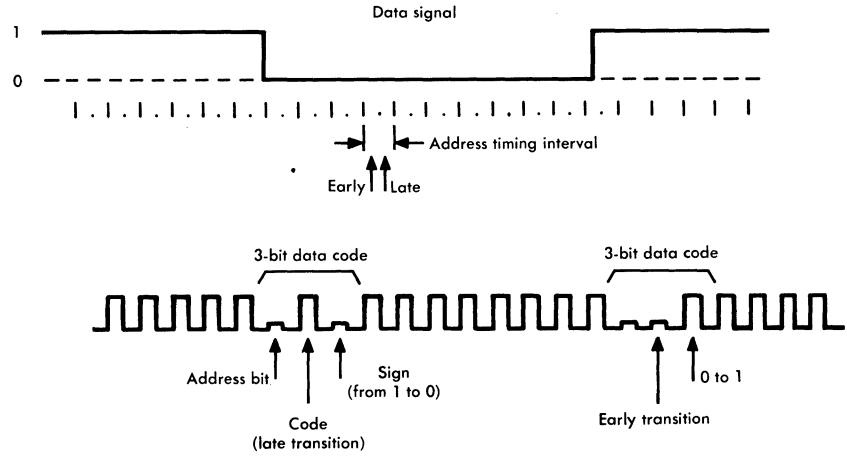


FIG. 24-7. Data coding format.

dicates a change of state, would be sufficient information. However, to prevent ambiguity in the event of errors, the sign bit provides a degree of error protection and limits error propagation.

The efficiency is about three transmitted bits for each transition or data bit. A 50-kb/s data signal thus displaces three 64-kb/s digital voice signals; a 250-kb/s data signal displaces 12 voice signals. Even at one-third efficiency, these data banks are more efficient in the substitution of data for voice signals than are analog systems, which require an entire supergroup of 60 message channels to transmit 250-kb/s data. Ultimately, data terminals can be designed which operate with better than 90 per cent efficiency.

Digital Multiplexers

As previously mentioned, digital multiplexers combine signals from several digital lines by the process of interleaving. Although digital signals at the same level of the hierarchy have the same nominal rate, they are not exactly synchronous. Multiplexers bring all signals to a synchronous rate, time division multiplex these signals, and process the resulting signal to a form suitable for application to the next level of the hierarchy.

Regenerative Repeaters

Regenerative repeaters are used at regularly spaced intervals along the digital transmission line to reconstruct the digital signal, thereby eliminating the effects of noise and distortion. A block diagram of a regenerative repeater consists of three functional parts: amplifier-equalizer, timing circuit, and regenerator, as shown in Fig. 24-8. The amplifier-equalizer shapes the incoming signal and raises its power level so that a pulse, no-pulse decision can be made by the regenerator circuit. The timing circuit provides the proper timing information for the regenerator to make its decision so as to minimize the chance of error. Spacing of the repeaters is designed to maintain an adequate signal-to-noise ratio for essentially error-free performance.

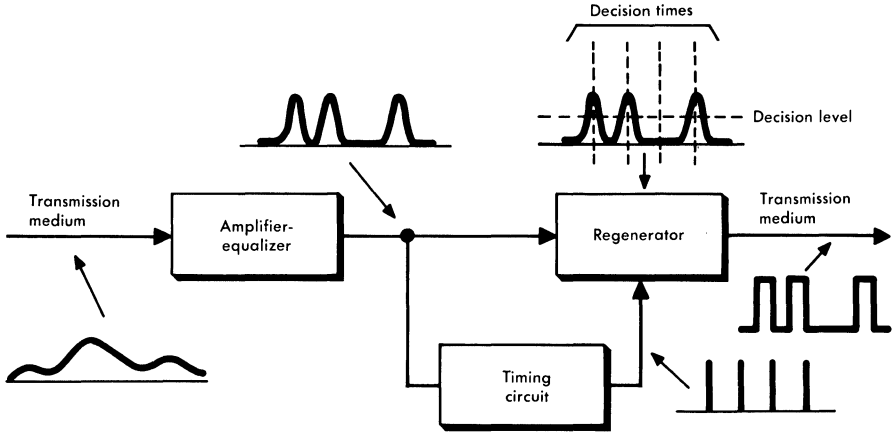


FIG. 24-8. Regenerative repeater block diagram.

24.3 ADVANTAGES AND DISADVANTAGES OF DIGITAL TRANSMISSION

A significant property of digital transmission is the separation of the noise generated in the terminal and noise encountered in the transmission line. Quantizing noise generated by the terminal equipment is unavoidable but can be made suitably small by the proper choice of the *codec* (coder-decoder) transfer characteristic. Because of the advantage of regeneration, line noise has little effect on the

message signal. Regenerative repeaters are designed such that infrequent errors caused by line noise have a negligible effect on terminal performance.

Signal to quantizing noise in dB increases linearly with the number of digits per sample and hence with the bandwidth utilized. When compared with FM where the signal-to-noise ratio in dB is only logarithmically proportional to bandwidth, this linear trade-off is more efficient.

Because the signal to line noise ratio requirement is lower in digital than in analog transmission, digital transmission often results in better utilization of noisy media even though the bandwidth is larger. In exchange cables consisting of tightly packed twisted pairs where crosstalk is a limiting factor, digital transmission can provide more channels at a lower cost than can analog transmission. In congested radio media, digital transmission can tolerate much greater external interference.

As devices become faster, circuits can be built that can be shared by more channels and thus lower the terminal cost. Time division multiplex is thus more economical than frequency division multiplex. It also has the property of treating all digitized message signals alike. This uniformity avoids the phase distortion and the resulting noise degradation that characterize the band edge effect in FDM transmission. In fact, since multiplexing digital signals imparts no impairment to these signals, the digital network is more flexible. Arrangement of facilities is not affected by the number of digital multiplex-demultiplex operations or the time position occupied by the digitized signal. The problem of allocating impairment to various parts of the telephone plant is greatly simplified in digital systems. In the future, switching may also be done digitally so that impairment through switches can also be eliminated.

Because all signals, including message, data, and video, are reduced to digital form, there is complete freedom to intermix these in a common facility. This flexibility is expected to make a digital system more useful than an analog system. Techniques for placing data and video signals on an analog facility usually result in a lower capacity of that facility as compared to a digital facility.

While the advantages are many, there are also a few disadvantages. One handicap is that a digital system must start in a telephone plant which is predominantly analog. Interface equipment, such as terminals to digitize mastergroup signals, is charged to the digital system. In media where noise and interference are low, e.g., the

coaxial cable, the ruggedness property is not as important and analog systems tend to have larger message capacity. Digital transmission over these media will probably become attractive when it is better established in other media or when the volume of data and video signals becomes significant.

Devices used in digital systems are valued for their speed in contrast to those used in analog systems, which are valued for their linearity. The relative advancement of digital and analog systems will depend on whether more improvement is made in switching speed or linearity. Digital transmission is a latecomer not because the concepts were not known [5, 6], but because the complex signal processing was not feasible with electron tube technology [7]. Semiconductor technology has made digital transmission not only possible but economical in many applications [8].

REFERENCES

1. Oliver, B. M., J. R. Pierce, and C. E. Shannon. "Philosophy of PCM," *Proc. IRE*, vol. 36 (Nov. 1948), pp. 1324-1331.
2. Hoth, D. F. "Digital Communications," *Bell Laboratories Record*, vol. 45 (Feb. 1967), pp. 38-43.
3. Fultz, K. E. and D. B. Penick. "T1 Carrier System," *Bell System Tech. J.*, vol. 44 (Sept. 1965), pp. 1405-1451.
4. Travis, L. F. and R. E. Yaeger. "Wideband Data on T1 Carrier," *Bell System Tech. J.*, vol. 44 (Oct. 1965), pp. 1567-1604.
5. Rainey, P. M. "Facsimile Telegraph System," U. S. Patent 1608527, assigned to Western Electric Company Nov. 30, 1926.
6. Reeves, A. H. "Electric Signaling System," U. S. Patent 2272070, assigned to International Standard Electric Corporation Feb. 3, 1942; also French Patent 852,183, Oct. 23, 1939.
7. Black, H. S. and J. O. Edson. "Pulse Code Modulation," *Trans. AIEE*, vol. 66 (1947), pp. 895-899.
8. Boxall, F. S. "New PCM Exchange Carrier System," *Proc. Nat. Electron. Conf.*, vol. 21 (1965), pp. 355-360.

Chapter 25

Digital Terminals

Digital terminals convert analog signals into a form suitable for transmission over digital facilities. Terminal outputs of the same rate can be processed identically both for application to transmission facilities and for further digital multiplexing. The different types of terminals have been described in Chap. 24. In this chapter detailed theory and practical realization of the major functions of digital terminals are discussed. These functions are sampling, coding, and framing.

25.1 SAMPLING

Crucial to the concept of digital transmission is the representation of a bandlimited signal by time samples. A proof of the sampling theorem previously stated in Chap. 5 is presented here followed by a discussion of a particular implementation of sampling.

Sampling Theorem

If a signal that is bandlimited is sampled at regular intervals and at a rate at least twice the highest frequency in the band, then the samples contain all of the information of the original signal [1, 2, 3]. A signal, $g(t)$, is said to be bandlimited if its Fourier transform vanishes outside a finite interval. Therefore $g(t)$ can be represented as

$$g(t) = \int_{-\frac{1}{2T}}^{\frac{1}{2T}} G(f)\, e^{j2\pi ft} df \tag{25-1}$$

where $G(f) = 0$ for $|f| > 1/2T$, and $1/2T$ is the bandwidth in hertz.

Since $G(f)$ need be defined only between $-1/2T$ and $1/2T$, it is possible to use a Fourier series representation over that interval

$$G(f) = \sum_{n=-\infty}^{\infty} C_n e^{j2n\pi fT} \qquad |f| < \frac{1}{2T} \tag{25-2}$$

where the coefficients are determined by

$$C_n = T \int_{-\frac{1}{2T}}^{\frac{1}{2T}} G(f) e^{-j2n\pi fT} df \tag{25-3}$$

Comparing this with Eq. (25-1) shows that C_n, except for constants, represents values of $g(t)$ at discrete points, or

$$C_n = Tg(-nT) \tag{25-4}$$

Thus, C_n, the samples of $g(t)$ taken at twice the highest frequency in the band, determines $G(f)$ and hence $g(t)$. Substituting Eq. (25-4) into Eq. (25-2) and then into Eq. (25-1) yields

$$g(t) = \sum_{n=-\infty}^{\infty} g(nT) \frac{\sin \frac{\pi}{T}(t - nT)}{\frac{\pi}{T}(t - nT)} \tag{25-5}$$

This equation shows how the original signal can be recovered from its samples by using as an interpolation function the familiar impulse response $h(t)$ of an ideal low-pass filter with gain T and bandwidth $1/2T$

$$h(t) = \frac{\sin \frac{\pi}{T} t}{\frac{\pi}{T} t} \tag{25-6}$$

In practice, ideal low-pass filters and sampling with pulses of zero width (impulse sampling) can only be approached. The consequences of these imperfections are called foldover noise and aperture effect.

To study these effects it is more convenient to think of sampling as product modulation of both a message function and an impulse train as was done in Chap. 5. When a signal is bandlimited by use of a practical low-pass filter, some high-frequency energy will invariably be present. The resulting spectrum after sampling, shown in Fig. 25-1, will have an overlap region between the baseband and the first sideband. To reconstruct the signal from this spectrum involves another practical low-pass filter which rejects most of the unwanted sidebands except near the band edge where finite roll-off permits unwanted signals to appear beyond the desired band edge. The unwanted spectrum is signal correlated and is called foldover distortion. As was true for AM, physical limitations on filter per-

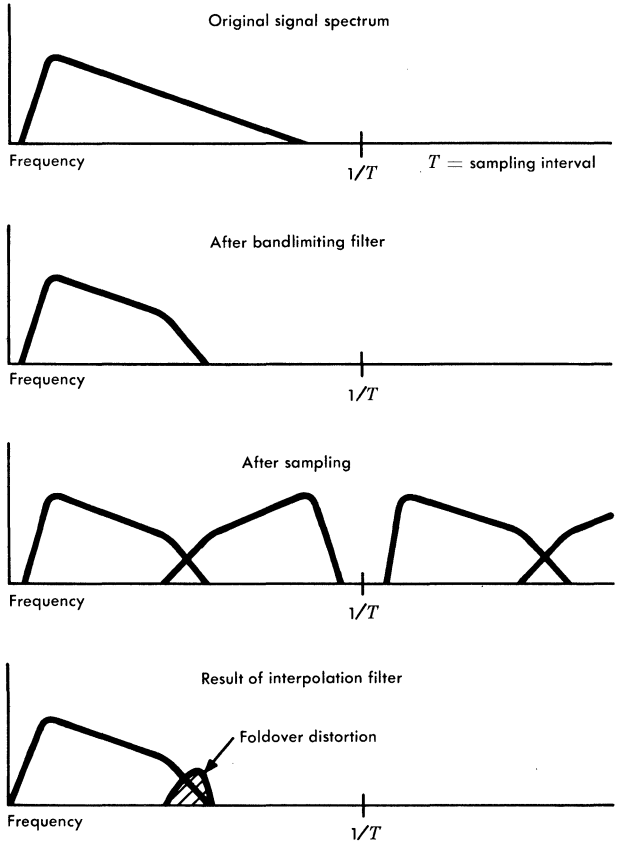
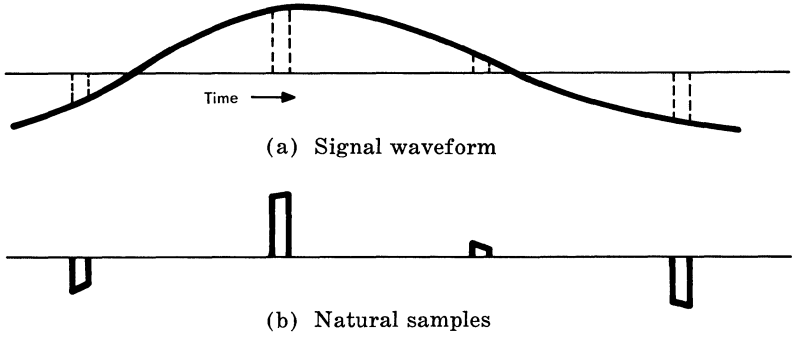


FIG. 25-1. Foldover distortion as a result of imperfect filtering.

formance make it necessary to leave a guard band near the half sampling rate to keep foldover distortion within requirements. Sampling a message channel at 8 kHz, for example, provides a usable bandwidth of 3.5 kHz.

Sampling with finite width pulses and later interpolation of finite width PAM pulses both result in frequency distortion of the signal. Sampling with finite width pulses results in natural samples as shown in Fig. 25-2. In some terminal designs, the area of these natural samples is used as a basis for coding. The net effect is equivalent to passing these natural samples through a filter whose impulse response is a rectangular pulse and then impulse sampling the signal at the filter output. Such a filter imparts its $(\sin x)/x$ roll-off to the original spectrum. The same effect occurs if the input to the interpolation filter is not a train of impulses proportional to the sample values but is a train of pulses of finite width. Both of these roll-offs are called aperture effect and can be compensated for by use of linear filters.



(a) Signal waveform

(b) Natural samples

FIG. 25-2. Sampling with finite width pulses.

For example, if the width of the sampling pulse is τ, the Fourier transform is

$$G(f) = \tau \frac{\sin \pi f\tau}{\pi f\tau}$$

and at the Nyquist frequency of $1/2T$, the response normalized to

that at zero frequency is

$$\frac{\sin\left[(\pi/2)\ (\tau/T)\right]}{(\pi/2)\ (\tau/T)}$$

where τ/T is the duty cycle of the sampling pulse. Aperture effect becomes negligible for duty cycles of less than 10 per cent.

In other terminal designs the natural samples are coded directly. Uncertainty of the coded word results because the pulse height is changing during the coding process. The net result is similar to additive noise.

Resonant Transfer

When a voltage or current is sampled by a transmission gate, the net energy after sampling is proportional to the duty cycle of the sampling pulse controlling the gate. As the sample width approaches zero, the energy transfer also approaches zero. An important method of sampling called resonant transfer overcomes this loss [4, 5]. Figure 25-3 shows a simplified schematic of a resonant transfer circuit. A low-pass filter with a terminating capacitor, C, is assumed. When the gate is closed, L, C, and the holding capacitor, C_H (normally made equal to C), form a series resonant circuit. If C has initial charge and C_H is discharged, switch closure results in the current and voltage waveforms as shown in Fig. 25-3. At the first zero in the current waveform, all the charge on C has been transferred to C_H. The gate is opened at this time, and the voltage on C_H is held constant for coding. The resonant frequency is determined by the duty cycle of the gate which can be made small, typically one to two per cent. There is no energy loss except for parasitic resistances in the circuit elements and the gate. A clamp is necessary to discharge C_H before the next sample.

25.2 CODING

The sampling process converts a continuous signal to one that is discrete in time. Digital transmission also requires that the signal be made discrete in amplitude so that it can be represented by a finite number of symbols. This is the function of coding. In this process each PAM sample is assigned a code word by a coder. Since the number of code words available is finite, a small range of amplitudes will use the same code word and thus all will be decoded into one particular amplitude at the receiving end.

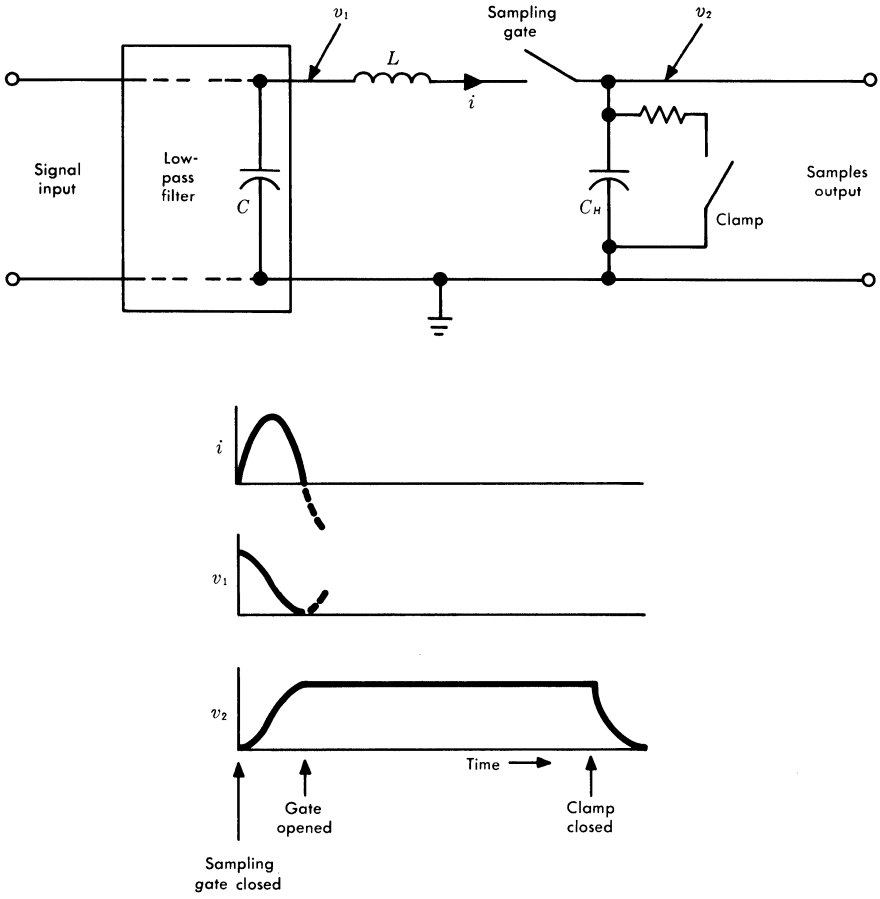
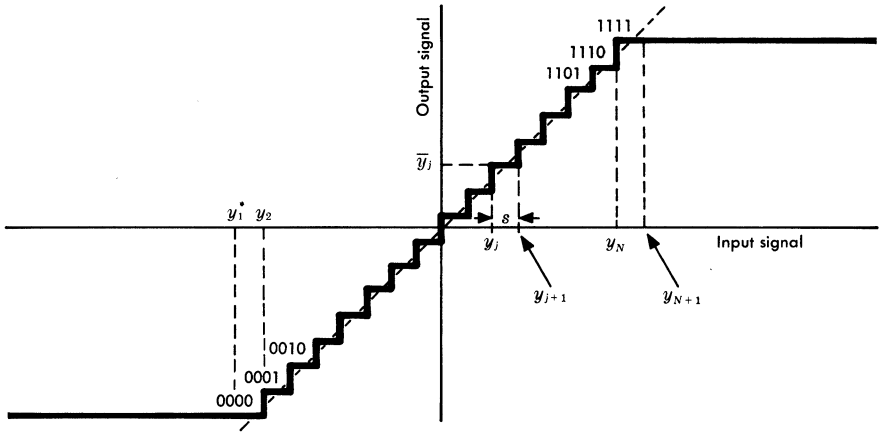


FIG. 25-3. Resonant transfer sampling.

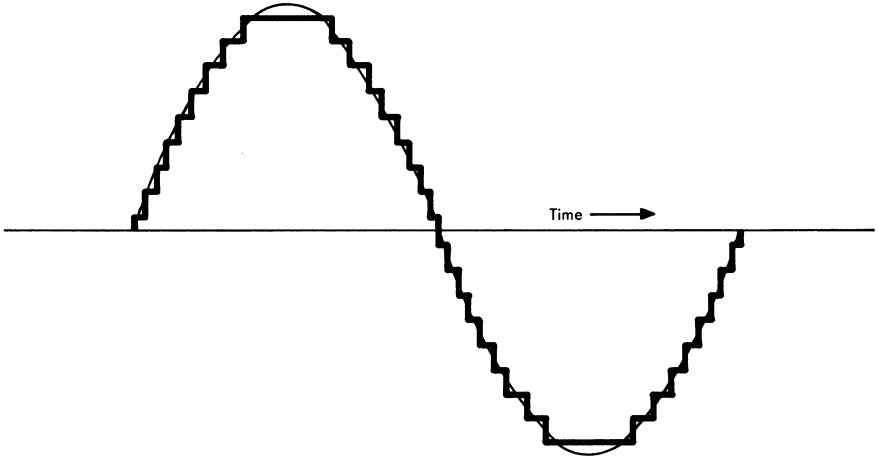
Quantizing

Errors introduced by quantizing can be described graphically by the composite transfer characteristic of a coder and decoder [Fig. 25-4(a)]. Here a uniform codec is shown. It is called uniform because the input amplitude range is divided into N steps of equal width, s, and the output levels are also uniformly spaced. The difference between the staircase characteristic and the straight line for distortionless transmission is the error characteristic plotted as a function of input amplitude as shown in Fig. 25-4(b).



(a) Uniform codec transfer characteristic

(b) Error characteristic



(c) Quantized full load sine wave

FIG. 25-4. Characteristics of a uniform codec.

If the amplitude distribution of the input samples is known, the mean square expected error can be computed by integrating the error characteristic over the amplitude distribution:

$$\overline{e^2} = \sum_{j=1}^{N} \int_{y_j}^{y_{j+1}} (y - \bar{y}_j)^2 \, p(y)\, dy$$

$$+ \int_{-\infty}^{y_1} (y - \bar{y}_1)^2 \, p(y)\, dy + \int_{y_{N+1}}^{\infty} (y - \bar{y}_N)^2 \, p(y)\, dy \quad (25\text{-}7)$$

where the intervals y_j to y_{j+1} are of length, s, and each interval is quantized to its midpoint, $\bar{y}_j = 1/2(y_j + y_{j+1})$, and $p(y)$ is the probability density function of the input signal. The last two integrals of Eq. (25-7) are overload terms which take into account the error introduced when the signal exceeds the bounds of the code range. Overload error is zero for sinusoidal inputs below full load. If the interval length, s, is small, $p(y)$ can be assumed to be uniform over that interval and replaced by a constant p_j which is the average probability density in the interval y_j to y_{j+1}. Furthermore, the error characteristic for all intervals is identical so that Eq. (25-7) evaluates to

$$\overline{e^2} = \sum_{j=1}^{N} \frac{(y_{j+1} - y_j)^2}{12} (y_{j+1} - y_j) p_j = \frac{s^2}{12} \qquad \text{(no overload)} \quad (25\text{-}8)$$

since $(y_{j+1} - y_j)p_j$ is the probability that the signal amplitude is in the interval y_j to y_{j+1} and the sum of probabilities is 1. If the number of binary digits in each code word is n, then there are 2^n different code words available. The peak-to-peak range of the codec is $2^n s$. The power of a sine wave spanning all 2^n code words [Fig. 25-4(c)] is $(2^n s)^2/8$, so that for a linear codec with n binary digits per sample, the ratio of full-load sine wave power to quantizing distortion power (S/D) is:

$$\text{S/D} = 10 \log \frac{(2^n s)^2/8}{s^2/12} = 20n \log 2 + 10 \log 1.5$$

$$= 6n + 1.8 \qquad \text{dB} \quad (25\text{-}9)$$

This demonstrates the linear relationship between the number of digits transmitted per sample and the signal-to-distortion ratio in dB. Each added binary digit increases the S/D ratio by 6 dB.

Companding. A codec whose transfer characteristic has uniform step sizes is not usually the best choice for message signals for two reasons [6]. First, the amplitude distribution of the message is not uniform. For a given talker volume, smaller amplitudes are more probable than larger amplitudes, and thus a better S/D ratio can be expected if the error characteristic is made smaller for the more probable amplitudes at the expense of larger errors for the less probable amplitudes. Second, message signals have a dynamic range of up to 40 dB. For a uniform codec, weak signals will experience 40 dB poorer S/D ratio than strong signals.

A nonuniform codec is a coder-decoder pair whose input amplitude range is divided into N steps of unequal widths, resulting in output levels of unequal spacing. A typical staircase characteristic for a nonuniform codec is shown in Fig. 25-5 along with its error characteristic.

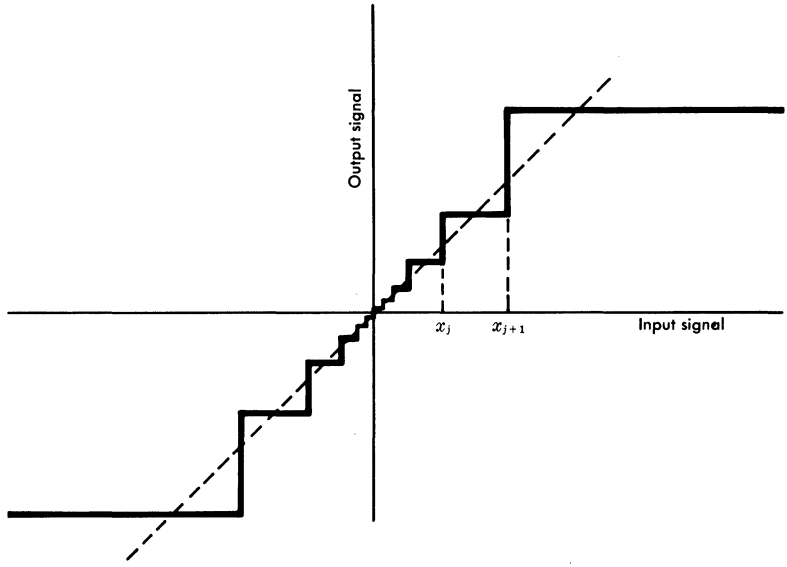
There are two ways to achieve nonuniform coding. One is to build a coder which assigns codes to input amplitudes in the desired nonuniform manner. At the receiving end the decoder must decode to the midpoint of these nonuniform intervals. Another method, shown in Fig. 25-6, is to predistort the samples of the input signal using an instantaneous compressor (as distinguished from slower acting syllabic compressors). The compressed signal amplitudes are then uniformly coded. Similarly, the receiving terminal can use a uniform decoder followed by an instantaneous expandor whose transfer characteristic is the inverse of that of the compressor. This system is called a *companded* codec. The difference between the two methods is only in implementation, and the second approach is used subsequently for purposes of analysis.

Mean square error of a companded codec can be calculated using Eq. (25-7) except that now the segment lengths $(y_{j+1} - y_j)$ are not all the same. With a uniform distribution within each interval and with overload neglected, the expression becomes

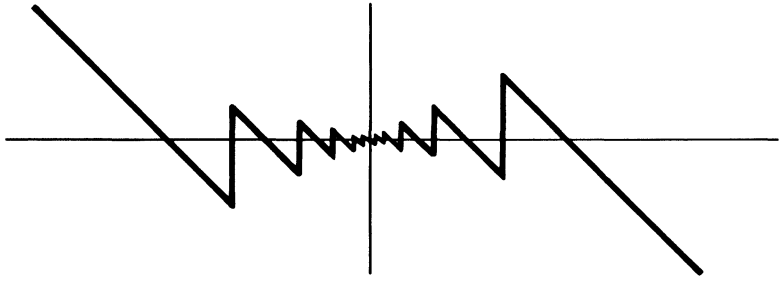
$$\overline{e^2} = \frac{1}{12} \sum_{j=1}^{N} (x_{j+1} - x_j)^2 (x_{j+1} - x_j) p_j \qquad (25\text{-}10)$$

where x is the input to the compressor. If the compressor characteristic is $y = F(x)$, the intervals in Eq. (25-10) can be related to those of Eq. (25-8) for a uniform quantizer. For each interval

$$(x_{j+1} - x_j) \approx s/F'(x_j) \tag{25-11}$$



(a) Nonuniform codec transfer characteristic



(b) Error characteristic

FIG. 25-5. Characteristics of a nonuniform codec.

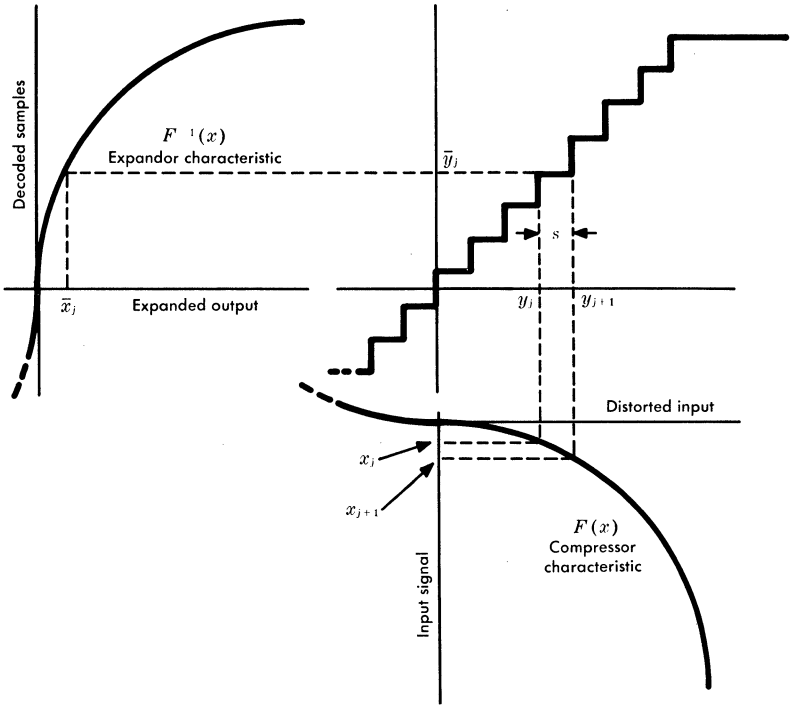


FIG. 25-6. Nonuniform codec using a compandor.

Substituting this into Eq. (25-10) and approximating the sum by an integral yields

$$\overline{e^2} = \frac{1}{12} \sum_{j=1}^{N} \frac{s^2}{[F'(x_j)]^2} (x_{j+1} - x_j) p_j$$

$$\approx \frac{s^2}{12} \int \frac{p(x)}{[F'(x)]^2} dx \tag{25-12}$$

By appropriate choice of $F(x)$, $\overline{e^2}$ can now be minimized. The optimum $F(x)$ is a function of the amplitude distribution. The reciprocal of the integral in Eq. (25-12) is known as the companding improvement

$$C_I = \frac{1}{\int \frac{p(x)}{[F'(x)]^2} dx} \tag{25-13}$$

Companding requirements are different for different signal distributions. For example, voice signals require constant S/D performance over a wide dynamic range, which means that the distortion must be proportional to signal amplitude for any signal level. To achieve this, a logarithmic compression law must be used. Of course a truly logarithmic assignment of code words is impossible because it implies both an infinite dynamic range and an infinite number of codes. Two methods for modifying the true logarithmic function have been used. In the first, called the μ-law [7], for the normalized coding range of $\pm$ 1,

$$F(x) = \text{sgn } (x) \frac{\ln (1 + \mu \mid x \mid)}{\ln (1 + \mu)} \qquad -1 \leq x \leq 1 \tag{25-14}$$

The compression laws for several values of μ are plotted in Fig. 25-7. For small x, $F(x)$ approaches a linear function, and for

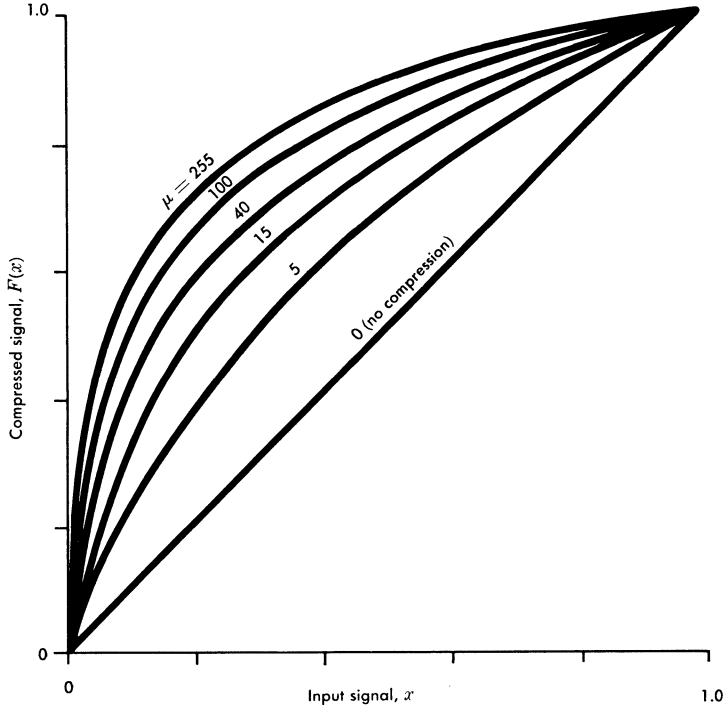


FIG. 25-7. Logarithmic compression characteristics.

large x it approaches a logarithmic function. The range of signal power over which S/D is relatively constant is determined by the parameter μ. For a relatively constant S/D ratio over a 40-dB dynamic range as shown in Fig. 25-8, μ should be greater than 100. Signal-to-distortion ratios can be calculated using the amplitude distribution of speech which is approximately Laplacian

$$p(x) = \frac{1}{\sqrt{2}\,\sigma} \exp\left(-\sqrt{2}\frac{|x|}{\sigma}\right) \tag{25-15}$$

where σ is the rms speech voltage.

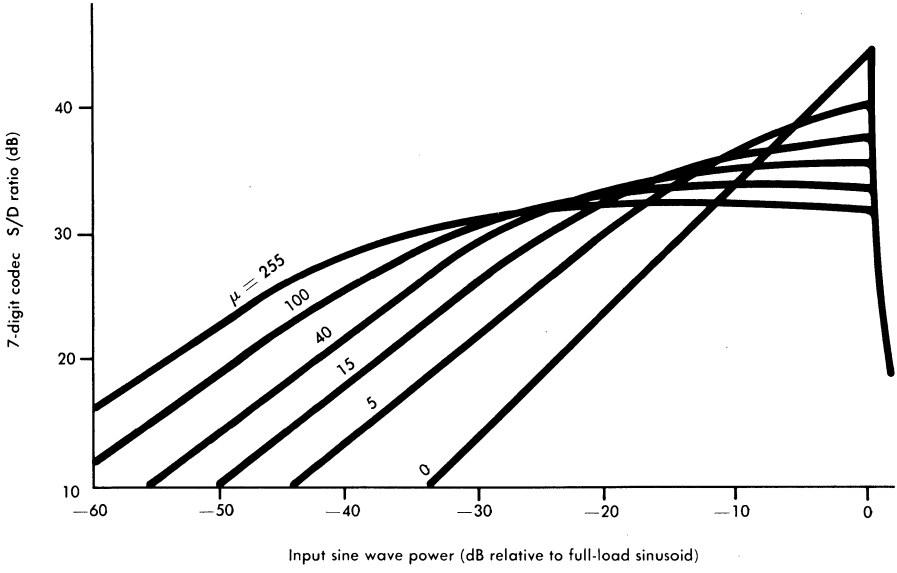


FIG. 25-8. Signal-to-distortion performance of logarithmic compandors.

In many instances, S/D performance of a codec is based on sinusoidal inputs. Graphical comparison of calculated S/D ratios for speech and sinusoidal inputs obtained from Eq. (25-7) showed good agreement for speech signals with normal loading range, Fig. 25-9. For strong speech signals, overload distortion occurs but is not as disturbing as quantizing distortion. Thus, measured and calculated S/D ratios based on sinusoids are often used in place of speech inputs as a measure of codec performance.

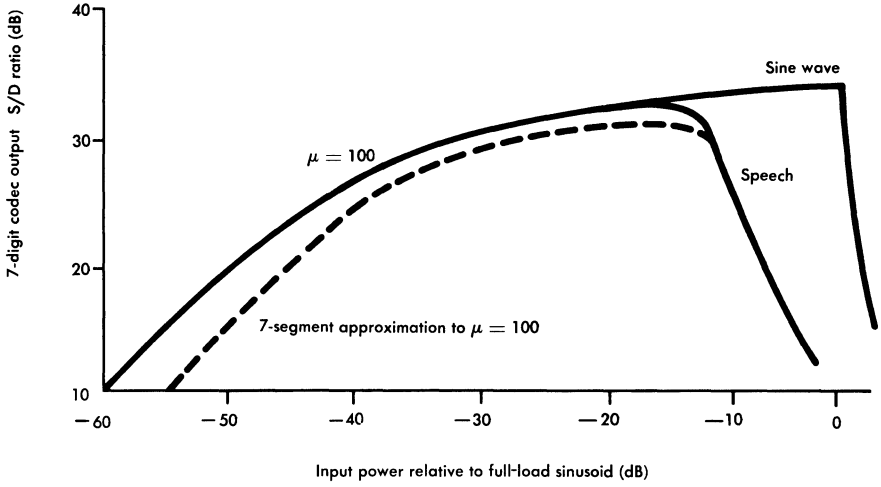


FIG. 25-9. Comparison of S/D ratios for speech and sine waves.

A second method of approximating the true logarithmic law is to substitute a linear segment to the logarithmic curve for small signals. It is called the A-law [8, 9] and is represented by:

$$F(x) = \text{sgn}\ (x)\ \frac{1 + \ln A\ |\ x\ |}{1 + \ln A} \qquad \frac{1}{A} \leq |\ x\ | \leq 1$$

$$F(x) = \text{sgn}\ (x)\ \frac{A\ |\ x\ |}{1 + \ln A} \qquad 0 \leq |\ x\ | \leq \frac{1}{A} \tag{25-16}$$

This curve is smooth at $x = 1/A$. The parameter A determines the dynamic range. Over the intended dynamic range it has a flatter S/D ratio than the μ-law, as shown in Fig. 25-11.

Companding improvement for these two logarithmic laws can be estimated easily for small signals since both are linear near the origin. For the μ-law, for example, companding improvement for small signals is given by

$$C_I = [F'(0)]^2 = \left[\frac{\mu}{\ln\ (1 + \mu)}\right]^2 \tag{25-17}$$

for $\mu = 100$, $C_I = 21.67^2$ or 26.7 dB. Another figure of merit is the

compression factor; it is the ratio of the largest to the smallest step sizes. For the μ-law it is $(1 + \mu)$.

Both the A-law and the μ-law tend to produce flat S/D ratios over a wide dynamic range. If instead of a flat S/D ratio, higher S/D ratios are desired for more probable talker volumes, hyperbolic compression laws are more attractive. These give better S/D ratios to average talkers at the expense of the smaller population of very weak and very loud talkers [10]. Flat S/D ratio curves are the only ones in widespread use however.

The logarithmic laws can be approximated by nonlinear devices such as diodes. These laws can also be implemented by a piecewise linear approximation using several segments. As an example, an approximation to a 7-digit μ=100 curve using eight segments is shown in Fig. 25-10. The vertices of the polygonal line lie on the curve being approximated and are spaced uniformly on the vertical axis. Four segments are on each side of zero with the center two

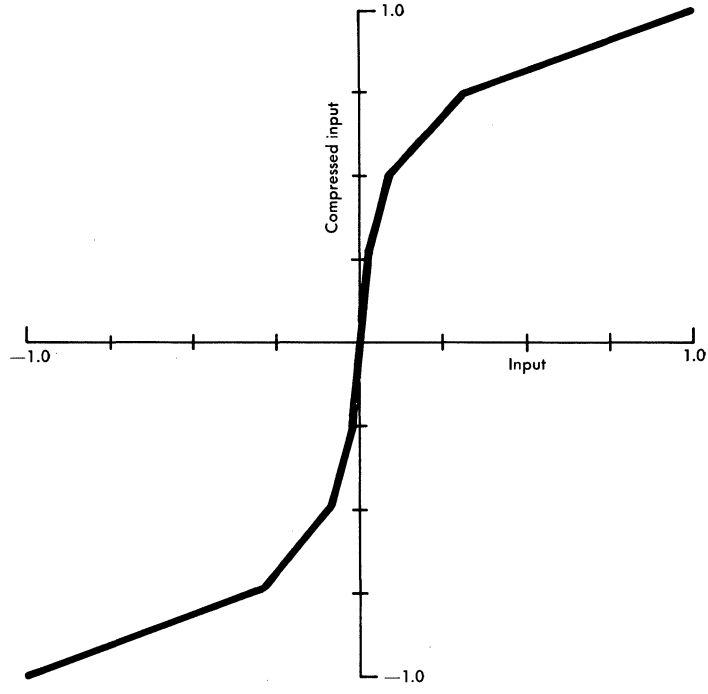


FIG. 25-10. Seven-segment piecewise linear approximation to $\mu = 100$ logarithmic compression law.

segments collinear.* According to this companding law the first three digits of the code word locate the signal in the eight original segments; the remaining four digits divide the segments into sixteen equal parts and further locate the signal. The S/D performance is shown in Fig. 25-9.

The digitally linearizable laws are an interesting class of piecewise linear compression laws. These laws are characterized by the property that the coding intervals, which are equal within each segment, are integral multiples of the size of a smallest coding interval.

Of particular interest for binary word coders are the cases where the coding intervals are related by powers of two, and each linear segment contains an equal number of coding intervals. For example, eight segments on each side of zero are commonly used. If the size of the coding intervals within each segment doubles for each segment outward from the center, a 15-segment digitally linearizable law results, which approximates the μ-law characteristic with $\mu = 255$. The ratio of largest to smallest coding interval (the compression factor) is $2^7 = 128$; the companding improvement for small signals is 30 dB.

If the center four segments of the eight original segments on each side of zero are made collinear, with the coding intervals of the remaining outer segments doubling in size as before, a 13-segment digitally linearizable compression law is achieved that has been adopted by many countries in Europe [11]. It approximates the A-law characteristic matching the slope at the origin with $A = 87.6$. The compression factor is $2^6 = 64$; the companding improvement for small signals is 24 dB. The S/D curves for the 15-segment and 13-segment laws are shown in Fig. 25-11.

A 7-digit 13-segment law can be implemented by starting with an 11 binary digit uniform coder. The first digit is the sign digit; the number of leading zeros in the remaining digits when subtracted from seven determines, in binary notation, the segment number on which the sample belongs. The first 1 following the leading 0's is skipped except for segment 0; the next three of the remaining digits are copied to determine one of the eight intervals in each segment. Figure 25-12 illustrates the translation. From the 7-digit compressed code an 11-digit linearized code can be obtained by filling the unknown lesser significant digits arbitrarily. Similarly, a 15-

*Because the center two segments are collinear, this is known as a seven segment approximation.

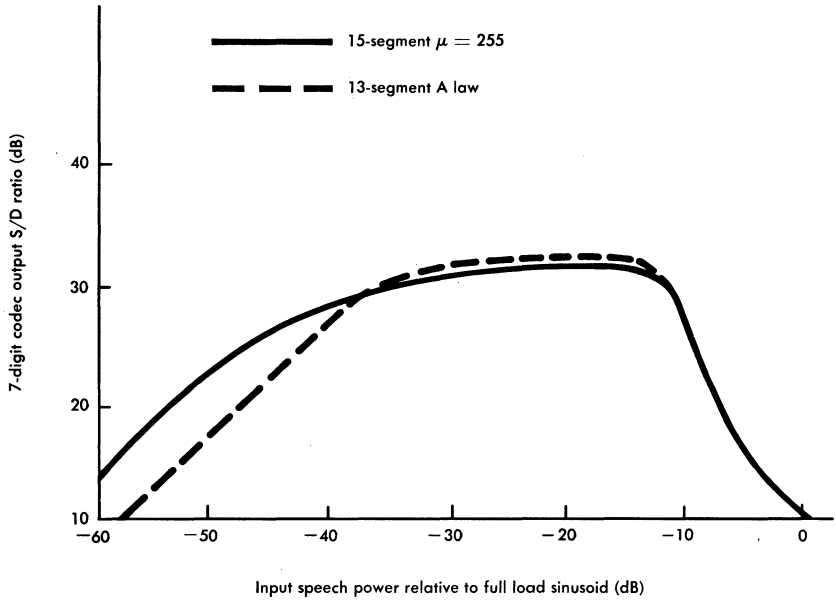


FIG. 25-11. Signal-to-distortion performance of two digitally linearizable laws.

11-digit linear code	7-digit compressed code
s 0 0 0 0 0 0 0 a b c	s 0 0 0 a b c
s 0 0 0 0 0 0 1 a b c	s 0 0 1 a b c
s 0 0 0 0 0 1 a b c x	s 0 1 0 a b c
s 0 0 0 0 1 a b c x x	s 0 1 1 a b c
s 0 0 0 1 a b c x x x	s 1 0 0 a b c
s 0 0 1 a b c x x x x	s 1 0 1 a b c
s 0 1 a b c x x x x x	s 1 1 0 a b c
s 1 a b c x x x x x x	s 1 1 1 a b c

Notes:

s = 1 for positive signals.

s = 0 for negative signals.

Digits a b c are copied from linear code to compressed code.

Digits x x . . . are ignored.

FIG. 25-12. Translation table from linear to compressed code for 13-segment law.

segment law can be implemented; it requires one more binary digit in the uniform coder.

Digitally linearizable compression also lends itself to digital signal processing. For example, to add two or more digitized message signals as would occur during a conference call, the uniform code equivalent of each signal is obtained, added numerically, and then converted back to a single nonuniform code word. This method of combining signals gives the same performance as decoding the several digital signals, combining them by analog methods, and recoding, since digital combining is subject to round-off errors and analog combining is subject to quantizing errors. More complicated processes can also be carried out digitally, such as multiplication to effect gain change and filtering to shape bandwidth.

Coding Methods

Methods used to quantize a PAM sample into 2^n levels and to assign an n digit binary code word to each sample can be classified according to whether the coding operation proceeds a level at a time, a digit at a time, or a word at a time.

Level-at-a-time coding requires 2^n sequential decisions to be made for each code word generated. Although coders using this method are simple, the large number of sequential operations restrict their use to low speed applications. In word-at-a-time coding, the entire code word is determined simultaneously. Coders designed for this method are more complex; however, they are suitable for high speed application. Digit-at-a-time coders provide a compromise between speed and complexity.

Level-at-a-Time Coding. This method of coding compares a PAM sample with a ramp waveform while a binary counter is advanced by a clock signal. When the ramp waveform equals or exceeds the PAM sample, the content of the counter is the resultant binary PCM code word. This method requires fast counting if the number of digits in the code word is large. For example, with an n binary digit coder, the counter must complete 2^n counts within the coding interval or word time, and a comparator circuit must react within the one-clock interval when the ramp is equal to or exceeds the input. Nonuniform coding can be achieved by using as the reference ramp a nonlinear function of time.

Digit-at-a-Time Coding. Digit-at-a-time coding determines each digit of the code word sequentially [12]. It is analogous in concept to a balance where known reference weights are used in combination to determine an unknown weight.

Feedback Coder. One kind of digit-at-a-time coder, called a network feedback coder, is illustrated in Fig. 25-13. This 7-bit coder consists of three major blocks: the resistor weighting network, the comparison circuit, and the logic circuit. The weighting network generates reference currents to be compared with the unknown current, I_x. Switches controlled by the logic circuit determine which of the currents are to be summed. Since the currents are related by factors of 2, there are 2^7 uniformly spaced reference currents (I_{ref}) that can be produced and compared with I_x.

Operation of the coder is as follows: switch S_1 is closed; if I_x is smaller than I_{ref}, S_1 is opened; otherwise, S_1 remains closed for the remainder of the coding process. Next, S_2 is closed and the procedure repeats for each stage until the seven digits are determined. At the end

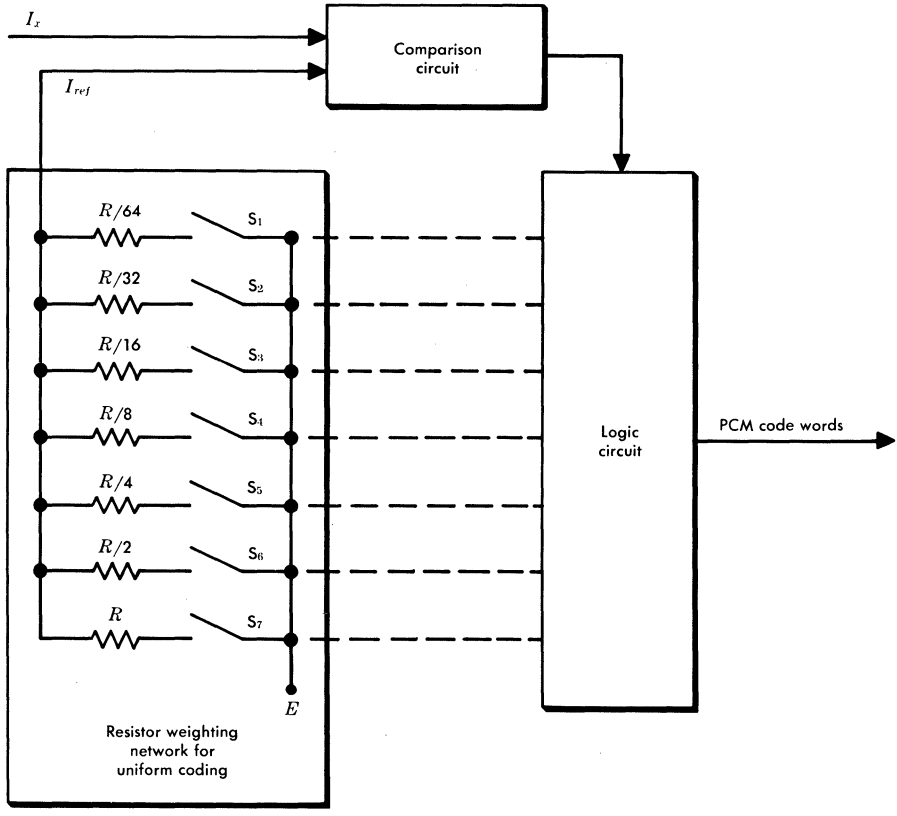


FIG. 25-13. Network feedback coder.

of the coding interval, the state of the switches indicates the resultant digital code word, and I_{ref} equals I_x to within one code step.

