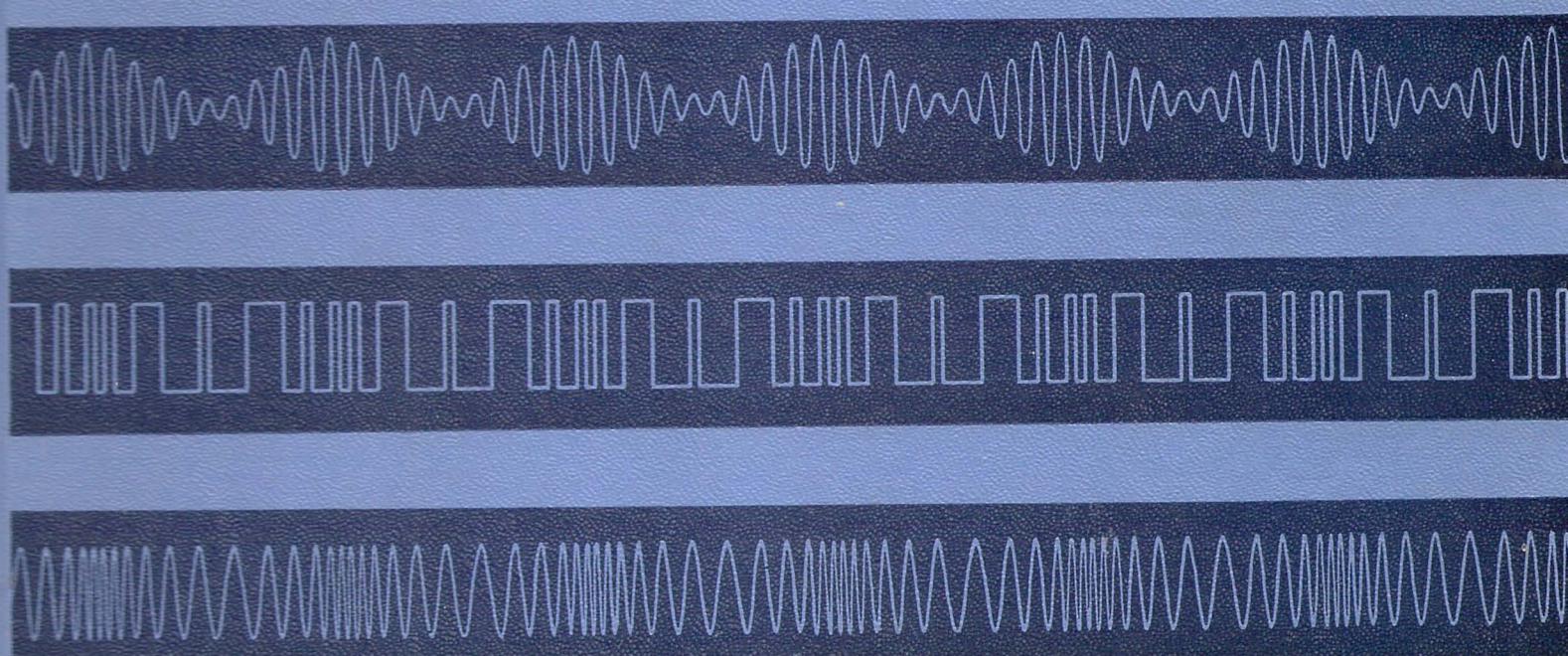
VOLUME II

TRANSMISSION SYSTEMS FOR COMMUNICATIONS



BELL TELEPHONE LABORATORIES

TRANSMISSION SYSTEMS FOR COMMUNICATIONS

By Members of the Technical Staff
Bell Telephone Laboratories

VOLUME II

Bell Telephone Laboratories, 463 West Street, New York, N. Y.

SUMMARY OF CHAPTERS

VOLUME 1

Chapter 1

Transmission System Environment

The composition of the Bell System plant is reviewed in terms of the types of transmission facilities used and the ways in which they are interconnected.

Chapter 2

Message Channel Objectives

The Bell System objectives for loss, noise, crosstalk, and echo in message circuits are stated, and the statistical nature and subjective foundation of these objectives are discussed.

Chapter 3

Voice Frequency Transmission

The voice frequency components of the telephone plant - subsets, loops, vf trunks and repeaters - are described. Voice frequency transmission characteristics, noise sources in the local plant, and crosstalk are discussed.

Chapter 4

Amplitude Modulation

Amplitude modulation and demodulation are analyzed, and various forms of AM signals are discussed. Emphasis is placed on the preparation of telephone message signals for transmission over carrier systems.

Chapter 5

Introduction to AM Carrier Systems

The building blocks of AM carrier systems are described. The chapter summarizes many of the important problems encountered in the design and engineering of these systems and serves as an introduction to the material that follows in Chapters 6-15.

Chapter 6

System Layout Terminology

This chapter collects important terminology used in Chapters 6 - 15 and introduces the reader to the problem of detailed system analysis.

Chapter 7

Random Noise

Sources of random or thermal noise in AM systems are discussed; formulae for computing tube noise, and methods of estimating noise figure of repeaters and the addition of noise in a string of repeaters are given.

Chapter 8

Modulation Distortion

Cross modulation between channels arising from non-linearity in an AM system is analyzed. The relation between the power series representation of the non-linear device, and the overall intermodulation performance of an AM multi-repeatered system is developed.