A nonuniform weighting network can be used in place of the uniform network, and the result is a nonuniform coder. For example, a 7-digit 15-segment digitally linearizable law can use the nonuniform weighting network of Fig. 25-14. With this network, S_1 is switched according to the polarity of the input current. The next three switches, S_2, S_3, and S_4, select one of the eight segments. The three remaining switches are part of a uniform resistor weighting network and are used to determine the code within the selected segment. The resultant code consists of a sign digit followed by a nonuniform binary code representing the magnitude of the PAM sample.

A nonuniform code can also be obtained by using a nonlinear device preceding a uniform coder. For example, the logarithmic voltage current relationship of semiconductor diodes can be used to approximate the μ-laws [13].

Tandem Binary Coders. An important class of digit-at-a-time coders is a tandem arrangement of identical coder stages where one stage is used for each digit as illustrated in Fig. 25-15 [14]. Each

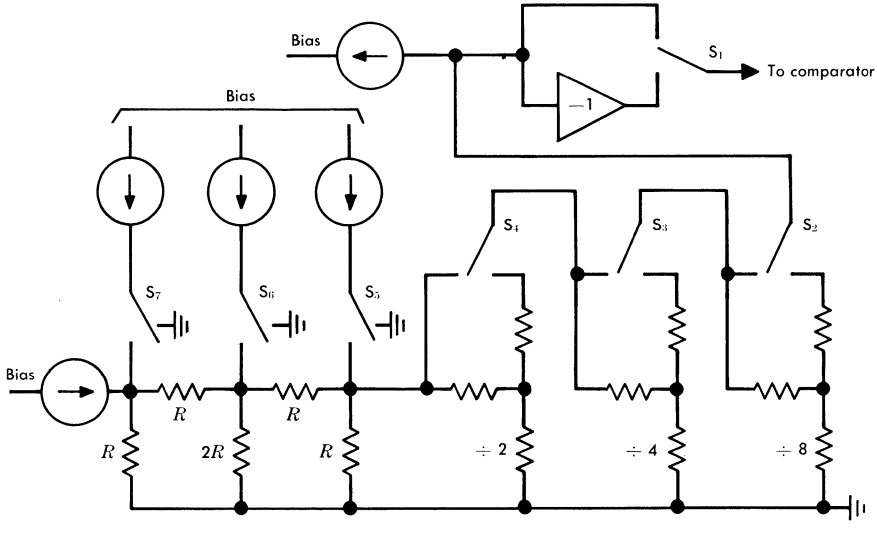


FIG. 25-14. Nonuniform weighting network for a 15-segment compression law.

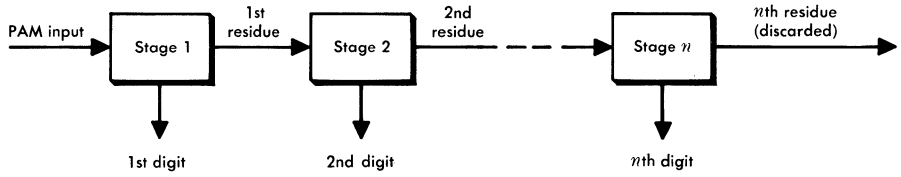


FIG. 25-15. Tandem stage coder.

stage has two outputs, a digit output and a residue output. The residue becomes the input to the next stage. The digit output and residue output characteristics required to generate the binary code are shown in Fig. 25-16. It should be noted that for the first stage the digit output corresponds to the most significant digit assignment of the binary code for all possible inputs, and the residue output determines the remaining code assignment. The binary code assignment is such that after the first digit is determined the remaining code can be divided into two identical halves which differ from each other only by their displacement on the input signal range as illustrated in Fig. 25-16.

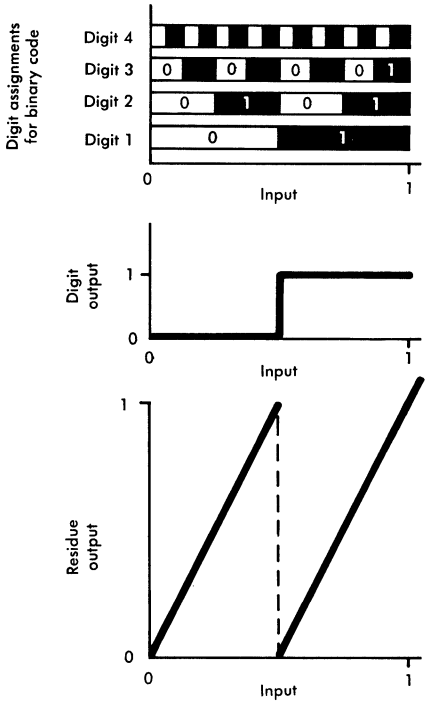


FIG. 25-16. Characteristic of one stage of a binary tandem stage coder.

Cascading of these stages produces a binary coder. Each stage refines the determination of the unknown voltage by halving the residue range of the previous stage. The residue and digit outputs of a binary three-stage coder plotted as a function of the input are illustrated in Fig. 25-17. In a practical situation, slowly varying inputs near a transition are likely to cause an abrupt change in digit output and a switching of reference current. Therefore, some form of

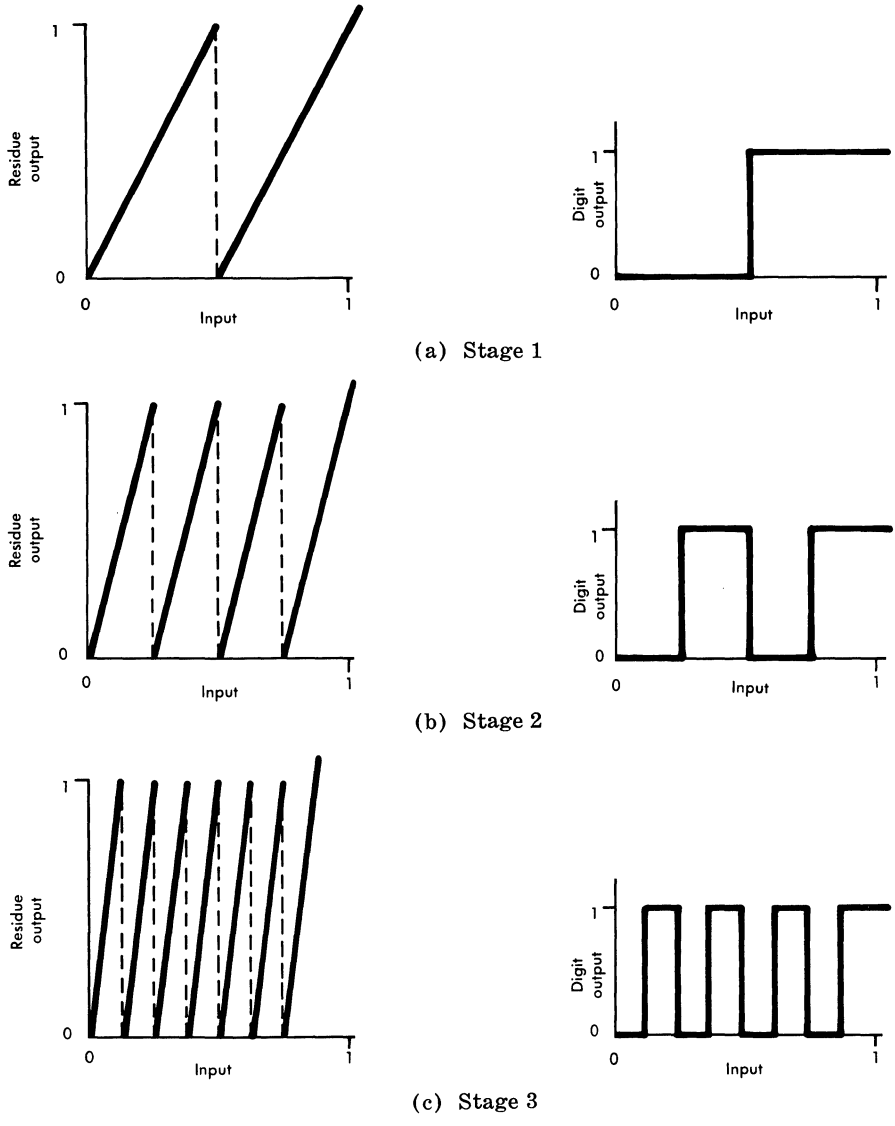


FIG. 25-17. Binary tandem stage coder transfer characteristic.

clocked output is required so that the digits are determined in the proper sequence and remain fixed for the remainder of the coding process.

The transfer characteristic of Fig. 25-16 can be generated by a circuit configuration of Fig. 25-18 [15]. Assuming a high gain in-

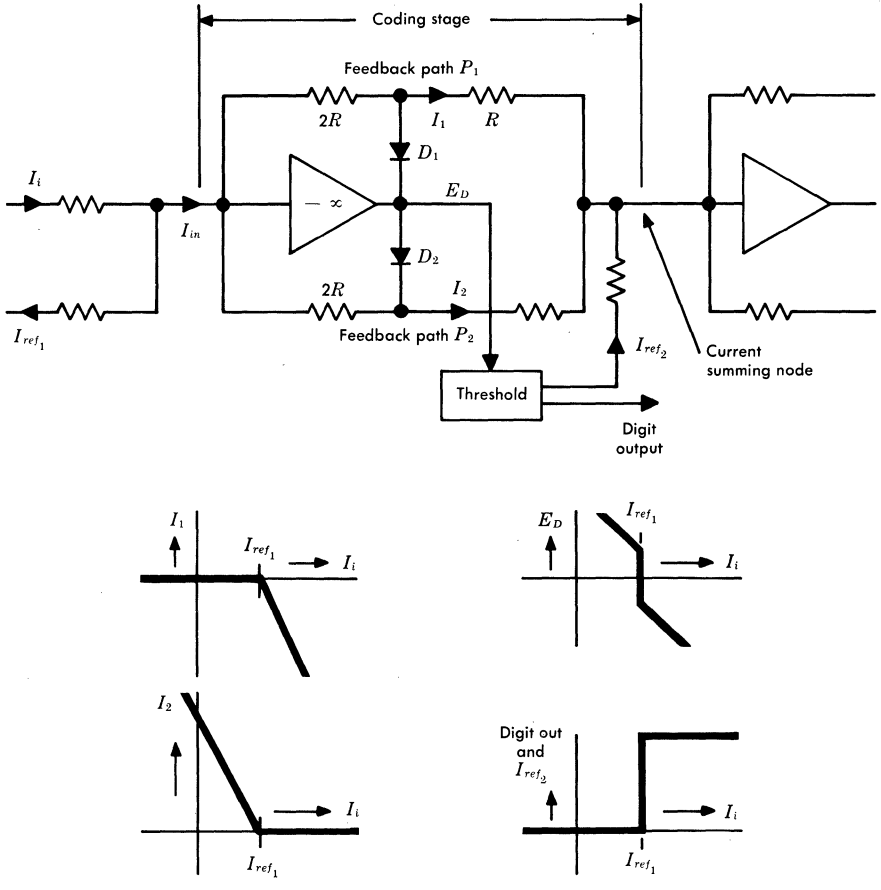


FIG. 25-18. Circuit of one stage of binary tandem coder.

verting amplifier and no reverse conduction in the diodes, any input current, I_i, must result in conduction through one of the two feedback paths, P_1 and P_2. Input currents greater than the reference current, I_{ref}, result in conduction through D_1, while currents less than the reference current result in conduction though D_2. The voltage and current relationships are:

$$\left.\begin{aligned} I_1 &= -2I_{in} \\ I_2 &= 0 \\ E_D &= -2I_{in}\,R - V_D \end{aligned}\right\} \quad I_{in} \geq 0$$

$$\left.\begin{aligned} I_1 &= 0 \\ I_2 &= -2I_{in} \\ E_D &= -2I_{in}R + V_D \end{aligned}\right\} \quad I_{in} \leq 0$$

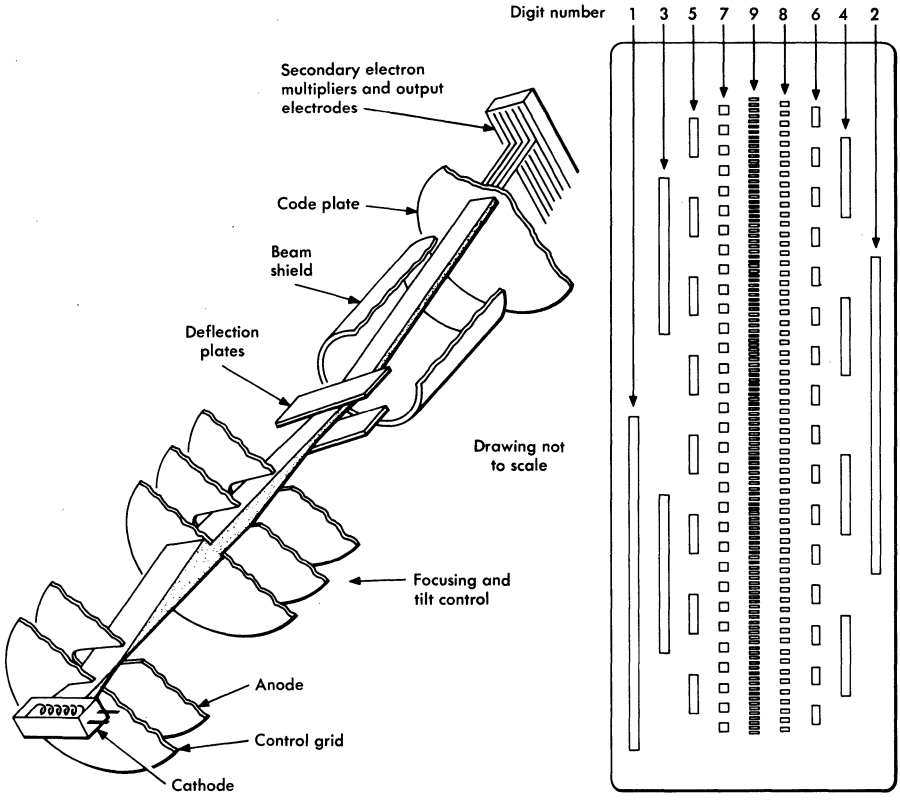
where $I_{in} = (I_i - I_{\text{ref}})$, and V_D is the diode forward voltage drop at a current of $3I_{in}$. The very steep slope of E_D at the vicinity of zero I_{in} enables a coarse threshold detector to sense the polarity very accurately and produce the one or zero code output. The threshold detector also controls the switching of a reference current which when added to I_1 and I_2 will produce the desired transfer characteristic for the residue.

The tandem stage binary coder can easily be adapted to nonuniform coding by making the gain of the two transmission paths, P_1 and P_2, of Fig. 25-18 unequal [16]. It can be shown that the gain ratio of the first stage of a μ-law coder is $(1 + \mu)^{1/2}$ and that of the second stage is $(1 + \mu)^{1/4}$, etc.

Of particular interest is the case for $\mu = 255$ which yields gain ratios of 16, 4, 2, $\sqrt{2}$, $2^{1/4}$, $2^{1/8}$. If the stages of this tandem coder are chosen instead to be 16, 4, 2, 1, 1, 1, a 15-segment piecewise linear approximation to the $\mu = 255$ compression characteristic results; since the gain ratios are in multiples of 2, the code is digitally linearizable. To produce the desired symmetric characteristic a rectifier stage must precede the unequal gain stages.

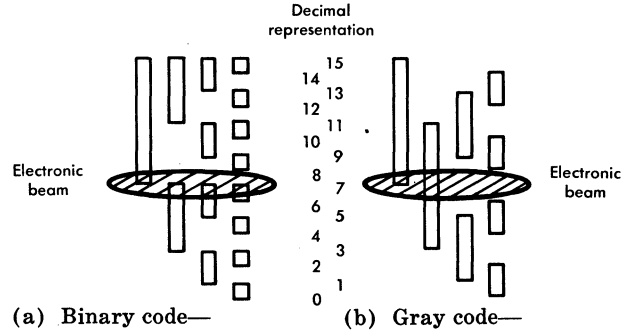
Word-at-a-Time Coding. Word-at-a-time coding is inherently the fastest of the three methods of coding. One method of word-at-a-time coding is to store all possible code words on a code plate in the form of hole or no hole. A beam coding tube illustrated in Fig. 25-19 generates a ribbon beam which is then deflected by the signal voltage. Collector wires behind the code plate sense the presence or absence of electron current, and this is translated into a code word [17, 18].

Word-at-a-time coders generally use the Gray code [19] rather than the usual binary code. For this code, only one digit changes between each pair of adjacent codes, whereas for the binary code several digits may change simultaneously. These codes are compared in Fig. 25-20. A binary coder is susceptible to gross errors of many levels when the PAM sample happens to lie between two code words and the digit transitions are interpreted inconsistently. The Gray coder, because only one digit can be in transition, can have at most an error of one level.



(a) Coding tube (b) 9-digit code plate

FIG. 25-19. Beam coding tube.



(a) Binary code— four digits in transition (b) Gray code— one digit in transition

FIG. 25-20. Comparison of binary and Gray codes for use in word-at-a-time coders.

Conversion from the Gray to binary code can be accomplished by the following serial algorithm

$$b_1 = g_1$$
$$b_k = b_{k-1} \oplus g_k \qquad k \geq 2$$

where b_1 and g_1 are the most significant digits of the binary and Gray codes, respectively, where b_k and g_k are the kth digit of the binary and Gray code words, and where $\oplus$ denotes modulo two addition.

Another implementation of word-at-a-time coding uses multiple threshold circuits. Logic circuits sense the highest threshold circuit triggered by the unknown sample voltage and produce the appropriate code word. Since the number of threshold circuits for an n digit coder is 2^n-1, this method is impractical for large n.

Both the beam coding tube and the multiple threshold coder can be made nonuniform by designing a nonuniform code plate for the beam coding tube and by suitable adjustment of the threshold detectors.

Tandem Stage Gray Coder. A tandem stage Gray coder eliminates the switched reference currents and the sequentially clocked digit outputs as needed in the tandem stage binary coder. If the residue output of a tandem stage coder is that shown in Fig. 25-21, the result is a Gray coder. The transfer characteristic is that of a negative full wave rectifier symmetrical about the center of the range of input amplitudes. This reflects the Gray code assignment which has the property that after determining the first digit, the re-

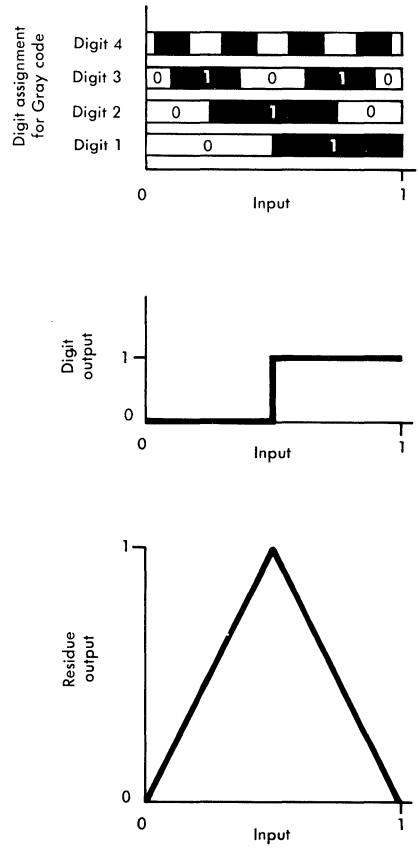


FIG. 25-21. Characteristic of one stage of a Gray tandem stage coder.

maining code is symmetrical about the center and each half is similar to the original code. This full wave rectifier characteristic can be generated by a circuit configuration shown as the first stage in Fig. 25-22 [20, 21]. Essentially, a copy of the input is summed with a half-wave characteristic of double amplitude and opposite polarity. Balanced outputs can be obtained from this stage. Subsequent stages of this tandem coder cannot use this configuration because it requires a voltage input but delivers a current output. Instead, a balanced-pair coder stage is used as shown as the second stage in Fig. 25-22.

Unlike the reference currents of the tandem stage binary coder, the reference currents in the Gray coder are constant quantities independent of the digit output of any stage. This fact and the continuous property of the transfer characteristic (i.e., a small change in input can not cause an abrupt change in the residue output of any coder stage) make this coder inherently fast in operation.

Decoding Methods

The process of decoding is usually a simpler task than coding. A network decoder can be constructed from the resistor weighting network discussed in conjunction with the digit-at-a-time network feedback coder. A decoder can also be built using tandem stages. Each decoding stage is the inverse of the corresponding coding stage. It has a residue input and a digit input . The transfer characteristic for a binary decoding stage is shown in Fig. 25-23(a) and for a Gray decoding stage in Fig. 25-23(b).

Differential PCM Coding

In discussing terminals for visual telephone signals, it was pointed out that for equivalent performance, differential PCM coding [22] could result in a lower digital rate than straight PCM coding. The advantage of DPCM is a result of the fact that video samples are highly correlated. For a bandlimited flat signal spectrum, there is no correlation between samples obtained at the Nyquist rate. For the same S/D ratio, therefore, DPCM results in a higher digital rate. This is because the difference between samples can be as large as $2V$ if the original signal has an amplitude range of $\pm V$. Sampling at a higher rate provides the needed correlation between samples, but since the digital rate is increased in proportion to the sampling rate the trade is an inefficient one. When the signal spectrum is not flat but

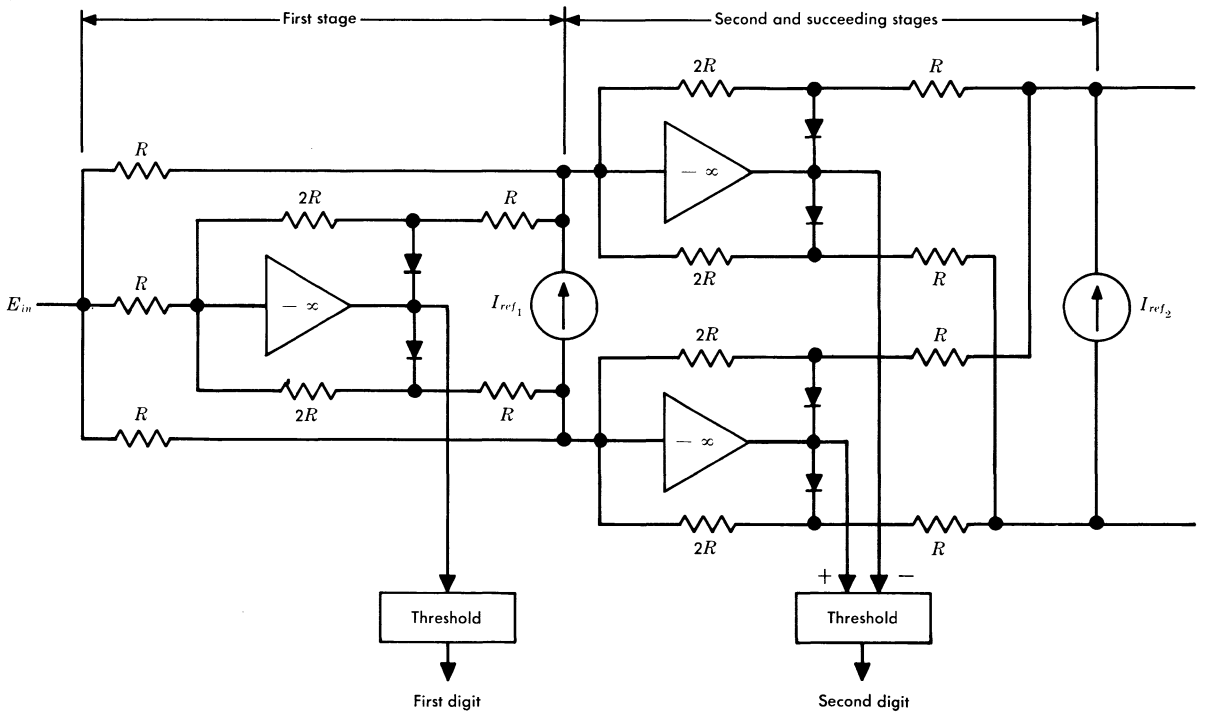
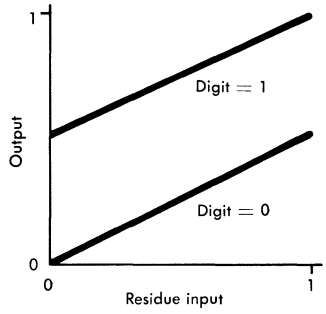
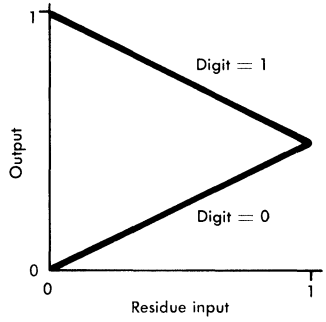


FIG. 25-22. Tandem Gray coder circuit.



(a) Binary decoding



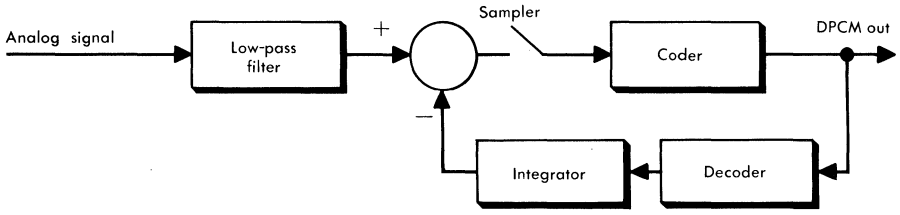
(b) Gray decoding

FIG. 25-23. Tandem stage decoding transfer characteristic.

rolls off gradually towards higher frequencies (as in monochrome video signals), there is correlation between samples and DPCM is attractive [23, 24].

A typical DPCM system is shown in Fig. 25-24. An incoming signal is first bandlimited by a filter to prepare it for sampling. Instead of direct sampling, the difference between the input signal and a prediction signal based on past samples is sampled and coded. For the simple case of zero order extrapolation, the decoded value of the last sample is compared with the input signal and the difference sampled. The prediction filter is therefore an integrator using the sum of all the past differences as the decoded sample value. Thus, in a DPCM system the transmitting section contains a decoder circuit identical to that used in the receiving section.

When the difference signal is coded into one binary digit, this DPCM system is known as delta modulation [25]. The output digits convey only the polarity of



(a) Transmitting section

(b) Receiving section

FIG. 25-24. DPCM terminal.

the difference signal. The decoding integrator will move a fixed increment either up or down to reduce the error between the decoded value and the incoming samples. Waveforms of this system are shown in Fig. 25-25(a). The decoded unfiltered output is a staircase signal. Because adjacent decoded samples can be different by only one step, large errors can occur when the slope of the input signal exceeds the slope of the staircase [26]. This slope is s/T, which is the product of the step size, s, and sampling rate, $1/T$. For an input

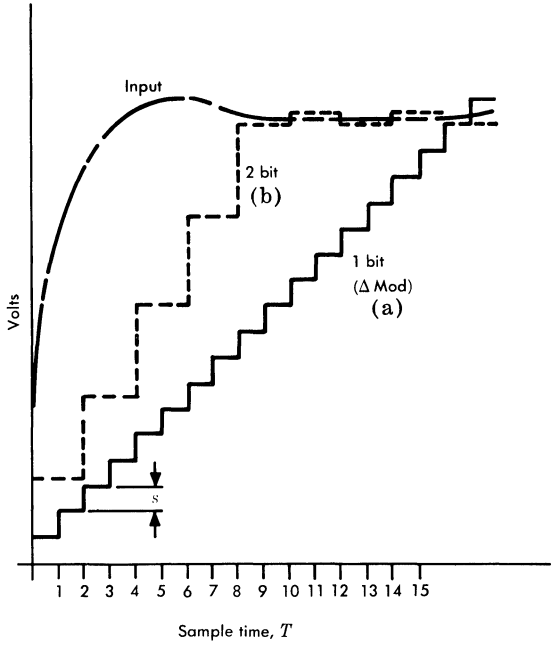
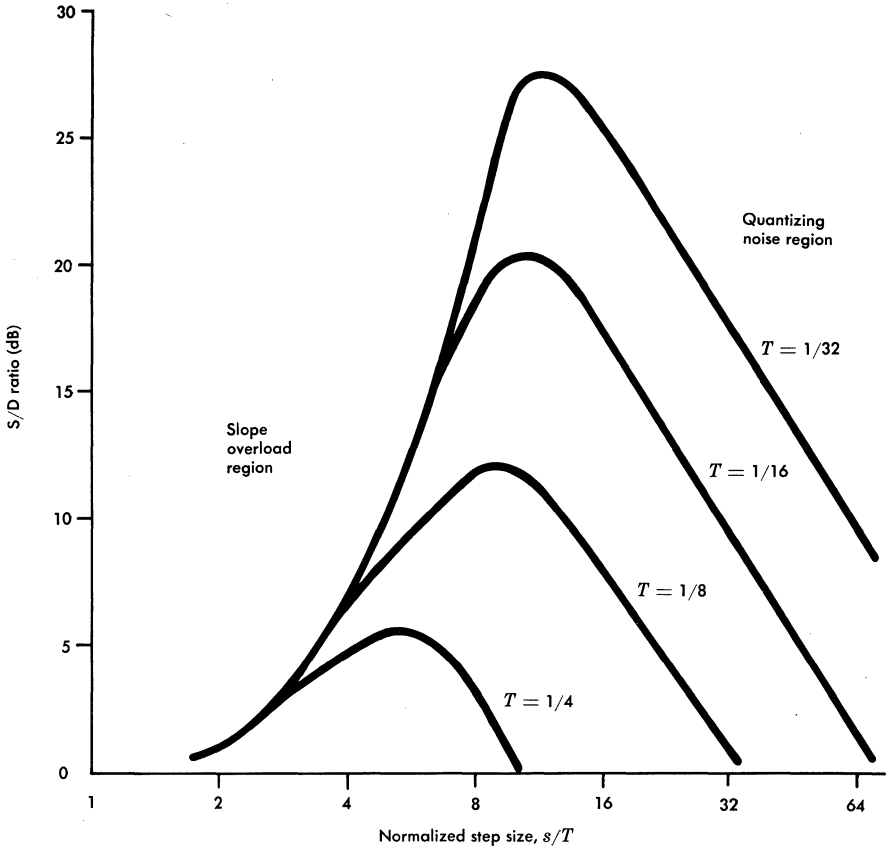


FIG. 25-25. Response of DPCM to a large change in signal amplitude.

slope less than s/T, the error resembles quantizing noise. As in straight PCM, optimum performance is obtained by adjusting the step size so that the sum of the slope-overload noise and quantizing noise is minimized. This is shown in Fig. 25-26, which also demonstrates that by doubling the output digital rate, signal-to-noise ratio is improved by 9 dB; whereas in PCM, with each additional digit in the code word, the ratio improves by 6 dB. For the same digital output rates, the S/D ratios of DPCM and PCM are shown in Fig. 25-27. For low rates DPCM has a larger ratio than that of PCM; for higher rates the reverse is true. The crossover depends on the signal spectrum.



Note: Input signal is gaussian bandlimited flat to $1/2T$ Hz with unit power.

FIG. 25-26. Δ-mod performance.

A system which codes the difference signal into two binary digits is called a two-bit DPCM system. The waveform of the decoder output is illustrated in Fig. 25-25(b). For a fair comparison the output is compared with a delta modulation system with twice the sampling rate so that the digital output rate is the same. The two-bit DPCM has two step sizes; for best performance the larger step size is made greater than twice the step size of the delta modulation system being compared, and the smaller step size is made less than the delta modulation step size. As shown, the two-bit system follows steep input slopes better and yields smaller errors for constant input. Because

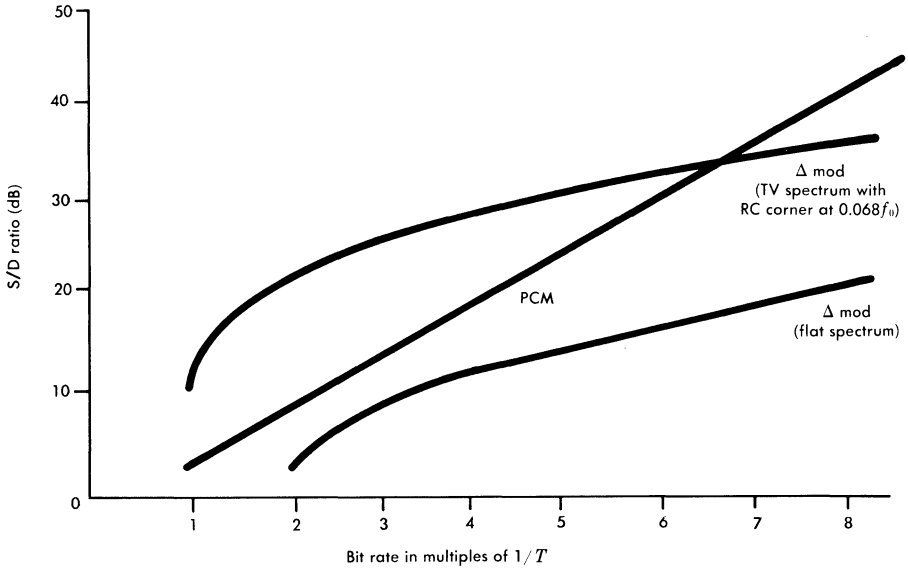


FIG. 25-27. Comparison of PCM and DPCM performance.

of the lower sampling rate, however, the two-bit system does not have a significantly different S/D ratio than a delta modulation system, Fig. 25-28. For video coding, S/D calculations cannot predict subjective reaction and the choice among delta modulation, two- or three-bit DPCM, or straight PCM rests on subjective evaluation.

Before leaving this subject, it should be pointed out that there are circumstances where delta modulation of speech becomes attractive. For example, when a single speech signal is to be coded, a delta modulation coder is much simpler than a PCM coder. Companding, which is necessary for speech, is achieved by varying the step size in response to the slope of the speech waveform.

Coding Impairments

Quantizing Distortion. The coders and decoders described so far are relatively complex and rely on a number of resistors, amplifiers, and decision elements for accuracy. It is difficult to evaluate the effect of practical component variations on the staircase coder characteristic. This job is best done with the aid of a computer. For a practical coder staircase characteristic, $f(x)$, the quantizing distor-

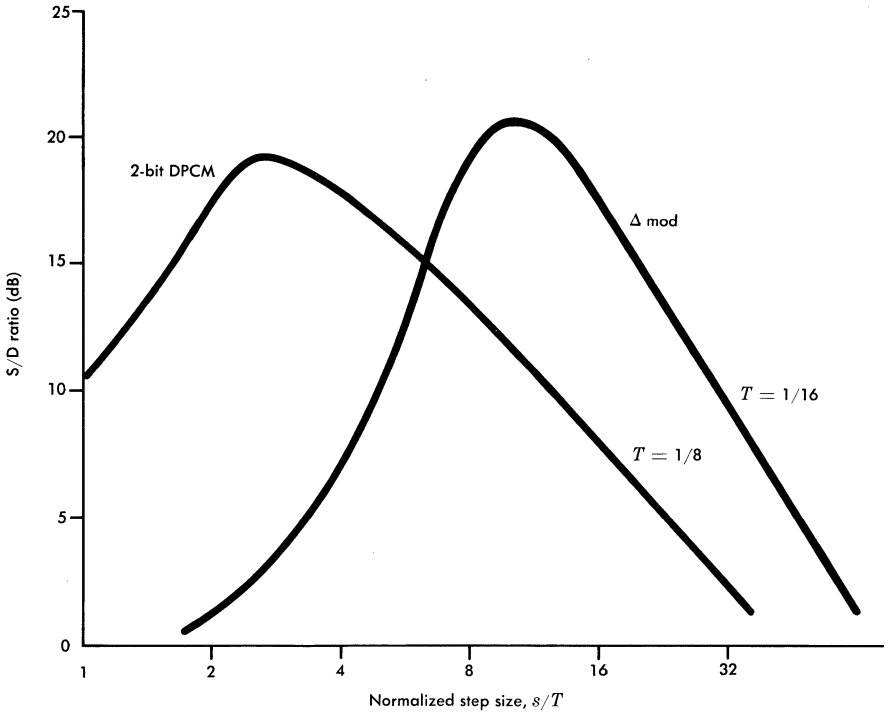


FIG. 25-28. Comparison of 1- and 2-bit DPCM (same bit rate).

tion introduced within the band from 0 to $1/2T$ Hz (where $1/T$ is the sampling rate) is given by

$$\overline{e^2} = \int_{-\infty}^{\infty} [x - Af(x) - B]^2 \, p(x) \, dx \qquad (25\text{-}18)$$

where A and B are selected to minimize $\overline{e^2}$. This minimization is necessary because A and B are gain and d-c offset factors which do not represent distortion. Due to practical component tolerances, the resultant quantizing distortion will be larger than the theoretical distortion given by Eq. (25-7).

Idle Circuit Noise and Interchannel Crosstalk. In the absence of speech input, PCM channels can exhibit enhancement of weak interference such as noise and crosstalk [27]. Although the enhancement varies with the amount of interference and d-c pedestal, it is most pronounced when an idle channel is biased at a decision level. Under

this condition, any minute interference changes the output code word. For this worst case the decoder output is a rectangular wave with peak-to-peak amplitude equal to s, and with random zero crossings. The a-c power of this wave is $s^2/4$. This is referred to as the system noise floor. For biasing at other than a decision level, the output noise power depends on both the bias and the noise amplitude. This is shown in Fig. 25-29 where the noise input power, $\overline{x^2}$, is normalized to the step size.

Crosstalk arriving at the coder can be evaluated in a similar manner. With a low level crosstalk signal exciting a coder that is biased at a decision level, the zero crossings of the decoder output will be governed by the crosstalk signal. The power in the fundamental frequency of the square wave is

$$P_x = \frac{1}{2}\left(\frac{2s}{\pi}\right)^2 \tag{25-19}$$

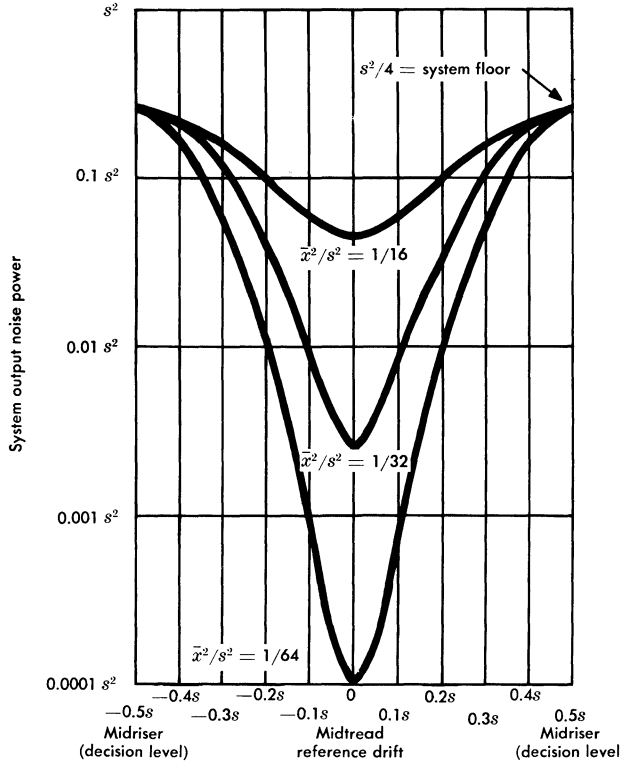


FIG. 25-29. Idle circuit noise versus reference drift.

which is independent of input crosstalk power. As with noise, P_x is referred to as the system crosstalk floor.

The crosstalk situation is improved with additive noise because the zero crossings tend to become scrambled and thus the recovered crosstalk power at the fundamental frequency is reduced. With sufficiently large noise power, crosstalk will not be enhanced, as shown in Fig. 25-30.

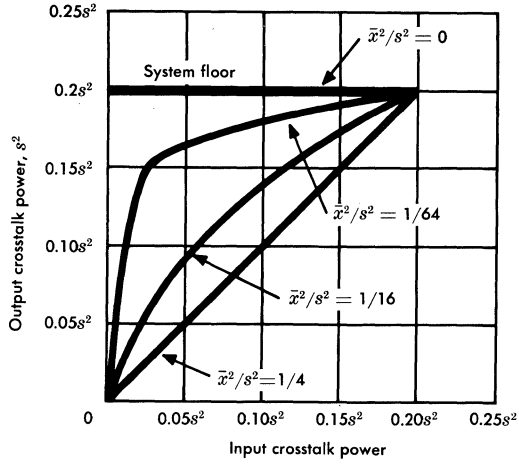


FIG. 25-30. Interchannel crosstalk when biased at a decision level.

25.3 FRAMING

Each pulse stream in a digital transmission system is composed of components from one or more signal sources. Usually the digitized signals are combined serially in a periodic fashion. This periodic structure is known as the line format, and one basic period is called a frame. A framing strategy is necessary to identify the various signal components and to bring the receiver into phase with respect to the line format so the digits can be properly assembled for decoding and demultiplexing. For single channel terminals, framing serves only to identify the PCM code word structure since only one signal is involved.

At the transmitter an identifiable pulse sequence associated with the periodic output format of the terminal, or some multiple of this format, is inserted into the digital stream. This framing character-

istic defines the frame length. For a frame length of N digits, there are N different phases that the receiver can assume, only one of which possesses the framing characteristic. The framing strategy consists of a procedure that searches through the N possible phases until the correct phase is found and verified. The receiver is then said to be in-frame.

Framing strategies utilize pulse patterns that are transmitted either by an inserted characteristic or by an intrinsic characteristic of the signal itself. In one method of inserting a framing pattern, called added digit framing, a dedicated digit position in a frame is used. In another method, called robbed digit framing, the characteristic pattern is transmitted by the insertion of framing digits in place of the information-carrying digits. In certain situations the statistical behavior of one digit position in a frame is significantly different from all the other digit positions. A strategy based on such an intrinsic characteristic is called statistical framing.

A framing strategy is evaluated by considering its effects on system performance as measured by impairments to the signal, reframe time, susceptibility of the framing circuit to line errors, and the required complexity of the framing circuitry in the terminal. For added digit framing the impairment to the signal is a reduced information rate for a given line digital rate. Furthermore, if the added framing digits interrupt the periodicity of the code groups, impairment arises from the residual jitter of the sample times even after smoothing. Of the three strategies mentioned, robbed digit framing has the greatest signal impairment because replacement of information digits with framing digits increases quantization noise. Statistical framing requires no dedicated digits and incurs no added quantization noise; however, as will be seen later it is not as reliable as the other strategies in maintaining frame.

Reframe time is the length of time required for the receiver to reestablish the in-frame position after inadvertent loss of frame. This involves searching through all possible pulse positions in order to locate the correct framing sequence. Requirements on reframe time depend upon the type of signal being transmitted and on the number of channels affected by the loss of frame.

Inadvertent loss of frame is usually caused by line errors which may temporarily alter the framing characteristic and cause the receiver to declare itself out-of-frame. Such an unnecessary reframe and attendant loss of information is called a misframe. The mean time between misframes even under conditions of high error rates

must be made long to prevent excessive loss of information. Statistical framing can misframe even without line errors. This occurs whenever the short term statistics of the signal being transmitted deviate from the expected norm.

All of these measures of performance can be made arbitrarily attractive at the expense of increased complexity of the framing circuitry. Since complexity increases terminal cost, the trade-off between these performance criteria must necessarily include circuit complexity.

Added Digit Framing

The most straightforward method of identifying a frame interval is to add a unique pattern to the line format [28, 29]. The receiver then searches for this pattern and locks onto it. The pattern chosen is usually an alternating one-zero sequence. An all 1's or all 0's pattern, although simple, can be easily confused with the binary code words of an idle circuit. The alternating framing pattern is verified by comparing the incoming framing digits with the framing digits generated internal to the receiver. If the pulse pattern of the time slot under examination is not the expected pulse sequence, the next position is searched. This procedure continues until a framing position is found where the one-zero pattern persists.

In order to provide long times between misframes and yet short reframe time, the receiving framing circuitry has two modes of operation. When the receiver is in-frame, a single framing digit error will not initiate a reframe search. It takes repeated violation of the expected framing pattern, such as three errors in seven frames or less, to cause the receiver to enter the out-of-frame mode. In this mode a single violation of the expected framing digit will cause a shift to the next position in the frame.

The total digital rate increase due to added digit framing depends on the desired reframe time. On the average, the receiving terminal dwells at a false framing position for two frame periods during a search, since 1's and 0's are statistically independent and equally likely. Searching through all possible positions requires N such tests, where N is the number of pulse positions in the frame. Therefore,

$$\text{Maximum average reframe time} = 2NT = 2N^2\tau \quad (25\text{-}20)$$

where T is the period of the frame and is equal to N times the digit period, τ. For example, in the D1 bank, N is 193 and T is 0.125 milli-

second, yielding a maximum average reframe time of 48 milliseconds. Equation (25-20) shows that reframe time increases as the square of the frame length.

In general, many digits can be assigned for framing in a given frame period, and these digits may be clustered together or distributed. The receiver can search every possible position in the frame simultaneously to achieve the fastest reframe, or it can search a portion of the frame as a compromise between fast reframe and complexity. Requirements on reframe time for message channels seldom require the rapid reframe provided by multiple framing digits.

Robbed Digit Framing

When the natural frame period is very short, inserted digit framing becomes inefficient. This occurs in single channel terminals where the frame is only the 9 digits of each PCM code word. Insertion of a framing digit with every word results in a 10 per cent loss in information capacity. A longer frame can be used, but inserting a framing digit destroys the periodicity of the code words. Since sampling and interpolation require periodic PAM pulses, both the transmitter and receiver must provide variable delay circuits to account for aperiodic transmission of the code words. Even then residual jitter will distort the signal. A simple solution to this framing problem is to replace the least significant digit of every kth code word by the framing digit [30]. The search strategy at the receiver will be the same as for added digit framing. The difference is that robbed digit framing introduces additional quantization impairment to the signal while added digit framing does not.

The parameter k is chosen as a compromise between reframe time and signal impairment. Since the number of digits in a frame, N, is proportional to k, Eq. (25-20) indicates that the reframe time is proportional to k^2. However, quantizing distortion decreases with increasing k. From Eq. (25-8) it can be seen that the resulting total noise is the weighted sum:

$$\overline{e^2} = \frac{k-1}{k} \cdot \frac{s^2}{12} + \frac{1}{k} \cdot \frac{(2s)^2}{12}$$

$$= \frac{s^2}{12}(1 + 3/k) \qquad (25\text{-}21)$$

Thus the impairment due to robbing is

$$I(k) = 10 \log (1 + 3/k) \quad \text{dB} \qquad (25\text{-}22)$$

If $k = 10$, the impairment is 1 dB. Equation (25-22) assumes that the receiver takes proper action to decode to the center of the doubled step sizes. If no action is taken, the framing digits will be treated as signal information and will result in digital errors. Equation (25-22) is also a good approximation for a nonuniform coder and for nonuniform amplitude probability distributions as long as the S/D ratio is high.

Statistical Framing

Framing information may be contained in the digital output of a coder without having to add or rob digits [31]. One of the easiest statistical framing methods is to monitor the one-zero probability of a digit position in the incoming digital stream [32]. When the Gray code is used, the second digit is a 1 in the central half of the code range and a 0 at the extremes, Fig. 25-20. Therefore, a centrally peaked amplitude distribution such as that of a mastergroup signal generates a high probability of a 1 in the second digit. At the loading that results in minimum distortion, the rms value of the mastergroup signal is one-quarter of the overload point of the coder, resulting in a probability of 0.95 of a 1 in the second digit. For all the other digits this probability is less than 0.5.

Digit statistics can be monitored by many methods. An RC integrating circuit followed by a threshold detector is perhaps the simplest. When a 0 in the second digit of a PCM code word is detected, a pulse is generated at the input to the RC integrator. The time constant and threshold are designed so that when the occurrence of second digit 0's is 5 per cent, the threshold is rarely exceeded, and when the occurrence increases towards 50 per cent, the threshold is exceeded in a few frame periods. For a mastergroup coder the misframe interval is longer than one day, and the maximum average reframe time is about 160 microseconds.

Although statistical framing has the attractive feature of no increase in digital rate, it must be used with caution. When the mastergroup transmission facility is tested with a sine wave or when the entire facility is to carry a signal with unknown statistics, the frequency of misframes may be intolerable.

25.4 TERMINAL PERFORMANCE MONITORING

To provide good quality service, the terminal performance should be monitored so that when a failure occurs or when the quantizing noise exceeds a maintenance limit, the terminal is removed from service. The maintenance limit and the time allowed for restoration of service depend on the required grade of service and the number of message channels affected. Generally, terminals that serve 600 or more channels must be restored rapidly which requires automatic switching to spares, while terminals serving fewer channels are allowed more time and service may be restored by manual connection to spares or by repair.

Framing information recovered at the receiving end provides a good indication of the performance of clock circuits at both the transmitting and receiving ends and of the performance of the digital transmission facility used. Continuous or frequent misframing is therefore used to initiate alarms at the terminal so that proper action can be taken.

Framing, however, cannot monitor the codec quantizing noise or other analog portions of the terminal. Other means must be used. Per channel monitoring is economically prohibitive for channel banks, and the philosophy is to monitor those parameters that can be monitored cheaply. For example, in the D2 channel bank the coder is made to code a known signal and the output code is monitored. Similarly, the decoder is given a known code word and the analog output is monitored. Such schemes are only partial solutions since they do not cover all possibilities of failure.

For the mastergroup terminal serving 600 message channels, monitoring is more inclusive. The in-band spectral nulls and the mastergroup pilot are monitored as a measure of the performance of the codec. Whenever the pilot levels or noise at the special nulls goes beyond prescribed limits, another terminal is substituted.

REFERENCES

1. Whittaker, E. T. "On the Functions Which Are Represented by the Expansions of the Interpolation Theory," *Proc. Royal Society*, vol. 35 (Edinburgh, 1914-1915), pp. 181-194.
2. Nyquist, H. "Certain Topics in Telegraph Transmission Theory," *Trans. AIEEE*, vol. 47 (1928), pp. 617-644.
3. Linden, D. A. "A Discussion of Sampling Theorems," *Proc. IRE*, vol. 47 (July 1959), pp. 1219-1226.

4. Haard, H. B. and C. G. Svala. "Means of Detecting and/or Generating Pulses," U. S. Patent 2718621, Sept. 20, 1955.
5. Cattermole, K. W. "Efficiency and Reciprocity in Pulse Amplitude Modulation," *Proc. IEE* (Dec. 1957), pp. 449-462.
6. Max, J. "Quantizing for Minimum Distortion," *Trans. IRE*, vol. IT (Mar. 1960), pp. 7-12.
7. Smith, B. "Instantaneous Companding of Quantized Signals," *Bell System Tech. J.*, vol. 36 (May 1957), pp. 653-709.
8. Purton, R. F. "Survey of Telephone Speech-Signal Statistics and Their Significance in the Choice of a PCM Companding Law," *Proc. IEE*, vol. B109 (London, Jan. 1962), pp. 60-66.
9. Cattermole, K. W. Discussion on the above paper by Purton, *Proc. IEE*, vol. B109 (London, Jan. 1962), pp. 485-487.
10. Kaneko, H. and T. Sekimoto. "Logarithmic PCM Encoding Without Diode Compandor," *IEEE Int. Conv. Rec.*, vol. 11, PT-8 (1963), pp. 266-281.
11. Richards, D. L. "Transmission Performance of Telephone Networks Containing PCM Links," *Proc. IEE*, vol. 115 (Sept. 1968), pp. 1245-1258.
12. Smith, B. D. "Coding by Feedback Methods," *Proc. IRE*, vol. 41 (Aug. 1953), pp. 1053-1058.
13. Mann, H., H. M. Straube, and C. P. Villars. "Companded Coder System for an Experimental PCM Terminal," *Bell System Tech. J.*, vol. 41 (Jan. 1962), pp. 173-226.
14. Smith, B. D. "An Unusual Electronic Analog-Digital Conversion Method," *IRE PGI*, vol. PGI-5 (June 1956), pp. 155-160.
15. Waldhauer, F. D. "PCM Coder," U.S. Patent 3161868, Dec. 15, 1964.
16. Dammann, C. L. "An Approach to Logarithmic Coders and Decoders," *NEREM Record* (1966), pp. 196-197.
17. Sears, R. W. "Electron Beam Deflection Tube for Pulse Code Modulation," *Bell System Tech J.*, vol. 27 (Jan. 1948), pp. 44-57.
18. Cooper, H. G., M. H. Crowell, and C. Maggs. "A High Speed PCM Coding Tube," *Bell Laboratories Record*, vol. 42 (Sept. 1964), pp. 266-272.
19. Gray, F. "Pulse Code Communications," U.S. Patent 2632058, March 17, 1953.
20. Waldhauer, F. D. "Analog-to-Digital Converter," U.S. Patent 3187325, June 1, 1965.
21. Edson, J. O. and H. H. Henning. "Broadband Codecs for an Experimental 224 Mb/s PCM Terminal," *Bell System Tech. J.*, vol. 44 (Nov. 1965), pp. 1887-1940.
22. Cutler, C. C. "Differential Quantization of Communications Signals," U.S. Patent 2605361, July 29, 1952.
23. O'Neal, J. B. "Delta Modulation Quantizing Noise—Analytical and Computer Simulation Results for Gaussian and Television Input Signals," *Bell System Tech. J.*, vol. 45 (Jan. 1966), pp. 117-141.
24. McDonald, R. A. "Signal to Noise and Idle Channel Performance of Differential Pulse Code Modulation Systems—Particular Applications to Voice Signals," *Bell System Tech. J.*, vol. 45 (Sept. 1966), pp. 1123-1151.
25. Deloraine, E. M., S. Van Merlo, and B. Derjavitch. French Patent 932-140, Aug. 10, 1946.
26. Protonotarious, E. N. "Slope Overload Noise in Differential Pulse Code Modulation Systems," *Bell System Tech. J.*, vol. 46 (Nov. 1967), pp. 2119-2162.

27. Shennum, R. H. and J. R. Gray. "Performance Limitations of a Practical PCM Terminal," *Bell System Tech. J.*, vol. 41 (Jan. 1962), pp. 143-171.
28. Gilbert, E. N. "Synchronization of Binary Messages," *IRE PGIT*, vol. IT-6 (Sept. 1960), pp. 470-477.
29. Rauch, L. L. "Considerations on Synchronization for PCM Telemetry," *IRE PGSET*, vol. Set-6, (Sept.-Dec. 1960), pp. 95-98.
30. Mayo, J. S. "Experimental 224 Mb/s PCM Terminals," *Bell System Tech. J.*, vol. 44 (Nov. 1965), pp. 1813-1941.
31. Mayo, J. S. and R. J. Trantham. "Statistical Framing of Code Words in a Pulse Code Receiver," U.S. Patent 3175157, Mar. 23, 1965.
32. Gray, J. R. and J. W. Pan. "Using Digital Statistics to Word-Frame PCM Signals," *Bell System Tech. J.*, vol. 43 (Nov. 1964), pp. 2985-3007.

Chapter 26

Digital Multiplexers

In the same way that various analog systems are used to transmit analog signals of different bandwidths, there are various digital transmission facilities designed to transmit digital signals of different rates. These facilities must be interconnected into a network that allows a digital signal to reach its destination using one or more of these facilities. Interconnections must be flexible enough to provide for alternate transmission paths in case of equipment failures, changing traffic patterns, and routine maintenance. Interconnection of facilities of the same digital rate involves either manual patching or automatic switching. Interconnection of different digital rates requires multiplexers that combine several signals to share a higher speed digital facility.

Time division multiplexing of several digital signals to produce a higher speed stream can be accomplished by a selector switch that takes a pulse from each incoming line in turn and applies it to the higher speed line. The receiving end will do the inverse of separating the higher speed pulse stream into its component parts and thus recover the several lower speed digital signals. The main problems involved are the synchronization of the several pulse streams so that they can be properly interleaved and the framing of the high-speed signal so that the component parts can be identified at the receiver end. Both of these operations require elastic stores, which constitute important parts of a multiplexer.

Elastic stores are also called data buffers. Information pulses arriving at the multiplexer must await their turn to be applied to a higher speed transmission system. Due to delay variations of the incoming lines and to the framing and synchronization operation of the multiplexer terminal, this wait is variable in time.

The framing problem is similar to that of digital terminals, and the same techniques are available. For reasons of flexibility, added bit framing is chosen for all multiplexers. This choice assumes no given input signal statistics and leaves the digits intact so that a multiplexer can handle any digital signal regardless of its source.

26.1 METHODS OF SYNCHRONIZATION

The major problem of multiplexer system design is synchronization. Digital signals cannot be directly interleaved in a way that allows for their eventual identification unless their pulse rates are locked to a common clock. Because the sources of these digital signals are often separated by large distances, their synchronization is difficult. Synchronization methods which can be used are: (1) master clock, (2) mutual synchronization, (3) stable clocks, and (4) pulse stuffing.

Master Clock

An obvious method of synchronization is the use of a master clock for timing the entire system [1]. One outstanding difficulty of this approach is the vulnerability of the system to failures of either the master clock or the transmission links. A system with enough redundancy and automatic protection against failures presents a difficult design problem and is expensive to establish.

Mutual Synchronization

In this method of synchronization (referred to also as phase averaging), each station or central office has its own clock whose frequency is the average of all the incoming frequencies and a local standard [2]. It can be shown that all stations will approach a common steady-state frequency. This method avoids one aspect of the master clock reliability problem since now no one clock or transmission path is essential. Further studies are necessary, however, to determine the optimum averaging algorithm for each station such that the network can grow gracefully from a few nodes to a large network and that minor disturbances such as protection switching of a transmission link will not cause the system frequency to swing beyond the design limits [3].

Stable Clocks

A third method of synchronization is the use of very stable clocks at each office containing digital terminals. Elastic stores are then used

to absorb the very slowly varying phase errors. Since their capacity is finite, these stores must be reset periodically with some loss of information. With atomic clocks stable to one part in 10^{12} and with large enough elastic stores, loss of information will be acceptably infrequent. A combination of the preceding three methods is likely to be used in the future. Each region will have stable clocks supplemented by mutual synchronization while the master clock scheme will be used within each region.

Pulse Stuffing

The final method of synchronization is pulse stuffing, presently being used in the design of multiplexers [4]. The concept is to have the outgoing digital rate of a multiplexer higher than the sum of the incoming rates by stuffing in additional pulses. All incoming digital signals are stuffed with a sufficient number of pulses to raise each of their rates to that of the locally generated clock signal, Fig. 26-1.

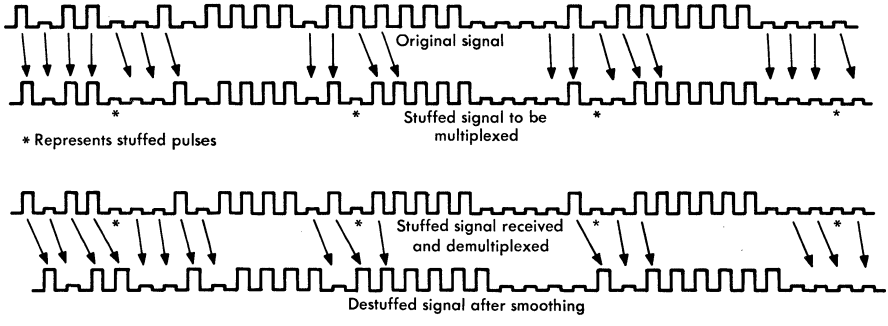


FIG. 26-1. Pulse stuffing synchronization.

Pulse stuffing is the least complex of the four methods proposed because it needs the least amount of buffer storage. In each of the other methods, even assuming that the clocks are perfectly synchronous, propagation delay variations may cause a surplus or deficit of pulses at any one location. For example, a 1000-kilometer coaxial cable carrying 3×10^8 pulses per second will have about one million pulses in transit, each occupying about one meter of the cable. A 0.01 per cent increase in propagation velocity, as would be produced

by a 1°F decrease in temperature will result in 100 fewer pulses in the cable; these must be absorbed by elastic stores in the multiplexer. With pulse stuffing, a deficit of one pulse is immediately made up by stuffing so that one cell of storage is all that is needed to handle such variations.

Pulse stuffing is done independently for each multiplexer, and this contributes to system reliability. The pulse rate on a particular transmission line is determined locally and will not influence any other clock in the system. Failure of one multiplexer or line will only affect those signals passing through that multiplexer or line.

By choosing pulse stuffing, network synchronization is not precluded. Asynchronous low-speed digital signals are stuffed in order to use a higher speed transmission line; they are destuffed and the original rate restored before they leave the transmission system. If digital switching is established in the future, the lower speed signals can be made synchronous and the higher speed transmission lines can still be used. Flexibility is thus preserved.

26.2 MULTIPLEXER SYSTEM DESIGN

In addition to pulse stuffing, other design choices must be made to characterize the family of digital multiplexers. When a pulse is stuffed, an additional communication channel is used to inform the receiving terminal of the location of the stuffed pulse. Either this channel can be provided separately for each signal being stuffed or a common data channel can be shared. Separate stuff channels would be more flexible but because shared equipment is more economical, stuffing information for all signals entering a multiplexer is processed and transmitted on one channel. This common data channel is multiplexed along with the information pulses for transmission over a higher speed digital line.

Another choice concerning the system design of the family of multiplexers is the structure of the digital hierarchy. Certain digital rates are designated as belonging to the hierarchy, e.g., 1.544 Mb/s is T1 rate, 6.312 Mb/s is T2, and 46.304 Mb/s is T3. Multiplexers accept signals of one rate and multiplex them only to the next higher rate. To combine several T1-speed digital signals into a T3 rate, it is necessary to go through two multiplexers. Multiplexers are named for the rates they bridge. Multiplexer M12, for example, is designed to combine several T1 signals into a T2 signal.

Signal Format

These design choices fix the general structure of the line formats of all multiplexers. Figure 26-2 illustrates a typical format, that of the M12 multiplexer [5]. Digit-by-digit interleaving of four signals at the T1 speed forms the fine structure of the format. After every 48 information time slots, 12 from each of the four T1 signals, a control digit is inserted by the multiplexer. Control digits labeled F are the main framing digits. Between F digits are control digits labeled M and C. Three successive C digits denote the presence or absence of a single stuffed pulse, and the corresponding M digit identifies in which of the four multiplexed T1 signals the stuff occurs. The M digits thus form secondary framing digits and identify four subframes. The subscripts of the M and F digits identify the digit as a 0 or a 1. Thus, F_1 is always a 1 and the next control digit is either an M_1 or M_0. The three C digits in the subframe following M_0 are stuffing indicators for the first T1 signal, three 1's for the presence of a stuffed pulse and three 0's for no-stuff. If the C digits indicate a stuff, the location of the stuffed pulse is the first information pulse position associated with the first T1 signal following the next F_1 pulse. The other sequences of C digits denote stuffing in the second, third, and fourth T1 signal. The use of three digits for a stuff indication provides a single digit error correction code.

The demultiplexer at the receiving M12 first searches for the $F_0F_1F_0F_1$ sequence. This establishes identity for the four T1 signals

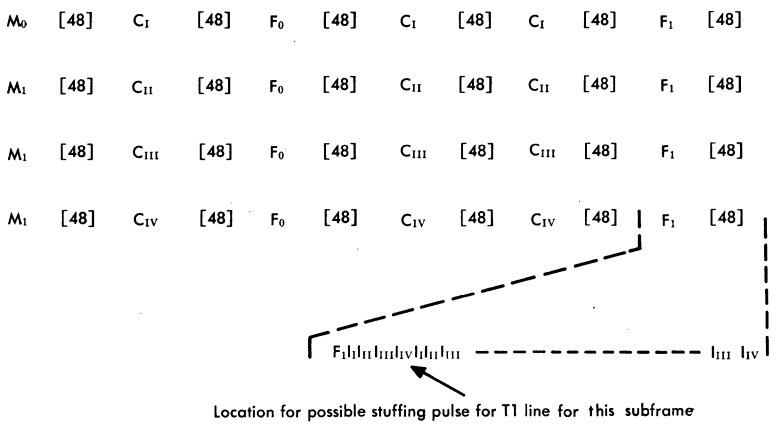


FIG. 26-2. M12 multiplexer format.

and also for the M and C control digits. From the $M_0M_1M_1M_1$ sequence, secondary framing of the C digits is established and the four T1 signals are properly demultiplexed and destuffed. This format has two safeguards. The first is framing. It is possible, although unlikely, that with just the $F_0F_1F_0F_1$ sequence, one of the T1 signals could contain a similar sequence. The receiver could then lock onto the wrong sequence. Presence of the $M_0M_1M_1M_1$ provides verification of the genuine $F_0F_1F_0F_1$ sequence. The second safeguard is the single error correction ability of the stuff indicators. Error rate objectives of digital transmission lines make double errors very unlikely.

The capacity of the M12 format to accommodate different input signal digital rates can be calculated from the format. In each M frame, defined as the interval containing one cycle of $M_0M_1M_1M_1$ digits, one pulse can be stuffed in each of four signals. Because each signal has $12 \times 6 \times 4$ or 288 positions in each M frame, it can be incremented by 1.544 Mb/s $\times$ 1/288 or 5.4 kb/s, which is much larger than the expected frequency tolerance of the incoming T1 signal. The local clock that determines the outgoing speed also determines the nominal stuff rate, which must be high enough to accommodate the highest expected incoming rate. In more precise formulation the nominal T2 rate is

$$1.544 \times 4 \times \frac{49}{48} \times \frac{288}{288 - S} \quad \text{Mb/s}$$

where 4 is the number of T1 signals, 49/48 is the ratio of total time slots to information time slots, and S is the ratio of nominal to maximum stuffing rates. Choosing an S of 1/2 gives the greatest frequency tolerance each way from nominal. A lower S reduces the waiting time jitter imparted to the signal as is explained subsequently. It is also desirable to select a nominal T2 clock frequency to be a multiple of 8 kHz. With these considerations, the T2 rate was chosen to be 6.312 megabits per second, and the resulting S is about 1/3.

Although it seems that the M12 format has far greater stuffing capability than that warranted by the long-term frequency stability of the terminal, there are reasons for the extra margin. One is that the instantaneous frequency of a signal as it reaches a multiplexer can deviate beyond terminal frequency limits. This deviation is caused by timing jitter of the digital repeaters. In the case of T1, the instantaneous frequency deviation is estimated to be ± 1 kHz. The main reason for the extra margin, however, relates to the

reframe time of the multiplexer. In Chap. 25, it was seen that reframe time is proportional to the square of N, the number of digits between framing digits. To ensure that reframe time is at least shorter than that of the channel banks, N is constrained to be between 100 and 200. The stuffing control digits M and C are then located in fixed positions between the F digits, which results in a simple framing circuit. These constraints lead to the M12 format which has some extra stuffing margin.

Designs with fewer control digits are possible, but the increase in cost of the multiplexer must be weighed against the decrease in required channel capacity.

System Block Diagram

A typical block diagram of a multiplexer is shown in Fig. 26-3. At the transmitting portion, each incoming pulse stream has a terminating repeater and code translator. This equipment converts the incoming multilevel pulse signal into binary digits and extracts the timing information. The binary information is then written into individual elastic stores under control of the derived timing signal. The information is read out of the elastic store under control of the local clock in the multiplexer which has a higher frequency than the incoming clock signal.

The occupancy of all the input elastic stores is sampled in turn by a common control circuit. The common control contains a local clock and countdown circuitry to time all its operations. When the occupancy of an elastic store is sampled, nothing is done if the state of depletion does not indicate a stuff; otherwise, a stuffing operation will take place as follows. First, a series of 1's will be transmitted in the C positions of the format, and then reading of the elastic store will be suspended for one time slot at a predetermined location in the frame. The common control then cycles to sample the next elastic store. The binary digits as assembled by the common control are then converted into an appropriate pulse stream for application to the transmission line.

The receiving portion of the multiplexer, Fig. 26-3, has a terminating repeater and code translator to convert the multilevel pulse signals from the high-speed line back to the binary form. The receiving common control contains a clock and countdown circuit, similar to that of the transmitter, for demultiplexing and destuffing operations. A framing circuit keeps the receiving common control in phase with the line format.

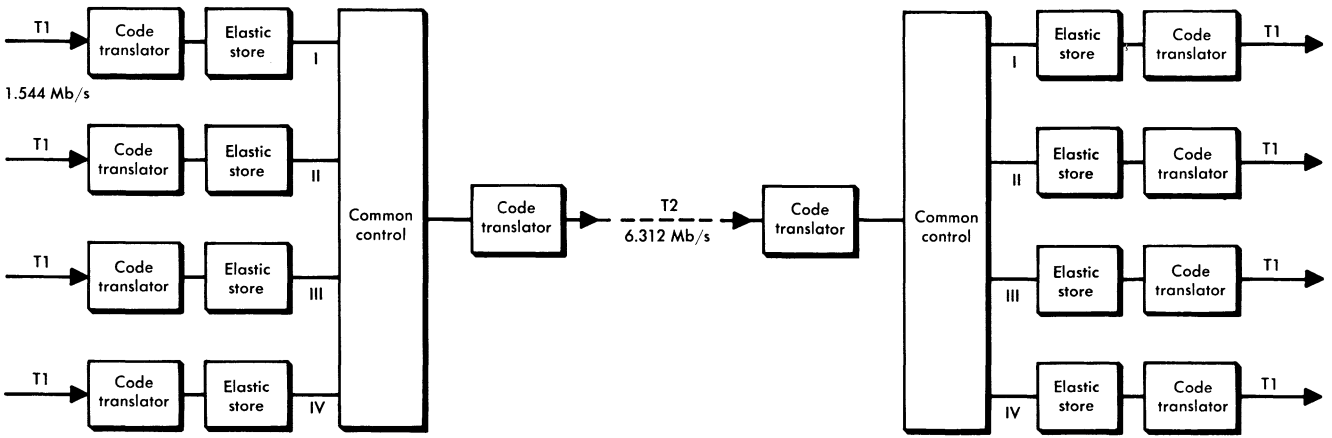


FIG. 26-3. M12 digital multiplexer block diagram.

The C pulses are accumulated in a register as they arrive. If at least two C's in a subframe are 1's, the next available time slot after the F_1 pulse for the appropriate T1 signal is assumed to contain a stuffed pulse that will not be written into the elastic store. Otherwise, information is demultiplexed and written into the respective elastic stores for each channel. Since information is put into the elastic stores periodically but with occasional interruptions due to destuffed pulses, a phase-controlled clock is used to read information out of the store as smoothly as possible. After the binary information is read out of the store, it is converted back to a T1 signal.

Elastic Stores

In a synchronous network, elastic stores compensate for delay variations, and it has been shown that the required size may become very large. In pulse stuffing, the elastic stores can be as small as one cell since single pulses are stuffed. However, other functions in the multiplex require some elastic delay, so elastic stores of four cells or more are usually used.

Incoming digits are written into the store under control of the incoming timing clock and read out under control of an independent local clock. Since the number of digits an elastic store can hold is limited by the number of cells provided in the design, the delay between writing and reading is bounded by the store size. Elastic stores can be designed in several ways [6, 7]; one design has the commutator analogy of Fig. 26-4. Segments of the commutator are connected to storage cells. One brush writes into the cells; another reads out of the cells. The angular velocities of the brushes correspond to the frequencies of the writing and reading clocks. If the reading clock is slower than the writing clock, then the reading brush will slowly lag behind and eventually be overtaken. When that happens, a block of digits equal to the store size will be lost. Conversely, if reading is faster, reading will overtake writing and a block of digits will be repeated if reading is nondestructive. In both of these situations, the elastic store is said to have spilled. Synchronization can be thought of as a technique that prevents indiscriminate spilling of elastic stores.

Pulse stuffing is in one sense controlled spilling that allows the eventual recovery of the correct sequence of digits. Two conditions are necessary. First, the reading clock must be faster than the writing clock and second, insertion of extra digits must be done at prearranged times to permit proper removal.

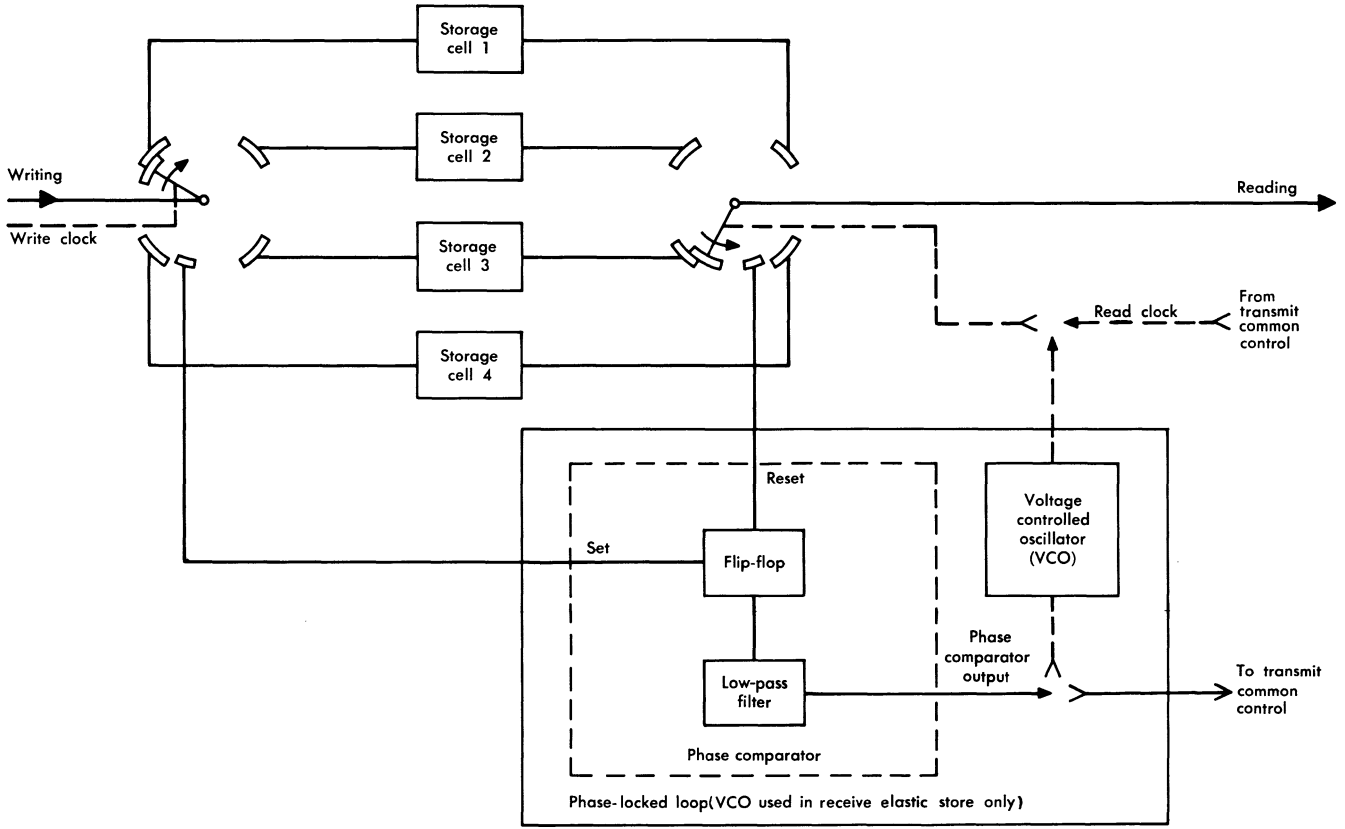


FIG. 26-4. An elastic store.

The first condition is satisfied by assignment of nominal frequencies and allowed frequency tolerances. The second condition is satisfied by periodically monitoring the delay between writing and reading. When the delay is below a given threshold, reading is caused to dwell at a cell for an additional time slot. This effectively simulates spilling of an ideal one-cell elastic store.

The measurement of the delay between writing and reading is made by a phase comparator as shown in Fig. 26-4. This can be implemented by a flip-flop that is set when the writing reaches a given cell and reset when the reading reaches the same cell. Under normal conditions when readings of each cell occur midway between writings, the flip-flop will generate a 50 per cent duty cycle square wave. As the delay of writing and reading changes, the duty cycle will change. The average output of this flip-flop is a measure of the duty cycle and thus of the store occupancy, Fig. 26-5.

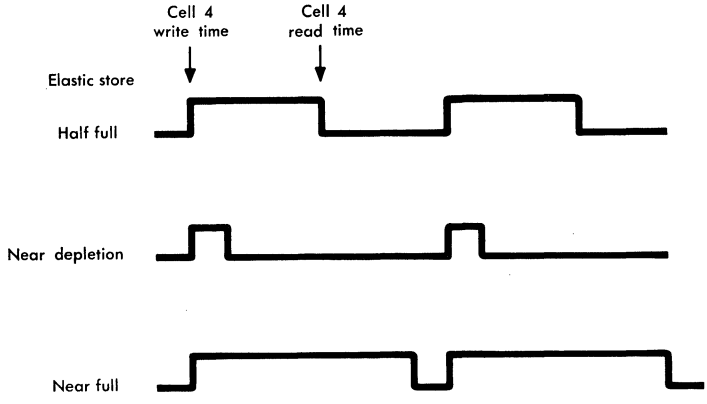


FIG. 26-5. Phase comparator output.

With pulse stuffing, even when the incoming information stream is uniformly spaced, the outgoing stream is not uniform. It will have occasional interruptions where the information pulses are interleaved with a stuffed pulse, Fig. 26-1. Thus, even after successful removing of stuffed pulses, the received information digits have jitter which must be smoothed before the digits can be processed further. Smoothing of this jitter is the function of a receiving elastic store and its phase-locked loop.

Phase-Locked Loop

In order to smooth a jittered digital stream, a read clock is needed that has a frequency equal to the average frequency of the incoming jittered clock. Since measuring the frequency of a jittered clock is difficult, the read clock can only make the best running estimate of the average input frequency. Such a read clock can be generated by a phase-locked loop.

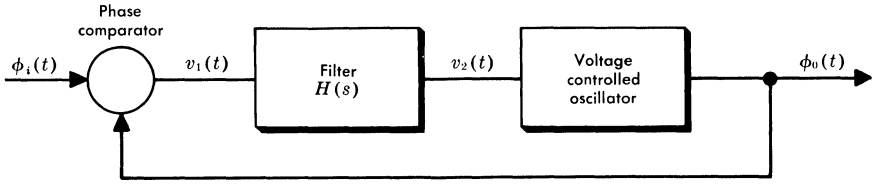


FIG. 26-6. Phase-locked loop.

A phase-locked loop, Fig. 26-6, consists of a voltage controlled oscillator (VCO) whose frequency is a function of a voltage, and a phase comparator whose output voltage is proportional to the phase difference between the incoming sinusoid and the sinusoid of the VCO. A filter is used between the phase comparator and the VCO to improve performance. The phase comparator is the same as that used in the transmitting portion to measure elastic store occupancy. The output voltage for the phase comparator is [8]

$$v_1(t) = \alpha_1 \phi_e(t) = \alpha_1[\phi_1(t) - \phi_0(t)] \qquad |\phi_e(t)| < N\pi \quad (26\text{-}1)$$

where ϕ_e is the adjusted delay between input and output clocks, and N is the number of cells in the elastic store. When the reading of a cell is exactly centered between writing times, ϕ_e is zero. When reading overtakes writing or vice versa, the store spills and N digits are gained or lost, producing a sawtooth characteristic for the phase comparator as shown in Fig. 26-7. The output phase is the in-

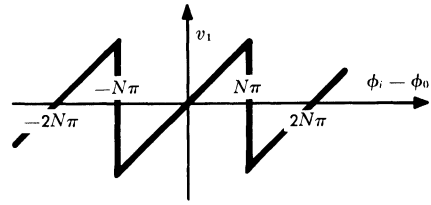


FIG. 26-7. Phase comparator characteristic.

tegral of the VCO frequency which is proportional to $v_2(t)$.

$$\Phi_0(s) = \frac{\alpha_3}{s} V_2(s)$$

$$= \frac{\alpha_3}{s} V_1(s)\, H(s) \tag{26-2}$$

where $\Phi_0(s)$, $V_1(s)$, and $V_2(s)$ are the Laplace transforms of $\phi_0(t)$, $v_1(t)$, and $v_2(t)$. Assume for the moment that the filter has gain only so that $H(s) = \alpha_2$. The total forward gain is $\alpha = \alpha_1\alpha_2\alpha_3$ which is a primary parameter of a phase-locked loop. With unity feedback, the system response is

$$\frac{\Phi_0(s)}{\Phi_1(s)} = \frac{\alpha}{s + \alpha} \tag{26-3}$$

which resembles a low-pass filter. Jitter at frequencies above α will be smoothed, while jitter at lower frequencies will still appear at the output.

If the VCO frequency with zero input voltage is not identical to the signal clock frequency, there will be a steady-state phase error. The comparator converts this error to a voltage that pulls the VCO frequency by an amount Δf to match the input frequency. From Eqs. (26-1) and (26-2) this steady state error is

$$\phi_e(t) = \frac{2\pi\Delta f}{\alpha} \tag{26-4}$$

Since ϕ_e cannot exceed $N\pi$, the phase-locked loop can only lock onto a finite range of input frequencies. From Eq. (26-4) this range is

$$|\Delta f| < \frac{N\alpha}{2} \tag{26-5}$$

Steady-state error is undesirable because it causes the loop and the elastic store to operate away from the desired half-full quiescent level. The margin against spilling will thus be reduced.

The steady-state phase error can be reduced by increasing α, but good filtering requires a small α. The trade-off situation can be improved by using a phase lag filter.

$$H(s) = \alpha_2 G(s) = \alpha_2 \frac{1 + s\tau_2}{1 + s\tau_1} \tag{26-6}$$

The loop response is now

$$\frac{\Phi_0(s)}{\Phi_1(s)} = \frac{\alpha G(s)}{s + \alpha G(s)} \tag{26-7}$$

and the time constants τ_1 and τ_2 afford additional freedom of design.

Without changing α, the output filter bandwidth can be reduced by suitable choices of τ_1 and τ_2. However, other undesirable effects appear and must be recognized in arriving at a compromise τ_1 and τ_2. The use of a filter can introduce enhancement of jitter in certain frequency bands. When many multiplexers are used in cascade, this enhancement accumulates. Another possible consequence of adding a filter is the overshoot of the VCO output frequency in response to rapid changes in input frequency. This can cause spilling of the elastic store and must be considered in the overall design.

The lock range of the loop is determined by Eq. (26-5). However, when a filter is present, Eq. (26-5) is valid only if the loop is initially locked. The range of mistuning for which lock is possible under any initial condition is smaller than that indicated by Eq. (26-5). This pull-in range depends on τ_1 and τ_2 and should exceed the expected frequency deviations of both the VCO and the incoming signal.

If pulses are stuffed one at a time, the maximum jitter is one time slot. To smooth this jitter an ideal one-cell store is needed. It has already been shown that because of constraints of phase-locked loop design, more than one cell is needed to accommodate frequency offset and possible overshoots. Digital signals arriving at a multiplexer from a long transmission line can have phase jitter, which produces instantaneous frequency deviations. The phase-locked loop can be designed either to smooth or to pass this line jitter. The first choice requires additional elastic storage capacity and small α, while the second choice requires a large α and smaller storage capacity.

Other functions of the demultiplexing operation also require their share of storage. Practical electronic circuits take finite time to operate and place minimum phase separation requirements between writing and reading clocks. Insertion and deletion of control pulses also require storage. Finally, there is waiting time jitter, which is treated in more detail in Section 26.3.

Taking all this into account, an estimated requirement on storage capacity can be made. For example, the following allocations for the M12 result in an eight-cell elastic store:

Read-write time	0.25 cells
Control pulses	0.25
Stuffing and waiting time jitter	2.00
Line jitter	5.50
Total	8.00 cells

26.3 DIGITAL MULTIPLEXER IMPAIRMENTS

Impairments arise when digital signals are processed by multiplexers. These impairments must be understood, and the multiplexers must be designed for negligible degradation to the original signal. Two significant impairments are waiting time jitter and multiplexer reframes.

Waiting Time Jitter

While it is obvious that pulse stuffing produces jitter of one time slot, there is a more subtle jitter caused by the fact that stuffing takes place only in certain allowed time slots [9]. This process can be demonstrated by a few simple examples.

A simple case occurs when the frequencies of the write and read clocks are such that stuffing takes place every third allowed time slot. This is called a stuffing ratio of 1:3. Consider the waveform of the phase comparator at the transmitting multiplex, and assume an ideal two-cell elastic store with writing immediately following the reading of a cell. The store is now full; however, it will slowly deplete because the read clock is faster. When the first opportunity for stuffing arrives, the store has depleted by only 1/3 of a cell; therefore, stuffing does not take place. At the second stuffing opportunity, the store has depleted by 2/3 of a cell, and stuffing still does not take place. At the third opportunity, the store has depleted one full cell so that stuffing can take place. After stuffing, writing immediately follows reading again and the cycle repeats. The phase comparator waveform, which is also the jitter imparted to the signal, is a sawtooth as shown in Fig. 26-8. For this case, a one-cell ideal store would be sufficient.

If the initial condition is such that writing occurs 1/6 time slot after reading, then at the second stuffing opportunity the store will be depleted by 5/6 of a cell. When depletion reaches one cell, stuffing

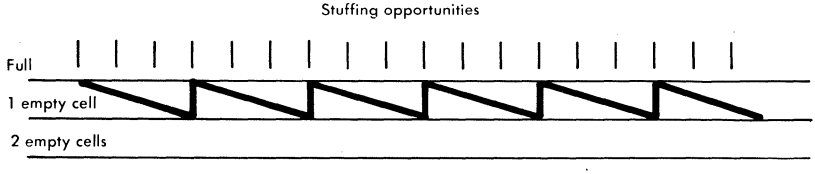


FIG. 26-8. Jitter due to stuffing at 1/3 of maximum rate.

cannot take place until the next allowed stuffing time. By that time, the depletion has reached 1-1/6 cells. The conclusion is that because of various initial conditions, at least 1-1/3 cells of elastic store are needed to stuff at one-third the maximum stuff rate.

Next, consider the case where the correct amount of stuffing is 5/14 of the maximum. The slope of the phase comparator is such that at the first stuffing opportunity, 5/14 of a cell is depleted. Initially, stuffing takes place at every third opportunity. However, since store depletion is slightly faster than 1/3 of a cell for each stuffing possibility, there is a gradual buildup of excess depletion which after 14 stuffing opportunities causes stuffing to be spaced by only two periods instead of three. This is shown in Fig. 26-9. Since stuffing has occurred five times in 14 intervals, the deficit has been made up exactly and the cycle repeats. What has been demonstrated is the phenomenon of waiting time jitter, which is a lower frequency jitter envelope superimposed on the faster jitter. It can be shown that if the nominal stuffing ratio, S, is n/m, where n and m are relatively prime, the period of the waiting time jitter is m stuffing intervals long. The ratios of two uncontrolled frequencies are, in general, irrational; therefore, the nominal stuffing ratios are irrational. Waiting time jitter is thus expected to have components down to zero frequency. The peak-to-peak waiting time jitter is S; thus a lower stuffing ratio results in lower waiting time jitter.

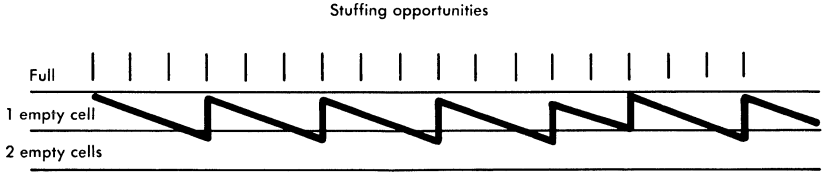


FIG. 26-9. Jitter due to stuffing at 5/14 of maximum rate.

The elastic store at the demultiplexer is designed to filter all jitter; however, since waiting time jitter extends to zero frequency, it will not be completely eliminated. For this reason pulse stuffing invariably imparts to the digital signal some low-frequency jitter which by design can be made insignificant. Usually, the low-frequency component of waiting time jitter tends to accumulate linearly with the number of tandem multiplexers. This will not cause any difficulty in digital transmission because low-frequency jitter is preserved by the stuffing operations, the phase-locked loops, and the digital repeaters.

This low-frequency jitter is passed on to the decoded baseband analog signal and, as long as the frequencies of the residual jitter are held low enough, negligible degradation to the signal results.

Multiplex Reframe

In a digital hierarchy, framing serves to hold the structure together so that each digit can be correctly demultiplexed. Since the impairment to the baseband signal due to loss of frame has no counterpart in analog transmission systems, it is important to recognize its effect in digital systems. If a channel bank loses frame, all channels served by the bank will have unintelligible signals until frame is reestablished. The requirement that a channel bank reframe within 50 milliseconds is based both on practicality and the impairment to message service, i.e., message signals and signaling. Reframe time is usually fast enough so that the impairments to message signals are inconsequential; however, some signaling may be affected. To minimize these impairments, either the reframe time must be short enough to have no effect or the incidents must be infrequent.

In a digital network the overall signal outage due to reframe anywhere in the network must be considered. Multiplexer reframes are particularly important for the following reasons:

1. Multiplexers serve a large number of circuits.
2. Multiplexer reframe causes loss of pulse stuffing information, which results in loss of frame for all individual signals.
3. Multiplexers are connected by higher speed digital lines which tend to be long and prone to protection switching and lightning hits.

In the worst case, each multiplexer waits until all of the higher speed multiplexers reframe and then incurs the maximum reframe time itself. The signal outage in this case becomes the sum of the maximum reframe times. To reduce total reframe time, two approaches can be taken. In the first, both multiplexers and channel banks are designed for a very short reframe time. In the second approach, the channel banks are allowed a long reframe time, but each multiplexer in the hierarchy is designed to aid the reframe of terminals connected to it and of lower speed multiplexers. As an example of the second approach, the M12 multiplexer, when it reframes, produces higher than the nominal pulse rate, which is equivalent to adding pulses to the bit stream. The assumed framing pulse position at a channel bank then occurs before the true framing pulse.

Normal search procedure of the channel bank will then test the next pulse positions for possible framing position. If the multiplexer reframe is fast and the number of pulses added is small, the channel banks need to search only a few positions before finding the true framing pulse.

26.4 MULTIPLEXER PERFORMANCE MONITORING

Since both inputs and outputs of a digital multiplexer are digital signals, precise monitoring of the performance of a multiplexer can be accomplished by putting another multiplexer in parallel and comparing the signals digit by digit. The second multiplexer is called a monitoring unit and is time shared by many multiplexers.

To monitor the transmitting part of the multiplexer, the higher rate digital output is demultiplexed by the monitoring unit and the resultant signal compared digit by digit with the input. To monitor the receiving part, the incoming digital signal is demultiplexed by the monitoring unit, and the resultant signal is compared with that of the multiplexer being monitored. An elastic store is used to adjust the relative delays.

The time necessary for the monitoring unit to cycle through each of the multiplexers in a bay determines the maximum time that a failure goes undetected. Failure of the monitoring unit itself is indicated by apparent failures in all the multiplexers.

REFERENCES

1. Darwin, G. P. and R. C. Prim. "Synchronization in a System of Interconnected Units," U.S. Patent 2986723, May 1961.
2. Runyon, J. P. "Reciprocal Timing of Time Division Switching Centers," U.S. Patent 3050586, August 21, 1962.
3. Pierce, J. R. "Synchronizing Digital Networks," *Bell System Tech. J.*, vol. 48 (Mar. 1969), pp. 615-636.
4. Graham, R. S. "Pulse Transmission System," U.S. Patent 3042751, 1962.
5. Bruce, R. A. "A 1.5 To 6 Megabit Digital Multiplex Employing Pulse Stuffing," *IEEE International Conference on Communications Record* (1969), pp. 34.1-34.7.
6. Kitamura, Z., K. Terada, and K. Asada. "Asynchronous Logical Delay Line for Elastic Stores," *Electronics and Communications in Japan*, vol. 50 (Nov. 1967), pp. 90-99.
7. Karnaugh, M. "Pulse Repeating System," U.S. Patent 3093815, June 11, 1963.
8. Byrne, C. J. "Properties and Design of the Phase Controlled Oscillator with a Sawtooth Comparator," *Bell System Tech. J.*, vol. 41 (Mar. 1962), pp. 559-602.
9. Witt, F. J. "Experimental 224 Mb/s Digital Multiplexer-Demultiplexer Using Pulse-Stuffing Synchronization," *Bell System Tech. J.*, vol. 44 (Nov. 1965), pp. 1843-1885.

Chapter 27

Digital Transmission Lines

The most important feature of digital transmission is the ability to reconstruct the transmitted pulse train after it has traveled through a dispersive and noisy medium. This process of reconstructing the pulse train is performed at intervals along the transmission path by regenerative repeaters. Three basic functions are performed by such repeaters: equalization, timing, and regeneration.

This functional division is depicted in Fig. 27-1, where a block diagram of a complete repeater section is shown. For purposes of illustration, it is assumed that the pulse train at the output of the

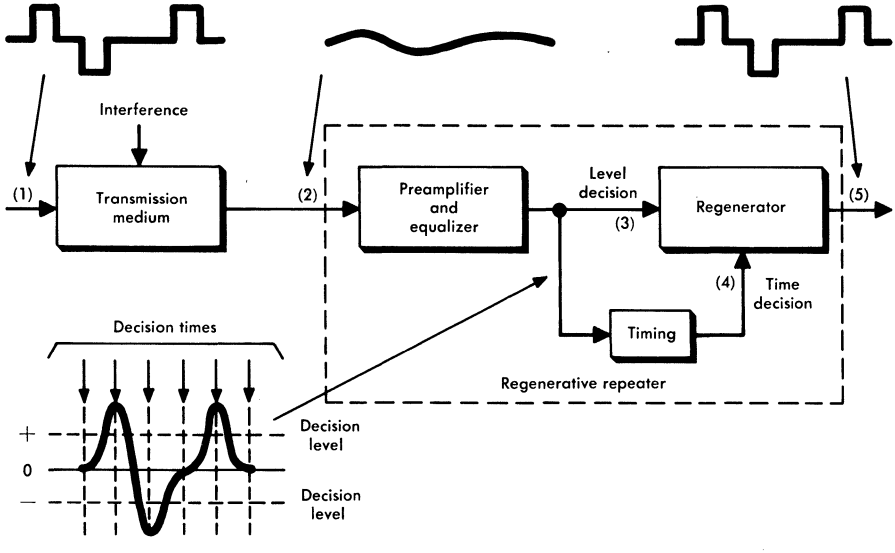


FIG. 27-1. Regenerative repeater section.

previous repeater, point 1 in the figure, consists of a series of positive pulses, negative pulses, and spaces. The pulses appear at point 2 and are distorted by the transmission characteristics of the cable as well as corrupted by additive interference.

The primary functions of the preamplifier and equalizer are to shape the pulses and to raise their level to the point where a pulse, no-pulse decision (level decision) can be made. Final reconstruction of the pulse train is accomplished by the simultaneous operations of timing and regeneration. The regenerator is enabled when the incoming pulse plus interference at point 3 exceeds the decision level (threshold) and when the timing signal at the output of the timing path, point 4, has the proper amplitude and polarity (decision time). The timing path provides a signal for the following purposes: (1) to sample the equalized pulse where the signal-to-interference ratio should be a maximum, (2) to maintain the proper pulse spacing, and (3) to turn off the regenerator at the proper time.

In the ideal situation, the reconstructed pulse train at point 5 would be an exact replica of the pulse train at point 1. In practice it departs from the ideal in three ways. First, if the interference is sufficiently large at the decision time, the wrong decision will be made and an error will occur. These errors introduce noise into the decoded analog signals. Second, if the spacing between pulses departs from its proper value, the resulting pulse position jitter introduces distortion and intermodulation noise. Finally, if the transmitted pulse shapes are not identical, the probability of making an error at the following repeater is increased.

27.1 ERROR RATE AND EYE DIAGRAMS

Performance of a regenerative repeater is measured by its error rate. The relationships between error rate and other system parameters such as signal-to-noise ratio, bandwidth, and number of transmitted levels are examined. The eye diagram, a technique for the quantitative evaluation of error rate, is introduced.

Error Performance With Gaussian Noise

To detect reliably the correct symbol at the input to the regenerator, some minimum signal-to-noise ratio is required. First, consider the case in which either positive or negative pulses (polar binary signals) of amplitude $+V_p$ *or* $-V_p$ with equal probability are received

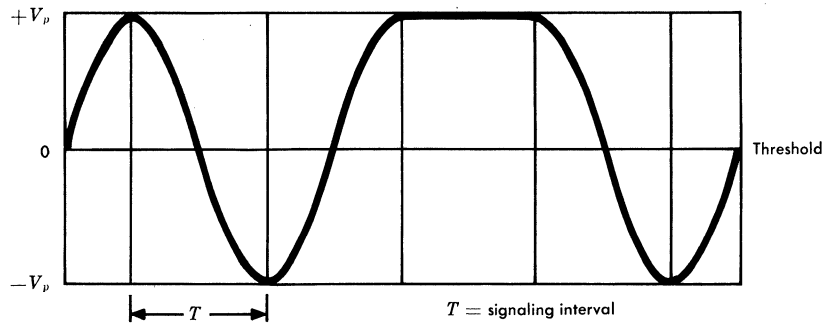


FIG. 27-2. Random polar binary pulses.

as shown in Fig. 27-2. For this situation, the decision threshold would be set at zero. At the decision time, if the signal plus noise is greater than zero, the regenerator would make the decision that a positive pulse had been transmitted, and if the signal plus noise is less than zero, a negative pulse would be regenerated. To find the probability of error, assume that the noise added to the signal is gaussian (discussed in Chap. 7). For a positive pulse an error occurs if at the decision time the noise is more negative than $-V_p$; for a negative pulse an error occurs if the noise is more positive than $+V_p$.

The error probability is

$$P_E = \frac{1}{2}\,\text{prob}\,(V_n > V_p) + \frac{1}{2}\,\text{prob}\,(V_n < -V_p)$$

$$= \frac{1}{\sqrt{2\pi}\,\sigma_n}\int_{V_p}^{\infty} e^{-V_n^2/2\sigma_n^2}\,dV_n$$

$$= \frac{1}{2}\,\text{erfc}\left(\frac{V_p}{\sqrt{2}\,\sigma_n}\right) \tag{27-1}$$

where σ_n is the rms value of the noise and erfc is the complementary error function. This relationship between probability of error for random polar binary and peak signal to rms gaussian noise (S/N) ratio is plotted in Fig. 27-3 as the curve labeled $m = 2$. In the range above 15 dB the probability of error decreases very rapidly with small increases in S/N ratio. This phenomenon is referred to as the threshold or cliff effect in digital transmission. To achieve a probability of one error in 10^{10} symbols, which is a typical repeater section requirement, an S/N ratio of approximately 16 dB is needed.

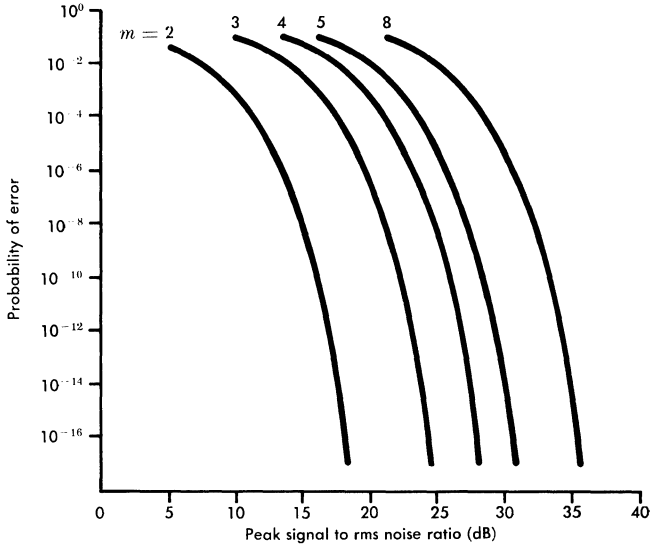


FIG. 27-3. Probability of error versus peak signal to rms gaussian noise for random m-level polar transmission.

Flexibility in the trade between S/N ratio and bandwidth (which is directly related to signaling rate in bauds) may be effected by using multilevel* transmission [1]. Consider the transmission of pulses which can take on with equal probability any of m rather than two amplitude levels. These levels are equally spaced from $+V_p$ to $-V_p$, as shown in Fig. 27-4. With this approach the information capacity of each transmitted symbol is $\log_2 m$ bits and therefore, for a constant information transmission rate, the bandwidth can be reduced by the same factor. The separation between levels now is $2V_p/(m-1)$ rather than the $2V_p$ for polar binary. Consequently, an error will be committed if the magnitude of the noise at the decision time is greater than $V_p/(m-1)$. Only at the extreme levels, $+V_p$ and $-V_p$, must the noise have the appropriate sign to cause an error. Thus the probability of error for polar m-ary is given by

$$P_E = \frac{m-1}{m} \operatorname{erfc}\left[\frac{V_p}{(m-1)\sqrt{2}\,\sigma_n}\right] \tag{27-2}$$

This expression is plotted in Fig. 27-3 for various values of m.

*Multilevel is used to mean multiple decision thresholds, which implies three or more transmitted levels.

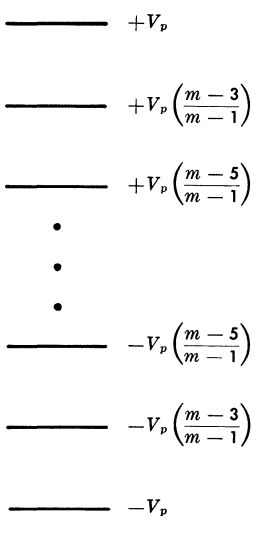


FIG. 27-4. Amplitude levels for polar m-ary.

As discussed in Chap. 8, the S/N requirement at each repeater is not a strong function of the number of repeaters or the length of the system. This characteristic is in sharp contrast to the noise accumulation laws of nonregenerative systems. The difference is illustrated in Fig. 27-5 for a digital transmission system with regeneration and for another system without regeneration, where the S/N requirement for each repeater section is given as a function of the total number of sections.

Eye Diagram

A convenient graphical technique for determining the effects of the practical degradations introduced into the pulses as they travel to the regenerator is the eye diagram [2]. This diagram of two signaling intervals duration is the result of superimposing all possible pulse sequences. Such an eye diagram is given in Fig. 27-6(a) for a ternary system in which the individual pulses at the input to the regenerator have the cosine squared shape illustrated in Fig. 27-6(b). The decision area or "eye" for each of the two decision levels is evident. In an m-level

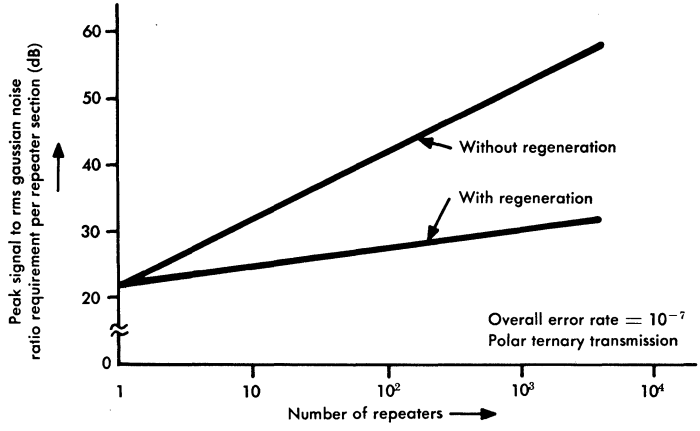
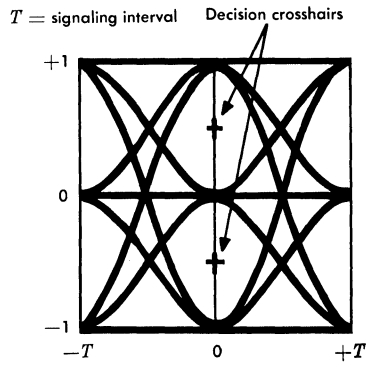


FIG. 27-5. Noise accumulation advantage of digital transmission.

system there will be $m - 1$ separate eyes. In Fig. 27-6(a), the horizontal lines labeled $+1$, 0, and -1 correspond to the ideal received amplitudes. The vertical lines, separated by the signaling interval, T, correspond to the ideal decision times.

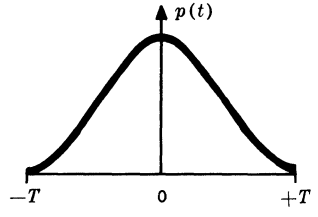
The decision making process in the regenerator can be represented by crosshairs in each eye as illustrated. The vertical hair represents the decision time, while the horizontal hair represents the decision level. To regenerate the pulse sequence without error, the eyes must be open, meaning a decision area must exist, and the decision crosshairs must be within the open area. The effect of practical degradations of the pulses is to reduce the size of the ideal eye. A measure of the margin against error is the minimum distance between the crosshair and the edges of the eye.



(a) Eye diagram for ternary reception

$$p(t) = \begin{cases} P \cos^2 (\pi/2 \cdot t/T) & |t| \leq T \\ 0 & |t| > T \end{cases}$$

$P = \{+1, 0, -1\}$



(b) Pulse shapes for eye diagram

FIG. 27-6. Eye diagram.

Error Rate With Nonideal Eyes

The various practical repeater section degradations shrink the ideal eyes. The additional S/N requirement to maintain the error rate is a function of the amount of degradation and the number of levels.

The margin against error is reduced by the various degradations added to the waveform as it travels to the regenerator and by the imperfections in the decision process itself. The first of these decreases the size of the eye, while the second moves the crosshair relative to the boundaries of the eye. It is more useful, however, to account for the latter by holding the crosshair fixed and shrinking the boundaries.

The degradations usually fall into the two categories of amplitude and timing, corresponding to vertical and horizontal displacement. To obtain the shrunken eye, the amplitude degradations such as intersymbol interference, echoes, regenerator output variations, and decision threshold uncertainties are summed. This sum is referred to as ΔA. The boundaries of the eye are then shifted vertically, as shown by the arrows in Fig. 27-7, to account for these amplitude degradations. Next, the timing degradations such as static decision time misalignment and jitter are summed. This sum is referred to as ΔT. The boundaries of the eye are then displaced horizontally, as shown in the figure. Finally, the placement of the crosshair is chosen so as to maximize the vertical distance between its center and the boundaries of the shrunken decision area. The only remaining degradation is noise, and to keep the probability of error unchanged from its value for the ideal system, the S/N ratio must be increased by

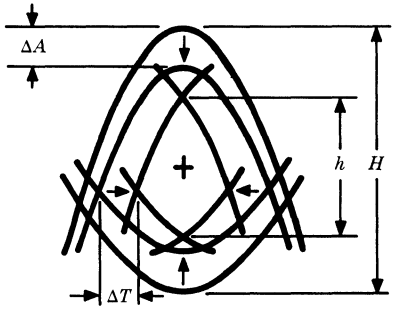


FIG. 27-7. Shrinking the eye to account for practical degradations.

$$\Delta \text{ S/N} = 20 \log \frac{H}{h} \qquad \text{dB} \tag{27-3}$$

where H and h are the vertical openings of the ideal and degraded eyes, respectively, as illustrated in the figure.

Since the degradations are usually related to the maximum value of a pulse rather than to the number of levels, it is useful to define a peak normalized eye degradation, D, as

$$D = \frac{H - h}{V_p} \tag{27-4}$$

Thus, Δ S/N for an m-level eye can be expressed as

$$\begin{aligned} \Delta \text{ S/N} &= -20 \log \left(1 - \frac{H - h}{H}\right) \\ &= -20 \log \left[1 - \frac{D}{2}(m - 1)\right] \end{aligned} \tag{27-5}$$

The relationship between S/N ratio, eye degradation, and number of levels is shown in Fig. 27-8. The figure presents the S/N requirement for random polar m-ary transmission and a 10^{-10} error probability as a function of D. The cliff phenomenon of digital transmission is again clearly evident.

For example, let $D = 0.6$. In this case, five-level transmission cannot be used since the five-level eye closes at $D = 0.5$. Four-level transmission would be quite difficult, requiring the S/N ratio to be 45.7 dB since the four-level eye closes at $D = 0.667$. Three-level transmission, however, would require only 30.1 dB.

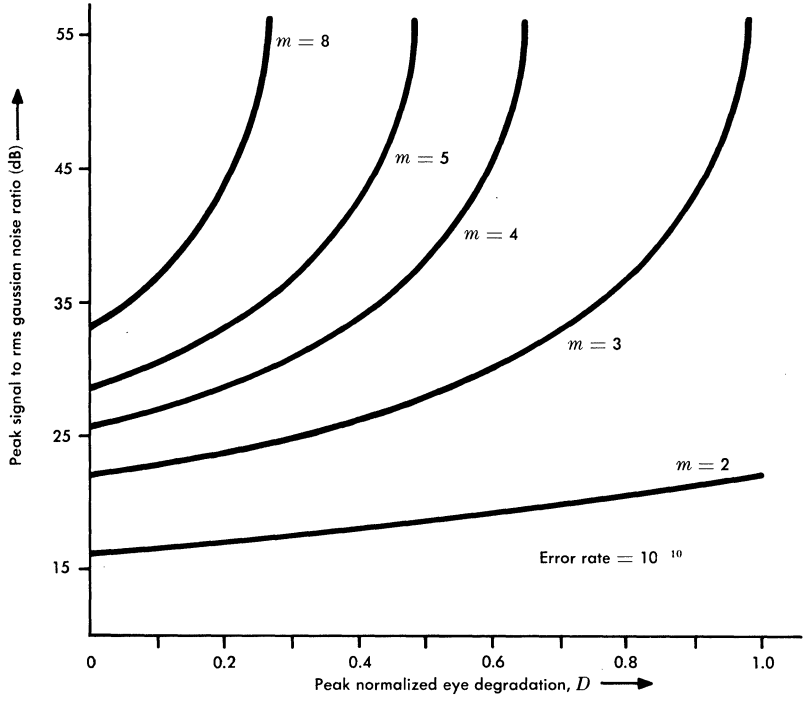
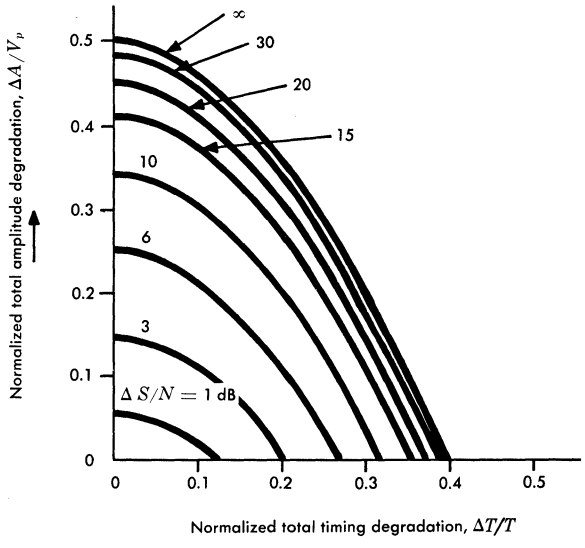
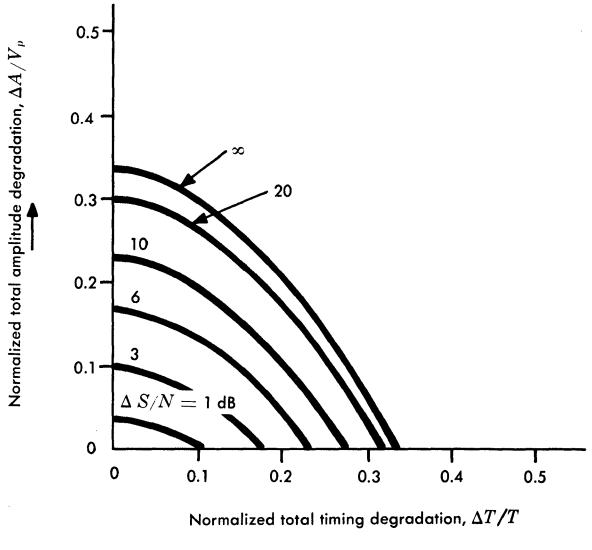


FIG. 27-8. Signal-to-noise requirements for random m-level transmission with degraded eyes.

An additional technique for evaluating degradations is the eye degradation plane. With ΔA and ΔT as coordinates, contours along which Δ S/N is a constant can be plotted. Figure 27-9 shows such contours for a polar ternary and a polar quarternary system in which the pulses at the input to the regenerator have the shape illustrated



(a) Polar ternary



(b) Polar quaternary

FIG. 27-9. Eye degradation contours.

previously in Fig. 27-6(b). The combinations of ΔA and ΔT which completely close the eye are given by the curve labeled Δ S/N $= \infty$.

27.2 CABLE MEDIA

One important medium for digital transmisson is cable, both paired and coaxial. The properties of cable media were discussed in Chap. 2. Here the effects of these properties on digital transmission are treated. The primary constants of PIC twisted pairs are plotted as a function of frequency in Fig. 27-10. It can be seen that those constants contributing to loss, R and G, increase with frequency. From Chap. 11 it is seen that crosstalk also increases with frequency. With repeater spacings of about one mile, these limitations restrict the signaling rate on PIC twisted pairs to about 10 megabauds.

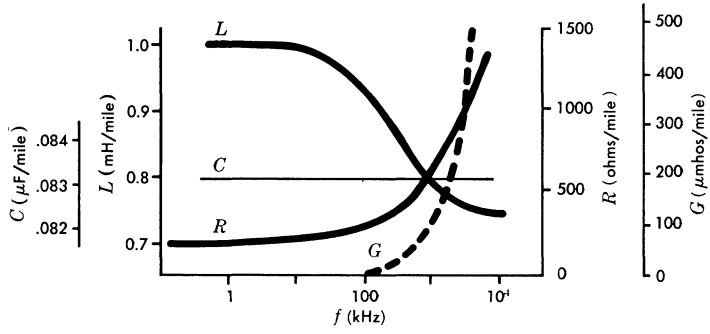


FIG. 27-10. 22-gauge PIC cable primary constants.

Coaxial cables are designed for low attenuation at carrier frequencies and have only the skin effect loss of the R term which increases as the square root of frequency. A gigabaud signaling rate with a half-mile repeater spacing is possible. At this high signaling rate, however, a significant source of digital signal degradation is the periodic structural irregularities that are introduced by the cable manufacturing process. These repetitive discontinuities produce multiple reflections of the applied signal and limit the transmission capability of coaxial lines. At higher signaling rates the polyethylene discs used for supporting the center conductor become important impedance discontinuities.

Propagation Characteristics

In Chap. 2 the secondary propagation constants have been derived from the primary constants as follows:

$$H(j\omega) = e^{-\gamma(j\omega)l} \tag{27-6}$$

where

$$\gamma(j\omega) = \alpha(\omega) + j\beta(\omega)$$

$$= j\omega\sqrt{LC}\left(1 + \frac{R}{j\omega L} + \frac{G}{j\omega C} - \frac{RG}{\omega^2 LC}\right)^{1/2}$$

Since the last three terms are much less than unity for digital transmission frequencies, a binomial approximation simplifies the above equation to

$$\gamma \approx j\omega\sqrt{LC}\left(1 - \frac{RG}{2\omega^2 LC} + \frac{R}{2j\omega L} + \frac{G}{2j\omega C}\right) \tag{27-7}$$

Skin effect can be accounted for by replacing the R term in Eq. (27-7) by a complex impedance term proportional to $\sqrt{j\omega}$. This accounts for the variation of both R and L due to skin effect [3]. The G terms can usually be neglected.

$$\gamma = A\sqrt{j\omega} + j\omega\sqrt{LC}$$

$$= \frac{A}{\sqrt{2}}\omega^{1/2} + j\frac{A}{\sqrt{2}}\omega^{1/2} + j\omega\sqrt{LC} \tag{27-8}$$

This expression is a good approximation for both coaxial cables and plastic-insulated paired cables. To find the cable response, the last term, a linear phase term contributing only to delay, is ignored. The inverse transform of Eq. (27-6), which is the impulse response of the cable, is

$$h(t) = \frac{1}{2\pi j}\int_{\sigma-j\infty}^{\sigma+j\infty} e^{-Al\sqrt{s}}\, e^{st} ds = \frac{Al}{2\sqrt{\pi t^3}} e^{-\left(\frac{A^2 l^2}{4t}\right)} \qquad t > 0 \tag{27-9}$$

This is plotted for several values of Al in Fig. 27-11.

Since Al increases with increasing cable length, the peak of a pulse in transit decreases and its base width widens as the cable lengthens. In order to successfully detect these pulses, the width must be compressed by means of pulse shaping networks at the

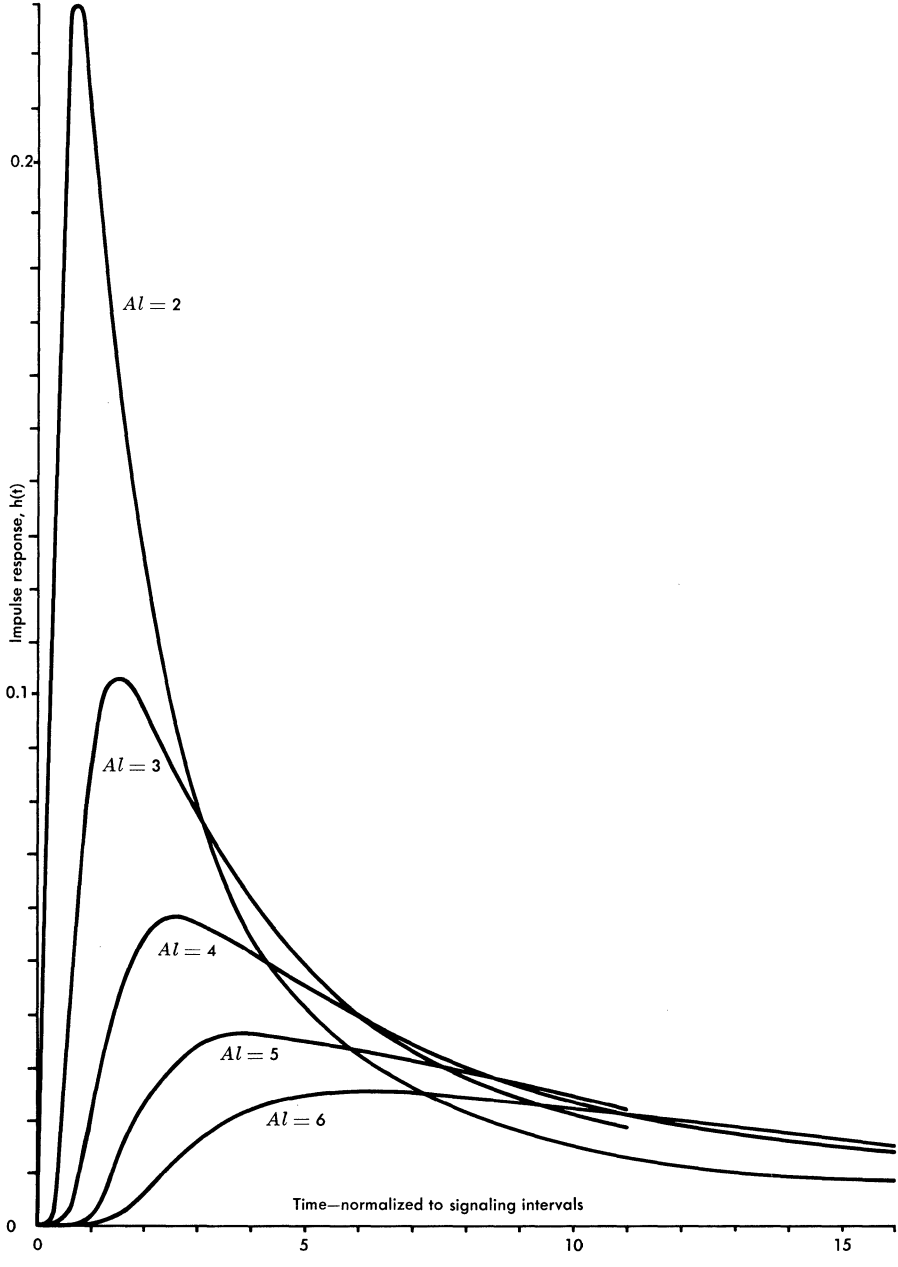


FIG. 27-11. Response of hypothetical cable to a unit impulse.

receiving end. Pulse shaping networks, also known as equalizers, not only must compensate for the length of the cable over which the pulse has traveled, but also must take into account the effects of temperature variations on cable constants.

Crosstalk

Crosstalk is an important limitation in the design of digital transmission systems for use over paired cables [4]. Chapter 11 discusses the characteristics of crosstalk in cable pairs. For systems with the two directions of transmission in the same cable sheath, near-end crosstalk (NEXT) is the major interference, and for systems where the two directions of transmission are isolated, far-end crosstalk (FEXT) is dominant. Other crosstalk paths can be controlled by frogging, repeater spacing, and other engineering rules.

Because regenerative repeaters are used in digital systems, crosstalk does not accumulate from one repeater section to the next. Averaging techniques for analog systems, discussed in Chap. 11, cannot be used in the design of digital systems. For example, as shown in Fig. 27-3, a single repeater section with a slightly lower than average S/N ratio will degrade the performance of the entire chain of repeater sections. Therefore, careful studies must be made of the distributions of crosstalk coupling loss.

The distribution of the pair-to-pair equal level coupling loss (ELCL) has been found to be log normal. Because excessively low ELCL will cause the cable to be rejected at the factory, the distribution will be truncated at the low loss end. A representative distribution of pair-to-pair ELCL for pulp cable measured at 3 MHz is shown in Fig. 27-12. Crosstalk coupling loss at frequencies other than 3 MHz can be inferred from the relation that crosstalk increases with frequency at 6 dB per octave for FEXT and 4.5 dB per octave for NEXT.

In a multipair cable, a given pair will receive crosstalk interference from many other energized pairs. The techniques used by Wilkinson [5] can be used to obtain the distribution of ELCL as the power sum of many interferers. The resulting distribution is again log normal. Figure 27-13 illustrates the distributions of NEXT and FEXT when 49 pairs of a 50-pair cable unit crosstalk into one pair. It should be noted that ELCL is not signal-to-interference ratio. Because of the effect of signal levels, NEXT is far worse than FEXT.

Figure 27-14 illustrates the relationship between crosstalk coupling loss and the number of interferers. Because of random addition, ELCL

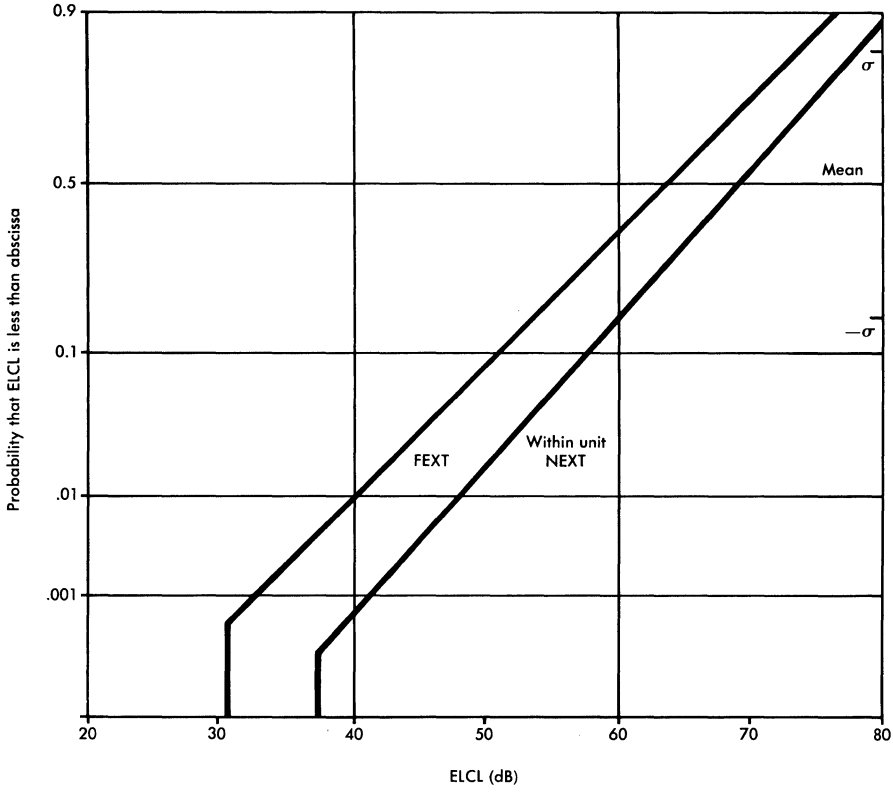


FIG. 27-12. Pair-to-pair ELCL at 3 MHz for 1000 ft of 50-pair unit, 22-gauge pulp cable.

decreases about 3 dB when the number of interferers doubles. The minimum FEXT shown represents the 99.9 per cent limit of the distribution and does not reflect the truncation of the original pair-to-pair ELCL distribution. The variation of FEXT with cable length is shown in Fig. 27-15. It verifies the assumption of random phase coupling along the length of FEXT exposure, which produces an interference that varies as the square root of the length of exposure.

Because crosstalk depends on the statistical nature of the manufacturing process, it is essential that extensive pair-to-pair coupling loss measurements be made on new designs of paired cable so that statistical distributions can be established for engineering digital transmission systems.

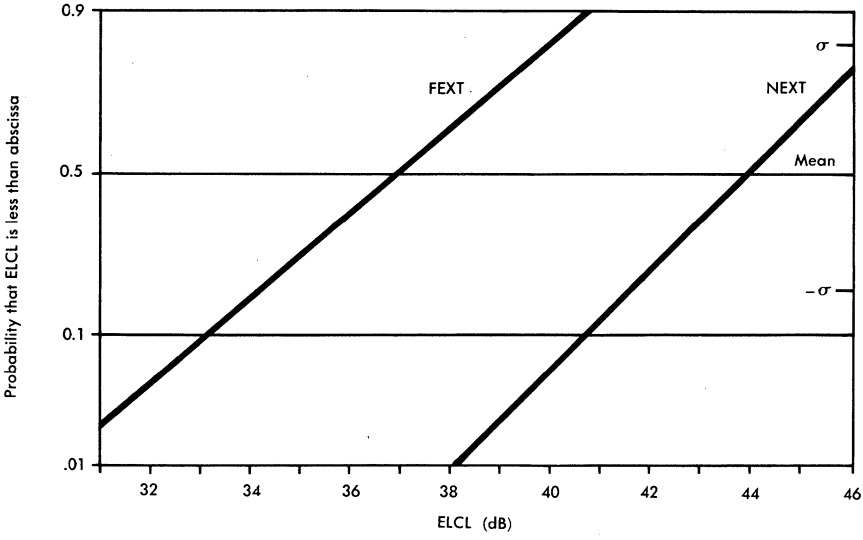


FIG. 27-13. Power sum of 49 interferers from data of Fig. 27-12.

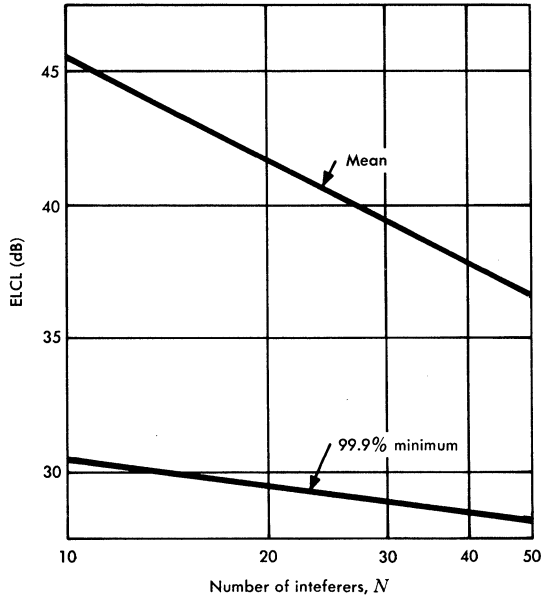


FIG. 27-14. FEXT power sum of interfering signals from data of Fig. 27-12.

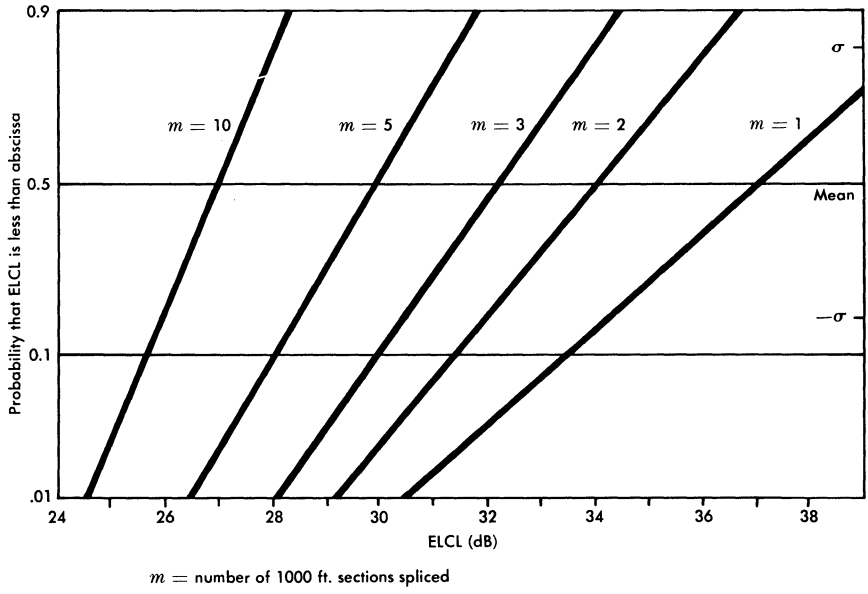


FIG. 27-15. FEXT power sum of 49 interfering signals.

Echo Interference

While crosstalk causes interference from one cable pair to another, echoes cause interference among pulses traveling in the same pair. It is a form of intersymbol interference. The periodic disc structure and manufacturing irregularities of coaxial cables have already been mentioned as sources of degradation for digital transmission. For paired cables, manufacturing techniques, field handling, cable structure, splices, and gas plugs all contribute to impedance discontinuities that give rise to echo interference. Two such causes of impedance discontinuities which affect repeater section design are gas plugs and splices.

Gas Plugs. Because most high capacity transmission cables are pressurized to prevent moisture accumulation, gas plugs are usually found in paired cables at points where they enter central offices and at maintenance boundaries. A gas plug is an airtight seal, usually formed from epoxy resin forced into the cable sheath, that allows application of pressure between plugs. The electrical effect of such

a plug is to add a lumped capacitance to each of the pairs in the cable, creating a discontinuity in the characteristic impedance. The magnitude of the discontinuity is frequently large enough to reflect 20 per cent of a transmitted pulse. Near gas plugs, repeater spacings are chosen short enough to accommodate the impairment.

Splices. One indirect effect of cable splicing is the reduction of the standard deviations of such cable properties as attenuation and crosstalk. There is a good chance that "bad" pairs in one length of cable will be spliced to "good" pairs in the next length, thus improving the worst pairs at the expense of impairing the best pairs. However, splices introduce impedance discontinuities and capacitance unbalance in the transmission paths and can thus enhance echoes and crosstalk. The enhancement depends upon the splicing procedure, and it is usually necessary to measure representative cables to characterize this effect.

Impulse Noise

Unlike thermal noise, impulse noise consists of extremely large amplitude peaks occurring in infrequent bursts against a relatively quiet background. In many practical situations, thermal noise is completely masked by the effects of impulse noise. The known sources of impulse noise include natural phenomena such as lightning and man-made sources such as switching transients occurring in telephone central offices. For digital systems operating on paired cables, high impulse noise levels are expected wherever the cable leaves a central office. Hence, preventive action is taken by shortening repeater spacings near central offices, thereby increasing the signal to impulse noise ratio.

The primary source of coupling of the impulse noise into the digital system is switched voice-frequency pairs that share a common cable sheath with the digital system. The coupling mechanism is pair-to-pair crosstalk coupling in the cable. Thus, many of the characteristics of central office impulse noise found in digital systems can be explained on the basis of the known properties of cable crosstalk. When necessary, dedicated cables or cable units are used for digital transmission near central offices, thereby eliminating much of the impulse noise.

It has been found empirically that impulse noise distributions on cables leaving the central office can be described quite accurately by

cumulative distributions of the form [6]

$$P_r(\,|\,X\,|\,>\chi) = F(\chi) = \left(\frac{a}{\chi+a}\right)^b \qquad (27\text{-}10)$$

where a and b depend on the particular office and the time of day. This equation may be used to calculate error rates where impulse noise is dominant.

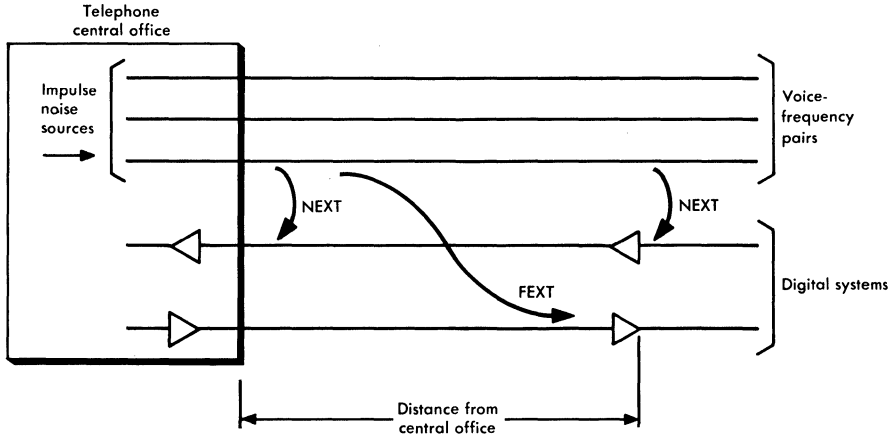


FIG. 27-16. Digital system entering and leaving a central office.

Figure 27-16 shows a typical digital line layout in a cable leaving a central office. Figure 27-17 shows qualitatively the manner in which the impulse noise power measured at the output of a repeater equalizer varies with distance from the central office. Noise coupled through near-end crosstalk paths decreases with distance because of its attenuation by the voice-frequency pairs over which it is propagated. Noise coupled through far-end crosstalk paths is also decreased by the voice-frequency pairs, but the crosstalk

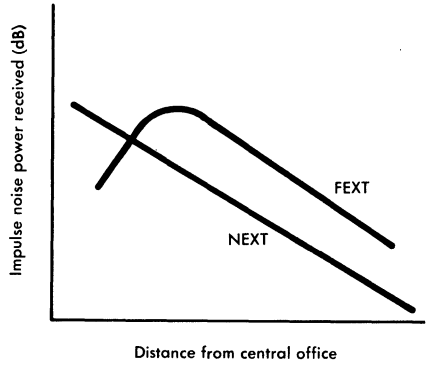


FIG. 27-17. Variation of impulse noise with distance from central office.

coupling coefficients themselves increase with length. Thus, there is a peak in the noise power at a certain distance.

Signaling Rate and Repeater Spacing

The effects of cable properties on digital systems may be summarized by obtaining the approximate repeater spacing limitations that they impose. A typical computed eye opening is shown in Fig. 27-18 [7]. Actual cable measurements were used for echo interference; computer simulations were used in determining intersymbol

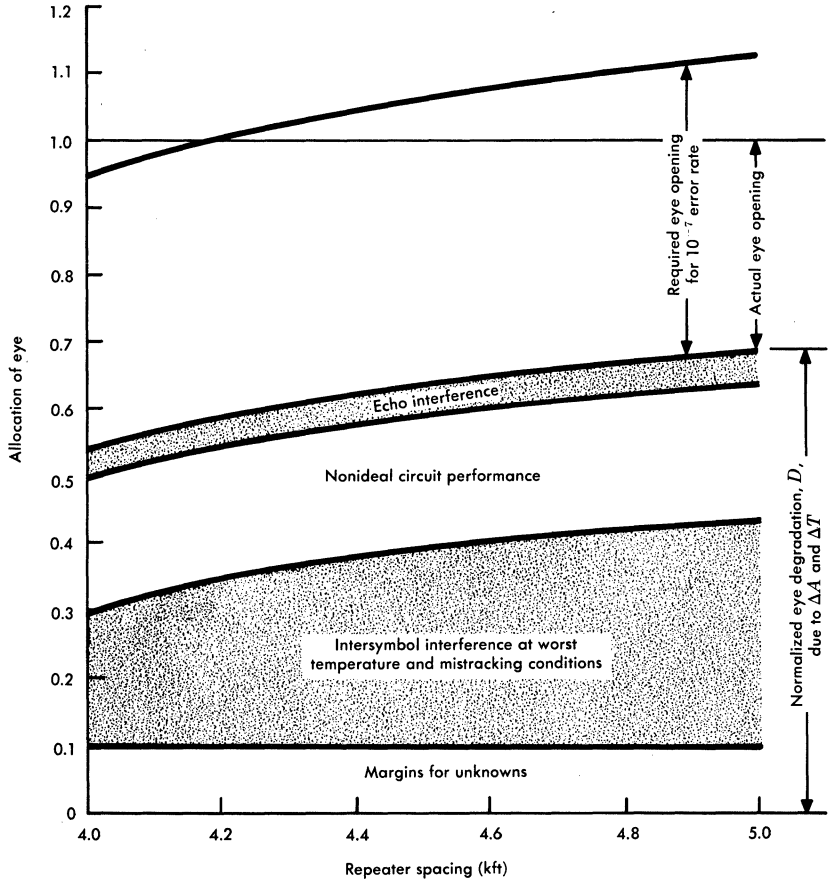


FIG. 27-18. Computed eye-opening versus repeater spacing for a two-cable spatially-frogged operation on 22-gauge pulp cable with a three-level 6.3-megabaud signal.

interference; and repeater measurements and calculations were used in determining the equivalent noise bandwidth and noise figure of the repeater. The error rate was calculated from Eq. (27-1) using the gaussian interference power given by

$$\sigma_T^2 = \sigma_n^2 + \sigma_f^2 \qquad (27\text{-}11)$$

where

σ_T^2 = mean square interference
σ_n^2 = mean square thermal noise
σ_f^2 = mean square far-end crosstalk interference

At 4100 feet, the system error rate objective is just satisfied. Impulse noise was not included in the calculations, and Fig. 27-18 refers to repeaters that are not adjacent to central offices.

If the eye closure due to impairments other than noise is assumed roughly independent of the signaling rate of the system, the propagation and crosstalk characteristics of the paired cable can be used to extrapolate repeater spacing calculations for a three-level 6.3-megabaud signal to systems operating at other signaling rates. Figure 27-19 shows approximate repeater spacing limits versus

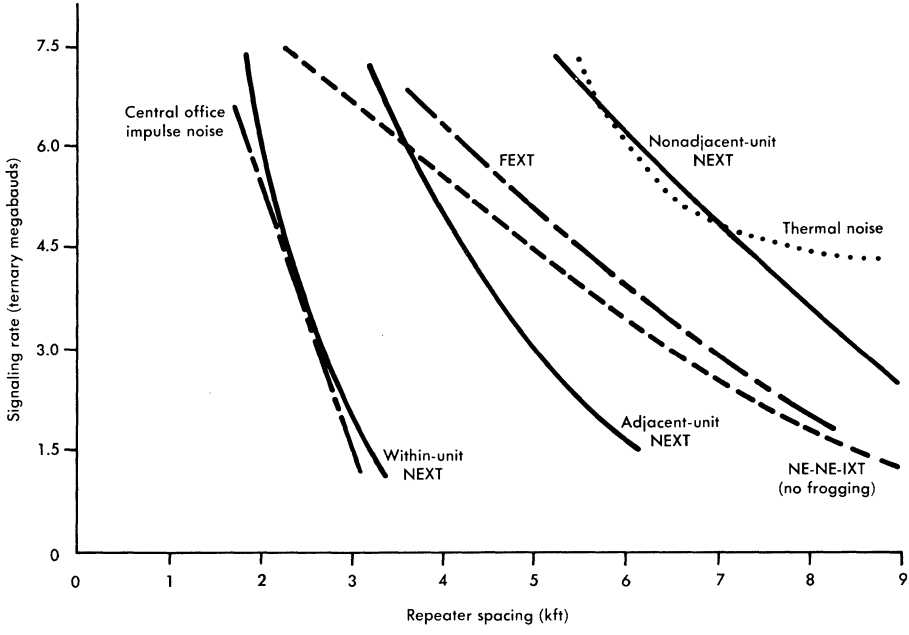


FIG. 27-19. Allowable repeater spacing when repeater is subjected to impairments shown (22-gauge pulp cable).

signaling rate due to various types of crosstalk, thermal noise, and impulse noise. To obtain reasonable repeater spacings, complex cable layouts must be used as the signaling rate is increased. For example, at 1.5 megabauds, with adjacent unit separations of opposite directions of transmission, only adjacent unit NEXT is limiting and a repeater spacing of 6000 feet is possible [8]. At 6.3 megabauds, on the other hand, both NEXT and NE-NE-IXT present severe spacing limitations. To circumvent these problems, opposite directions of transmission are placed in separate cables or in nonadjacent units of a large cable to remove NEXT paths, and spatial frogging is used to eliminate the NE-NE-IXT paths. Except in the vicinity of central offices, where impulse noise is severe, spacing is limited by a combination of FEXT and thermal noise.

Figure 27-19 also suggests that signaling rates much greater than 6 megabauds are not attractive on 22-gauge pulp-insulated cables. However, new cables are being developed that have more desirable attenuation and crosstalk properties. These may be used to increase repeater spacings with a given signaling rate or to allow higher signaling rates at repeater spacings similar to those used with pulp-insulated cables. Of course, at some signaling rate it is appropriate to abandon the paired cable approach in favor of coaxial cables [9]. For comparison, the repeater spacing curve for 0.375-inch coaxial cables is shown in Fig. 27-20, where only thermal noise contributes to signal interference.

27.3 PULSE SHAPING

After transmission through a length of cable, the high-frequency content of each pulse is severely attenuated and, as shown in Fig. 27-11, the pulse is spread over many signaling intervals or time slots. A PCM train consisting of such pulses cannot be easily regenerated, and it is the function of the equalizer in the digital repeater to compensate for the distortion introduced by the transmission medium and to shape the pulse into an acceptable form for regeneration. Generally, due to noise and other circuit considerations, it is impractical to provide equalization which narrows the pulse to the extent that it is confined to one signaling interval. The shaped pulse will therefore still have some spillover into adjacent signaling intervals. In a given signaling interval in the composite pulse stream, the summed contribution of tails and precursors arising respectively from pulses in preceding and following signaling intervals is called

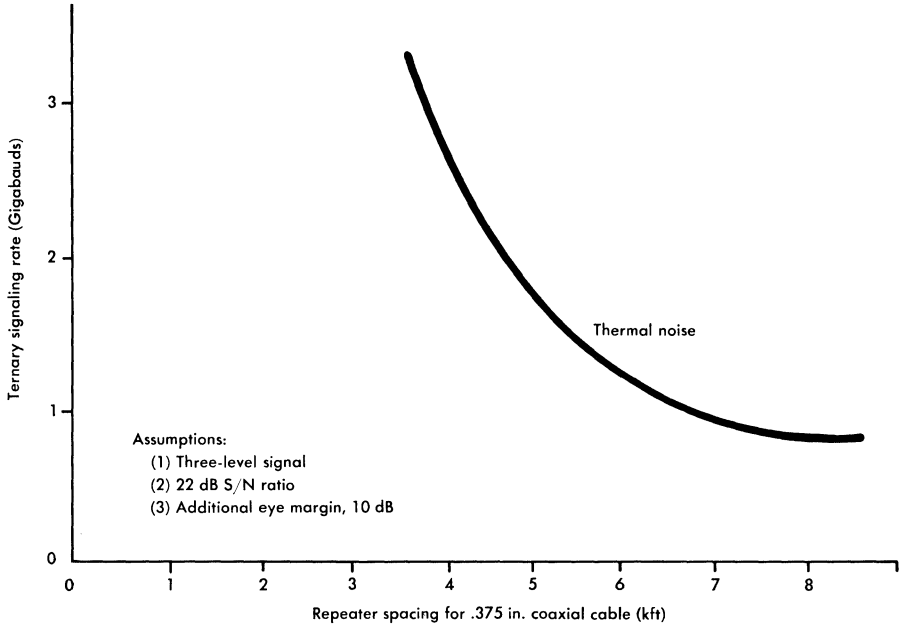


FIG. 27-20. Signaling rate versus repeater spacing.

intersymbol interference. Equalization is designed so that the maximum intersymbol interference is small enough to allow for regeneration with low error probability even when noise, timing jitter, and other system degradations are included.

Theoretical Considerations

Consider the generalized model of a single repeater section depicted in Fig. 27-21. A sequence of m-level data symbols, $\{a_n\}$, is transmitted as the pulse stream $\sum_n \alpha_n \delta(t - nT)$ through a signal shaping filter, a transmission medium, and a receiving filter. Generally, equalization refers to both the signal shaping filter and receiving filter. In practice, however, the transmitted signal shape is assumed to be fixed and equalization refers only to the receiving filter. In the repeater, the equalized pulse train is sampled by comparing the sampled value to a given set of thresholds and a decision is made by the regenerator as to which symbol in the set of m levels was transmitted at time $t = nT$. Degradations caused by intersymbol interference, noise, and timing jitter obscure the decision process so that the regenerated sequence, $\{b_n\}$, differs from the transmitted

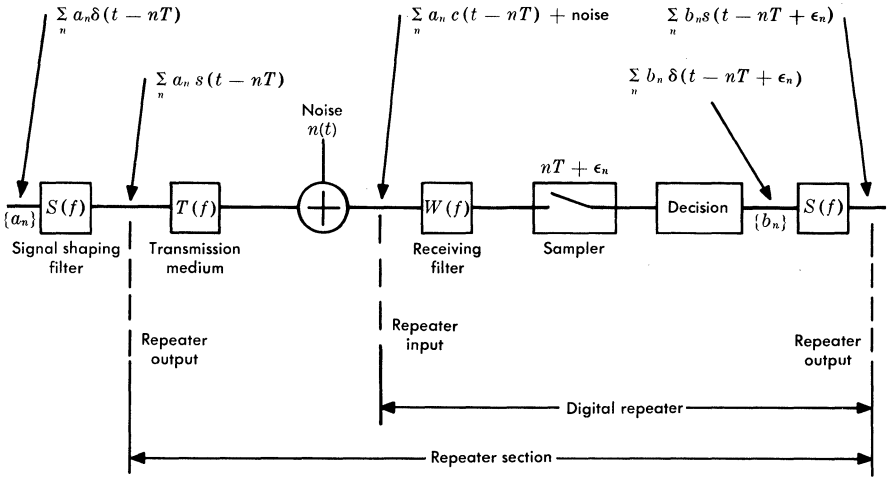


FIG. 27-21. Model of a single repeater section.

sequence, $\{a_n\}$. The optimum design of a repeater link involves the appropriate choice of transmitting and receiving filters to minimize the error rate contributed by that link in the presence of system degradations.

As an introduction, generalized channel shaping that results in zero intersymbol interference at the nominal sampling times is considered. Subsequently, the influence of noise and timing jitter upon choice of shaping is discussed.

Characteristics for Zero Intersymbol Interference. Consider the channel frequency characteristic given by $R(f) = S(f)T(f)W(f)$ in Fig. 27-21. The corresponding time response function, $r(t)$, is said to belong to the class of Nyquist I signals [10] if $r(0) = r_0$ and $r(kT) = 0$ for all integer k not equal to 0. Such a pulse shape results in a zero intersymbol interference contribution at the nominal sampling times. Hence, the amplitude value of the signal at the nominal sampling time in a given signaling interval is due entirely to the message symbol transmitted in that interval.

To determine the generalized frequency characteristics corresponding to Nyquist I signal shapes [11], consider the pulse, $r(t)$, given by the inverse Fourier transform of $R(f)$.

$$r(t) = \int_{-\infty}^{\infty} R(f) e^{j2\pi ft} df \qquad (27\text{-}12)$$

At the sampling times, $t = kT$, Eq. (27-12) may be written as

$$r(kT) = r_k = \sum_{n=-\infty}^{\infty} \int_{(2n-1)/2T}^{(2n+1)/2T} R(f)\, e^{j2\pi fkT} df$$

$$= \int_{-\frac{1}{2T}}^{\frac{1}{2T}} \left[\sum_{n=-\infty}^{\infty} R\left(u + \frac{n}{T}\right) \right] e^{j2\pi ukT} du \qquad (27\text{-}13)$$

where the infinite frequency interval is divided into subintervals of width $1/T$ and where the last equation is obtained by substituting $(u + n/T)$ for f.

Note that r_kT is the k^{th} coefficient of an exponential Fourier series expansion of

$$\sum_{n=-\infty}^{\infty} R\left(u + \frac{n}{T}\right)$$

in the interval $|u| \leq 1/2T$. Hence, imposing the Nyquist I constraint on r_k yields the requirement in the frequency domain given as

$$\sum_{n=-\infty}^{\infty} R\left(u + \frac{n}{T}\right) = T \sum_{n=-\infty}^{\infty} r_k e^{-j2\pi ukT} = r_0 T \qquad (27\text{-}14)$$

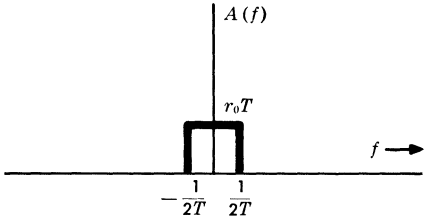
If $A(f)$ and $\alpha(f)$ are defined respectively to be the amplitude and phase of $R(f)$, Eq. (27-14) may be written in the form

$$\sum_{n=-\infty}^{\infty} A\left(u + \frac{n}{T}\right) \cos \alpha\left(u + \frac{n}{T}\right) = r_0 T$$

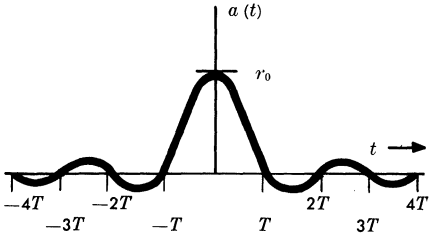
$$\sum_{n=-\infty}^{\infty} A\left(u + \frac{n}{T}\right) \sin \alpha\left(u + \frac{n}{T}\right) = 0 \qquad (27\text{-}15)$$

Equation (27-15) clearly indicates that there are an infinite number of channel shapes, not necessarily bandlimited, which satisfy the Nyquist I constraints and that there are possible trade-offs between amplitude and phase shaping. The following discussion is confined to a more familiar class of Nyquist I channel shapes where the phase, α, is linear with frequency and where the band is limited to the interval $|f| \leq 1/T$. Under these restrictions, the Nyquist equations reduce to

$$A(u) + A\left(u - \frac{1}{T}\right) = r_0 T \qquad 0 \leq u \leq \frac{1}{T} \qquad (27\text{-}16)$$



(a) Channel shape



(b) Time response

FIG. 27-22. Minimum bandwidth Nyquist I channel.

All possible channel shapes pass through $r_0T/2$ at $f = 1/2T$, and there is symmetry about this frequency.

The minimum bandwidth channel which satisfies the Nyquist I constraint of Eq. (27-16) is the flat channel extending across the interval $|f| \leq 1/2T$ [Fig. 27-22(a)]. The corresponding time response given by

$$a(t) = r_0 \frac{\sin \frac{\pi}{T} t}{\frac{\pi}{T} t}$$

is shown in Fig. 27-22(b). As expected, the time function has zeros at the desired time points. However, the strong ripples in the time response result in an eye pattern with zero width. Since some deviation from the ideal sampling time must be expected, wider bandwidth channels are required for useful digital transmission.

One particular class of useful channel shapes that satisfies the Nyquist I requirements is the family of cosine roll-off channels given by

$$R(f) = r_0T \qquad 0 \leq |f| \leq (1-b)\frac{1}{2T}$$

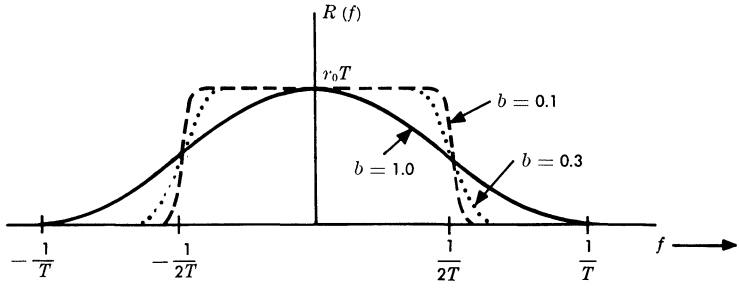
$$R(f) = \frac{r_0T}{2}\left[1 - \sin\left(\frac{T}{2b} 2\pi |f| - \frac{\pi}{2b}\right)\right] \qquad (1-b)\frac{1}{2T} \leq |f| \leq (1+b)\frac{1}{2T}$$

$$R(f) = 0 \qquad (1+b)\frac{1}{2T} < |f|$$

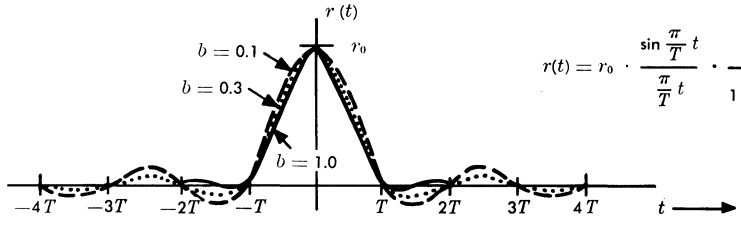
(27-17)

where b is chosen between 0 and 1. The frequency spectra corresponding to various values of b are shown in Fig. 27-23(a) and the corresponding time responses in Fig. 27-23(b).

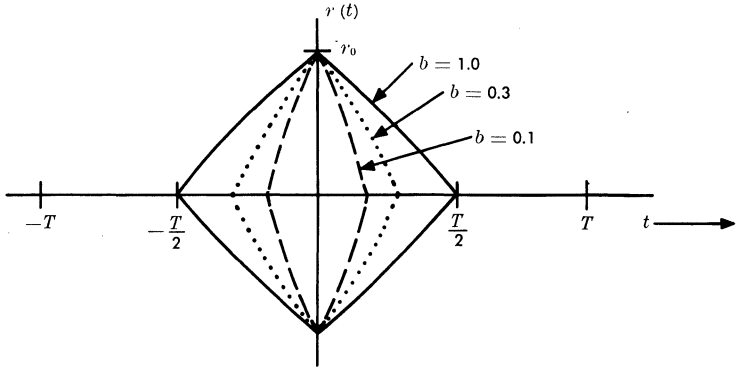
From the signal shapes shown in Fig. 27-23, it may be seen that as b increases (wider bandwidth channel), the ripples in the time response decrease. The worst-case eye diagrams corresponding to



(a) Channel shapes



(b) Time responses



(c) Worst-case eye diagrams

FIG. 27-23. Cosine roll-off Nyquist I channel.

these channel shapes are shown in Fig. 27-23(c). The eyes become wider as the channel bandwidth is increased and thus allow for greater tolerance to sampling jitter.

The preceding discussion has shown the advantages of wider band channel shaping to limit the increasing intersymbol interference with sampling offset. Wider channel bandwidths, however, require that the equalizer gain be large at high frequencies to compensate for the increase of cable loss with frequency. This leads to an increase in the noise detected at the input to the sampler, and thus degrades the system error performance. The curves of Fig. 27-24 illustrate this point for the family of cosine roll-off channels. For a constant peak received equalized pulse, the increase in noise relative to the flat channel shape is plotted as a function of b. Note that the detected noise relative to the flat channel at first decreases as b increases. This is because less gain is required at frequencies just

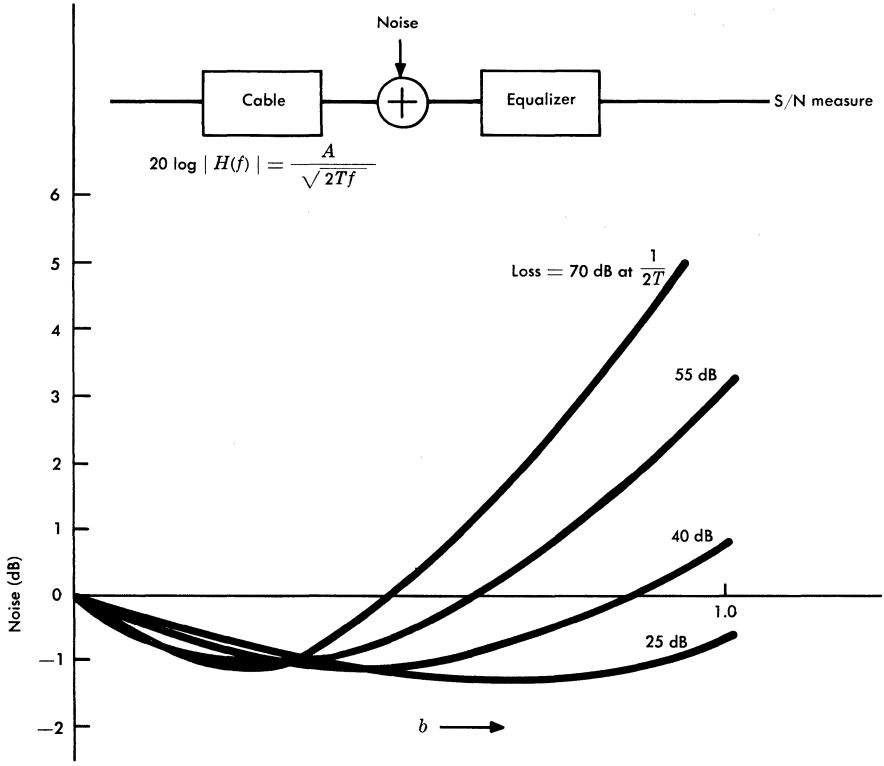


FIG. 27-24. Noise in cosine roll-off channel relative to flat channel.

below the frequency $1/2T$ as the channel shape begins to deviate from the flat shape. The result is a corresponding decrease in noise, which overcomes the noise increase due to the wider bandwidth. Eventually, however, a point is reached where the noise begins to increase rapidly with increases in b. The region of rapid noise increase with bandwidth is generally the operating region of digital systems. The choice of channel bandwidth in this region represents a compromise between noise considerations requiring smaller channel bandwidths and jitter considerations requiring wider channel bandwidths.

Analytical Methods for Optimization. Nyquist I signal shaping is not necessarily optimum when noise performance, timing jitter, average and peak signal energy constraints, and other system factors are considered. The more general problem, taking the above system factors into account, is to design the jointly optimum transmitting and receiving filters whose resulting channel shape leads to minimum error probability in the regenerator decision process. Analytical solutions have been found for certain situations [12, 13, 14, 15]. The resulting optimum channel shape is generally not Nyquist I, but the deviation from Nyquist I is not great for cases of interest.

Practical Considerations

These analytical methods for the optimization of equalizers have various shortcomings in practical design situations. One is that the equalizer amplitude and phase characteristics derived by analytical methods must be approximated by the transfer function of a physical network of limited complexity. When the equalizer is thus approximated, there is no assurance that the limited degrees of freedom available in the physical network are fully utilized. Another shortcoming is that analytical methods of optimization become intractable when practical impairments are taken into consideration. For example, normal variations in repeater spacing and cable temperature lead to significant pulse shape degradations which require compensation. Since only certain of the equalizer singularities can be varied to change the equalizer shape, only partial compensation is achieved. A compromise equalizer design over the given length and temperature range does not lend itself to analytical methods.

An approach is needed that results in a physically realizable equalizer network that achieves minimum error probability under practical situations. One such approach makes use of computers to take into

account (1) the unequalized pulse shape at various cable lengths and temperatures, (2) noise and crosstalk, (3) vertical eye degradation of the regenerator, and (4) sampling offset and jitter of the timing circuit [16]. By use of an optimization strategy [17], the equalizer parameters are adjusted until the error probability based on the computed equalized pulse response is minimized.

Equalizer Example. Equalization of the three-level 6.3-megabaud signal over 6000 feet of PIC cable is accomplished by applying the received pulse through a fixed section and then an adaptive section of an equalizer, as shown in Fig. 27-25. The fixed section of the equalizer consists of two or three cascaded constant resistance bridged-T networks. The purpose of the fixed section is to provide gross compensation of the cable characteristic.

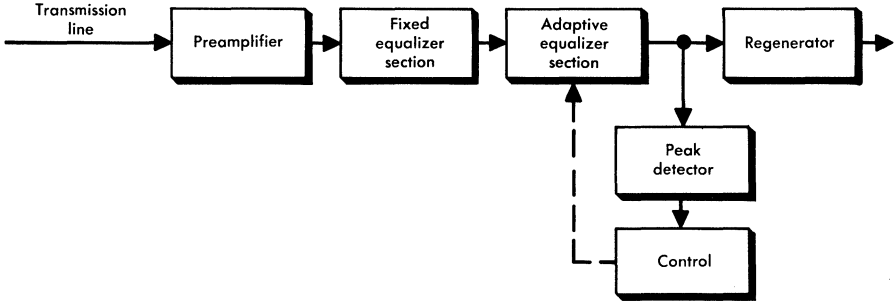


FIG. 27-25. Block diagram of T2 repeatered line equalization.

The adaptive section, shown in Fig. 27-26, has a transfer function characterized by the positions of two real poles plus gain. The variable resistors adjust the output pulse peak to a constant reference. The adaptive section is meant to compensate for cable temperature variations and for variations in cable length in a given range. Such variations translate into level changes in the received pulse peak. The adaptive equalizer is designed to insert a frequency-shaped loss, corresponding to an effective addition or subtraction of an appropriate cable length, which keeps the overall channel characteristics approximately constant over the total range of variation. In effect, the adaptive section builds out the channel characteristics; thus it is also called the automatic line build-out (ALBO). Since the ALBO adjustment is peak dependent, any changes in the received pulse level due to effects other than cable variations (such as flat gain variations) can cause misequalization of the received pulse.

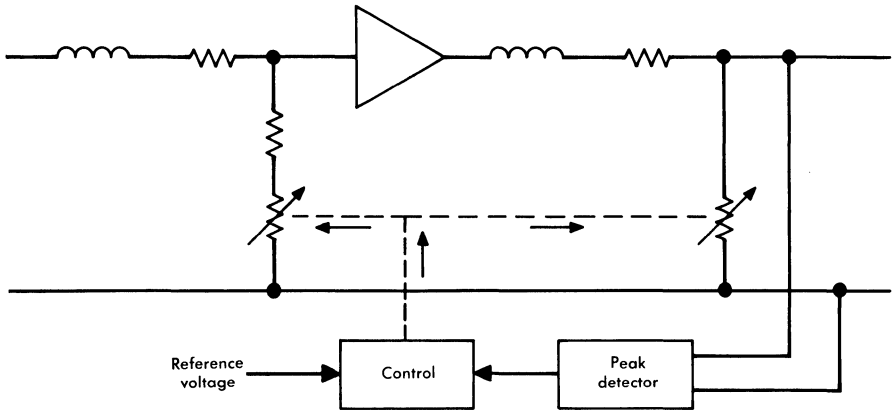


FIG. 27-26. Schematic diagram of an adaptive equalizer.

Low-Frequency Cutoff

Low-frequency cutoffs arise from transformer or capacitor coupling used in the repeater. This a-c coupling is desirable for a number of reasons; e.g., it permits the repeater to be powered by direct current carried on the signal leads, and it isolates the repeater from low-frequency noise on the line. The effect of a single low-frequency cutoff on an isolated rectangular pulse is baseline wander.

One technique for dealing with this is to restrict allowable pulse patterns by use of coding or scrambling. Another technique is to use quantized feedback together with d-c restoration [18, 19]. Quantized feedback is an arrangement whereby the low-frequency components removed from the signal by the coupling networks are replaced by the low-frequency components in the regenerator output. Its use is illustrated in Fig. 27-27. Compensation for a single pulse is indicated where the low-frequency cutoff is characterized by a single pole. The coupling network and quantized feedback network responses are matched so that $C(s) + Q(s) = 1$, and the pulse tail is therefore canceled. Clearly, to maintain matched conditions requires precise control of both the regenerator output pulse area and the gain and pole position of the quantized feedback filter, $Q(s)$. Also, repeater errors can cause improper feedback leading to error enhancement. To aid the operation of quantized feedback, d-c restoration can be employed to establish the baseline at times when the d-c level is known, e.g., when a known symbol is inserted periodically. The baseline can then be set at the known symbol to bring the direct

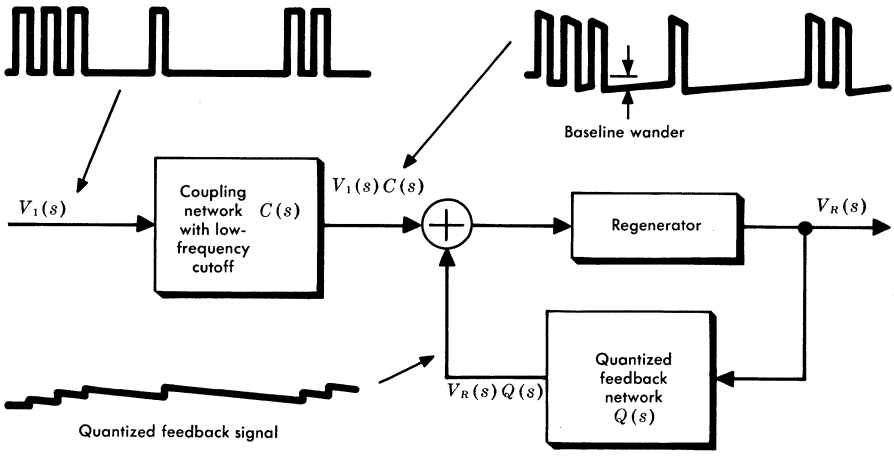


FIG. 27-27. Quantized feedback.

current to the correct level. With this arrangement, quantized feedback serves to correct for baseline variations within the symbol block and the d-c restoration serves to correct for the residual long-term baseline wander.

27.4 TIMING

After equalization, the pulse train in the repeater is in a form suitable for regeneration. The timing circuits in the repeater control the regeneration process by providing a clock signal to (1) sample the equalized pulse train near the center of the eye, (2) maintain proper pulse spacing at the regenerator output, and (3) assure correct pulse width.

Timing is extracted from the pulse train after equalization and amplification. It is important to note that in general no discrete signaling rate component is present in the spectrum of the transmitted signal. This requires that timing information be extracted from the equalized pulse train by nonlinear means.

A generalized block diagram of the repeater timing path is shown in Fig. 27-28. The equalized pulse train first undergoes nonlinear processing (rectification and clipping) which introduces a discrete component at the signaling rate. The timing extractor, which is a high Q circuit tuned to the timing frequency, extracts the desired sinusoidal component. The timing component is then amplified and

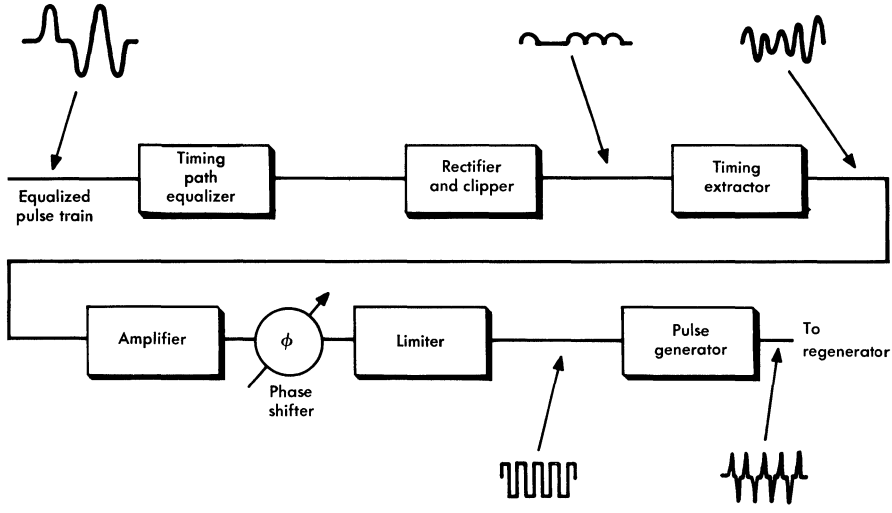


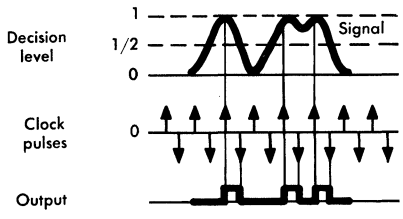
FIG. 27-28. Repeater timing path model.

limited to produce an approximate square wave at the signaling rate. The resulting signal controls a clock-pulse generator which yields narrow positive and negative clock pulses at the zero crossings of the square wave. A phase shifter in the timing path adjusts the phase of the timing pulse so that it occurs at the maximum eye opening. The clock signal obtained in this fashion is termed forward acting timing and the digital repeater is said to be self-timed [4].

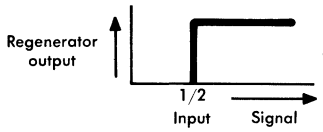
The narrow clock pulses generated at the positive-going zero crossings of the timing wave are used to gate the incoming equalized pulse train to the regenerator. Similarly, the clock pulses generated at the negative-going zero crossings are used to turn off the regenerator, thus controlling the width of the regenerated pulses. This action, known as complete timing with pulse width control, is depicted in Fig. 27-29.

Sources of Timing Jitter

The random phase modulation, or timing jitter, introduced at each repeater can be shown to accumulate in a repeater chain and may lead to crosstalk and distortion in the reconstructed analog signal. The sources of timing jitter may be classified as systematic or nonsystematic according to whether or not they are related to pulse pattern. Systematic jitter sources yield effects which degrade the



(a) Complete retiming with pulse width control



(b) Output versus input characteristic for ideal regeneration

FIG. 27-29. Timing methods.

pulse train in the same way at all repeaters in a repeater chain. Examples of such sources include intersymbol interference, finite pulse width, and clock threshold offsets [20]. Nonsystematic jitter sources such as mistuning and crosstalk result from timing degradations which are random from repeater to repeater. Thermal and impulse noise are not serious contributors to timing jitter because, if the total noise is small enough to permit a low regenerator error rate, the narrowband timing wave extractor will be affected even less [21]. In a long repeater chain the total accumulated jitter is dominated by components produced by systematic sources.

Finite Pulse Width and Pattern Effects. When the pulses exciting the high Q tuned circuit of Fig. 27-28 are not impulses or 50 per cent duty cycle pulses, the zero crossings at the output of the tuned circuit are perturbed from their nominal positions [22]. This is because the pulses driving the tuned circuit are not zero at the points when the timing wave goes through zero. These deviations depend on the pulse pattern and result in amplitude-to-phase conversion.

Intersymbol Interference. Imperfect equalization due to temperature and other systematic cable and repeater variations can affect the equalized pulse shape in the same way in a large number of repeaters. The result is that the position of the pulse peak in a given time slot depends upon the surrounding pulse pattern. The shift in peak results in phase variations in the output of the tuned circuit.

Clock Threshold Offsets. In the timing path shown in Fig. 27-28, clock pulses are generated at the zero crossings of the limiter output signal. If the threshold for generating the clock pulses is not exactly at zero, then the phase of the clock signal depends on the timing wave amplitude since the slope at the zero crossing is proportional to that amplitude. This is another pulse pattern dependent effect.

Mistuning. Both static and dynamic timing deviations can result from mistuning of the timing extractor. Static phase shift, $\Delta\phi_0$, is directly proportional to Q and to fractional mistuning, $\Delta f_0/f_0$ [23]. Thus,

$$\tan \Delta\phi_0 = 2Q \frac{\Delta f_0}{f_0} \tag{27-18}$$

Mistuning also causes dynamic phase fluctuation due to the randomness of the incoming pulse train. The rms value of this jitter is proportional to the product of the square root of Q and the fractional mistuning [24]. It is closely related to the pulse density of the signal and decreases substantially as the pulse density increases.

If the assumption is made that all timing tanks are mistuned in the same direction, it can be shown that the systematic propagation of jitter in a long repeater chain is limited to approximately twice the jitter arising in a single repeater.

Crosstalk. Signal crosstalk from other digital systems can introduce phase shift in the repeater timing wave. This phase shift is generally nonsystematic because of the differences in crosstalk coupling from repeater to repeater.

Timing Jitter Accumulation

Timing jitter accumulation in a chain of digital repeaters is caused primarily by systematic sources related to the pulse pattern. Several sources contribute significantly. As a result, a simple model is more useful than a more complex, precise model which considers each jitter source separately. The model [25] for jitter accumulation employed in this approximate analysis is shown in Fig. 27-30. The digital signal is not represented since it acts only as a carrier for the timing

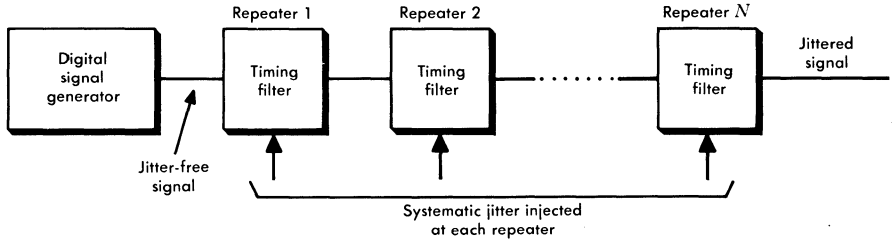


FIG. 27-30. Model for timing jitter accumulation.

wave. Delays between repeaters are eliminated since they do not affect the manner of jitter accumulation. A number of assumptions have been made to make the model tractable. These are:

1. The same jitter is injected at each repeater. This assumption is based on the fact that the same pulse pattern appears at each repeater.
2. All significant jitter sources at the input of each timing extractor can be represented by a step source for transitions between two fixed patterns and by a white noise source for random patterns.
3. Jitter adds coherently from repeater to repeater since the same pulse pattern appears at each repeater.
4. The timing extractor in each repeater is a single-tuned circuit tuned to the pulse repetition frequency, f_0.
5. For phase modulation, the tuned circuit can be represented as a low-pass filter with a single pole corresponding to the half-bandwidth of the tuned circuit.

Jitter Analysis. The consequence of the low-pass nature of the the tuned circuit to phase modulation is that low-frequency jitter is unchanged by the timing filter, while higher frequency jitter is attenuated and shifted in phase. The jitter at the end of a chain of N repeaters will be the sum of the jitter introduced by the last repeater and operated on by one tuned circuit, the jitter introduced by the next to last repeater and operated on by two tuned circuits in cascade, and so on, back to the first repeater. Since the jitter waveforms introduced in each repeater are identical, the accumulated jitter is

$$\Theta_N(s) = \sum_{n=1}^{N} \Theta(s) \left[\frac{1}{1 + s \Big/ \left(\frac{\pi f_0}{Q}\right)}\right]^n \tag{27-19}$$

where $\Theta(s)$ is the transform of the equivalent jitter source in each repeater.

The right side of Eq. (27-19) is the sum of a geometric series and is given by

$$\Theta_N(s) = \Theta(s) \frac{\pi f_0}{Qs} \left\{1 - \left[\frac{1}{1 + s \Big/ \left(\frac{\pi f_0}{Q}\right)}\right]^N\right\} \tag{27-20}$$

Neglecting short-time (high-frequency) phenomena,

$$\frac{s}{\left(\frac{\pi f_0}{Q}\right)} << 1$$

$\Theta_N(s)$ may, for large N, be approximated by

$$\Theta_N(s) \approx \Theta(s)\,\frac{\pi f_0}{Qs}\left\{1 - \exp\left[\frac{-Ns}{\left(\frac{\pi f_0}{Q}\right)}\right]\right\} \tag{27-21}$$

For a transition from fixed pulse pattern i to another fixed pulse pattern j at $t = 0$, with associated steady state phase shifts θ_i and θ_j, respectively, the injected jitter source $\Theta(s)$ corresponds to the transform of a step of height $\theta_j - \theta_i$. Hence, from Eq. (27-21), the phase transition is linear as shown in Fig. 27-31. Note that the phase variation is directly proportional to the number of repeaters, but the phase transition slope (instantaneous frequency deviation) is independent of the number of repeaters and approximately equal to $(\pi f_0/Q)\ (\theta_j - \theta_i)$. The worst-case jitter for both phase and frequency deviations is characterized by the maximum value of $(\theta_j - \theta_i)$ for all possible pairs of repetitive patterns. This value can therefore be used as a figure of merit for the regenerator. The smaller this value, the less will be the maximum phase excursion for a given length repeater chain and the less will be the frequency deviation.

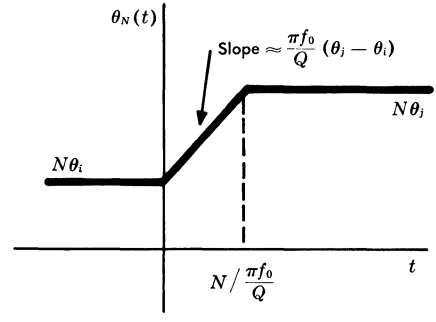


FIG. 27-31. Jitter at end of N repeaters due to transition from pattern i to pattern j.

A typical photograph of transitions between two fixed patterns obtained from measurements on a chain of repeaters is shown in Fig. 27-32. The photograph agrees quite well with the curve of Fig. 27-31. The linear increase of phase shift with the number of repeaters was also demonstrated by measurements.

Analogous to Eq. (27-20) for the case of random pulse pattern, the jitter power density at the end of a chain of N repeaters is

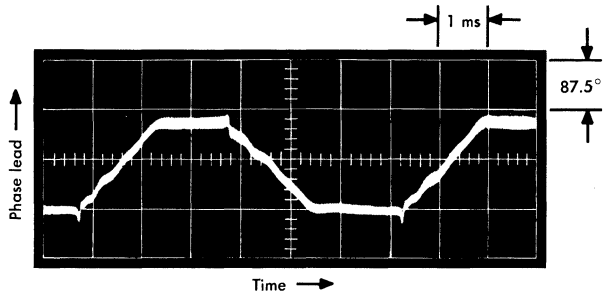


FIG. 27-32. Photograph of phase shift of timing signal due to a transition from 1/8 pattern (top) to an 8/8 pattern (bottom), with 84 repeaters; $f_0 = 1.544$ MHz and $Q \approx 100$.

given by

$$\Phi_N(f) = \left(\frac{f_0}{2Qf}\right)^2 \left| 1 - \left(\frac{1}{1 + j\frac{2Qf}{f_0}} \right)^N \right|^2 \Phi \qquad (27\text{-}22)$$

where Φ is a constant equal to the value of the flat power spectral density (two-sided) of the jitter source in each repeater. The jitter spectrum is plotted in Fig. 27-33 for a range of values of N. At very low frequencies, the jitter spectrum is flat and the power density is proportional to the square of N. For higher frequencies, the spectrum falls off as the inverse square of frequency. As a result of the spectral shaping, the bulk of the jitter energy for large N is at low frequency.

The mean-square value of jitter, $\bar{\Phi}_N$, can be found by integrating Eq. (27-22) over all frequencies. The result of the integration can be shown to be

$$\bar{\Phi}_N = \Phi \frac{\pi f_0}{Q} P(N) \qquad (27\text{-}23)$$

where $P(N)$ is given by the expression

$$P(N) = \int_{-\infty}^{\infty} \frac{1}{(2\pi f)^2} \left| 1 - \left(\frac{1}{1 + j2\pi f} \right)^N \right|^2 df$$

$$= \left\{ N - \frac{1/2\,(2N-1)\,!}{4^{(N-1)}[(N-1)\,!]^2} \right\} \qquad (27\text{-}24)$$

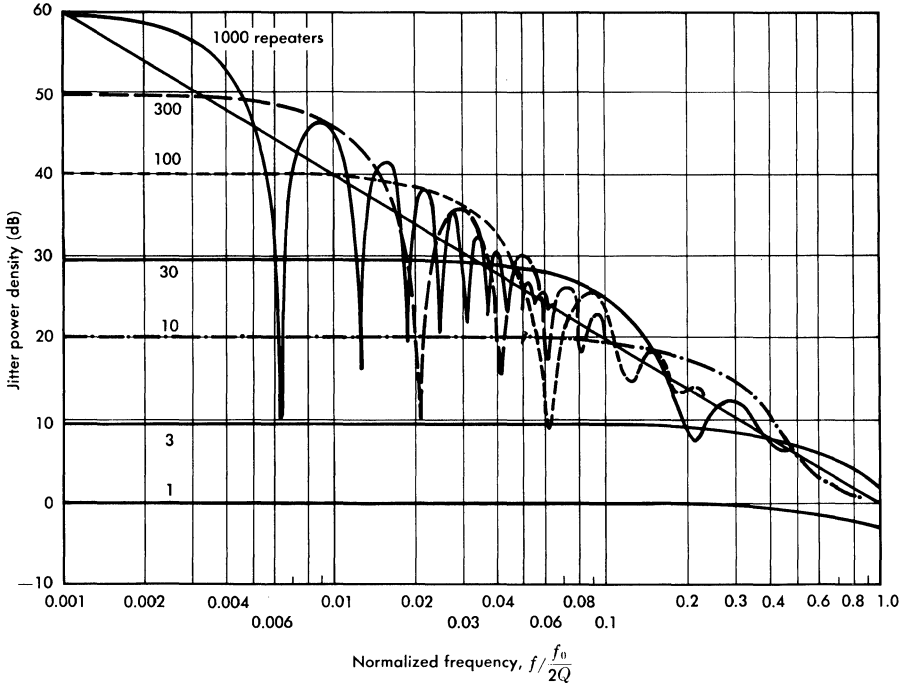


FIG. 27-33. Jitter spectrum due to a random pattern (computed from the model).

For large N, $P(N) \approx N$ so that in a long repeater chain the accumulated mean-square value of the timing jitter is given by the approximate expression

$$\overline{\Phi}_N \approx \Phi \frac{\pi f_0 N}{Q} \tag{27-25}$$

Therefore, the mean-square value of jitter in a long chain increases with N, and the rms value of jitter increases with the square root of N. Also, the jitter is proportional to the timing filter bandwidth. This confirms that high Q tuned circuits in the repeaters will reduce jitter.

Timing Extractor. Important considerations in the design of a timing extractor are (1) mistuning, (2) jitter, and (3) simplicity and low cost. The simplest circuit for timing extraction is the single-tuned LC circuit. At the output of the timing extractor shown in Fig. 27-28, the signal is a sinusoid at the pulse repetition frequency.

The Q of the tuned circuit must be large enough to provide adequate suppression to timing jitter and small enough to meet the stability requirements. One serious disadvantage of the tuned circuit, the amplitude variation with changing pulse pattern, can be overcome by the use of a phase-locked oscillator which provides constant voltage output under all pattern conditions.

Effects of Jitter. The ultimate reconstruction of the transmitted message signal in the digital receiving terminal is accomplished by converting the PCM code words to PAM pulses and then interpolating with a low-pass filter. The PAM pulses are representations of periodic samples of the original signal. At the transmitter, the time interval between PAM samples is determined by the frequency of the transmitted clock. The clock at the receiver is slaved to the frequency of the incoming digital signal. Due to the accumulation of timing jitter along the digital line, the instantaneous received frequency may differ appreciably from the transmitted frequency. Thus, since the pulse position of the PAM pulses at the receiver is controlled by the receiver clock, the sample positions will vary about a nominal spacing and give rise to a PAM pulse sequence with undesired pulse position modulation. This can lead to crosstalk and distortion in the recovered analog signal as shown by the following simplified analysis.

Consider an analog signal $v(t)$ which at the receiver is represented by the phase-modulated train of PAM impulse samples defined by

$$y(t) = \sum_{n=-\infty}^{\infty} v_n \delta(t - nT - \epsilon_n) \tag{27-26}$$

where T is the nominal spacing between samples and ϵ_n is the time deviation of the nth sample from its nominal time. The PAM pulse amplitudes are given by the sequence $\{v_n\}$. The spectrum of this impulse train is determined by taking the Fourier transform of Eq. (27-26) term by term yielding the expression

$$Y(f) = \sum_{n=-\infty}^{\infty} v_n e^{-j2\pi fnT} e^{-j2\pi f\epsilon_n} \tag{27-27}$$

When the maximum jitter amplitude is small relative to the reciprocal of the highest frequency of interest at the filter output, the approximation $e^{-j2\pi f\epsilon_n} \approx 1 - j2\pi f\epsilon_n$ is valid. If the interpolating filter has transmission, $T(f) = 0$ for $|f| > 1/2T$, the spectrum at the system output is $Y(f) \cdot T(f) = G(f)$. Under these conditions $G(f)$ reduces to approximately

$$G(f) \approx T(f)\left[V(f) - j2\pi f\int_{-\infty}^{\infty} V(x)\, E(f-x)\, dx\right] \quad (27\text{-}28)$$

where $E(f)$ is the spectrum of the jitter. This yields the original signal spectrum $V(f)$ plus a differentiated bandlimited convolution of the signal and jitter spectra. The latter gives rise to noise and distortion.

A more complete analysis of the effects of jitter would consider the finite width of the pulses in the PAM train and the spectral density and correlation function of the jitter. Based on this analysis, terminal requirements on timing jitter may be derived. An example of terminal requirements for a PCM system transmitting coded mastergroups is shown in Fig. 27-34. The curve reflects the consideration of message channel distortion and crosstalk. High-frequency jitter (> 4 kHz) must be kept small since it causes inter-

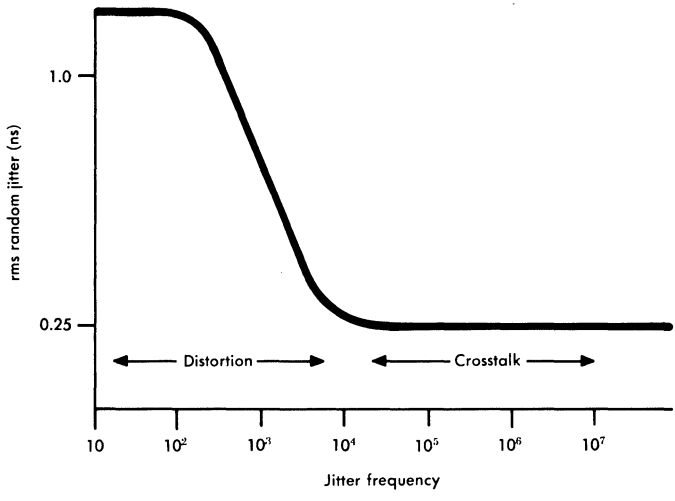


FIG. 27-34. Jitter requirement for mastergroup terminal.

channel crosstalk whereas low-frequency jitter causes only distortion within a channel.

Control of Jitter Accumulation. For a given timing path design, the jitter accumulated through a chain of N repeaters is given by Eq. (27-23). This jitter is reduced by elastic stores in conjunction with high Q phase-locked loops which may be inserted for this purpose.

27.5 LINE CODING

For transmission over a digital repeatered line, the binary information generated within a terminal must be coded into a sequence of symbols. Conceptually, the simplest transmission code is unipolar. In this format, the binary 1's and 0's are coded for transmission as presence and absence of pulses, respectively. If the pulses are independent, the power spectral density of the unipolar pulse train is [24]

$$P(f) = |G(f)|^2 \left[\frac{p(1-p)}{T} + \frac{p^2}{T^2} \sum_{n=-\infty}^{\infty} \delta\left(f - \frac{n}{T}\right) \right] \quad (27\text{-}29)$$

where $G(f)$ is the Fourier transform of the pulse shape; p is the probability of a pulse; and $1/T$ is the signaling rate. This spectrum is plotted in Fig. 27-35 for a rectangular pulse.

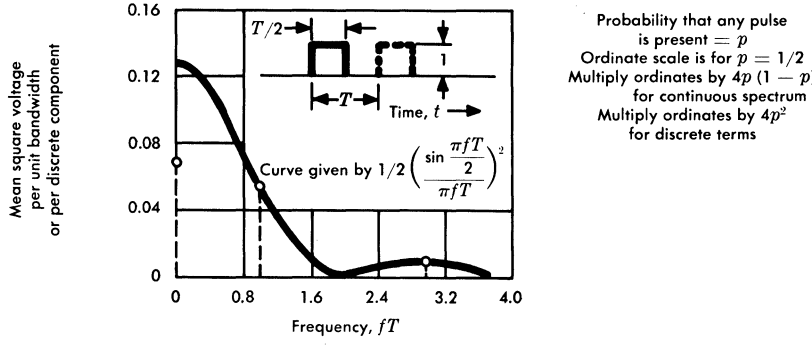


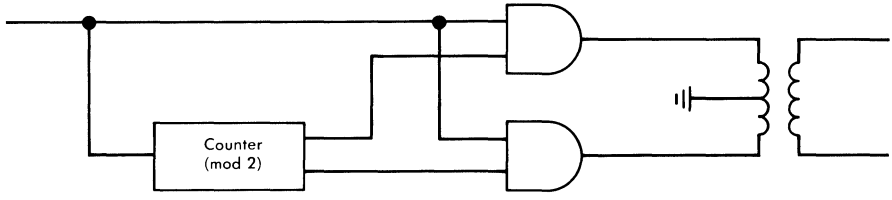
FIG. 27-35. Continuous and line spectral density components for binary on-off signaling.

There are three significant practical problems associated with this unipolar code. First, transmission of long sequences of 0's results in a reduction of the discrete timing component, which in turn yields poor jitter performance. Second, since the unipolar spectrum contains significant energy at low frequencies, the pulse stream is subject to d-c wander when a-c coupling is used. Third, in-service performance monitoring of the line error rate is impossible unless source statistics are known.

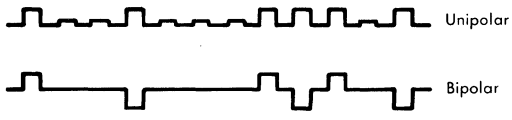
These three problems can be overcome by introducing redundancy into the coding process. Two possible techniques for introducing redundancy are (1) the transmitted signaling rate can be made greater than the input binary rate, and (2) selected forms of multilevel signals can be transmitted over the line while keeping the signaling rate equal to the input binary rate.

Bipolar Coding

Bipolar coding is a form of ternary coding where the signaling rate is equal to the input binary rate [28].* Binary 0's are coded as absence of pulses; binary 1's, however, are alternately coded as positive and negative pulses with the alternation taking place at every occurrence of a 1. The block diagram of a circuit for performing the coding is shown in Fig. 27-36, along with an example.



(a) Unipolar to bipolar converter



(b) Example of bipolar coding

FIG. 27-36. Bipolar coding.

*Strictly speaking, the binary information rate is slightly less than the signaling rate in order to meet the pulse density requirements for timing recovery.

The power spectrum of a bipolar sequence with unequal magnitudes of positive and negative pulses contains both discrete and continuous components. For balanced positive and negative pulses, however, the discrete spectrum vanishes and the continuous portion reduces to

$$P(f) = \frac{2p(1-p)}{T} \mid G(f) \mid^2 \frac{1 - \cos 2\pi fT}{1 - 2(2p-1)\cos 2\pi fT + (2p-1)^2} \tag{27-30}$$

where $G(f)$ is the Fourier transform of the positive pulse, and p is the probability of a 1 in the original binary sequence. Part of the spectrum is plotted in Fig. 27-37, normalized to $G(f) = 1$. There are nulls in the spectrum at zero frequency and at integer multiples of the signaling rate because successive pulses, independent of intervening spaces, alternate in sign. The reduction of low-frequency energy eases the d-c wander problem.

The pulse alternation property also provides a means of accomplishing in-service performance monitoring. Any isolated error, whether it deletes a pulse or creates a pulse, will cause a violation of this property.

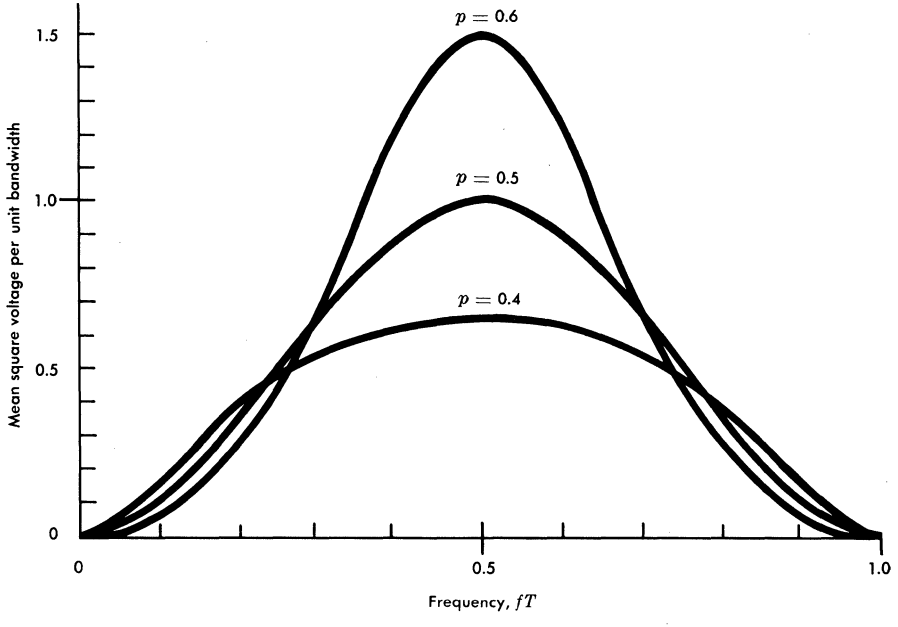


FIG. 27-37. Power spectra of bipolar coding.

In bipolar coding, however, all discrete components which could have been used for timing have been suppressed. Rectification of a bipolar train produces a unipolar train which has a discrete frequency at the signaling rate. The energy for this frequency comes from cross-modulation of frequencies located symmetrically about half the signaling rate where the original spectrum has the greatest energy. While bipolar coding solves the d-c and performance monitoring problems, it does not solve the timing problem occasioned by a long string of 0's. This problem can be minimized by restricting the terminals to produce a signal which contains a minimum density of 1's.

Another solution to the pattern density problem is to substitute a sequence of symbols having a special characteristic in place of a long string of 0's. Such codes are called bipolar with N zeros substitution (BNZS) codes.

A fundamental requirement for the substituted word is that it have an easily identified characteristic. In a bipolar stream, one such characteristic is a violation of the sign alternation property. Furthermore, the substituted words should be d-c balanced, and they should avoid confusion of the receiver by bipolar violations arising from randomly occurring line errors.

A B6ZS code which embodies the preceding considerations substitutes either $0+-0-+$ or $0-+0+-$ for blocks of six consecutive zeros, depending on the polarity of the last pulse transmitted [27]; if the last pulse is a $+$, then $0+-0-+$ is used. An example of B6ZS coding is shown in Fig. 27-38.

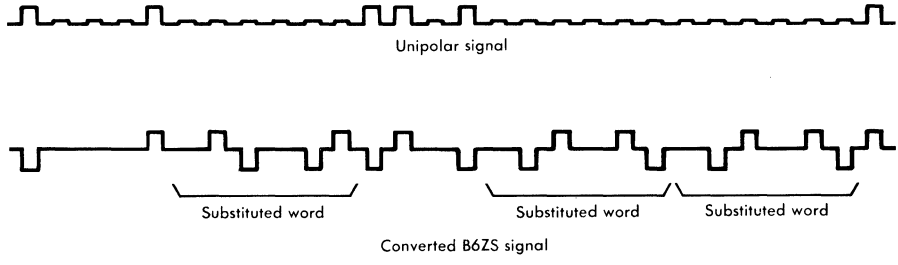


FIG. 27-38. Conversion of a unipolar signal to B6ZS.

Part of the power spectrum is plotted in Fig. 27-39 normalized to $G(f) = 1$. There are many other generalizations of bipolar codings.

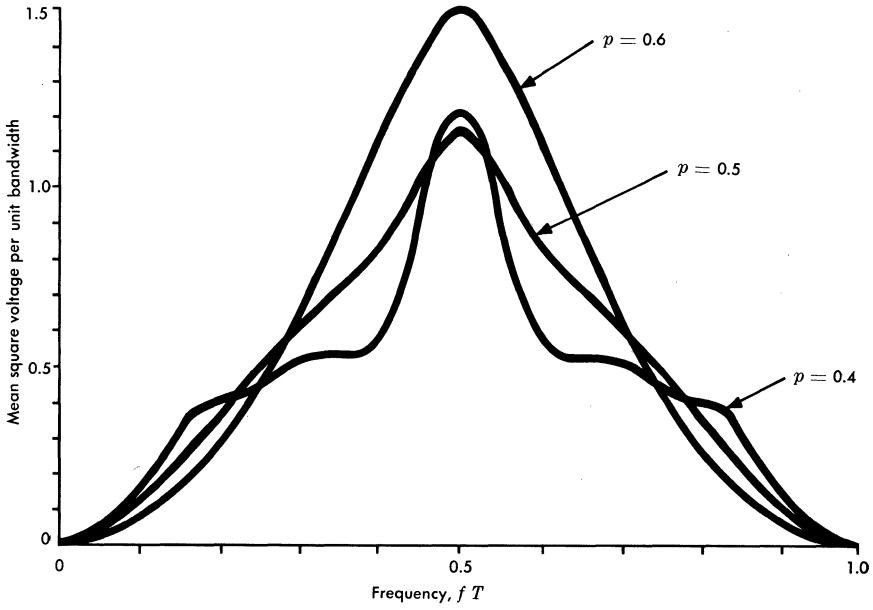


FIG. 27-39. B6ZS spectra.

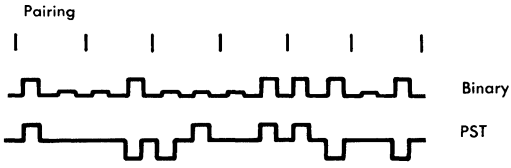
Most of these have power spectra that achieve low energy at frequencies other than zero or the signaling rate. The usefulness of such spectrum shaping is limited however.

Paired Selected Ternary Coding

So far, coding plans in which one binary symbol at a time is translated into a ternary symbol have been considered. By translating in blocks of two or more binary symbols at a time, the three practical problems of d-c wander, timing, and performance monitoring can be solved. The binary sequence to be transmitted is framed into pairs and translated into a ternary format according to the coding table given in Fig. 27-40(a). Two modes are employed, and the mode is changed after each occurrence of a zero-one or a one-zero binary pair. This is known as paired selected ternary (PST) coding [28]. An example of PST coding is given in Fig. 27-40(b). With this code, d-c wander is controlled by the alternating mode which is equivalent to the alternating polarity of bipolar coding. Pairs of 0's in the binary sequence are translated into pulses so that timing

Binary	PST	
	+Mode	−Mode
1 1	+−	+−
1 0	+0	−0
0 1	0+	0−
0 0	−+	−+

(a) PST translation table



(b) Example of PST coding

FIG. 27-40. PST coding.

information is assured. Finally, the properties of the PST sequence, of which mode alternation is one, provide the error monitoring capability.

The comparison of the power spectra of PST, bipolar, and B6ZS coding is shown in Fig. 27-41. More power is transmitted with PST coding than with B6ZS at almost all frequencies. This additional power, while desirable for timing purposes, accentuates crosstalk problems in paired cables.

Framing of the received PST sequence is essential in order to correctly recover the original binary sequence. Statistical framing can be used because the occurrence of any unused ternary pairs (++, −−, and 00) indicates an out-of-frame condition.

However, if the original binary sequence contains only 1's or only 0's, an alternating sequence of +−+− . . . is transmitted; therefore, the out-of-frame condition cannot be recognized. This problem is minimized by a modified PST code shown in Fig. 27-42. For this code, the binary sequence which may be decoded incorrectly due to lack of framing information has been changed to 101010 which is a sequence less likely to be transmitted than the all 1's or all 0's sequence.

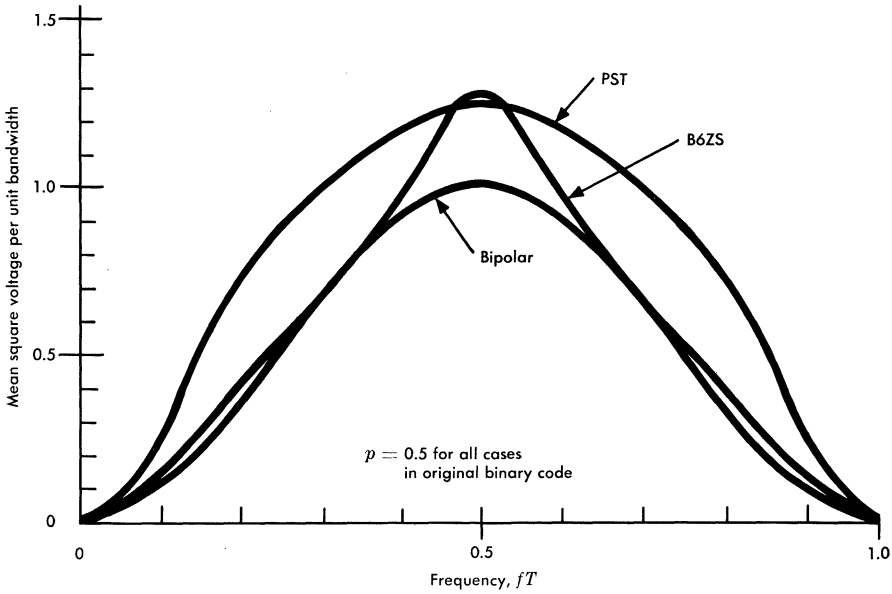


FIG. 27-41. Comparison of bipolar, B6ZS, and PST spectra.

Binary	MPST	
	+Mode	−Mode
1 1	+0	0−
1 0	+−	+−
0 1	−+	−+
0 0	0+	−0

FIG. 27-42. Modified PST coding.

All of the codes discussed so far have a signaling rate equal to the input binary rate. A ternary symbol has a theoretical information capacity of $\log_2 3 = 1.585$ bits. Consequently, the efficiency of these codes is only $1/\log_2 3 = 0.63$. Concepts employed in PST coding can be used for more efficient selected ternary codes. Block codes in which each N bit binary word is coded into a selected M digit ternary word where $M < N$ have been designed with good d-c wander and timing performance [29].

These concepts may also be generalized to the design of codes for other than three-level digital transmission. A balanced binary code has been developed for two-level transmission, which translates each input ternary symbol into selected binary words of length 2 [30].

Another Coding Approach

To this point, the use of redundancy and block coding to minimize the dependence of the line performance on the statistics of the signal source has been discussed. Other approaches are possible. Considering separately the three basic problems—repeater timing, d-c wander, and in-service performance monitoring—results in another coding approach.

The repeater timing problem arises because the timing component in the input spectrum is subject to large variations as the signal statistics change. A burst of symbols of alternating sequence can be inserted into the information sequence at a periodic rate to provide timing information. In each repeater, the timing burst can be identified by its specific characteristics and periodic occurrence. Hence, each repeater can derive timing from the burst rather than from the information sequence.

The d-c wander problem arises because all possible signal sequences can be emitted from the terminal, thereby causing a large variation in the low-frequency energy. Pulse scramblers can be used to reduce the variation in the low-frequency energy of the transmitted signal so that line sequences that contain long strings of 1's or 0's rarely occur. If this can be achieved, quantized feedback and d-c restoration can be used to control the remaining d-c wander. If, instead of baseband transmission, frequency shift keying on radio or waveguide is used, d-c wander is not a problem. To provide in-service performance monitoring, a parity check bit can be added as part of the timing burst.

Choice of Coding Method

In addition to the three fundamental problems, there are other factors which must be considered when choosing a code for digital transmission. These factors include crosstalk and economics.

Crosstalk. In paired cables, crosstalk not only reduces the signal-to-noise ratio at the input to the regenerator, it also reduces the signal-to-noise ratio at the timing extractor, thereby increasing the

timing jitter. As an example, B6ZS is preferable to PST in a crosstalk limited environment since it contains less energy than the PST signal.

For crosstalk-induced jitter, a bipolar format is preferable to a unipolar format because with the former the energy for repeater timing is derived from frequency components near the half-signaling rate, while with the latter this energy comes from a discrete component at the signaling rate.

With many existing paired cables, some of the pairs are in use for other types of transmission. Crosstalk coupled interference from the new to the existing systems can prevent the application of the new system to the remaining cable pairs. A choice of code format with low spectral energy in the frequency band of the existing systems may be necessary.

Economics. In the final analysis, the optimum code format for a digital transmission line is the one which provides suitable performance at minimum cost. The major economic considerations are those associated with code efficiency, repeater spacing, cable cost, and code translator cost. As an example, for short-haul applications, the higher efficiency attainable from coding that translates four binary symbols into three ternary symbols usually does not justify the additional code translator cost. This situation is reversed, however, for long-haul applications.

27.6 LINE MONITORING AND FAULT LOCATION

Digital transmission lines can be continuously monitored in service so that a spare can be substituted immediately in the event of a failure. To facilitate repair, means are provided to isolate the failure to one repeater section.

For the transmission codes discussed, violation of the coding sequence is used by monitors to detect line errors. Monitors can be located along the digital transmission route at maintenance offices. The lines between such offices are called maintenance spans. The monitors remove all violations detected so that violations do not propagate beyond the maintenance span, hence the name violation monitor and removal (VMR) unit. Removal of a violation may either remove the error or create a second error since the codes only have error detection capability. Failures that result in complete loss of signal must also be detected. In this case, a substitute signal is generated by the VMR to prevent alarms in other spans.

After a failed line is removed from service, the fault must be further isolated. For the codes that are d-c balanced, an effective technique for fault location is to introduce intentional violations of the code format. Since the repeater operation depends on absence of direct current, the net effect of violations is to move the regenerator crosshair relative to the input eye and to create a d-c component at the regenerator output. The regenerator outputs of all repeaters at a given site are coupled to a tuned circuit whose resonant frequency is identified with that site. The outputs of tuned circuits at all sites along the maintenance span are coupled to the maintenance office through a common pair. When the violation density is varied at a slow rate, a low-frequency component will appear at the regenerator output. If this frequency corresponds to that of the tuned circuit, it will be detected at the maintenance office and will indicate that all repeater sections of the cable pair being tested up to this site are operative. The magnitude of the violation density (and hence the amount of direct current) that the repeater can tolerate and remain operative is related to the distance from the crosshair to the eye edge and is a measure of the regenerator margin.

REFERENCES

1. Pierce, J. R. "Information of a Coaxial Cable with Various Modulation Systems," *Bell System Tech. J.*, vol. 45 (Oct. 1966), pp. 1197-1207.
2. Mayo, J. S. "Bipolar Repeater for Pulse Code Modulation Signals," *Bell System Tech. J.*, vol. 41 (Jan. 1962), pp. 25-97.
3. Wigington, R. L. and N. S. Nahman. "Transient Analysis of Coaxial Cables Considering Skin Effect," *Proc. IRE*, vol. 45 (Feb. 1957), pp. 166-174.
4. Aaron, M. R. "PCM Transmission in the Exchange Plant," *Bell System Tech. J.*, vol. 41 (Jan. 1962), pp. 99-141.
5. Nassell, I. "Some Properties of the Power Sums of Truncated Normal Random Variables," *Bell System Tech. J.*, vol 46 (Nov. 1967), pp. 2091-2110.
6. Mertz, P. "Model of Impulsive Noise for Data Transmission," *IRE Trans. Comm. Sys.*, CS-9 (June 1961), pp. 130-137.
7. Davis, J. H. "T2: A 6.3 Mb/s Digital Repeatered Line," *IEEE International Conference on Communications Record* (1969), pp. 34.9-34.16.
8. Cravis, H. and T. V. Crater. "Engineering of T1 Carrier System Repeatered Lines," *Bell System Tech. J.*, vol. 42 (Mar. 1963), pp. 431-486.
9. Dorros, I., J. M. Sipress, and F. D. Waldhauer. "An Experimental 224 Mb/s Digital Repeatered Line," *Bell System Tech. J.*, vol. 45 (Sept. 1966), pp. 993-1043.
10. Bennett, W. R. and J. R. Davey. *Data Transmission* (New York: McGraw-Hill Book Company, Inc., 1965).
11. Gibby, R. A. and J. W. Smith. "Some Extensions of Nyquist's Telegraph Transmission Theory," *Bell System Tech. J.*, vol. 44 (Nov. 1965), pp. 1487-1510.

12. Tufts, D. W. "Nyquist's Problem—The Joint Optimization of Transmitter and Receiver in Pulse Amplitude Modulation," *Proc. IEEE*, vol. 53 (Mar. 1965), pp. 248-259.
13. Aaron, M. R. and D. W. Tufts. "Intersymbol Interference and Error Probability," *Trans. IEEE Information Theory*, vol. IT-12 (Jan. 1966), pp. 26-34.
14. Tufts, D. W. and T. Berger. "Optimum Pulse Amplitude Modulation—I: Transmitter-Receiver Design and Bounds from Information Theory," *Trans. IEEE Information Theory*, vol. IT-13 (Apr. 1967), pp. 196-208.
15. Tufts, D. W. and T. Berger. "Optimum Pulse Amplitude Modulation—II: Inclusion of Timing Jitter," *Trans. IEEE Information Theory*, vol. IT-13 (Apr. 1967), pp. 209-216.
16. Kuo, F. F. and J. F. Kaiser. *System Analysis by Digital Computers* (New York: John Wiley and Sons, Inc., 1966).
17. Nelder, J. A. and R. Mead. "A Simplex Method for Function Minimization," *The Computer Journal*, vol. 7 (Jan. 1965), pp. 308-313.
18. Bennett, W. R. "Synthesis of Active Networks," *PROC. Symposium on Modern Networks Synthesis* (1955), pp. 45-61.
19. Aaron, M. R. and M. K. Simon. "Approximation of the Error Probability in a Regenerative Repeater with Quantized Feedback," *Bell System Tech. J.*, vol. 45 (Dec. 1966), pp. 1845-1847.
20. Aaron, M. R. and J. R. Gray. "Probability Distributions for the Phase Jitter in Self-Timed Reconstructive Repeaters for PCM," *Bell System Tech. J.*, vol. 41 (Feb. 1962), pp. 503-558.
21. De-Lange, O. E. and M. Pustelnyk. "Experiments on the Timing Regenerative Repeaters," *Bell System Tech. J.*, vol. 37 (Nov. 1958), pp. 1487-1500.
22. Rowe, H. E. "Timing in a Long Chain of Regenerative Binary Repeaters," *Bell System Tech. J.*, vol. 37 (Nov. 1958), pp. 1543-1598.
23. De-Lange, O. E. "The Timing of High-Speed Regenerative Repeaters," *Bell System Tech. J.*, vol. 37 (Nov. 1958), pp. 1455-1486.
24. Bennett, W. R. "Statistics of Regenerative Digital Transmission," *Bell System Tech. J.*, vol. 37 (Nov. 1958), pp. 1501-1542.
25. Byrne, C. J., B. J. Karafin, and D. B. Robinson. "Systematic Jitter in a Chain of Digital Regenerators," *Bell System Tech. J.*, vol. 42 (Nov. 1963), pp. 2679-2714.
26. Andrews, F. T. "Bipolar Pulse Transmission and Regeneration," U. S. Patent 2996578, Aug. 15, 1961.
27. Johannes, V. I., A. G. Kaim, and T. Walzman. "Bipolar Pulse Transmission with Zero Extraction," *Trans. IEEE Communications Technology*, vol. Com-17 (Apr. 1969), pp. 303-310.
28. Sipress, J. M. "A New Class of Selected Ternary Pulse Transmission Plans for Digital Transmission Lines," *Trans. IEEE Communications Technology*, vol. Com-13, no. 3 (Sept. 1965), pp. 366-372.
29. Franaszek, P. A. "Sequence State Coding for Digital Transmission," *Bell System Tech. J.*, vol. 47 (Feb. 1968), pp. 143-157.
30. Neu, W. and A. Kundig. "Project for a Digital Telephone Network," *Trans. IEEE Communications Technology*, vol. Com-16 (Oct. 1968).

Chapter 28

Syllabic Companding and TASI

This chapter discusses the uses of two special techniques and their transmission effects on multichannel analog system design. The techniques covered are syllabic companding and time assignment speech interpolation (TASI).

Two characteristics of the speech signal allow the use of these techniques to enhance the capability of transmission systems. Basically, the telephone speech signal has a wide volume range with a high peak factor and is intermittent in nature. Companding takes advantage of the wide volume range, and TASI utilizes the intermittent characteristic of speech.

28.1 SYLLABIC COMPANDORS

Syllabic compandors [1, 2, 3] are voice-operated electronic devices used to improve the telephone channel speech-to-noise ratio. The word *compandor* is coined from a contraction of compressor and expandor. A compressor is associated with each channel in the transmitting terminal, and likewise a complementary expandor is associated with the receiving terminal. These devices are generally part of the voice-frequency channel portion of such systems as *N*-type carrier, where the use of inexpensive cable media is permitted by the S/N advantage realized by companding. In such applications, terminal costs are increased; however, this is justified by the low cost of the transmission line.

A compandor provides noise advantage for speech transmission in several ways. The compressor raises the weaker speech signals and reduces the volume range before the signals are exposed to the

noise of the transmission system. Thus, in the transmitting medium, the speech-to-noise ratio has been improved. The compressor acts as an amplifier of the input speech signals, but the gain decreases as the speech volume increases. Typically, the output from the compressor varies 1 dB for every 2-dB change in input signal. The expandor performs the opposite function, inserting loss which increases with decreasing speech volumes. During a syllable, the gain of the compressor and the loss of the expandor are dependent upon the average power of the talker. During quiet periods, the loss of the expandor is constant and very high. Thus, noise during the latter period is heavily attenuated. This action provides the major part of the noise advantage realized by companding.

The level diagram of Fig. 28-1(a) shows the action of an N system compandor for sine-wave signals. The ideal input-output characteristics of the compressor and expandor which produce this level diagram are shown in Fig. 28-1(b) and can be expressed as follows:

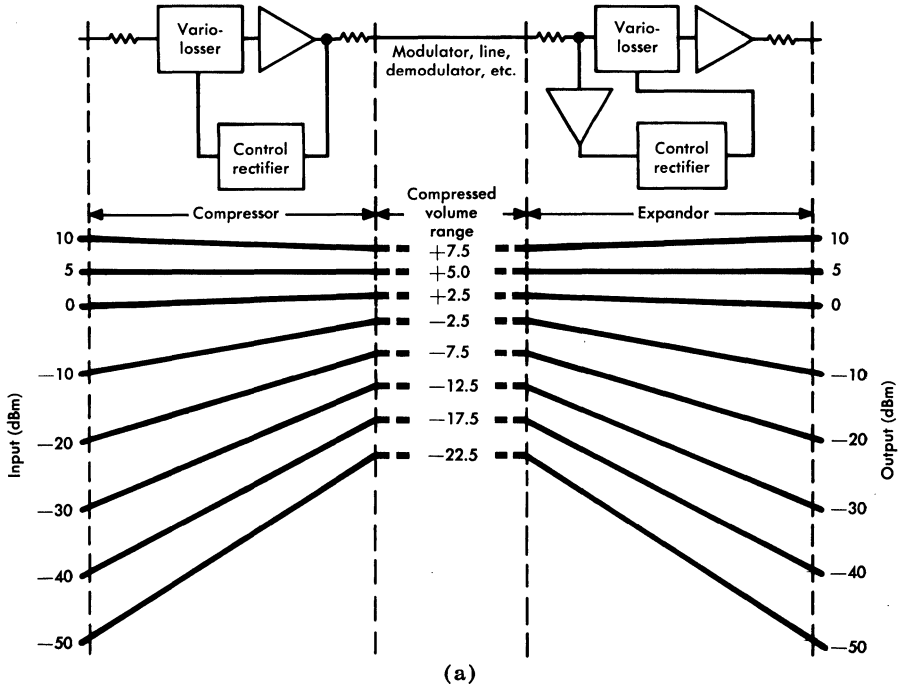
$$\text{Compressor power out (dBm)} = \frac{\text{power in (dBm0)}}{K} + A \tag{28-1}$$

$$\text{Expandor power out (dBm0)} = K\ [\text{power in (dBm)}] - KA$$

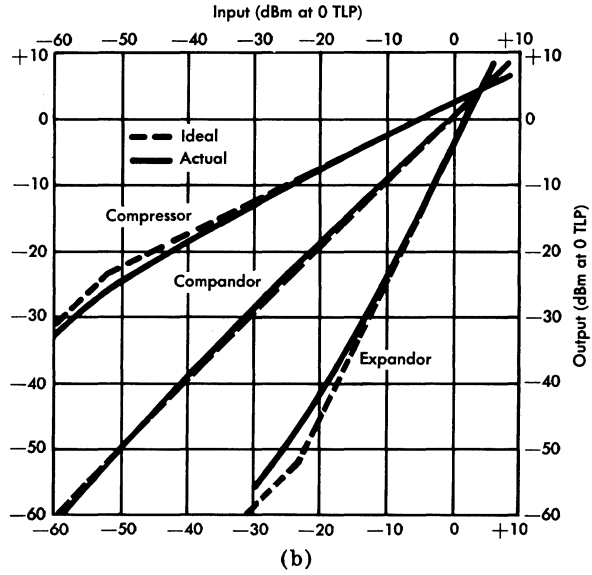
where K and A depend upon the compandor design. The value of K, the compression ratio, is determined by the compressor characteristic, and $KA/(K-1)$ is the crossover level of the compressor and expandor.

The values of K and A should be as high as possible for maximum noise advantage. From practical considerations, however, the value of K is made 2 and KA is typically 5.0 dBm. The value of K is limited in practice because the loss variations in the medium between the compandor and expandor are multiplied by K. Using higher values of K would require increasingly tighter tolerances for the loss variations and for the relative frequency response across the voice-frequency band of the transmission facilities. In addition, keeping the matching or tracking error of the compressor and expandor characteristics to a small value over a wide input range becomes more difficult. The value of KA is restricted by the peak power that can be transmitted.

In the absence of speech energy Fig. 28-1(b) shows that noise signals of about −22 dBm or less at the expandor input will be attenuated by 28 dB. In the presence of speech energy, there is no



(a)



(b)

FIG. 28-1. Input-output characteristics of N system compressors and expandors.

relative noise reduction since speech will control the action of the expandor. However, in this case the noise is generally masked by the speech and is much less annoying than in the absence of speech. The overall result is to permit a system to operate with considerably more noise, crosstalk, or interference up to the input to the expandor than could be otherwise tolerated. There is no noise advantage for data transmission, and the data signal energy level must be constant and continuous (like an FM or PM data set) to eliminate possible errors due to compandor transient action.

Another important characteristic of syllabic compandors is the response to suddenly applied signals such as speech bursts and pulsed signaling tones. For example, a sudden increase in input signal will appear initially as a sudden increase in the output signal of the compressor, which will immediately start dropping toward its compressed value. The converse action takes place for a decrease in input signal. These actions have been called the "attack" and "recovery" times and are under the control of the designer. If these times are too fast, the output signal will require wider bandwidths for faithful transmission. If the attack time is too slow, the duration of signal overload may be excessive. If the recovery time is too slow, noise between syllables could become objectionable. Recommendations of CCITT [4] specify attack time as equal to or less than 5 milliseconds and recovery time as equal to or less than 22.5 milliseconds.

The compressor changes the amplitude distribution of the speech signal and, as a result, the effects of the compressed signal on the multichannel loading and on modulation noise must be considered. To understand the nature of these effects, consider the compandored system of Fig. 28-2. Let A and B be fixed level points before the application of compandors. The multichannel load, P_s, at point A can found from Eq. (9-22).

$$P_s = V_0 - 1.4 + 0.115\sigma^2 + 10 \log \tau_L + 10 \log N + \Delta_c \quad \text{dBm0} \qquad (9\text{-}22)$$

After companding, P_s will be changed because the compressor changes the values of V_0, σ, and Δ_c. Assume that the mean talker power, $V_0 - 1.4$, is -16 dBm0 and that σ is 5 dB. From Eq. (28-1) the mean talker power becomes $-16/2 + 2.5 = -5.5$ dBm0, an increase of 10.5 dB. The standard deviation, σ, is reduced by the factor of $K = 2$ or from 5 to 2.5 dB, and the $0.115\sigma^2$ term is reduced by 2.2 dB. The value of Δ_c is a function of σ for constant τ_L and N

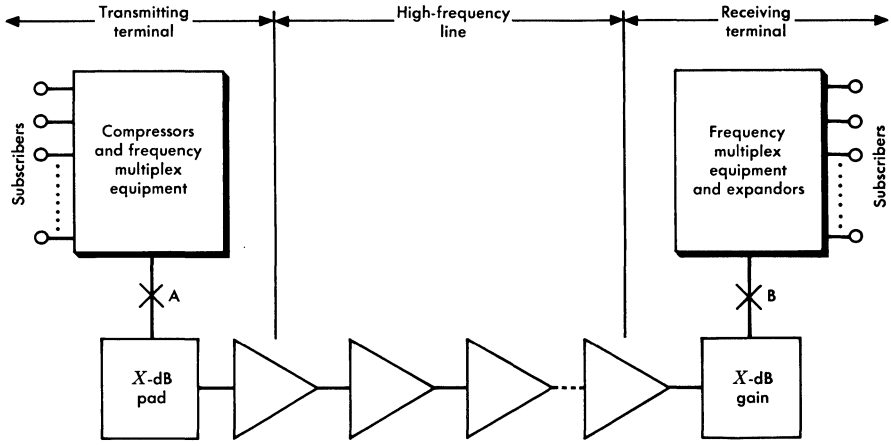


FIG. 28-2. System with compandors—block schematic.

as indicated by Eq. (9-21), and values of Δ_c are tabulated in Fig. 28-3. For a 100 channel system, Δ_c decreases by 2.2 dB. Thus, for this system, companding increases the multichannel load, P_s, by $+10.5 - 2.2 - 2.2 = 6.1$ dB. It is therefore usually necessary to lower the line level accordingly with a pad (labeled X on Fig. 28-2) to keep the same peak load. This reduces the companding advantage correspondingly since gain is required at the receiving end to restore the proper level.

The thermal noise at point B is independent of the input signal and can be computed by methods developed in previous chapters.

Number of channels	Δ_c for $\sigma = 2.5$ (dB)	Decrease in Δ_c for $\sigma = 2.5$ vs 5.0 (dB)
10	18.4	3.4
20	16.9	2.9
50	14.6	2.5
100	13.0	2.2
200	11.9	1.8
500	11.0	1.2
1000	10.4	1.0

FIG. 28-3. The multichannel load factor for compressed voice signals.

Modulation noise at point B can be computed by using the new values of V_0 and σ of the compressed load in Eq. (10-27). It will be found that the increase in modulation noise is usually greater than the increase in P_s.

The relationship between the noise at point B and at the expandor output must be considered. Figure 28-1(b) indicates that the noise between syllables will not be expanded unless it exceeds —22 dBm at the expandor input. Below this power it will be attenuated by 28 dB in passing through the expandor. Not all of this loss can be taken as an advantage, however, since the expandor action is not instantaneous, and because noise during syllables is not attenuated relative to the signal. Listening tests indicate that a 5-dB allowance should be made for these effects. Thus, the net compandor advantage for thermal noise in dB is 28 — 5 — x where x is the increase in P_s due to compression. The advantage for modulation noise in dB will be 28 — 5 dB less the increase in modulation noise found above.

28.2 TASI

TASI, an abbreviation of time assignment speech interpolation, is a voice-operated switching and channel assignment system to interpolate additional talkers onto communication facilities [5]. In the past, TASI systems have usually been used with submarine cable voice channel facilities. New systems may operate with mixed satellite and cable voice channels.

In a normal telephone conversation each subscriber speaks less than half of the time. The remainder of the time is composed of listening, gaps between words and syllables, and pauses. Measurements on working transatlantic circuits show that speech energy as detected by the TASI speech detectors is present on the average for only about 40 per cent of the time the circuit is in a "busy" state at the switchboard. This is the TASI activity factor and should not be confused with the speech activity factor used in modulation and load computations. Since long distance circuits use separate paths for the two directions of transmission, each one-way circuit is, on the average, free about 60 per cent of the time. TASI takes advantage of this free time by detecting the presence of a talker's speech and assigning him an unused channel. He will keep that channel until he is silent and his channel is needed for another talker. During low traffic periods the system does no switching and the talker may keep the same channel throughout a conversation.

During high traffic periods, however, successive speech spurts may be assigned and switched to different channels. The assignment and switching processes are made in milliseconds so that little of the initial syllable is lost. The increase in circuit capacity (TASI advantage) is dependent upon the TASI activity factor, the allowable "freezeout" percentage (percentage of time that the number of active talkers exceeds the number of channels), and the number of channels available. For example, with TASI B, 96 speech channels in the medium can serve 235 customers.

The implementation of TASI involves expenditures for switching and logic control equipment, but when compared with the cost of additional long cable or satellite facilities, the system may provide a substantial reduction in the per-circuit cost. The problems involved in recognizing speech energy, coordinating the connections at both ends, and controlling the system to make the subjective effects of freezeout negligible are complex. This requires what is essentially a special purpose computer for each terminal.

So far as the design of the transmission system between TASI terminals is concerned, the application of TASI is merely an increase in the telephone load activity factor, τ_L, to a new value

$$\tau_L' = \frac{N_T}{N_C}\tau_L$$

where N_T is the number of off-hook talkers and N_C is the number of available channels. The factor τ_L' must be used to compute the modulation noise. There will be two effects on the multichannel load: the average value will increase, and Δ_c will decrease. To find the new value of Δ_c, curves such as shown in Fig. 9-4 may be generated using τ_L'. In a practical case, however, a simpler approximation is to compute the multichannel load for N_T talkers. For example, with 96 channels about 235 talkers can be served. Thus, the multichannel load for the 96 channels would be computed for $N_T = 235$, $\tau_L = 0.25$, with V_0 and σ as measured with the given talkers.

REFERENCES

1. Dickieson, A. C., D. Mitchell, and C. W. Carter, Jr. "Application of Compandors to Telephone Circuits," *Trans. AIEE*, vol. 65 (Aug. 1946), pp. 1079-1086.
2. Rizzoni, E. M. "Compandor Loading and Noise Improvement in Frequency Division Multiplex Radio Relay Systems," *Proc. IRE* (Feb. 1960).

3. Boyd, R. C., et al. "N2 Carrier System," *Bell System Tech. J.*, vol. 44 (May-June 1965), pp. 731-822.
4. The International Telegraph and Telephone Consultative Committee (CCITT). *Blue Book*, vol. 3, Third Plenary Assembly (Geneva: May 25-June 26, 1964), Recommendation G-162, p. 62.
5. Fraser, J. M., D. B. Bullock, and N. G. Long. "Overall Characteristics of a TASI System," *Bell System Tech. J.*, vol. 51 (July 1962), pp. 1439-1454.

Chapter 29

Television and Visual Telephone Transmission

The Bell System has in operation many thousands of route miles of wideband transmission circuits which serve the television broadcasting industry. The television signals transmitted over these circuits are described in this chapter, and system objectives for the transmission circuit are discussed. In addition, a brief description of the visual telephone system now under development is included.

29.1 CHARACTERISTICS OF THE TELEVISION SIGNAL

The television signal [1] contains information in electrical form, from which a picture can be recreated with fidelity. A still monochrome picture may be expressed as a variation in luminance over a two-dimensional field. In a moving picture, however, the luminance function also varies with time. The moving picture, therefore, is a luminance function of three independent variables.

The electrical signal consists of a current or voltage amplitude which is a function of time. At any instant, the signal can represent the value of luminance at only one point in the picture. It is necessary, therefore, in the translation of a complete picture into an electrical signal, that the picture be scanned in a systematic manner. If the scanning pattern is sufficiently detailed and conducted rapidly, a satisfactory reproduction of picture detail and motion is obtained. The basic system consists of a series of scans in nearly horizontal lines from left to right, starting at the top of the image field. When the bottom of the field is reached, the process is started again from the top with alternate fields interlaced. This scanning process is illustrated in Fig. 29-1.

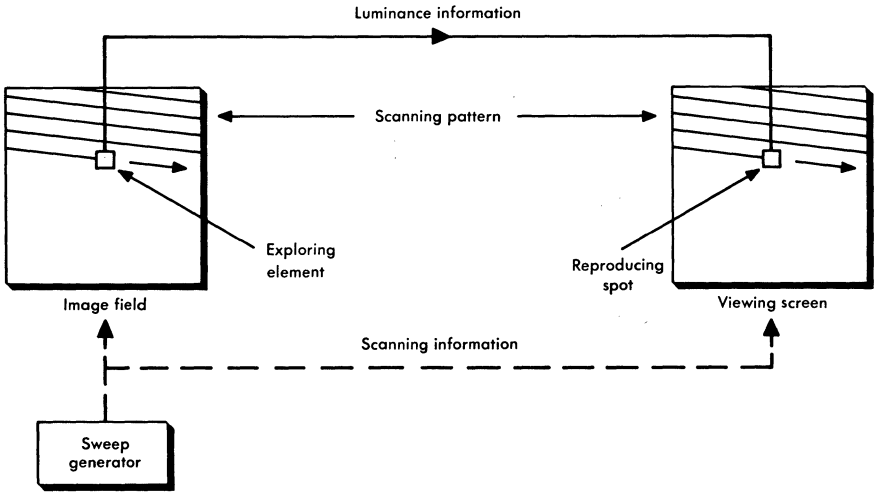


FIG. 29-1. Television scanning process.

For the successful decoding of the signal into a picture at the receiver, it is necessary to transmit a key to the scanning pattern. In the standard signal, this consists of frequent short-duration pulses (synchronizing or sync pulses), indicating characteristic points in the course of the scanning pattern such as the beginning of scanning lines and fields. This is coupled with the condition that the motion of the scanning spot between pulses is uniform with the time in the field of view.

The synchronizing pulses must be distinguishable from the picture signal. This is accomplished by time and polarity separation. An illustration of a portion of the signal is shown in Fig. 29-2.

The picture signal is interrupted during retrace time and replaced by a black signal known as a blanking pulse. Because of this, the return trace will not be visible in the picture. The line synchronizing pulses are superimposed on the blanking pulses and occupy the amplitude range from b to c. Since the picture signal extends from white at a to black at b, this region is "blacker than black," and the sync pulses do not register in the picture.

The line pulses synchronize the individual horizontal scanning lines. Similarly, it is necessary to synchronize field (vertical) scans. This is done by a longer pulse in the same amplitude range, as shown in Fig. 29-3. Serrations in this pulse maintain horizontal line synchronization, and the horizontal and vertical sync pulses are identified

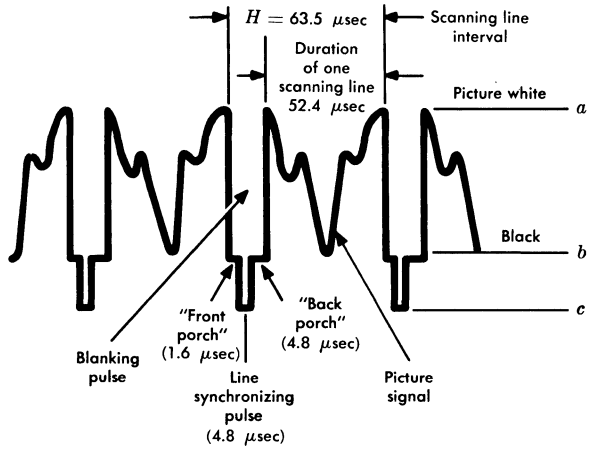


FIG. 29-2. Portion of a television signal showing line synchronizing pulses.

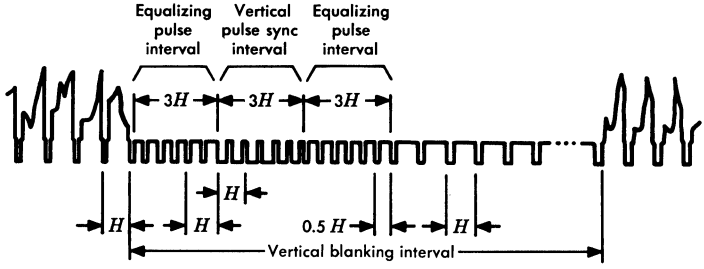


FIG. 29-3. Portion of television signal showing field synchronizing pulses.

at the receiver by their greatly different durations. During the field retrace time, the picture is again blanked.

Waveforms

Figures 29-2 and 29-3 show portions of the normal waveforms for both line and field synchronizing pulses of the National Television Systems Committee (NTSC) 525-line television signal. The time interval from the start of one line to the start of the next, *H*, is 63.5 microseconds (15.750-kHz line frequency). The picture is scanned vertically at a rate of 60 fields per second. Each complete frame consists of two interlaced fields giving the frame rate of 30 per second. For proper interlace, alternate fields start at the midpoint of a horizontal line. This is accomplished by using equalizing pulses at twice

the line rate during the vertical blanking interval. Of the 525 horizontal scanning lines per frame, only about 93 per cent are visible because of blanking during the vertical blanking interval.

The most direct method available for measuring the amplitude and time relationships between the various components of a video signal is to observe the waveforms on an oscilloscope. The method of measurement and the specifications and adjustments of amplitude relationships were standardized by the IRE (now IEEE). The IRE standard scale is shown in Fig. 29-4.

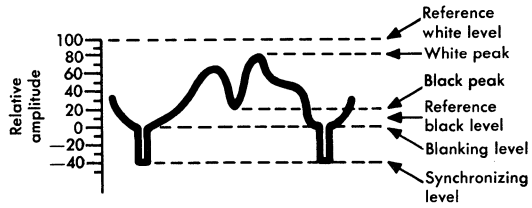


FIG. 29-4. Television signal amplitude relationships.

Vertical Interval Test Signals

The television signal received from the major television networks may also include test signals for use in measuring certain transmission characteristics. The vertical interval test signals (VITS) are inserted by the broadcaster in the vertical blanking interval on lines 18 and 19. These signals, patterned after full frame signals used for out-of-service testing, allow an approximate transmission evaluation without interfering with the picture seen by the viewer. Commonly introduced signals are shown in Fig. 29-5 and include:

1. A multiburst signal (a burst of white and separate bursts of six sine wave frequencies) to permit frequency response measurement.
2. A sine-squared pulse and bar signal to provide transient response information.
3. A stairstep signal with a 3.58-MHz carrier burst superimposed on each step to provide differential gain and phase information.

Additional test signals intended particularly for color transmission are under consideration.

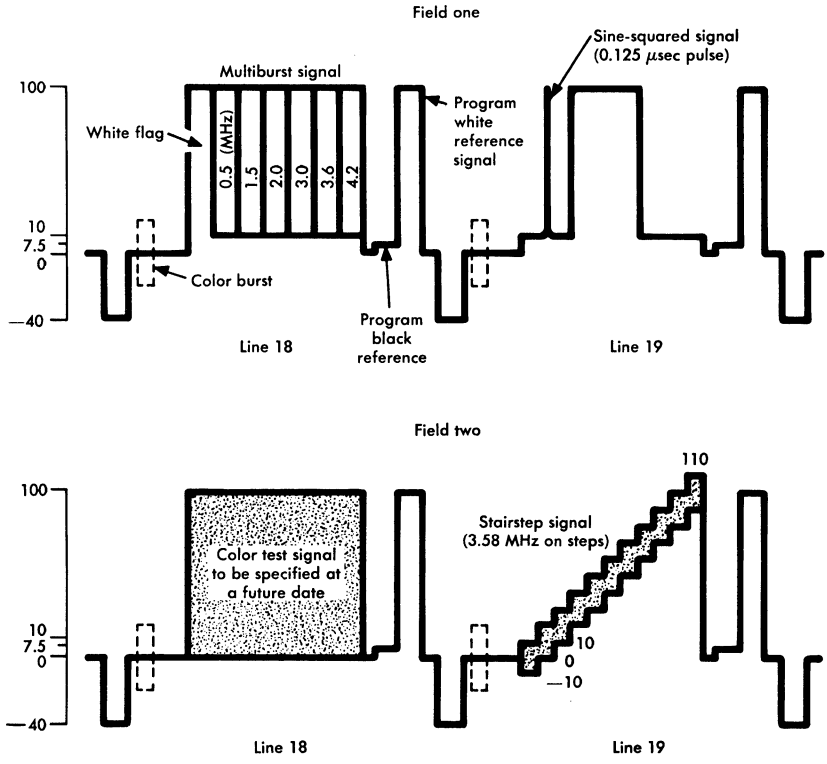


FIG. 29-5. Vertical interval test signals.

Bandwidth

The bandwidth occupied by a television signal is a function of the frame rate and the fineness of detail to be transmitted. Important components will appear as low as the 30-Hz frame scanning rate. It is necessary, therefore, to transmit frequencies almost down to zero to maintain good phase response at 30 Hz. Subjective viewing tests indicate that an upper frequency limit of about 4 MHz results in very little degradation; a greater bandwidth results in an improvement which analysis indicates would not be economically worthwhile. A reduction to a 3-MHz bandwidth, on the other hand, results in a noticeable degradation for monochrome transmission. For color, this bandwidth is unusable since the color carrier is at 3.58 MHz.

It is instructive to compute the required bandwidth, making certain simplifying assumptions. The standard television picture in the

United States uses a frame rate of 30 Hz with 525 lines per frame. The ratio of picture width to height (aspect ratio) is 4:3.

Vertical Resolution. Because of the loss of lines during blanking time between fields, the vertical resolution is reduced to about 93 per cent. Further loss of resolution is inherent in the scanning process. Since the scanning lines are discrete and of finite width, the relative position of the scanning lines and any horizontal lines in the scanned original will affect reproduction. For example, if the original consists of alternate black and white lines of the same width as the scanning lines and perfectly coincident with them, reproduction will be accurate. If these same scanned lines are centered on the boundary between scanning lines, however, they will produce a gray picture. Experimental study indicates that for typical pictures this effect decreases vertical resolution to about 70 per cent. The resulting number of vertical elements which can be resolved is then $n_v = 525 \times 0.7 \times 0.93 = 342$.

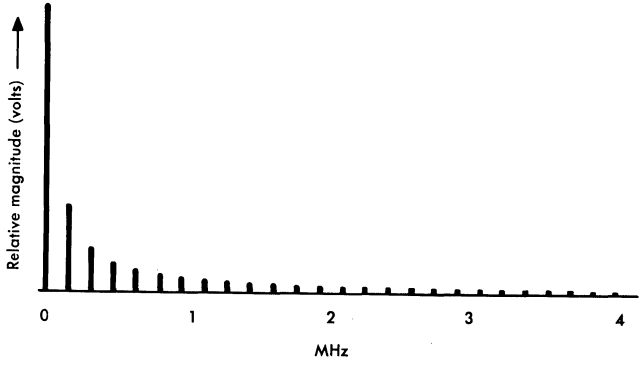
Horizontal Resolution. Horizontal resolution is determined by the highest frequency component which can be resolved along a line. If the simplifying assumption is made that a simple sinusoid will generate a series of alternate black and white spots, it is easy to see that the finest detail which can be reproduced will be determined by the highest frequency sinusoid that can be transmitted (assuming that spot size is not the limiting factor, of course). Subjective tests indicate that satisfactory results are obtained if horizontal resolution is made approximately equal to vertical resolution. Thus, since the aspect ratio is 4 : 3, the desired number of horizontal picture elements per line scan would be $n_h = 4/3 \times 342 = 456$.

Required Bandwidth. A sinusoid which would generate 456 alternate black and white spots would go through 228 cycles along a scanning line. As shown on Fig. 29-2, the duration of the scanning line is 52.4 microseconds. The top transmitted frequency must then be $f_{max} =$ (228 cycles per line)/(52.4 microseconds per line), or about 4.3 MHz. In spite of all the simplifying assumptions made, this result does not differ substantially from the generally used value of 4.2 MHz.

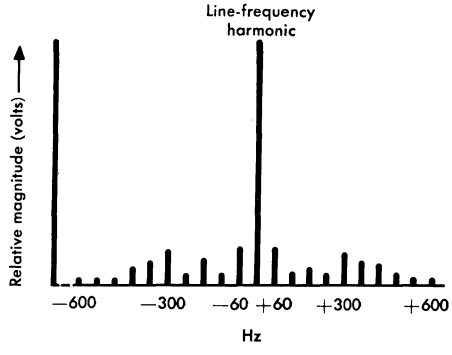
The significant thing to note in this discussion is that the upper frequency cutoff of the channel bandwidth determines the horizontal resolution. If this cutoff is lowered, the horizontal detail suffers.

Spectrum

The scanning process determines the basic distribution of energy in the signal band. Line scanning of picture information concentrates the signal energy into harmonics of the line-frequency. The 60-Hz field scan gives rise to 60-Hz sidebands on each line-frequency harmonic. The television signal, therefore, consists of a number of fixed frequencies which vary in phase and amplitude at a slow rate only as a result of motion in the picture. Each of these frequencies can be considered as a carrier and the effect of motion as sidebands added around the carrier. The net result is a signal frequency composition similar to that shown in Fig. 29-6.



(a) Typical spectrum showing every tenth line-frequency harmonic



(b) Typical sidebands around each line-frequency harmonic

FIG. 29-6. Signal frequency composition of monochrome TV spectrum.

Figure 29-6(a) illustrates the entire 4-MHz bandwidth, indicating the relative amplitude of line-frequency harmonics for a typical signal. Only every tenth harmonic is shown and the 60-Hz components near zero frequency have been omitted for clarity. A small section of Fig. 29-6(a) magnified to illustrate the presence of the 60-Hz sidebands that cluster about each line-frequency harmonic is shown in Fig. 29-6(b).

Color Signal

The NTSC color television system [2] is based on the principle that color may be adequately defined in terms of three characteristics: luminance, hue, and saturation. Luminance is defined as intensity or brightness and is the basis on which the present monochrome system operates. Hue is the color in terms of whether it is red, blue, green, yellow, etc. Saturation is the degree to which the hue is mixed with white. For example, pink is a low saturation red. A high saturation red would be a brilliant crimson.

The color signal, therefore, must contain luminance, hue, and saturation information. The color system uses the same type of signal to transmit luminance information as is used in the monochrome system. To this is added the saturation and hue information which is the basic difference between the monochrome signal and a color signal. The necessity of transmitting three pieces of information instead of one, simultaneously and without interaction or distortion, imposes new requirements on the transmission facilities. This situation is analogous to the transmission of two or more voice signals simultaneously in a carrier telephone circuit. If the circuit is perfectly linear, there is no difficulty in separating the various voice channels at the receiving end. If the circuit is not linear, the channels modulate with one another and crosstalk occurs.

The saturation and hue information is added to the luminance signal in the form of a new broadband signal which is modulated on the 3.58-MHz color subcarrier. The amplitude of this signal represents the saturation of the color. A large amplitude represents high saturation or brilliant color. Distortion of color saturation will occur if the gain of the transmission system at the color carrier frequency is a function of the amplitude of the luminance signal. This variation in the amplitude transmission of the color signal caused by variation in the amplitude of the luminance signal is called *differential gain.* The presence of differential gain in a system used to transmit color

television may result in a picture in which some colors may appear dim or washed out while others may appear oversaturated.

The instantaneous phase relationship of the color subcarrier to a reference synchronizing signal of the same frequency (color burst) determines the hue of the color. The color burst consists of approximately 9 cycles of the color carrier frequency placed on the back porch of the horizontal blanking signal, as shown in Fig. 29-7.

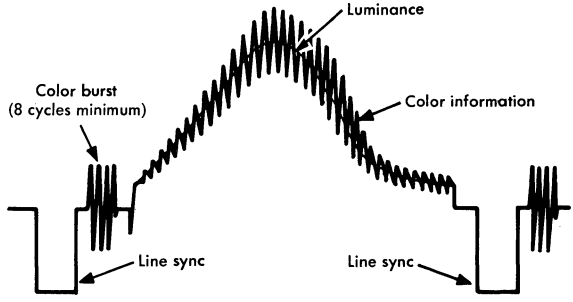


FIG. 29-7. Color signal waveform.

Distortions of hue will occur if the instantaneous phase shift of the transmission system at color carrier frequency is a function of the amplitude of the luminance signal. This variation in color carrier phase shift caused by variations in amplitude of the luminance signal is called *differential phase* and results in a distortion in the hue of the colors.

The frequency of the color subcarrier, 3.579545 MHz, is an odd multiple of half-line frequency (7867 Hz for color television). The effect of this is to interleave the components of the chrominance signal spectrum between the luminance signal components. The frequency composition of a typical NTSC color signal is shown in Fig. 29-8. The smaller chrominance components on either side of the subcarrier are produced by the scanning as in the case for luminance, and they vary in amplitude and phase in accordance with the hue and saturation information being transmitted.

29.2 TELEVISION TRANSMISSION IMPAIRMENTS AND OBJECTIVES

Television objectives, like telephone objectives, are based on the results of many subjective tests. Extensive viewing tests have been

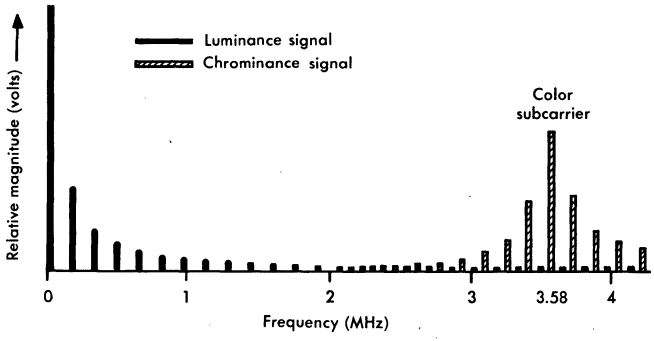


FIG. 29-8. NTSC color TV spectrum.

made in which various impairments were added to the picture and the result judged in terms of seven preworded comments ranging from "not perceptible" to "extremely objectionable." The data for all observers were combined to determine the curve for a median observer. In general, the objective for a particular effect has been set so that the median observer will find the picture impairments to be "just perceptible." Picture impairments are primarily due to the following:

1. Bandwidth limitations.
2. Transmission deviations.
3. Crosstalk.
4. Random noise.
5. Single-frequency interference.
6. Nonlinear effects.

Bandwidth Impairment

An unimpaired test signal or test pattern as reproduced on a television monitor is shown in Fig. 29-9. Note that the vertical lines in the pattern can be resolved into the central circle.

To illustrate an impairment due to reduced bandwidth, assume a low-pass network having a transmission characteristic which is flat to 1 MHz and which then falls off at the rate of 6 dB per octave is inserted between the picture source and the monitor. The test pattern will then be reproduced as shown in Fig. 29-10. The picture is no longer crisp; the vertical bars cannot be resolved close to the central circle (see the lower wedge).

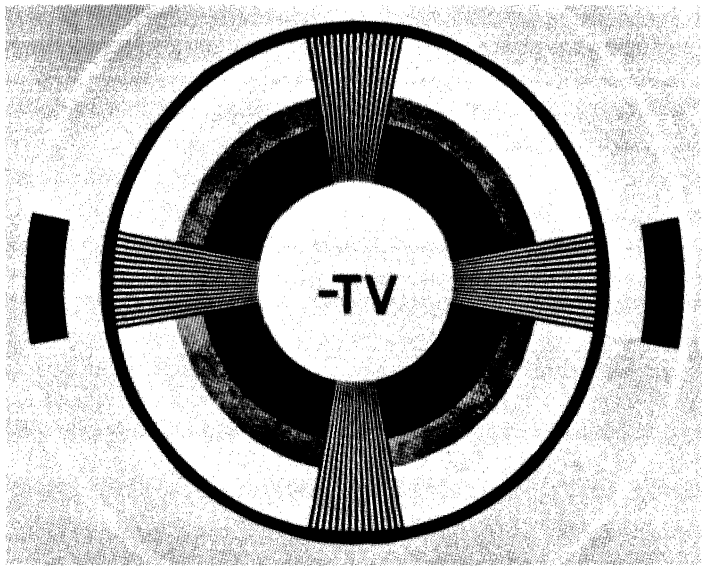


FIG. 29-9. Unimpaired signal.

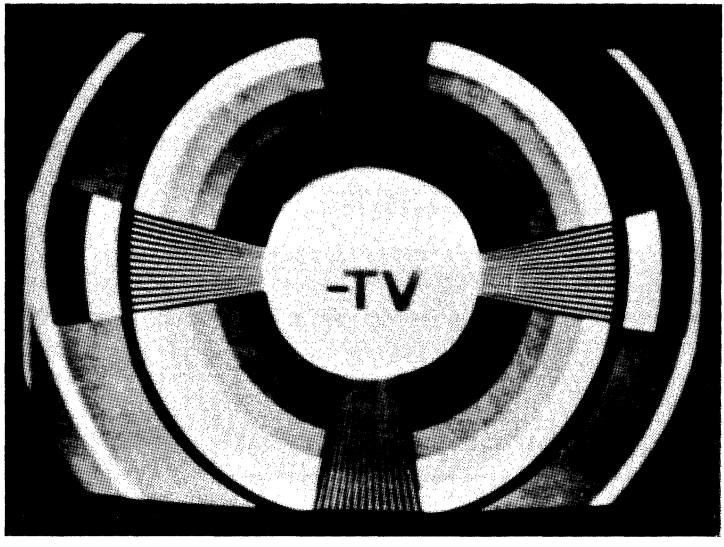


FIG. 29-10. 1-MHz roll-off.

Transmission Deviations

Departures from flat gain and delay characteristics over the required bandwidth may result in a number of picture impairments. These impairments, such as streaking, smearing, overshoot, ringing, and echoes, are described qualitatively.

Streaking and Smearing. Streaking is caused by transmission distortions in the lower frequency range up to about 200 kHz. Smearing is generally caused by distortions at somewhat higher frequencies. Streaking and smearing affect both color and monochrome signal transmission. Amplitude and phase distortion tolerances at the low end of the frequency band (below 5 kHz) are relatively less critical because of the use of electronic circuits called clampers. Clampers restore low-frequency signal components which were not faithfully transmitted. They permit at least a 35-dB relaxation of gain and phase distortion at 60 Hz, but their effectiveness decreases as frequency increases. All Bell System television networks include clampers [3].

Overshoot. In a television signal, an overshoot is an excessive response to a sudden change in signal. A sharp overshoot is commonly referred to as a spike and is generally caused by excess gain at high frequencies. There is a black outline to the right of white objects and a white outline to the right of black objects.

Ringing. Multiple overshoot, or ringing, generally results from the transmission of sudden transitions over a system that has a finite passband with a sharp cutoff at the upper end of the frequency range. It may also result from a marked transmission irregularity at some frequency below cutoff. When a signal containing a sudden transition is applied to such a circuit, damped oscillations or ringing occurs at approximately the frequency of cutoff or other discontinuity. The duration of the ringing depends upon the sharpness of the irregularity. Ringing is accentuated by a rising gain characteristic preceding the discontinuity.

Echoes. An echo signal [4], or ghost, can be defined as an approximate duplicate of the original video signal displaced horizontally from the original signal. Echoes are due to impairments in the transmission circuit, which cause the signal pulses to reach the viewer at two or more discrete times. The impairment effect of the

echo picture not only varies with echo signal amplitude but also with the time offset and the nature of the original video signal. As a practical matter, echo signals are generally not true reproductions of the original signal since the deviations in a transmission system are usually not continuous throughout the band.

Objectives for Transmission Deviations

The several types of impairments that have been considered have one thing in common—they are all caused by transmission deviations and can be reduced by providing gain and phase equalization. Objectives may be placed on gain and phase deviations, but the overall transmission objective is usually stated in terms of *echo rating.*

Gain and Phase. Objectives for gain and phase characteristics are given in terms of coarse structure and fine structure deviations in the frequency domain.

Figure 29-11 shows an illustrative steady-state phase deviation versus frequency after the linear component has been subtracted. Widely spaced variations like that indicated by the broken line are known as coarse structure. Such variations are also described as having low periodicity in the sense that they would represent a slowly changing function to an observer who scanned the transmitted band. Closely spaced variations as shown by the solid line are known as fine structure or high periodicity variations. Quantitatively, a deviation is fine structure if Δf is much less than 555 kHz,

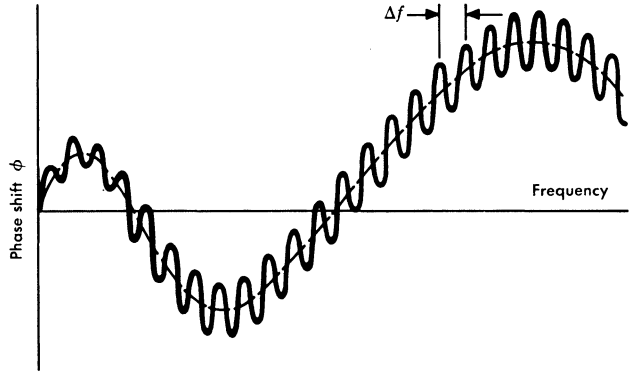


FIG. 29-11. Steady-state phase curve illustrating coarse and fine structure transmission deviations.

coarse structure if Δf is much more than 555 kHz. The definition is obviously a somewhat arbitrary one and is applied to gain deviations versus frequency as well.

The coarse structure objectives for monochrome television from 7875 Hz (half-line frequency) to the upper cutoff are given as ± 0.025 microsecond for envelope delay and ± 1.4 dB for gain. However, for color transmission the gain at 3.6 MHz (color carrier) should be very nearly the same as at low frequencies.

The fine structure objectives for a sinusoidal ripple across the band are ± 0.8 degree for phase, and ± 0.12 dB for gain (each corresponding to -40 dB echo rating). These objectives are for loss and phase (or delay) deviations occurring separately. Where they occur together, the limits must be divided by $\sqrt{2}$. When the deviations occur only over part of the band, the objectives can be relaxed. The allowable deviations at the upper end of the band, except near the color subcarrier, may be appreciably larger than those at the low end.

Echo Rating. Transmission deviations can also be expressed in terms of echoes having various amplitudes and delay times with reference to the main signal. The effects of these echoes have been evaluated by subjective testing in respect to the echo time delay and to the frequency of a deviation that occurs over a part of the transmission band. The time weighting and the frequency weighting and the manner in which multiple echoes add in their interfering effect provide a means for evaluating a circuit in terms of an echo rating.

It can be shown that a small cosinusoidal gain deviation over the transmission band produces a pair of symmetrical echoes of the same polarity, one leading and one lagging the signal. The relationship between the periodicity, Δf, of the transmission deviation (gain or phase) and the time displacement of the echo which it produces is $T = 1/\Delta f$. If the deviation of the gain-frequency characteristic has a coarse structure, the resulting echo displacement, T, will be small. For example, a coarse cosinusoidal gain ripple with $\Delta f = 2$ MHz will result in a pair of echoes 0.5 microsecond away from the signal, or about 0.13 inch on a television screen 17 inches wide with the standard scanning rates specified previously. A fine structure gain ripple (high periodicity) with $\Delta f = 200$ kHz would produce echoes displaced 5 microseconds, or about 1.3 inches from the signal.

Similarly, it can be shown that a small sinusoidal phase deviation over the band results in a pair of echoes, one leading and one lagging the signal, but of opposite polarity.

The examples that have been discussed are, of course, simple cases of the general problem. Usually, the transmission characteristic does not exhibit continuous sinusoidal deviations versus frequency, but is more complex. However, these more complex characteristics can usually be analyzed by Fourier methods, whereby a complex gain or phase transmission pattern may be separated into a number of sinusoidal deviations over the whole frequency band, each one of which will produce a pair of echoes.

The remaining step is to obtain an overall "rating" of the circuit. This is done by taking the root sum square value of all the weighted individual echoes to determine the magnitude of a single equivalent echo. It has been found by subjective tests that the impairment of a number of echoes, however complex, is the same as a single, well displaced (greater than 10 microseconds) echo of appropriate amplitude. The ratio expressed in decibels of the amplitude of this equivalent echo to the amplitude of the total composite video signal is defined as the echo rating of the transmission circuit.

Time Weighting. Subjective tests have demonstrated that the visual impairment due to an echo is a function of the time spacing between signal and echo. Close-in echoes tend to be masked by the signal. Those further out stand more by themselves and tend to be more annoying. The time weighting function shown in Fig. 29-12 is applicable to either leading or lagging echoes. For example, an echo 0.5 microsecond away from the signal is about 11 dB less annoying than one of the same amplitude spaced 10 microseconds

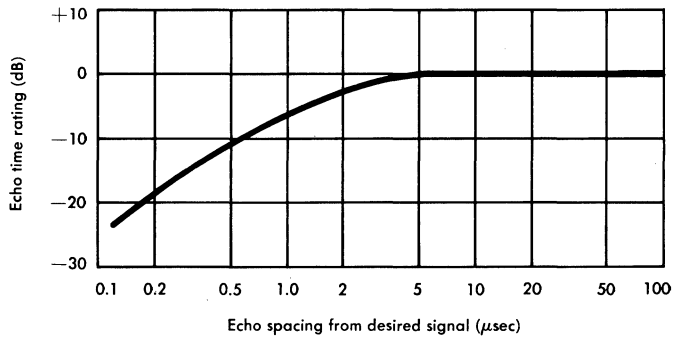


FIG. 29-12. Echo time weighting.

away. The time weighting must be taken into consideration when deriving an echo rating.

Frequency Weighting. When the transmission deviation extends only over a part of the band, the echo effect is reduced by the ratio of the bandwidths. The effect is more severe when these deviations occur at the lower video frequencies. It produces an echo which is a low definition replica of the picture, whereas the same amplitude deviation occurring at high frequencies produces echoes consisting of narrow spikes corresponding to each sharp vertical line of the picture. The spikes tend to be masked by noise and are much less objectionable than the replica. These effects are taken into account by the frequency weighting curve of Fig. 29-13. Here it can be seen, for example, that a deviation centered at about 1 MHz is about 14 dB less severe than the same deviation at 50 kHz. The absolute values of the weighting function are chosen to make the area under the curve (0 to 4 MHz) have a value of unity so that a deviation over the whole band would have 0 dB weighting. For color, however, the frequency weighting is not applicable in the vicinity of the color subcarrier frequency because of increased sensitivity in this area.

To apply the frequency weighting curve to a given sinusoidal gain or delay ripple centered at a frequency f_1 and having a bandwidth B, two factors are involved: the bandwidth factor, 10 log $B/4.2$, and the weighting factor given by Fig. 29-13. Thus, if such a

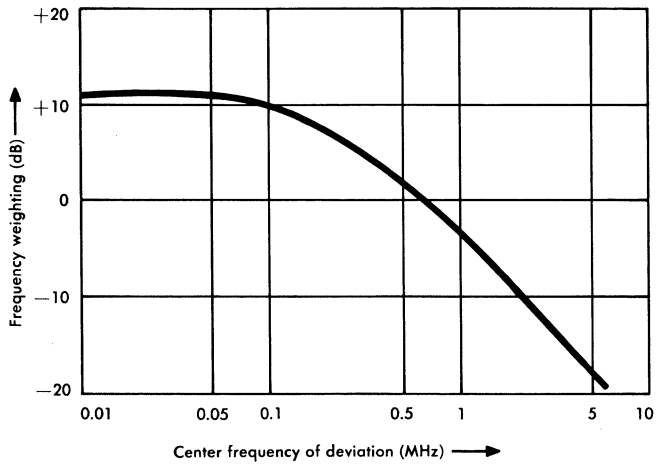


FIG. 29-13. Frequency weighting for monochrome television.

deviation is centered at 2 MHz and extends over a band B of 0.42 MHz, the bandwidth factor would be —10 dB and the frequency weighting factor would be —10 dB. The sum, —20 dB, is the amount by which the impairment due to this deviation is less severe than one which has the same amplitude and periodicity but extends over the entire bandwidth.

Echo Rating Objective. The Bell System echo rating objective for a 4000-mile television network including microwave radio and coaxial cable toll circuits, video circuits, and all switching facilities is —40 dB.

Crosstalk

Video crosstalk [5] is an important consideration when two or more video transmission systems operate on adjacent cable facilities. If coupling between systems is excessive, crosstalk from one system will seriously impair the picture transmitted by the other.

Video crosstalk produces an image of the unwanted signal moving erratically back and forth across the wanted picture. This motion occurs because of the lack of synchronism between independent video sources. As the crosstalk image moves across the main picture, it appears to be framed. This frame is formed by the horizontal and vertical blanking and is much more noticeable than any feature in the interfering image. At the threshold of interference, there is no semblance of frame or image, and only a slight flicker appears as the frame moves across some gray area of the main picture. When the coupling has a rising gain versus frequency characteristic, a differentiated crosstalk image will appear in bas relief.

The crosstalk coupling loss objective for monochrome signals between equal-level points is plotted in Fig. 29-14, in which coupling loss required at 4 MHz is plotted against the loss slope in dB per octave of the coupling path. If the coupling path from one video circuit to another is flat versus frequency, the loss required at all frequencies is 58 dB. If, however, a high loss is obtained at low frequencies, the 4-MHz objective is eased. For example, if the coupling path loss decreases at 6 dB per octave, a loss of 23 dB at 4 MHz is satisfactory; the corresponding loss at 20 kHz will then be 69 dB.

The curve of Fig. 29-14 represents lumped rather than distributed coupling. When the coupling is distributed over a long distance, as

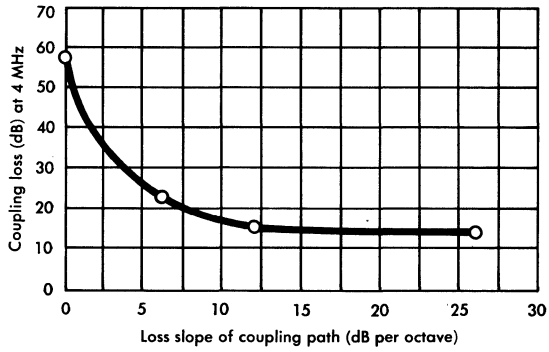


FIG. 29-14. Crosstalk objective for monochrome signals—required coupling loss at 4 MHz versus loss slope of coupling path.

in adjacent pairs, the crosstalk image will be less clear, and probably somewhat less severe requirements would be imposed.

Despite the easing of the 4-MHz coupling loss objective in going from flat to sloping coupling, it is more often the latter requirements that are the hardest to meet. This is particularly true in instances of near-end crosstalk in cable circuits carrying video signals in opposite directions. As the length of cable to the nearest repeater or terminal is increased, two things occur: (1) the magnitude of the incoming signal is decreased, and (2) the steepness of the equalization slope of the receiving amplifier is increased. The first increases the effective coupling between circuits, and the second increases the high-frequency transmission of the crosstalk signal.

Random Noise

In a video transmission system, wideband random noise is generated primarily in the input amplifiers of each repeater or receiving terminal. The subsequent equalizers and amplifiers shape this noise spectrum. Figure 29-15 illustrates the overall noise shapes for several systems used to transmit television.

Subjective tests have been made to determine the interfering effects of broad, narrow, and mixed bands of random noise distributed throughout the television band [6]. Figure 29-16 gives noise versus a frequency weighting characteristic obtained by the testing. This weighting is applicable for both monochrome and color facilities

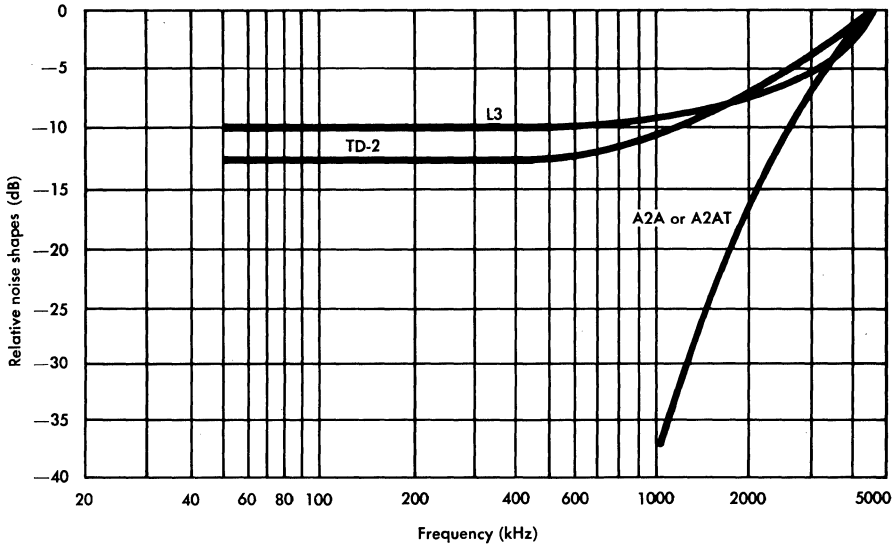


FIG. 29-15. Typical system noise shapes.

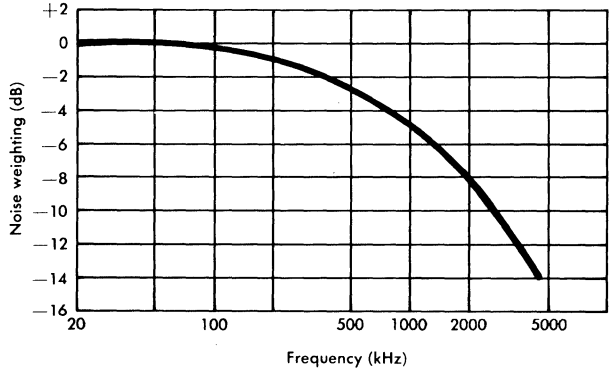


FIG. 29-16. Noise weighting characteristic.

and is used in conjunction with a simple power summing device, such as a power meter, for direct measurement of weighted noise.

The general principles derived from these subjective tests have been summarized as follows:

1. Low-frequency noise is more interfering than high-frequency noise of equal power.

2. A given amount of noise power is more objectionable if it is concentrated in a narrow band than if it is spread out over a wider band in the same frequency region.
3. Human vision in combination with present television monitors does not precisely add weighted noise powers in arriving at an overall assessment of the interfering effect of random noise bands. A reasonable compromise, however, can be obtained with weighting applied to a power meter.

In order to measure the actual signal to weighted noise ratio of a television system, a network having the loss shape of the weighting curve of Fig. 29-16 can be interposed between the output of the system and a wideband rms meter. The meter reading will then give the total weighted noise in rms volts, E_N. If the normal television signal at this point is E_S volts peak-to-peak, then the signal to weighted noise ratio in dB is defined as 20 log E_S/E_N. Here a peak-to-peak voltage, which includes the sync pulse as well as the picture signal, is compared with an rms voltage. These voltages are the values most readily measured, however, and for the video signal the peak-to-peak voltage is well defined.

The overall signal to weighted noise objective for a 4000-mile system as determined from subjective tests and use of the weighting curve of Fig. 29-16 is 53 dB.

The objective for a component video section is considerably more stringent, of course. The current allocation of the noise objective allows about 57 dB for all of the local video systems in an overall television network. If five of the local video sections are of maximum line length and hence absorb all the requirements, the objective for each maximum length section becomes $57 + 10 \log 5 = 64$ dB.

Single-Frequency Interference

The addition of an interfering sine wave to the television signal superimposes a bar pattern on the picture. If the frequency is a rational multiple of either the line scanning or field scanning fundamental, the bar pattern will be stationary. If the interfering frequency is not a rational multiple of the scanning frequency, the interfering pattern will move.

The visibility of a bar pattern, either vertical or horizontal, depends on the amplitude of the interference and on the angle which the bars subtend at the eye of the viewer. For a given viewing distance, therefore, a 4-MHz pattern is more difficult to see than a

1-MHz pattern. A high-frequency interference pattern which is nearly synchronous with a line scanning component is much more disturbing than an interfering pattern that results from a frequency which falls midway between line scanning components. This fact influenced the choice of the color subcarrier frequency of 3.579545 MHz which falls midway between the 227th and the 228th harmonics of the horizontal frequency. The interfering effect of a bar pattern is also a function of the picture background; the effect is most easily observed in the grays.

Threshold observations have been made to determine an objective for this type of interference. A complete plot of the results would show a large number of maxima and minima as the interfering frequency approaches or recedes from synchronization with multiples of the line scanning frequency. If only the most stringent threshold points are plotted, the curve of Fig. 29-17 is obtained.

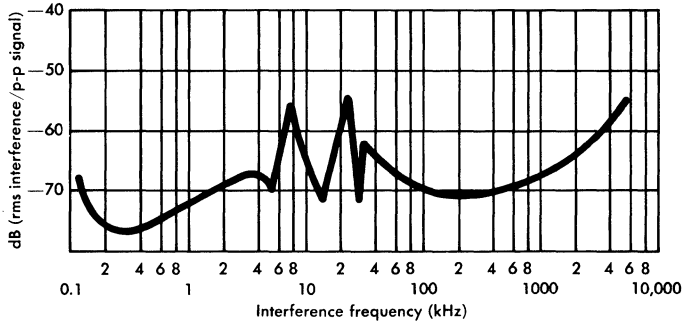


FIG. 29-17. Single-frequency thresholds—monochrome signal viewed on monochrome monitor.

Flicker. When an interfering signal [7] below 100 Hz is superimposed on a television picture (a normal scene containing highlights, shadows, and various values of gray) and the interference is just visible, it may not be noticed as a horizontal bar pattern at all, but rather as a flicker in some areas of the picture. The rate of flicker will be the beat frequency between the interfering signal and the 60-Hz field signal. This flicker is much more noticeable and disturbing than the brightness distortion caused by an interfering frequency which is synchronized with some component of the field frequency.

Viewing tests have been made to determine the tolerable level for low-frequency interference. Here again the signal-to-interference

ratio which a median observer finds to be "just perceptible" is taken as the objective. This curve has been reproduced in Fig. 29-18, and shows that the most sensitive flicker rate is in the vicinity of 5 Hz.

Another source of moving bar patterns and flicker is the modulation of a television signal (transmitted either at video or carrier frequencies) by power frequency voltages in the repeaters of a transmission system. Objectives on transmission systems with respect to this effect have been greatly eased by the use of clampers at the TV receiving terminals.

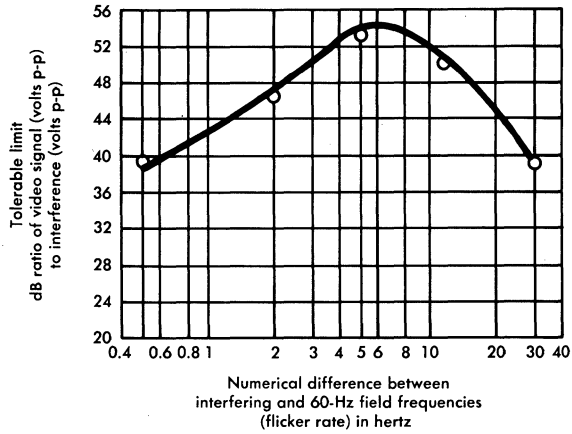


FIG. 29-18. Tolerable limits for low-frequency interferences.

Effects Due to Nonlinearities

Nonlinear distortion is contributed by the system amplifiers. It is usually evaluated in amplifier design work by measuring second and third order modulation. The odd order terms, particularly the third, contribute a fundamental frequency term which can be interpreted as a compression term. As the overload point of the amplifier in question is approached, there is no longer a 1:1 relationship between input and output, and the output is said to be compressed. In television transmission this may result in compression of the synchronizing signal and the video signal (monochrome or color). The controlling requirements on compression for the color television signal, however, are determined by the differential phase and gain performance of the transmission system. It has been established that a critical observer can detect hue changes due to differential phase

distortion of about 5° to 10° and for differential gain distortion of about 2 dB to 4 dB. These judgements are strongly dependent on picture content and shape of the distortion.

As before, these values are established for an overall transmission system. When color transmission is satisfactory, monochrome transmission will be better than just adequate as far as compression is concerned.

Summary of Objectives

The various objectives for an overall 4000-mile system for satisfactory transmission of television signals may be summarized as follows:

1. The *bandwidth* should be approximately 4.2 MHz, preferably with gentle roll-off above that frequency.
2. *Gain and delay* objectives are given in terms of the echo rating, which should not be worse than −40 dB.
3. The *weighted rms noise* should be 53 dB below the peak-to-peak video signal. Figure 29-16 gives the weighting curves for color or monochrome television.
4. *Single-frequency interference* objectives are the threshold values given in Figs. 29-17 and 29-18.
5. *Differential gain and phase* objectives are 2 dB and 5 degrees, respectively.
6. The objective for *crosstalk* flat with frequency is 58 dB. For non-flat coupling, Fig. 29-14 applies.

29.3 VISUAL TELEPHONE

Plans are under way to provide a video adjunct to the voice service now offered telephone subscribers [8, 9, 10]. This PICTUREPHONE service is primarily intended for face-to-face communications, however a means for transmitting limited resolution graphics has been included to fill business needs. Initial application will be in downtown areas of large cities, i.e., areas with high business telephone concentration. The capability of the visual telephone switched network initially will include computer access and wideband data services; eventually it may be extended to include such services as compatible color and high resolution graphics. Since this service is relatively new, the information contained in the following brief description may be subject to change.

The Video Signal

Since the video characteristics of PICTUREPHONE service are based on providing acceptable picture quality at reasonable cost over limited bandwidth transmission links, many compromises in design have been made. The picture size, 5-1/2 by 5 inches, has been chosen to display a head and shoulders view of a user with space allocated for side to side movement such as might occur during normal conversation. Two-to-one interlace is used, as in broadcast TV, with an approximate frame rate of 30 frames per second and a field rate of approximately 60 fields per second. Each frame consists of about 250 lines. To transmit this signal, approximately 1 MHz of bandwidth is required. The actual bandwidth of the video channel, excluding the station set, is flat from essentially 0 to 1 MHz.

Figure 29-19 shows a portion of the video waveform consisting of odd and even fields with synchronizing pulses and vertical blanking intervals. Special signal shaping is employed to reduce noise and interference effects incurred in the transmission network. The amplitude of the synchronizing pulse is used to set the gain of the receiver automatically to correct for variations in the low-frequency loss of the transmission channel.

Video Transmission

The general network plan for PICTUREPHONE service is shown in Fig. 29-20. Starting with the baseband output signal from the PICTUREPHONE station set, the signal remains in analog form for transmission in the local plant and is converted into digital form for transmission in the long-haul plant. Special care must be taken in the design of the local plant since the video signal in analog form is vulnerable to such impairments as deviations in gain and phase characteristics, crosstalk, interference, and noise. Keeping these impairments to acceptable levels at reasonable cost is one of the major design challenges. In allocating impairments between the loop and the toll plant, the larger portion is allowed for video loops to minimize the cost of these facilities. The allocation to short-haul trunks limits the length of baseband analog transmission at each end of the connection to about 20 miles. When these trunks exceed this length, digital transmission, which is relatively immune to transmission impairments other than the coding process, will be required to prevent further degradation of the video signal. The allocation to digital trunks allows only one encoding and decoding of the video

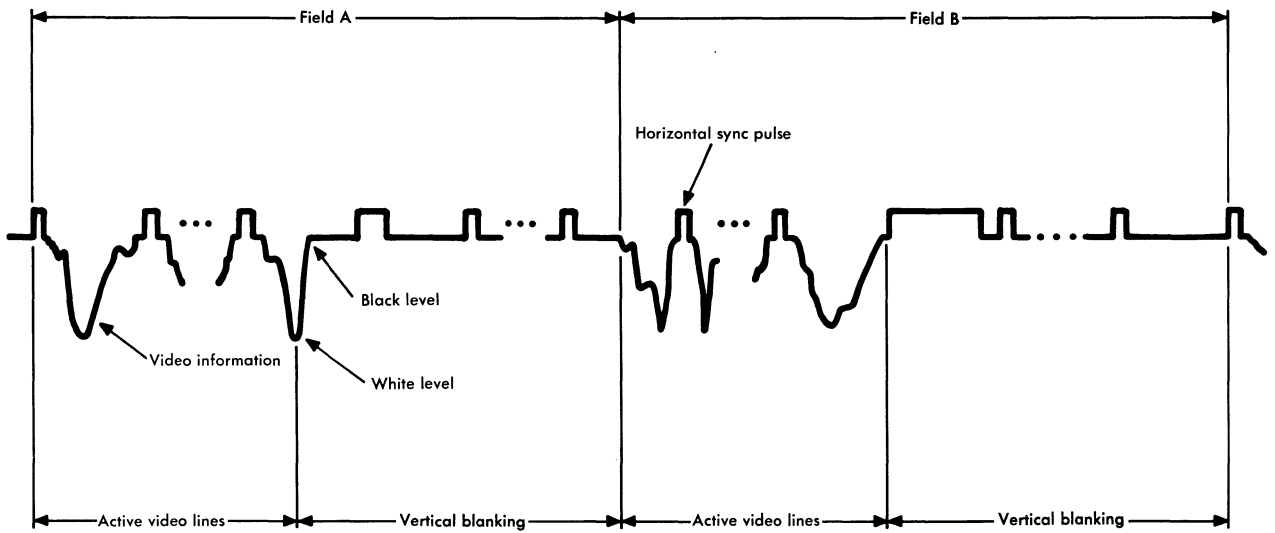


FIG. 29-19. PICTUREPHONE video signal.

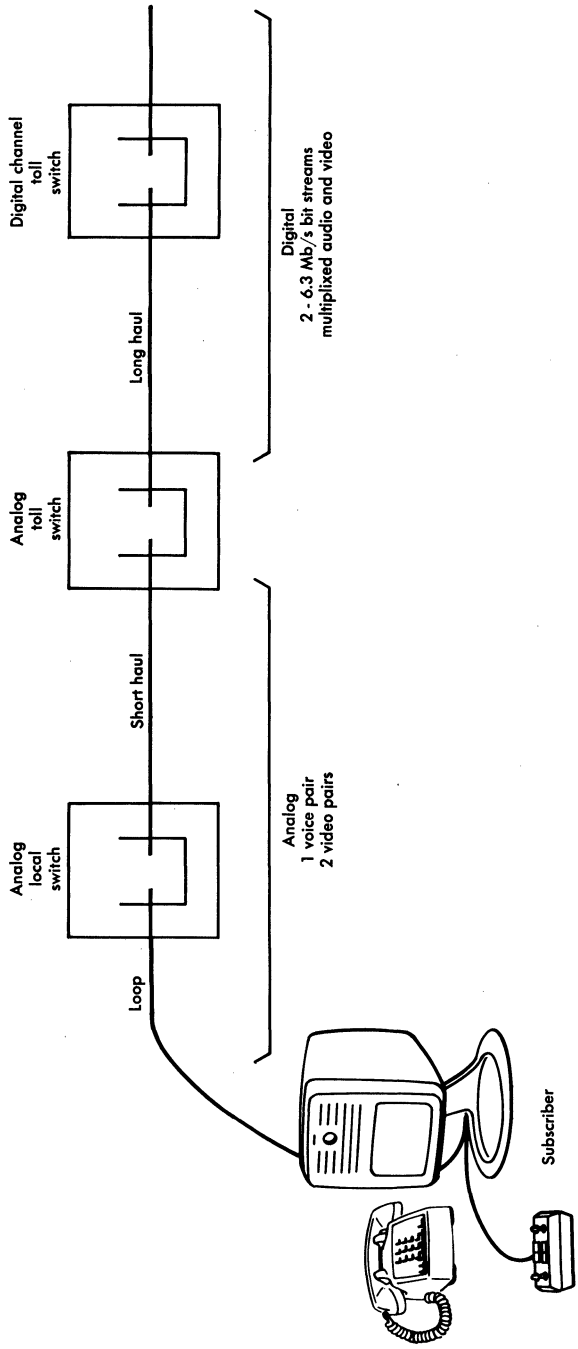


FIG. 29-20. PICTUREPHONE transmission network.

signal in an end-to-end connection. Hence, once in digital form, it is necessary to continue with this transmission mode until the signal can be delivered close enough to the distant station set so that it can be transmitted over analog trunks and loops. This digital transmission is employed in the higher levels of the switching hierarchy. The plan results in analog switching at the local office and digital bit stream switching in the toll plant. In the analog portion of the video network, nonloaded cable pairs of 19, 22, 24, or 26 gauge will be used with special equalizers spaced at about 1-mile intervals. Automatic temperature regulation will be required on all trunks and on long loops. Three cable pairs are required: one pair for each direction of video transmission and a third pair for two-way voice with the usual signaling and supervision features. Dialing will be accomplished over the voice pair, and the video will be switched at local offices by means of a video switching adjunct slaved to the voice switch.

For the digital portion of the network, the audio will be coded and multiplexed with the digital video signal to make a 6.3 Mb/s data stream by means of a digital video telephone terminal discussed in Chap. 24. One 6.3 Mb/s channel is required for each direction of transmission. Initially, short-haul digital transmission will be over T2 carrier facilities and long-haul digital transmission via TD2 radio and L4 carrier. Eventually, waveguide and PCM coaxial systems are expected to be used.

The voice transmission plan for PICTUREPHONE service will take advantage of the fact that the major portion of the telephone connection will be via the digital channel. A new plan has been developed which results in a design for loops identical to that in the DDD network and in a design for toll-connecting and intertoll trunks which unlike the VNL concept used in the DDD network produces a fixed loss from end-office to end-office, independent of the distance and number of trunks in a connection. Thus, the average loss of a voice connection on the PICTUREPHONE network is significantly lower, and the variation around this average is minimal. Echo suppression, when needed, will be accomplished by either digital or analog suppression of the audio path, depending on the type of trunk involved.

REFERENCES

1. Fink, D. G. *Television Engineering Handbook* (New York: McGraw-Hill Book Company, Inc., 1957), chapters 1 and 2.
2. Wentworth, J. W. *Color Television Engineering* (New York: McGraw-Hill Book Company, Inc., 1955).

3. Doba, S., Jr. and J. W. Rieke. "Clampers in Video Transmission," *Trans. AIEE*, pt. I, vol. 69 (1950), pp. 477-487.
4. Mertz, P. "Influence of Echoes on Television Transmission," *J. Society of Motion Picture and Television Engineers*, vol. 60 (May 1953), pp. 572-596.
5. Fowler, A. D. "Observer Reaction to Video Crosstalk," *J. Society of Motion Picture and Television Engineers*, vol. 57 (Nov. 1951), pp. 416-424.
6. Barstow, J. M. and H. N. Christopher. "Measurement of Random Video Interference to Monochrome and Color TV," *Trans. AIEE*, pt. I, vol. 81 (Nov. 1962), pp. 312-320.
7. Fowler, A. D. "Observer Reaction to Low-Frequency Interference in Television Pictures," *Proc. IRE*, vol. 39 (Oct. 1951), pp. 1332-1336.
8. Hall, A. D. "Experiments with PICTUREPHONE Service," *Bell Laboratories Record* (Apr. 1964) pp. 114-120.
9. Carson, D. N. "The Evolution of PICTUREPHONE Service," *Bell Laboratories Record* (Oct. 1968), pp. 282-291.
10. *Bell Laboratories Record*, Special Issue, (May/June 1969).

Molnar, J. P. "PICTUREPHONE Service—A New Way of Communicating," pp. 134-135.

Dorros, Irwin. "PICTUREPHONE," pp. 136-141.

Davis, C. G. "Getting the Picture," pp. 142-147.

Harris, J. R. and R. D. Williams. "Video Service for Business," pp. 148-153.

Korn, F. A. and A. E. Ritchie. "Choosing the Route," pp. 154-159.

Nast, D. W. and I. Welber. "Transmission Across Town or Across the Country," pp. 162-168.

Andrews, F. T., Jr. and H. Z. Hardaway. "Connecting the Customer," pp. 169-173.

Ekstrand, S. O. "Devices—The Hardware of Progress," pp. 174-180.

Spencer, A. E. "Maintenance—Keeping the System in Trim," pp. 181-185.

Chapter 30

Wideband Data Transmission

The network of Bell System communication facilities is designed primarily to interconnect telephones, but it is capable of transmitting other than voice signals. Data signals have been transmitted over the telephone network for many years, ranging from supervisory signals and dial pulses through data signals occupying the entire message channel bandwidth.

The network can also be adapted to transmit signals occupying bandwidths wider than a message channel to provide the transmission capability referred to as *wideband services.* This chapter deals with a specific type of wideband service known as *wideband data service.* Wideband data signaling rates range from tens of kilobits to multimegabits per second.

Some signals are continuous *(analog)* while others are quantized into discrete form *(digital).* It is interesting to note that information supplied for human interpretation is usually quantized into discrete form, for example, language in the form of speech sounds or written characters. The electrical signal generated by a telephone instrument is analog, however, as are many forms of purely machine-oriented information. Examples of the latter are continuously variable electrical currents generated by temperature or pressure transducers. Certain signals have characteristics that are both analog and digital in nature. For example, some facsimile signals are analog in time but are quantized into discrete levels.

The majority of data signals are quantized in both amplitude and time. As a result, the terms *data signals* and *digital signals* are often used synonymously. This chapter deals exclusively with the problems of transmitting wideband data signals that are quantized in amplitude but may or may not be quantized in time.

Just as signals can be classified as analog or digital, transmission systems can also be classified as analog or digital. Analog transmission systems are those whose transmission properties are predominantly linear; i.e., any nonlinearities are inadvertent by-products of physical realizability and are held to tolerably small values. Such systems comprise the overwhelming majority of present transmission facilities. Digital transmission systems are designed to transmit a finite set of symbols and classify the received signal only in terms of members of that set. Hence, these systems handle only quantized information and usually incorporate certain nonlinear characteristics.

Four permutations of signal and transmission system types may be considered for study. They are

1. Analog signals on analog transmission systems.
2. Analog signals on digital transmission systems.
3. Digital signals on digital transmission systems.
4. Digital signals on analog transmission systems.

The first two of these are discussed in previous chapters of this book. The third and especially the fourth items are principal subjects of this chapter.

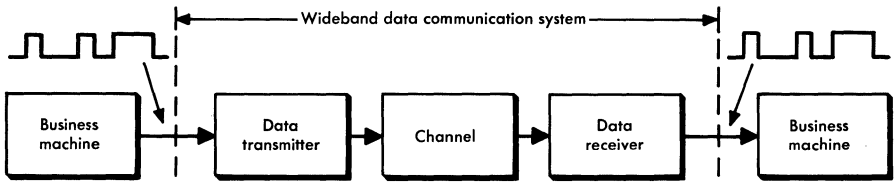


FIG. 30-1. Block diagram of data communication system.

A wideband data communication system can be represented by the block diagram of Fig. 30-1. The principal objective of the data communication system is to deliver to the receiving business machine the same time domain signal wave shape that was accepted from the transmitting business machine. The only imperfections in signals delivered to the distant business machine are probabilistic; i.e., there is a small probability that some of the binary digits will be in error. The signals are free of noise and any other form of signal distortion. (Transitions in the signal, however, may depart somewhat from ideal timing.) This objective for data transmission represents a departure from that for speech transmission in which the time domain wave

shape is often altered when it reaches the output of the receiver by bandlimiting and by phase distortion, neither of which significantly affects speech communication [1]. Similar distortions of digital signals must be controlled sufficiently to permit the receiver to reconstruct a replica of the original time domain wave shape. Since the present transmission facilities were designed principally to handle voice-frequency transmission, further conditioning is required to accommodate wideband data signals.

30.1 BANDWIDTH RESTRICTIONS

The spectral content of rectangular-shaped data signals extends over an unlimited frequency band. In analog transmission systems, however, bandwidth is a valuable resource and should be conserved. It is therefore important to minimize the required bandwidth while permitting accurate reconstruction of the data signal at the receiver. In practice, the available bandwidths are already established by the nature of the existing analog carrier systems. The problem, therefore, becomes one of determining the optimum transmitting and receiving filter shapes which will maximize the signaling rate for a given signal power, channel bandwidth, noise spectral density, and performance criterion [2].

There are numerous data signal formats which can be considered for wideband data transmission. A very efficient data signal format, and the one most commonly used, is pulse amplitude modulation (PAM). In this format the information is encoded into signal amplitudes or data symbols, a_k, which are transmitted at uniform time intervals t_0 seconds apart. A PAM signal, $x(t)$, can be represented by

$$x(t) = \sum_{k=-\infty}^{\infty} a_k f(t-kt_0)$$

where $f(t)$ is the pulse shape. After the signal, $x(t)$, is transmitted through a bandlimited channel, $H(\omega)$, the received signal, $y(t)$, becomes:

$$y(t) = \sum_{k=-\infty}^{\infty} a_k g(t-kt_0)$$

where

$$g(t) = \mathfrak{F}^{-1}[F(\omega)\, H(\omega)] \text{ and } F(\omega) = \mathfrak{F}\,[f(t)]^*$$

The receiver samples the signal at time jt_0 to determine which of the allowed symbols, a_j, was transmitted. Let $y(jt_0)$ be an estimate for the jth digit, a_j.

$$y(jt_0) = \sum_{k=-\infty}^{\infty} a_k g(jt_0 - kt_0)$$

$$= a_j g(0) + \sum_{k \neq j} a_k g(jt_0 - kt_0) \qquad (30\text{-}1)$$

The second term in Eq. (30-1) represents *intersymbol interference* which arises from the overlapping tails of adjacent pulses. The intersymbol interference will be zero for all j if, and only if, $g(it_0) = 0$ for $i \neq 0$. To satisfy this condition, the Nyquist criterion requires that the sum of the translates of the Fourier transform, $G(\omega)$, of the received pulse, $g(t)$, be a real constant [3].

$$\sum_{n=-\infty}^{\infty} G\left(\omega - \frac{2\pi n}{t_0}\right) = \text{real constant}$$

If the input pulse shape, $f(t)$, is a unit impulse, then $H(\omega) = G(\omega)$, and the minimum bandwidth channel which satisfies the Nyquist criterion consists of an ideal low-pass filter, $H_l(\omega)$, where

$$H_l(2\pi f) = 1 \qquad |f| \leq \frac{1}{2t_0}$$

$$H_l(2\pi f) = 0 \qquad \text{otherwise}$$

Impulses can be transmitted through this ideal filter at intervals of t_0 seconds without intersymbol interference, provided the signal is sampled at the peaks of the received pulses. The *symbol rate* is $1/t_0$ symbols per second, or $1/t_0$ *bauds*. The frequency $1/2t_0$ is referred to as the *Nyquist frequency*. In practice, the ideal low-pass filter can

* $\mathfrak{F}$ indicates the Fourier transform and $\mathfrak{F}^{-1}$ the inverse transform.

only be approximated by use of complicated networks. Furthermore, the performance of such a channel would be very sensitive to small distortions. As a result, the Nyquist frequency is generally selected to be less than the bandwidth of the channel in order to permit gentle shaping at the band edge. In order to satisfy the Nyquist criterion under this condition, the frequency response of the channel must consist of an ideal low-pass function plus an arbitrary function whose real part has odd symmetry and whose imaginary part has even symmetry about the Nyquist frequency. A frequently used class of Nyquist channels utilizes a *raised-cosine* amplitude characteristic [4] which is of the form:

$$H_{rc}(2\pi f) = 1 \qquad |f| \leq \frac{1-\alpha}{2t_0}$$

$$H_{rc}(2\pi f) = \frac{1}{2}\left[1 - \sin\left(\frac{t_0\pi|f|}{\alpha} - \frac{\pi}{2\alpha}\right)\right] \qquad \frac{1-\alpha}{2t_0} \leq |f| \leq \frac{1+\alpha}{2t_0}$$

$$H_{rc}(2\pi f) = 0 \qquad \text{otherwise}$$

The corresponding impulse response is

$$g(t) = \frac{\sin(\pi t/t_0)}{\pi t}\,\frac{\cos(\alpha\pi t/t_0)}{1-(2\alpha t/t_0)^2} \qquad 0 \leq \alpha \leq 1$$

The parameter α establishes the width of the *roll-off* band (that band in excess of $1/2t_0$) as illustrated in Fig. 30-2. As α decreases, less

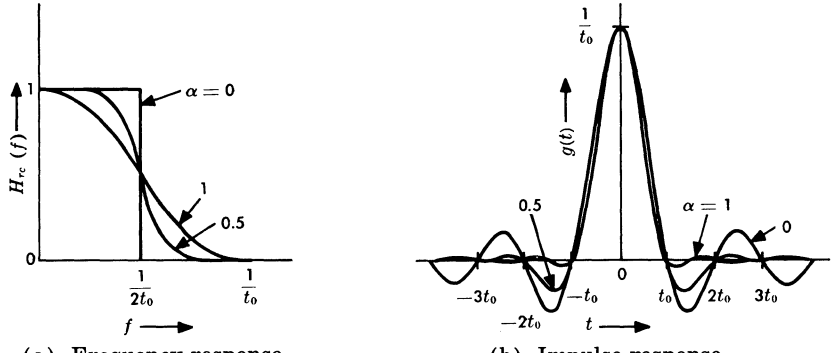


FIG. 30-2. Raised cosine pulse shaping.

bandwidth is required for a given rate of transmission; however, the tails of the response, $g(t)$, become larger, and the problems of timing and channel equalization become more difficult. An optimum choice of roll-off bandwidth and roll-off band shape can only be made after consideration of all signal distortions together with an appropriate measure of performance. One example of bandwidth optimization is discussed in Chap. 27.

If the input pulse shape, $f(t)$, is rectangular and of width t_a, the pulse spectrum is

$$F(\omega) = \frac{F(0)\sin(t_a\omega/2)}{(t_a\omega/2)} \qquad t_a \leq t_0$$

The received pulse shape, $g(t)$, may be kept the same as for the transmission of impulses in order to satisfy the Nyquist criterion by modifying the transmission characteristic of the channel to have the form $H(\omega)/F(\omega)$ instead of $H(\omega)$.

30.2 SIGNAL LEVEL AND NONLINEAR DISTORTION

The accuracy of data transmission is enhanced by maximizing the signal-to-noise (S/N) ratio in addition to optimizing channel shaping. For example, an increase in S/N ratio by as little as 1 dB can improve the error rate performance more than an order of magnitude when the controlling interference is gaussian noise. The maximum average power level at which the data signal can be transmitted depends upon the nonlinear characteristics of the analog carrier transmission system and upon the spectral characteristics of the signal. A typical channel imposes three independent restrictions on the signal power:

1. Peak signal amplitude—controlled by system overload.
2. Power spectral density—controlled by intermodulation distortion which generates modulation noise.
3. Maximum single-frequency amplitude—controlled by intermodulation distortion which generates crosstalk.

The causes and consequences of these phenomena have been discussed previously in this text. In practice, the wideband signal is generally transmitted at approximately the same average power spectral density as the speech signals displaced. The amplitude distribution for random data signals after modulation is considered no worse than that for many uncorrelated voice signals (gaussian) so that peak signal and modulation noise effects are comparable.

The single-frequency restrictions are imposed by the risk of intelligible crosstalk. For example, a single-frequency signal must be limited to −14 dBm0 in the L-type multiplex facilities if the frequency falls at or near a multiple of 4 kHz because intermodulation of a message signal with such a sinusoid may appear as undistorted speech in another channel. The requirement becomes more lenient at other frequencies because the product is shifted in frequency, which reduces the inband power and distorts intelligibility. High level sinusoids can arise from repeated pulse patterns. Special randomizing circuits can be included in data terminals to reduce the risk of intelligible crosstalk to a negligible factor. One such circuit is described on page 742.

30.3 WIDEBAND CHANNEL CHARACTERISTICS

The analog carrier systems which derive a multiplicity of wideband channels by frequency division multiplex require bandpass filters to isolate individual channels. These filters provide a steep rise in attenuation at both edges of the band which contributes large inband phase distortion. In addition, minor systematic attenuation distortion in each channel filter can accumulate to intolerable proportions in a 4000-mile circuit.

Deviation from linear phase in analog transmission systems constitutes the most damaging distortion to wideband data signals. For successful wideband data transmission, this distortion must be equalized by the addition of corrective networks. These equalizers frequently contain many network sections, which leads to residual attenuation and phase ripple across the band. If the ripple is sinusoidal, the resulting data system performance can be estimated by classical echo theory [5]. In general, simultaneous ripples in both gain and phase or ripples having several frequencies result in interactive terms which complicate the analysis. Simple addition of the echoes from the separate ripple components can give a good insight into the system performance, however. Assume that residual deviation from the desired characteristic is given by

$$H(\omega) = A(\omega)e^{-j\Phi(\omega)}$$

where $A(\omega)$ is the amplitude characteristic, and $\Phi(\omega)$ is the phase characteristic. The distorted output signal resulting from these deviations is given by

$$g(t) = \mathfrak{F}^{-1}[F(\omega)A(\omega)e^{-j\Phi(\omega)}] \qquad (30\text{-}2)$$

where $F(\omega)$ is the bandlimited input signal. Since it is only necessary to evaluate residual distortion after equalization, the magnitude of deviation from linear phase and from uniform amplitude will be small, permitting the use of certain simplifying approximations. Let

$$A(\omega) = 1 + a \cos \omega\tau$$

$$\Phi(\omega) = \omega T + b \sin \omega\tau$$

The constant, τ, is equal to the number of cycles of sinusoidal ripple across the equalized band divided by the bandwidth in hertz. For $b << 1$, Eq. (30-2) can be written as

$$g(t) \approx \mathfrak{F}^{-1}[F(\omega) \cdot (1 + a \cos \omega\tau)\ (1 - jb \sin \omega\tau)\ e^{-j\omega T}]$$

The term, $e^{-j\omega T}$, which results from the linear phase component, represents a constant delay for all frequency components and can be neglected. For a and $b << 1$, the output signal becomes

$$g(t) \approx \frac{1}{\pi} \int_0^\infty F(\omega) e^{j\omega t} d\omega - \frac{b}{2\pi} \int_0^\infty F(\omega) \cdot (e^{j\omega\tau} - e^{-j\omega\tau})\ e^{j\omega t} d\omega$$

$$+ \frac{a}{\pi} \int_0^\infty F(\omega) \cos \omega\tau\ e^{j\omega t} d\omega$$

or

$$g(t) \approx f(t) + \frac{(a-b)}{2} f(t+\tau) + \frac{(a+b)}{2} f(t-\tau)$$

A channel that exhibits amplitude ripple with linear phase ($b=0$) produces an output signal that consists of the input signal plus two echoes of amplitude $a/2$. The echoes are displaced $\pm\tau$ seconds from the input and are of the same polarity as indicated by the solid lines in Fig. 30-3. This symmetrical distortion does not alter the peak time of the pulse. For a channel with phase ripple and uniform amplitude ($a=0$), the two echoes have an amplitude $b/2$ and are of opposite polarity. In contrast to amplitude ripple, the distortion is antisymmetrical so that nearby echoes tend to shift the peak time of the pulse. For equal amplitude and phase ripple ($a=b$), the output consists of the input signal plus a single lagging echo which is twice the amplitude of the echoes which would result from the corresponding amplitude or phase ripples alone. High order ripples

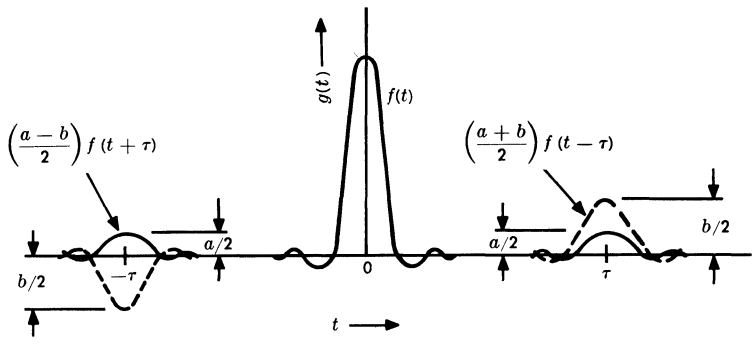


FIG. 30-3. Echoes resulting from amplitude and phase ripples.

result in distant echoes; slowly varying ripples lead to nearby echoes which may overlap the main pulse to form a single distorted pulse.

Usually, phase distortion is measured in terms of the envelope delay, which is defined as the first derivative of the phase function with respect to angular velocity, $d\Phi/d\omega$. If the phase function is

$$\Phi(\omega) = b \sin \omega\tau$$

then the envelope delay, $D(\omega)$, is

$$D(\omega) = d\Phi/d\omega = b\tau \cos \omega\tau$$

A pair of echoes of amplitude $b/2$ is associated with a sinusoidal phase deviation of amplitude b. However, the corresponding deviation in the envelope delay is dependent on the location of the echoes as well as their amplitude. For a given magnitude echo, the tolerable envelope delay distortion will be larger for remote echoes resulting from many cycles of deviation across the equalized band. Because of the ease of making envelope delay measurements rather than phase deviation measurements, most wideband channels are characterized and specified in terms of envelope delay distortion*, i.e., the departure of the envelope delay from a constant value. This is satisfactory providing the importance of the associated characteristic shapes is not ignored.

*The envelope delay is measured by applying a low-frequency amplitude-modulated carrier which is varied in frequency over the band of interest. The variations in phase of the low-frequency envelope are measured, which is equivalent to measuring the slope of small segments of the phase characteristic.

There is a practical limit as to how well a particular analog channel can be equalized with fixed networks designed on the basis of average channel amplitude and phase distortion. Further improvement in equalization can be realized, however, by the addition of various types of adjustable equalizers as described in Chap. 15. A most useful form of adjustable network for wideband data signals is the tapped delay line transversal equalizer. It can be used for the specific purpose of minimizing the probability of error of a data signal, although this goal may be somewhat at odds with the traditional one of flattening the amplitude and linearizing the phase characteristics.

This adjustable equalizer consists of a delay line tapped at intervals of t_0 seconds where t_0 is the symbol period. Each tap is connected through a variable gain element ($| G | \leq 1$) to a summing bus as indicated in Fig. 30-4. The signal is available at the center tap unaltered in amplitude, while appropriate adjustments of the other tap settings introduce echoes of the signal. Thus, in the frequency domain the tapped delay line is a finite Fourier series synthesizer that compensates for the deviation of the overall channel characteristic from the desired characteristic. The equalizer is quite effective in correcting all types of amplitude and phase deviations; it is particularly effective in correcting ripples which result in echoes that fall within $n/2$ symbol intervals of the main pulse, where

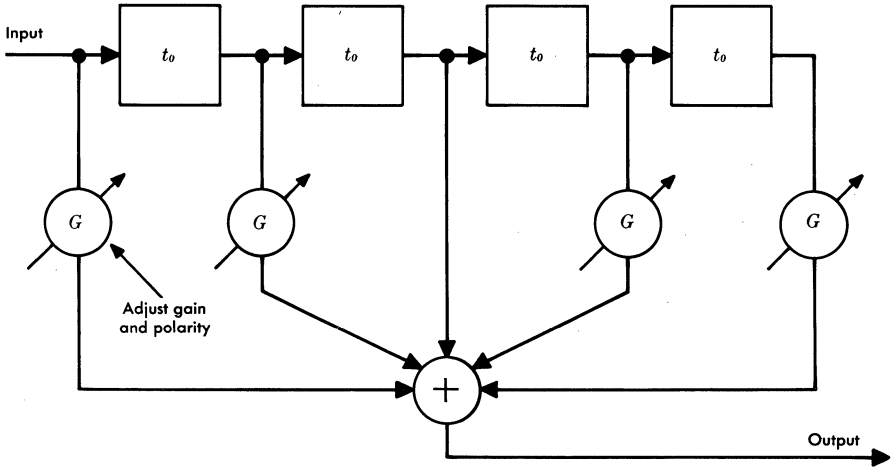


FIG. 30-4. Tapped delay line transversal filter.

n is the number of delay sections in the equalizer. The tapped delay line transversal equalizer readily lends itself to automatic adaptive operation [2].

30.4 SYSTEM NOISE

Noise imposes a fundamental limitation to the transmission of information by electrical signals. The electrical noise found in an analog communication channel generally falls into three distinct categories: gaussian, impulse, and single-frequency. Gaussian noise, because of its omnipresence and mathematical tractability, has been established as a reference impairment with which all the other impairments are compared. The effect of intersymbol interference is to make received signals more vulnerable to noise, and thus a higher S/N ratio is required for equivalent performance. The change in S/N ratio, measured in dB, which is required to give the same performance with and without the distortion is the S/N impairment.

Most of the noise power in a long-haul carrier system is approximately gaussian, but impulse noise is often controlling in a wideband data channel. The occasional large impulses which have amplitudes comparable to the signal will almost always cause errors. As a result, the system may have to be designed so that the ever present gaussian noise does not significantly add to the error rate resulting from impulse noise. Impulse noise has many natural and man-made sources, such as lightning, deep microwave fades, switching transients, and transients caused by system additions and maintenance. Such noise is not readily characterized in that the probability of encountering large amplitudes is not predictable nor is it simply related to the average noise power. Practical engineering evaluation of a particular facility is generally based on direct measurement of the number of times a critical threshold is exceeded in a test interval.

Single-frequency low-amplitude signals are often found in wideband facilities derived by frequency division multiplex. They arise from crosstalk couplings or intermodulation within the multiplex, with the fundamental source being the numerous carriers used for frequency translation. Since the frequencies of all such carriers are integral multiples of 4 kHz, the frequencies of the interferences are also 4-kHz multiples. These interferences are of little or no consequence in message service because they always correspond to zero frequency or 4 kHz in voiceband channels. Such interferences

are seldom of sufficient amplitude to effect two-level wideband data signals, but they may be quite troublesome in multilevel data transmission.

30.5 TRANSMISSION VARIATIONS WITH TIME

Amplitude and Phase

The problems associated with the design of equalizers to compensate for wideband channel amplitude and phase deviations are compounded by short-term and long-term variations in these characteristics. For example, the amplitude characteristics of transmission cables are dependent upon ambient temperature. Many long-haul cable systems require periodic adjustment to maintain transmission within specified limits. Microwave radio propagation is subject to large daily and seasonal variations. Multiple propagation paths may cause deep fades and severe phase distortion for short intervals. If residual variations of amplitude and delay distortion are a problem, some form of automatic equalization for the wideband channels may be necessary, utilizing, for example, the transversal equalizer.

Quadrature Distortion

Carrier telephone systems commonly employ single-sideband suppressed carrier amplitude modulation. The demodulating carrier at the receiving end may differ slightly in frequency from the modulating carrier. To avoid significant speech distortion, the maximum frequency offset must be less than approximately 10 Hz; however, most systems are designed for a maximum offset of 2 Hz. This *phase precession* makes the channel a time-varying system and can lead to a variable phase intercept and quadrature distortion. When there is a non-zero phase intercept at zero frequency in the baseband signal, the harmonic relationship between the signal components is affected, resulting in a distortion of the wave shape. Typical wideband data receivers incorporate phase coherent detectors which recover the proper frequency and phase of the original modulating carrier. However, a small phase error often remains which is proportional to the frequency offset of the channel. The sensitivity of the particular data signal to a constant carrier phase error sets the requirements for the design of the carrier recovery circuit.

Phase Instability

Noise interference in the generation of transmission system carrier frequencies may cause phase jitter of the received data signal. This jitter, or *incidental* FM, appears as a low-index frequency modulation. The most satisfactory method of minimizing frequency jitter has been to reduce it at the source as much as possible. Low-frequency jitter can often be mitigated, however, by proper design of the coherent detector bandwidth.

Large abrupt changes in phase may occur on wideband channels as a result of switching two carrier supplies not in phase, or switching to alternate transmission facilities having different propagation times. These changes may cause all data to be in error for the period of time required for the phase coherent detector to recognize and correct the phase of the recovered carrier. Thus, the transient response of the detector can be an important factor in overall performance.

Noise

A large proportion of long-haul facilities is derived from radio systems, and the noise may increase for short intervals due to radio propagation fading. Fading radio circuits often establish the minimum S/N ratio which may confront the wideband signal. Studies of fading phenomena indicate that about one per cent of the time noise increases as much as 11 dB in a 4000-mile system before diversity switching acts to prevent further performance degradation.

Signal Level

Virtually all modern carrier systems employ some form of gain control to maintain the flat loss of the system within specified bounds. Generally, the system requirements which are dictated by speech signal considerations are not adequate for wideband data signals. Impairments due to signal level variations may arise in two ways: (1) if the signal level is allowed to drop at the input of a link, the S/N ratio in that link will be reduced; (2) a signal level variation at the input of a threshold detector may be equivalent to a variation in the threshold level and hence decrease the immunity to noise. Each wideband digital receiver, therefore, must incorporate a precision automatic gain control circuit.

30.6 PROBABILITY OF ERROR CRITERION

Consider the polar signal $x(t)$* shown in Fig. 30-5 having peak-to-peak amplitude, A, and with a threshold detector adjusted to operate at 0 volts. The waveform must be distorted in excess of $\pm A/2$ before an error is made. If gaussian noise is present and is the only form of impairment, the *probability of error* is related to the probability of the noise component exceeding $\pm A/2$ at the threshold detector.

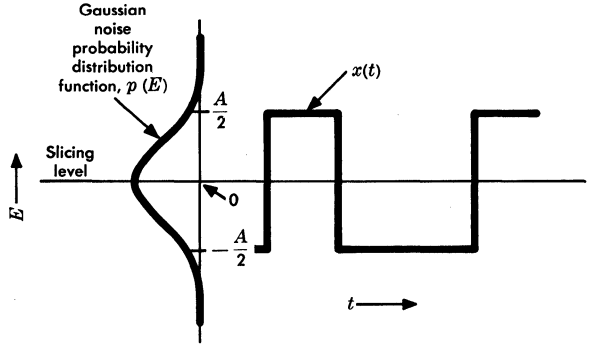


FIG. 30-5. Polar signal with gaussian noise.

The probability of making an error for a two-level signal, p_2, is found by integrating the probability distribution function for gaussian noise, $p(E)$,

$$p_2 = \int_{-\infty}^{-A/2} p(E)\,dE = \frac{1}{2}\left[1-\text{erf}\left(\frac{A}{2\sigma_n\sqrt{2}}\right)\right]$$

where $\text{erf}\,(x) = \frac{2}{\sqrt{\pi}}\int_0^x e^{-z^2}dz$, and σ_n is the rms noise voltage (Chap. 27). The error function is tabulated in many handbooks [6]. Note that the expression for p_2 also applies to a bandlimited two-level signal satisfying the Nyquist criterion if only the sampling instants are considered.

*A polar signal is one which has equal positive and negative amplitudes.

For equally probable signal levels, the average probability of error for an n level signal is [7]

$$p_n = \left(1-\frac{1}{n}\right)\left[1-\text{erf}\left(\frac{A}{2(n-1)\sigma_n\sqrt{2}}\right)\right] \qquad (30\text{-}3)$$

This expression represents the probability of error in correctly identifying a particular received level (symbol) and does not take into account any consideration of bit error rate which depends on the particular coding of the levels.

Echo Intersymbol Interference

Consider the effect of the echo distortion on two-level random data in the presence of gaussian noise. An objective for maximum tolerable amplitude of echo could be derived by recognizing that, depending on polarity, the echo may either reinforce or weaken the signal. In view of the wide variety of possible signals and controlling interferences, it seems conservative to assume that the presence of echo will always reduce the margin for distinction between pulse and no pulse, which leads to the following probability of error expression for a two-level signal with echo amplitude $a/2$:

$$p_2 = \frac{1}{2}\left[1-\text{erf}\left(\frac{A(1-a)}{2\sigma_n\sqrt{2}}\right)\right]$$

The corresponding signal-to-noise impairment in dB is:*

$$\text{Impairment} = 20\log\left(\frac{1}{1-a}\right)$$

Gaussian Intersymbol Interference

Solution of the probability of error expression for all but the most simple types of distortion is quite complex [8]. Approximate solutions can be found if certain simplifying assumptions are made. For example, it may be assumed that the probability distribution for intersymbol interference is gaussian when the sources are many and varied. There is substantial evidence to suggest, however, that this assumption invariably leads to a pessimistic result. A calculation of this type therefore represents an interesting upper bound on error

*The parameter, a, is also referred to as *eye closure.*

rate. Of course, if the system is primarily noise limited, the computed performance is largely independent of the assumed intersymbol interference distribution.

The theoretical probability of error, assuming the presence of both gaussian noise and gaussian intersymbol interference distribution, can be found in the following manner. The mean-square intersymbol interference is given by Lucky, Salz, and Weldon [9] as:

$$\sigma_{ISI}^2 = \frac{A^2\xi^2(n+1)}{12(n-1)}$$

where A is the peak-to-peak amplitude of the signal at the decision point, and n is the number of signal levels. The term ξ^2 represents the mean-square distortion in the channel impulse response and can be written as

$$\xi^2 = \sum_{i\neq 0} g^2(it_0)$$

which follows from Eq. (30-1). Thus, in Eq. (30-3), the rms noise voltage, σ_n, is replaced by σ_T where

$$\sigma_T = (\sigma_{ISI}^2 + \sigma_n^2)^{1/2}$$

and the probability of error becomes

$$p_n = \left(1-\frac{1}{n}\right)\left\{1-\text{erf}\left[8(n-1)^2\left(\frac{\sigma_n}{A}\right)^2 + \frac{2\xi^2(n^2-1)}{3}\right]^{-1/2}\right\} \tag{30-4}$$

This expression is plotted in Fig. 30-6 as a function of peak signal to rms noise ratio ($20 \log A/2\sigma_n$) for $\xi = 0$ and $\xi = 0.04$ and for various values of n.

To evaluate the performance of an existing data transmission system, it is necessary to measure the occurrence of errors since no other criterion is as meaningful. The signal can be quite badly distorted and yet still be received essentially error free if the decision thresholds are not violated. For this reason, special test equipment has been designed and manufactured to count errors directly. A pseudo-random sequence of pulses is generated at the transmitter for transmission over the channel. An identical pseudo-random

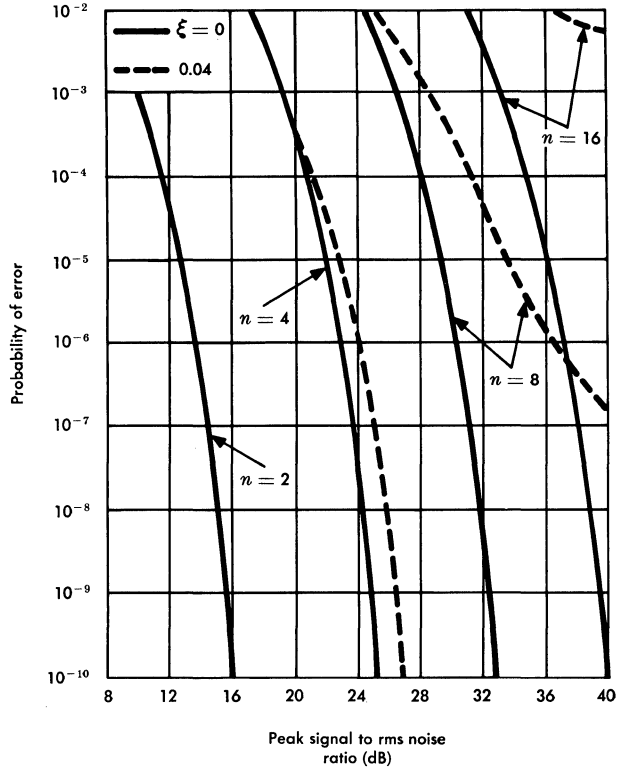


FIG. 30-6. Symbol error rate of random n-level signals with gaussian noise and gaussian intersymbol interference.

sequence or test word is generated at the receiver and it is synchronized with the received signal. Comparator circuits make a bit-by-bit error check and record the number of received bits that differ from the locally generated ones. Gaussian noise is often added to permit margin checking. The test word must have characteristics sufficiently close to a truly random sequence to give a performance measurement that is representative of normal transmission.

30.7 DATA SIGNAL CHARACTERIZATION

Most digital data signals are *synchronous;* i.e., the time scale is quantized into equal symbol intervals, and the voltage in each interval represents a specific quantity of information. Such synchronous signals have the distinct advantage of permitting intermediate signal

regeneration to prevent the accumulation of transmission distortions from one link to another. Other signals, such as black and white facsimile, are *nonsynchronous* and require the freedom to send at any time transitions which are not necessarily spaced at multiples of uniform time intervals.

Multilevel data signals permit an increase in the information rate over the two-level signals for a specified bandwidth. The maximum number of permitted levels is determined by the average noise and intersymbol interference in the channel as demonstrated by Eq. (30-4). Each level is often coded into a unique binary word. The information rate, or *bit rate,* will increase from $1/t_0$ for two-level data to a maximum of $(1/t_0)\log_2 n$ for a multilevel signal, where n is the number of levels and t_0 is the sampling interval. The bit rate is sometimes less than the maximum to gain other advantages of special coding techniques, e.g., *partial response* [10].

Data signals are generally random in nature and must be characterized in terms of their statistical properties. A property of particular interest is the power spectral density which can be computed by formulating a mathematical model for the signal.

Facsimile Signal

A model for a random facsimile signal, $x(t)$ (Fig. 30-5), is formulated by assuming a sequence of points with a random distribution in time according to a Poisson process and with an average density λ. At each point a transition may occur in the polar signal. In the successive intervals between the points, the signal may have amplitude $+A/2$ with a probability p_1, or $-A/2$ with a probability $p_2 = 1-p_1$. The values of the signal in these intervals are assumed to be statistically independent. The power spectral density is obtained by taking the Fourier transform of its autocorrelation function, $\mathcal{R}_x(\tau)$. The autocorrelation function for the random facsimile signal is [11]

$$\mathcal{R}_x(\tau) = \frac{A^2}{4} - A^2 p_1(1-p_1)\,(1-e^{-\lambda|\tau|}) \tag{30-5}$$

The power spectral density of $x(t)$ is therefore

$$\begin{aligned} S_x(\omega) &= \int_{-\infty}^{\infty} \mathcal{R}_x(\tau)\, e^{-j\omega\tau} d\tau \\ &= \frac{2\pi A^2}{4}[1-4p_1(1-p_1)]\,\delta(\omega) + \frac{2A^2\lambda p_1(1-p_1)}{\omega^2+\lambda^2} \end{aligned} \tag{30-6}$$

The probability, p_{tr}, of a transition at a potential transition point equals the probability that $x(t)$ was $+A/2$ and changed to $-A/2$ plus the probability that $x(t)$ was $-A/2$ and changed to $+A/2$

$$p_{tr} = 2p_1(1-p_1)$$

Since the average number of potential transition points per unit time is λ, the average number of actual transitions per unit time is

$$\overline{N} = \lambda p_{tr} = 2\lambda p_1(1-p_1)$$

If A is in volts and the signal is impressed on a 1-ohm line, then $S_x(\omega)$ has the dimensions power per unit bandwidth. The first term in Eq. (30-6) includes a delta function and represents the d-c power in the signal, while the second term represents the a-c power distributed over the spectrum. Note that the d-c term in any data signal contains no useful information and hence need not be transmitted.

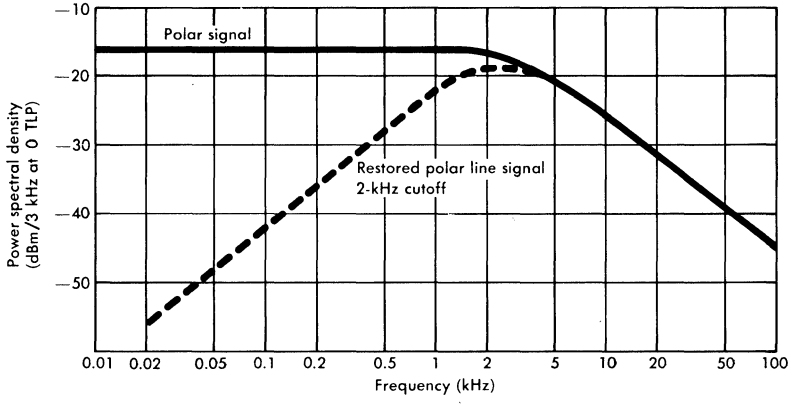


FIG. 30-7. Power density spectrum of facsimile signal.

Figure 30-7 shows the spectrum of a typical polar facsimile signal which could be transmitted in an L-type multiplex group band. The following assumptions have been made:

1. The average transition rate, $\overline{N}$, is 4000 per second.
2. Pages are 10 per cent black and 90 per cent white ($p_1=0.1$).
3. Signal power spectral density at $\omega << \lambda$ is -16 dBm0/3-kHz band.

The envelope of the power spectral density has a corner frequency at $f_c = \lambda/2\pi = 3.537$ kHz, followed by a 6 dB per octave asymptote. The total power, P_x, in the signal is

$$P_x = \frac{1}{2\pi}\int_{-\infty}^{\infty} S_x(\omega)\,d\omega$$

$$= \frac{A^2}{4}\,[1-4p_1(1-p_1) + A^2p_1(1-p_1)] = \frac{A^2}{4}$$

which is equivalent to $\Re_x(0)$.

When the facsimile signal is bandlimited by a Nyquist channel, the data receiver cannot accurately resolve transitions that occur at intervals less than t_0 seconds, where $1/2t_0$ is the Nyquist frequency. Therefore, the Nyquist bandwidth is generally selected to provide a resolution that is comparable to that of the business machine.

Synchronous Binary Signal

The following model for a random synchronous binary signal, $y(t)$, is similar to that for the facsimile signal with the exception that the points distributed in time are uniformly spaced. A transition may occur in the signal only at intervals of t_0 seconds. Let the probability that the amplitude is $+ A/2$ equal p; hence, the probability that the amplitude is $- A/2$ is $1-p$. The value of the signal in any one interval is assumed to be statistically independent of the values in all other intervals. The autocorrelation function for this signal is similar in form to that for the facsimile signal and is given by [12]

$$\Re_y(\tau) = \frac{A^2}{4} - A^2p(1-p)\,[1-q(\tau)] \qquad (30\text{-}7)$$

where $q(\tau)$ is a triangular function of width $2t_0$ given by

$$q(\tau) = 1 - \frac{|\tau|}{t_0} \qquad |\tau| \leq t_0$$

$$= 0 \qquad |\tau| > t_0$$

The power spectral density of $y(t)$ is

$$S_y(\omega) = \frac{2\pi A^2}{4}[1-4p(1-p)]\delta(\omega) + A^2 t_0 p(1-p)\frac{\sin^2\left(\frac{t_0\omega}{2}\right)}{\left(\frac{t_0\omega}{2}\right)^2} \tag{30-8}$$

Figure 30-8 shows the power spectrum for a polar signal corresponding to the parameters $t_0 = 20$ microseconds and $p = 0.5$. The total power, P_y, in the signal is

$$P_y = \mathcal{R}_y(0) = \frac{A^2}{4}$$

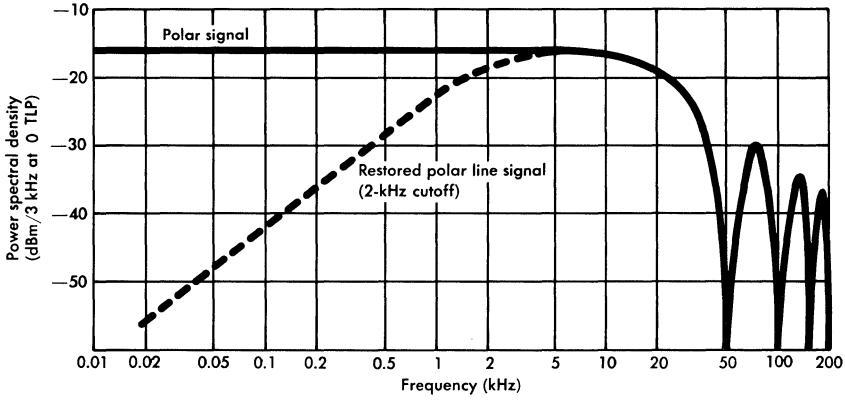


FIG. 30-8. Power density spectrum of synchronous polar signal.

Restored Polar Line Signal

It is common practice to alter the baseband signal in such a way as to remove the low-frequency components. This relieves the difficult requirement of providing transmission channels with good response down to zero frequency. The design of local loop baseband repeaters and line equalizers can be greatly simplified by the use of line coupling transformers. Suppressing the low-frequency energy in the data signal also facilitates the design of the coherent detector in the receiver of a carrier channel by providing a guard space about the carrier frequency. A low-level carrier pilot can be transmitted

so as to provide the necessary reference information to permit recovery of the correct frequency and phase of the demodulating carrier. In addition, the low-frequency reduction can greatly reduce the maximum signal power on the line (after modulation). This is particularly true for facsimile signals in which long periods of all black signal might correspond to a continuous transmission of the carrier frequency. The addition of a high-pass filter approaches the desirable condition of transmitting transitions only, and the percentage of the copy that is black or white is no longer significant. The line power is now related to the time-density of transitions. The name *restored polar* has been used to describe this modified baseband signal. The missing low-frequency components can be restored in the receiver by quantized feedback as indicated in Fig. 30-9. The d-c reinsertion circuit is a low-pass RC configuration having a cutoff frequency to match the one at the transmitter [13]. The principle applies equally well to bandlimited or nonbandlimited signals. The slicer output is a two-level signal similar to that produced by the business machine at the transmitter.

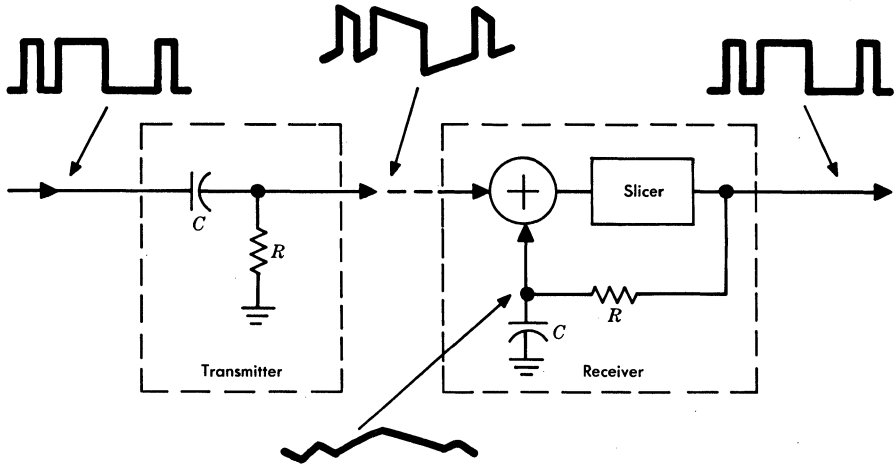


FIG. 30-9. Principle of d-c restoration.

The effects of high-pass filtering on the random facsimile signals and synchronous binary signals characterized previously are derived as follows. Let $x(t)$, the random facsimile signal with power density spectrum $S_x(\omega)$, Eq. (30-6), be passed through the linear single-stage

RC filter having the transfer function:

$$H(\omega) = \frac{j\omega}{\alpha + j\omega}$$

where $\alpha = 1/RC$ radians/second. Then the power density spectrum of the restored polar signal, $x_0(t)$, is:

$$S_{x_0}(\omega) = S_x(\omega) \mid H(\omega) \mid^2 = \frac{2A^2\lambda\omega^2 p_1(1-p_1)}{(\omega^2+\alpha^2)(\omega^2+\lambda^2)} \qquad (30\text{-}9)$$

The accompanying autocorrelation function is

$$\Re_{x_0}(\tau) = \frac{A^2 p_1\lambda(1-p_1)(\alpha e^{-\alpha \mid \tau \mid} - \lambda e^{-\lambda \mid \tau \mid})}{\alpha^2 - \lambda^2}$$

Thus, the average power of the signal is

$$P_{x_0} = \Re_{x_0}(0) = \frac{A^2 p_1(1-p_1)\lambda}{\alpha + \lambda}$$

whereas from the second term of Eq. (30-5) the average power of the a-c component of input signal $x(t)$ is

$$P_{x_{ac}} = A^2 p_1(1-p_1)$$

Therefore the power loss in dB due to the filter is

$$\text{dB loss} = -10 \log \frac{P_{x_{ac}}}{P_{x_0}} = -10 \log(1+\alpha/\lambda)$$

Equation (30-9) is plotted in Fig. 30-7 for the case where $\alpha = 4\pi \times 10^3$ radian/second.

For the synchronous binary signal, $y(t)$ is characterized by Eqs. (30-7) and (30-8). Where $p = 1 - p = 0.5$, Eq. (30-7) becomes

$$\Re_y(\tau) = A^2/4\, q(\tau)$$

and Eq. (30-8) becomes

$$S_y(\omega) = \frac{A^2 t_0}{4} \frac{\sin^2\left(\frac{t_0\omega}{2}\right)}{\left(\frac{t_0\omega}{2}\right)^2}$$

Passing $y(t)$ through high-pass filter $H(\omega)$ yields the restored polar signal $y_0(t)$, whose characteristics are:

$$\mathcal{R}_{y_0}(\tau) = \frac{A^2}{4\alpha t_0}\left[e^{-\alpha|\tau|} - \frac{1}{2}\left(e^{-\alpha|\tau+t_0|} + e^{-\alpha|\tau-t_0|}\right)\right] \tag{30-10}$$

and

$$S_{y_0}(\omega) = \frac{A^2 t_0 \omega^2 \sin^2\left(\frac{t_0\omega}{2}\right)}{4(\omega^2+\alpha^2)\left(\frac{t_0\omega}{2}\right)^2} \tag{30-11}$$

From Eq. (30-7) the average a-c power in $y(t)$ is

$$P_{y_{ac}} = \frac{A^2}{4}$$

and from Eq. (30-10) the average power in $y_0(t)$ is

$$P_{y_0} = \frac{A^2}{4\alpha t_0}\left(1-e^{-\alpha t_0}\right)$$

The dB power loss due to the filter is therefore

$$\text{Loss} = -10 \log \frac{P_{y_{ac}}}{P_{y_0}} = -10 \log \left(\frac{\alpha t_0}{1-e^{-\alpha t_0}}\right)$$

Equation (30-11) is plotted in Fig. 30-8 for $\alpha = 4\pi \times 10^3$ radians/second as before.

30.8 SYSTEM DESIGN

The following design example illustrates many of the data transmission principles discussed previously in this chapter. The assumed channel characteristics of the short-haul and long-haul carrier facilities are somewhat idealized in order to simplify the computations.

System Objectives

The design objectives listed below are patterned after existing half-group, group, and supergroup standard wideband systems.

1. The data system will utilize the 48-kHz group band of the long-haul carrier and should be capable of transmitting synchronous or nonsynchronous, two-level data at any bit rate up to 50 kb/s.
2. Transmission facilities will consist primarily of existing analog systems; however, the T1 digital system is available and is included in the example.
3. Introduction of wideband data services must not degrade the performance of other services sharing the same facility.
4. Coordination or control information required between business machines must also be handled.
5. The average error rate should not exceed one error per million bits of random data 95 per cent or more of the time.

Figure 30-10 shows a possible coast-to-coast wideband connection, although all carrier systems shown will not necessarily appear in every circuit. This connection will be used as a model to define the allocation of transmission impairments in order to yield a high standard of performance for a large percentage of all possible connections. The circuit shown does not include the maximum lengths at which satisfactory transmission is desired, but rather a length beyond which that kind of transmission system would seldom be required for wideband service.

The data set transmitter converts the two-level input signal from the business machine into a restored polar format. In addition to having desirable spectral characteristics, this signal format permits either synchronous or nonsynchronous operation as specified above. For the local loop, a baseband repeater system is needed to transmit the signal over ordinary telephone cable pairs at distances up to 20 miles. It is needed to span distances from customers' premises to the nearest telephone office having wideband access to the carrier transmission facilities. The N carrier transmission system was designed to accommodate 12 or 24 voice-frequency channels and is used on short-haul routes up to 200 miles. The T1 carrier system utilizes pulse code modulation and time division multiplexing techniques to provide short-haul transmission of 24 voice-frequency signals. The lengths of T1 carrier systems range from about 10 to 25 miles, although longer routes are possible. The long-haul transmission plant is made up of L-type multiplex (LMX) terminals together with repeatered coaxial cables and microwave radio channels.

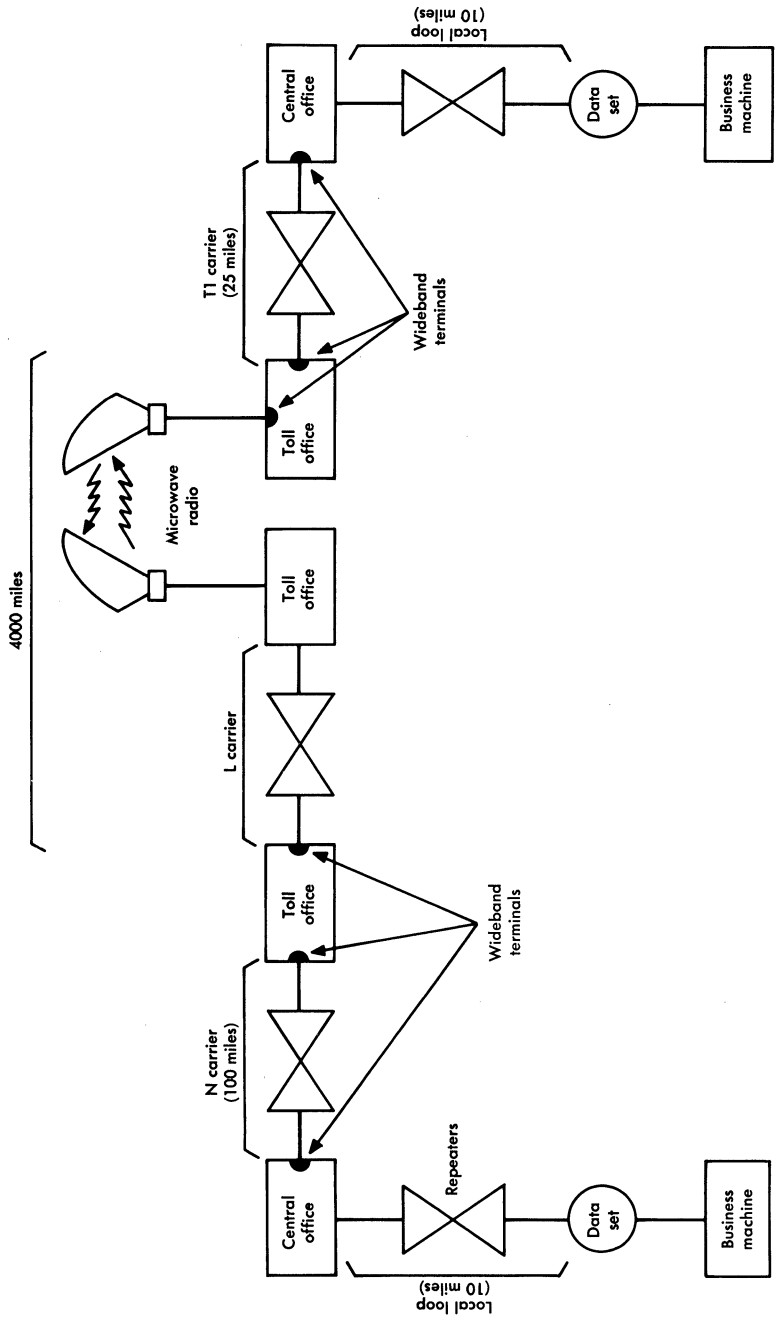


FIG. 30-10. Wideband transmission network model.

The baseband signal must be recovered at the output of each of the transmission facilities depicted in Fig. 30-10. Thus, the data set will be able to operate directly into any one of them, and furthermore, the order of interconnection of the various systems will be unimportant.

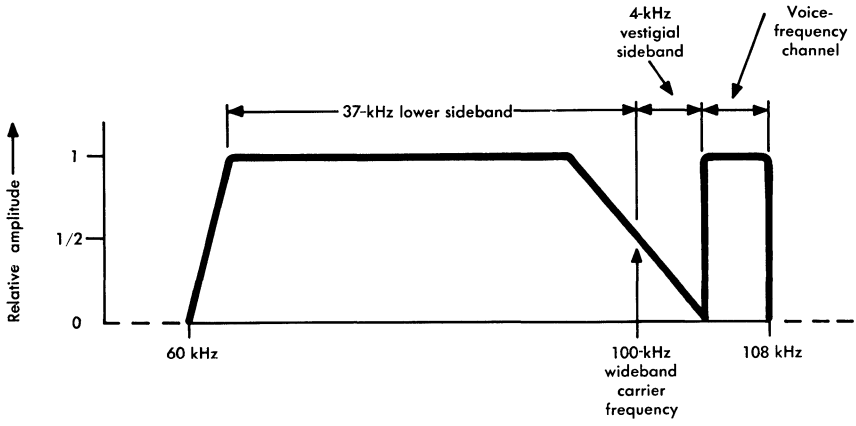
L-Type Multiplex Wideband Modem

In this example, the wideband modem for the LMX can utilize vestigial sideband, suppressed carrier AM, and coherent demodulation. Vestigial sideband operation permits optimum utilization of the available bandwidth. Suppressing the carrier results in maximum signal-to-noise performance since the long-haul transmission systems impose limitations on both total and single-frequency power. The basic group band extends from 60 to 108 kHz. The available bandwidth is somewhat restricted by a 104.08-kHz pilot which is used for automatic gain regulation of the LMX broadband terminals. The frequency band between the pilot and the 108-kHz band edge is available, however, for a voice-frequency coordination channel. The voice-frequency channel will permit coordination and control of business machine operation by voice communication or by termination of the voice channel in a data set to handle automatic control signals. A wideband modem carrier frequency of 100 kHz will divide the band into a 4-kHz vestigial upper sideband and a 37-kHz lower sideband, as indicated in Fig. 30-11(a).

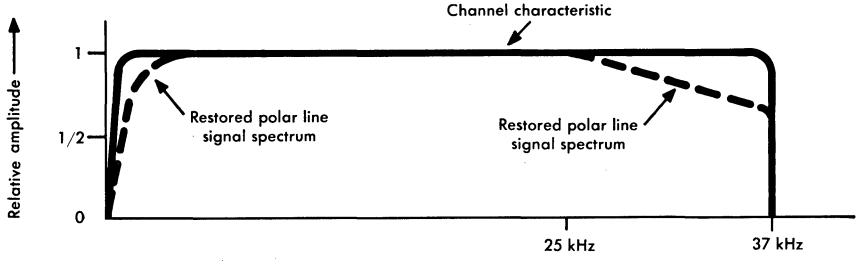
The resulting baseband channel characteristic is shown in Fig. 30-11(b). The 37-kHz bandwidth will permit a raised cosine roll-off band of 12 kHz. The square band characteristic exhibited by the modem contributes no inband shape and therefore will permit several carrier links to be connected in tandem as required by the network model of Fig. 30-10.

N Carrier Wideband Modem

The N carrier system is not single-frequency power limited; therefore, in this example, the wideband modem can utilize vestigial sideband, transmitted carrier AM, and envelope detection. The wideband signal will consist of a 40-kHz lower sideband and a 12-kHz vestigial sideband, as indicated in Fig. 30-12. In addition, there is sufficient bandwidth for two double-sideband voice-frequency signals, together with their transmitted carriers.



(a) Group frequency



(b) Baseband

FIG. 30-11. Frequency allocation of 50-kb transmission system.

T1 Carrier Wideband Terminal

The T1 digital transmission system with a bit capacity of 1.544 Mb/s provides opportunities for the transmission of a variety of high-speed data signals. A T1 repeatered line should have a much better error rate than the nominal requirement of one error per 10^6 pulse positions for wideband data transmission. Of primary concern is the determination of the minimum sampling rate necessary to reproduce the wideband data signal adequately. This rate is established by the maximum amount of time quantization jitter that can be tolerated. Thus, the sampling rate must be set sufficiently high to provide adequate performance for facsimile, which is the data signal most affected by time quantization since jitter in a

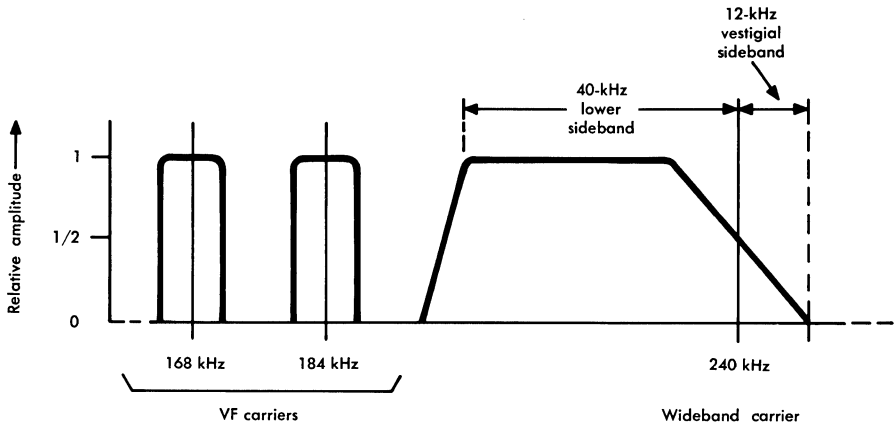


FIG. 30-12. N-carrier, basic high group 50 kb/s channel.

synchronous signal can be removed by regeneration. Acceptable performance requires that the maximum timing error due to quantization be about 10 per cent or less of the minimum data bit interval which corresponds to a sample rate about five times the maximum data bit rate. Subjective evaluation of an asynchronous facsimile signal has substantiated that this sampling rate is sufficient.

Signal Level

An important limitation in the overall performance of the long-haul analog transmission system is the intermodulation noise that is proportional to total signal loading. To avoid degradation of other services, the wideband data system should transmit no more average signal power than the message load displaced, which, for the group band, corresponds to approximately —5 dBm0. This can be allocated between the wideband signal at —5.5 dBm0 and the carrier pilot at —14.5 dBm0.

Discrete frequencies appear in the transmitted signal as a result of repetitive sequences of input data. In synchronous transmission, the amplitude of single-frequency components can be reduced to negligible proportions by a scrambler. This can be accomplished by modulo 2 addition of a quasi-random bit sequence and the data train [14]. At the receiving end, the descrambler will perform the inverse operation to recover the original data train.

The method used to synchronize scrambler and descrambler is shown in Fig. 30-13. The scrambler and descrambler both consist of an n-stage shift register with feedback from stages j and n. It can be seen that the bit sequences entering the shift registers at the scrambler and descrambler are identical. Consequently, the descrambler becomes synchronized with the scrambler as soon as the descrambler shift register is filled.

The data train sequences may be expressed as follows:

$$A_m = (B_m \oplus A_{m-j} \oplus A_{m-n})$$

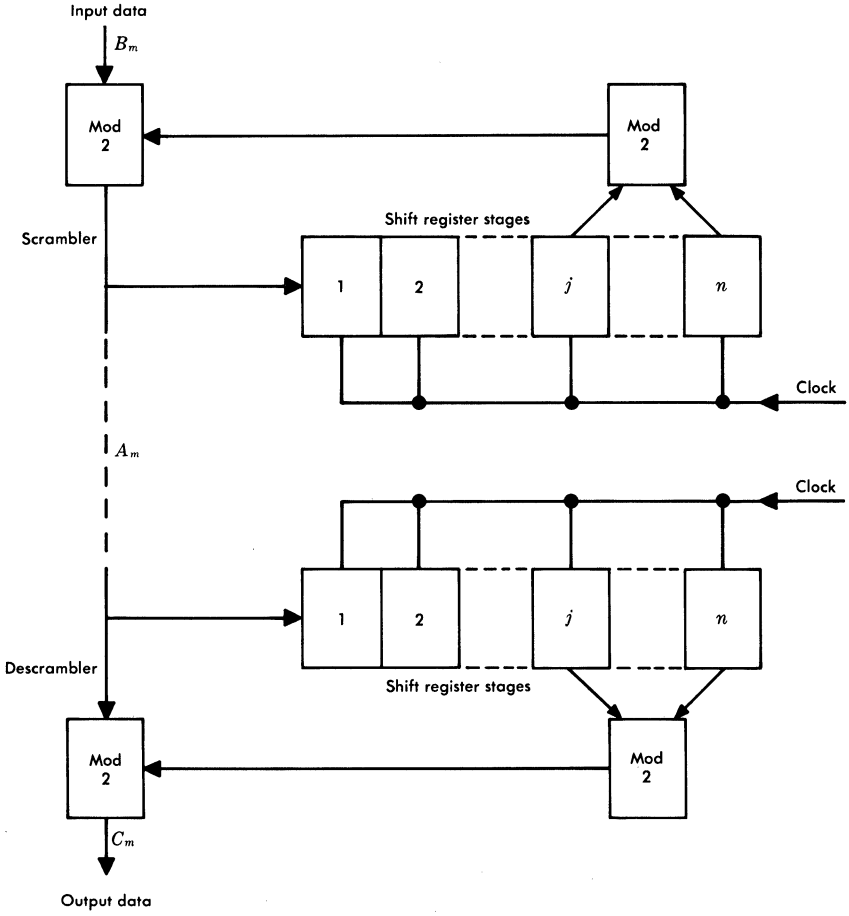


FIG. 30-13. Scrambler and descrambler block diagram.

where B_m is the data input, A_m is the scrambled line sequence, and $\oplus$ indicates modulo 2 addition. The descrambled data train, C_m, is then shown to be identical with the original data train, B_m:

$$C_m = (A_m \oplus A_{m-j} \oplus A_{m-n})$$
$$= (B_m \oplus A_{m-j} \oplus A_{m-n} \oplus A_{m-j} \oplus A_{m-n}) = B_m$$

Even though most commonly used repetitive input sequences are properly broken up, there are still a few inputs that result in undesirable line sequences. A data input sequence of alternating 0's and 1's, for example, if started at a time when the scrambler shift registers are in alternate 1 and 0 states, would produce a sinusoid at the Nyquist frequency with a power approximately equal to the power of a random data signal.

Figure 30-14 shows a monitor logic circuit that can be applied to the scrambler to guard against most of the troublesome cases. The NAND gate A senses an all 0's state in the shift registers and forces

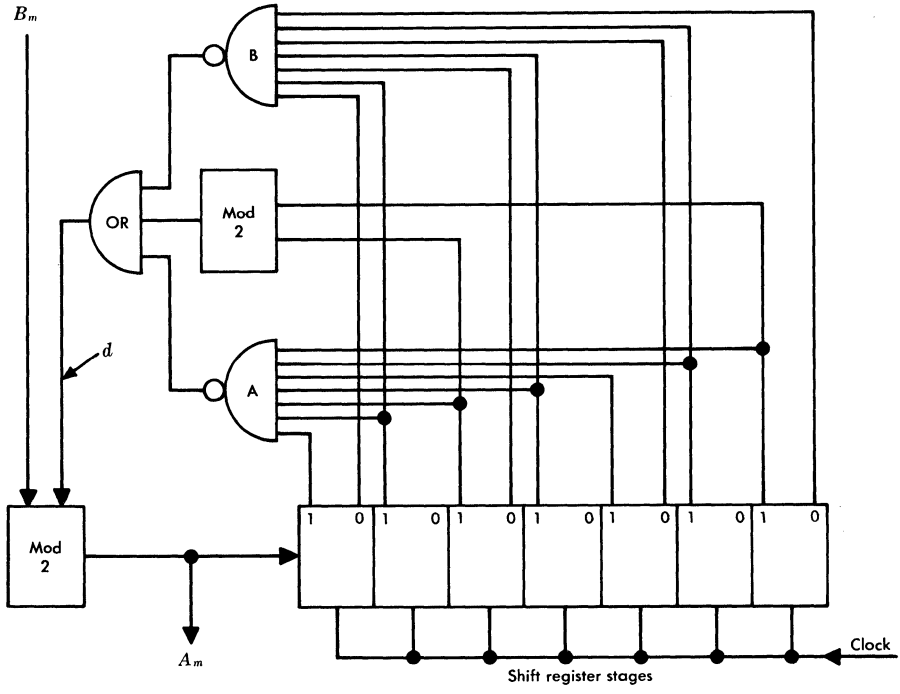


FIG. 30-14. Scrambler with monitor logic.

a 1 at d to prevent the possibility of an all 0's line signal. Similarly, NAND gate B senses an alternate 1, 0 state in the registers and forces a 1 at d to avoid the possibility of a prolonged alternate 1, 0 train for A_m. The descrambler contains the identical monitor logic and, consequently, provides appropriate descrambling. (Additional circuitry is required if it also becomes necessary to guard against the all 1's condition.)

Signal-to-Noise Margin

For this example, the mean value of noise is assumed to be —45 dBm0 per 3-kHz band for a 4000-mile system. The data signal displaces 37 kHz, resulting in an average noise power of —45 +10 log 37/3, or —34 dBm0. The S/N ratio is therefore 34 — 5.5, or 28.5 dB. On the assumption that impulse noise may often be controlling, it is desirable that errors caused by random noise be an order of magnitude less than the objective for error rate. A probability of error less than 10^{-7} for two-level data requires at least a 14.4-dB peak signal to rms noise ratio as indicated in Fig. 30-6.

The probability of error for an n-level signal with optimum filtering and raised cosine roll-off is [15]

$$p_n = (1-1/n)\left\{1-\text{erf}\left[\frac{3P_S}{(n^2-1)P_N}\right]^{1/2}\right\}$$

where P_S is the average signal power from the line and P_N is the average noise power measured in a bandwidth equal to the symbol rate. Therefore, from Eq. (30-3) for a two-level signal,

$$\frac{P_S}{P_N} = \left(\frac{A}{2\sigma_n\sqrt{2}}\right)^2$$

and

$$10 \log \frac{P_S}{P_N} = \left(20 \log \frac{A}{2\sigma_n}\right) - 3 \qquad \text{dB}$$

Thus, the P_S/P_N ratio which will also give a 10^{-7} error rate is 11.4 dB. Since the effective bandwidth of the long-haul facility is 0.75 times the symbol rate, the actual requirement for the S/N ratio becomes 11.4 + 10 log 4/3, or 12.7 dB. The margin for other impairments is thus 28.5 — 12.7, or 15.8 dB.

Signal-to-Noise Impairments

Since a large proportion of long-haul facilities is derived from radio systems, it is important to consider the increased noise that may prevail for short intervals due to radio propagation fading. Studies indicate that about one per cent of the time, noise may increase by as much as 11 dB in a 4000-mile system; this corresponds to 2.3σ if a normal distribution is assumed. In order to maintain the desired error rate performance at least 95 per cent of the time as stated in the system objectives, a minimum S/N margin of 7.8 dB must be maintained after all other impairments are considered. The other impairments, therefore, must not degrade the system more than 15.8—7.8, or 8.0 dB.

Laboratory tests indicate that a restored polar data set will contribute in the order of 1.5 dB of S/N impairment. This results from imperfect quantized feedback in the slicer and minor errors in the roll-off band shape and timing recovery circuitry. An additional 1.5 dB of impairment will result from the jitter in the time of data transitions introduced by the T1 system.

Net loss variations in most individual links will be within 1 dB of nominal. Since noise will generally accumulate uniformly with distance, the net effect on noise variations is about half that of signal deviation, or 0.5 dB. The signal-to-noise impairment can thus be characterized as having a mean value of 0 and a σ of 0.5 dB. In order to assure satisfactory operation at least 95 per cent of the time, an S/N impairment of 1 dB is assumed.

In the worst case the impairments from independent sources might add linearly, but in other instances the impairments could be complementary and cancel. The simplest method of dealing with them is to assume that they add in quadrature, i.e., on a root sum square basis. This is consistent with the goal of having satisfactory performance for 95 per cent of all situations. Any impairment measured in dB at a particular probability of error can be associated with an equivalent eye closure, a, where the impairment is 20 log $[1/(1 - a)]$. The total effective eye closure, a_t, for n independent sources becomes

$$a_t = \left(\sum_{i=1}^{n} a_i^2 \right)^{1/2}$$

Thus the combined eye closure for data set, T1 carrier, and net loss becomes 0.25, and the corresponding impairment is 2.5 dB. If the

remaining impairment is allocated to amplitude and phase distortion, it can be divided equally among the five transmission system links given in Fig. 30-10. The maximum eye closure per link that can be tolerated becomes 0.18 which corresponds to approximately 1.7 dB of impairment. Since deviations in similar systems are often systematic, impairments in the two local loops can be added linearly as can impairments in the two links of L carrier. An objective for the degree to which envelope delay distortion should be equalized in the L carrier facilities can be determined by considering the residual distortion after application of corrective networks. The primary source of delay distortion is the group connector filters. The addition of a delay equalizer to each connector with, for example, 13 delay sections results in fine-grain attenuation and delay ripple which has therefore 13 cycles within the 48-kHz band. Paired echo theory can be used to relate the allotted 1.7 dB of S/N impairment in the time domain to limitations on residual distortion in the frequency characteristic. The resulting requirements for an echo amplitude of 0.09 are approximately ± 0.8 dB of attenuation ripple or ± 25 microseconds of envelope delay ripple per L carrier link.

Multiple section delay equalizers are not required in the local loop or N carrier facilities; however, distortion can arise in the form of amplitude and delay slope. Echo theory computations for close-in echoes due to this type of distortion can be misleading. An echo which is displaced a symbol period or less primarily affects the pulse that produced it. The effect is to alter the apparent net loss and flat delay of the circuit, both of which can often be compensated by minor circuit adjustments. Thus, the presence of a close-in echo is not as damaging as a remote echo of equal amplitude. As a result, the expected S/N impairment for linear delay and amplitude distortion often must be determined with the aid of a digital computer.

This design example points up the important sources of impairments in a wideband network and also their relative effect on probability of error performance although the transmission characteristics and computations are somewhat simplified. As a result, the impairments that have been allocated to amplitude and phase distortion are larger than those normally encountered in the actual analog transmission systems. Additional analyses would also be required to establish the effects of other parameters such as timing error, incidental FM, and roll-off bandwidth.

REFERENCES

1. Lummis, R. C. "The Secret Code of Hearing," *Bell Laboratories Record* (Sept. 1968).
2. Lucky, R. W., J. Salz, and E. J. Weldon, Jr. *Principles of Data Communication* (New York: McGraw-Hill Book Company, Inc., 1968).
3. Nyquist, H. "Certain Topics in Telegraph Transmission Theory," *Trans. AIEE*, vol. 47 (Apr. 1928), pp. 617-644.
4. Lucky, R. W., J. Salz, and E. J. Weldon, Jr. *Principles of Data Communication* (New York: McGraw-Hill Book Company, Inc., 1968), p. 50.
5. Wheeler, H. A. "The Interpretation of Amplitude and Phase Distortion in Terms of Paired Echoes," *Proc. IRE*, vol. 27 (June 1939), pp. 359-385.
6. Abramowitz, M. and I. A. Stegun. *Handbook of Mathematical Functions* (National Bureau of Standards, 1964).
7. Bennett, W. R. and J. R. Davey. *Data Transmission* (New York: McGraw-Hill Book Company, Inc., 1965), p. 115.
8. Saltzberg, B. R. "Intersymbol Interference Error Bounds with Application to Ideal Bandlimited Signaling," *IEEE Transactions on Information Theory*, vol. IT-14, no. 4 (July 1968).
9. Lucky, R. W., J. Salz, and E. J. Weldon, Jr. *Principles of Data Communication* (New York: McGraw-Hill Book Company, Inc., 1968), p. 76.
10. Kretzmer, E. R. "Generalization of a Technique for Binary Data Communication," *IEEE Trans. on Communication Technology*, vol. COM-14 (Feb. 1966).
11. Franks, L. E. "Power Spectral Density of Random Facsimile Signals," *Proc. IEEE*, vol. 52, no. 4 (Apr. 1964).
12. Franks, L. E. *Signal Theory* (Englewood Cliffs, N. J.: Prentice Hall, 1969), Chapter 8.
13. Bennett, W. R. "Synthesis of Active Networks," *Proc. Polytech. Inst. Brooklyn Symp. Series*, vol. 5, *Modern Network Synthesis* (Apr. 1955), pp. 45-61.
14. Caldwell, S. H. *Switching Circuits and Logical Design* (New York: John Wiley and Sons, 1965), pp. 667 and 668.
15. Bennett, W. R. and J. R. Davey. *Data Transmission* (New York: McGraw-Hill Book Company, Inc., 1965), p. 116.

INDEX