Chapter 9

Load Capacity, Gains and Losses

System load and overload are defined in terms of an equivalent single frequency sinusoid. Equality of repeater section transmission path loss and repeater gain is shown to be an important objective.

Chapter 10

System Layout and Analysis

The material developed in Chapters 6 through 9 is used to illustrate the problems of setting repeater spacing, system levels, and analyzing system performance.

Chapter 11

Misalignment

The problem of systematic misalignment - all repeaters slightly too high or all repeaters slightly too low in gain - is analyzed and the necessarily adverse effect on signal-to-noise ratio studied.

Chapter 12

Overload and Modulation Requirements

Methods of deriving overload and intermodulation requirements for a system from a knowledge of the speech load are studied. It is shown that the peak value of the voltage wave corresponding to a telephone multiplex signal can be expressed in terms of a sine wave having the same peak voltage; this concept is also made use of in FM systems later. Methods of computing modulation noise developed here are similarly adaptable to FM system problems.

Chapter 13

Feedback Repeater Design

The problems of working through a feedback repeater design from its initial conception to its final form, and estimating the repeater performance throughout the design process, are discussed as an example of the interdependence of device development objectives, circuit design and system performance.

Chapter 14

Regulation and Equalization

Requirements on the transmission-frequency characteristic for telephone and television transmission are discussed, and methods for equalizing and regulating to meet these requirements are described. The frequently unexpected impact of the equalization plan on other aspects of system performance illustrates the complex nature of the system problem.

Chapter 15

Shaped Levels, Feedback, Compandors, TASI

The effect of shaped feedback and pre-emphasis of the telephone multiplex load on repeater noise, intermodulation, and overload is discussed. The problems and advantages of compandors are described. The principle of time-sharing of channels is introduced.

VOLUME 2

Chapter 16

Television Transmission

The nature of the television signal, its sensitivity to interference, and the resulting requirements on transmission systems for this signal are discussed.

Chapter 17

Introduction to Microwave Systems

The building blocks of a radio system are described. Some similarities and differences between radio and wire systems are discussed.

Chapter 18

Radio Propagation

Antenna gain and path loss relations are analyzed. Characteristics of typical antennas and the problems of fading and absorption are discussed.

Chapter 19

Properties of the Frequency Modulated Signal

The spectrum of a carrier which is phase or frequency modulated by one or more sinusoidal signals is derived. The spectrum resulting from angle modulation by a band of random noise representing a telephone multiplex signal is given.

Chapter 20

Random Noise in FM and PM Systems

The method of analyzing the noise performance of an FM or PM system is given. The noise advantage of FM over AM systems is derived, and shown to be an example of the principle of trading bandwidth for signal-to-noise ratio.

Chapter 21

Use of the Fourier Transform for Transmission System Analysis

The Fourier Transform is reviewed at this point to serve as a tool for analyzing subsequent FM and PCM material.

Chapter 22

Effect of Transmission Deviations in PM and FM Systems

The methods of analyzing the effects of transmission deviations in an FM or PM system are presented.

Chapter 23

Frequency Allocation

The factors effecting choice of baseband width and the mechanisms of interchannel interference are discussed. Frequency allocations of present radio systems are illustrated.

Chapter 24

Illustrative Radio Systems Design Problem

The material in the previous chapters is summarized by applying it to the analysis of a short haul 100 channel system.

Chapter 25

The Philosophy of Pulse Code Modulation Systems

A general introduction is given to the principles of message sampling, quantizing, coding, decoding, and reconstruction. Time division multiplex and the trading of bandwidth for signal-to-noise ratio are examined for a PCM system, and the results are related to previous discussion of AM and FM systems.

Chapter 26

Preparation and Processing of Signals in PCM

The spectrum of a sampled message is examined to introduce the problem of filter requirements. This is followed by a description of the terminal equipment and a discussion of estimated noise performance of a 24 channel system.

Chapter 27

Pulse Transmission and Reshaping

High-end shaping and transmission deviations are analyzed in terms of error rate. Methods of compensating for the effects of low frequency suppression in transmission systems are discussed.

Chapter 28

Regeneration and Retiming

Ideal vs. partial regeneration and retiming are studied in terms of the system error rate. The advantages of a regenerative system over a conventional AM or FM system are discussed.

Chapter 29

Signal Processing

The nature of speech is discussed, and methods which have been devised to extract and transmit only the information content of the message are examined.

Chapter 16

TELEVISION TRANSMISSION

The wave shapes and frequency spectra of the monochrome and color television signals are described. There is a discussion of the sensitivity of the signals to such impairments as bandwidth limitations, transmission deviations, crosstalk, random noise, single frequency interference and non-linearity. The transmission system requirements on these impairments are outlined. The chapter concludes with a discussion of the vestigial sideband method of transmission.

The Bell System has constructed many thousands of route miles of wide band transmission circuits in recent years to serve the rapidly expanding television broadcasting industry. The associated companies make use of coaxial and microwave intercity systems, television switching arrangements at toll centers, and video transmission systems for relatively short intracity links. The magnitude of these operations may be appreciated by noting that the Bell System's investment in equipment capable of transmitting television exceeds the aggregate investment in plant of all the broadcasters. Hence, television transmission is important to the Bell System, and the designer of broadband transmission systems must have a knowledge of television transmission requirements. We shall begin with an examination of the television signal itself before discussing system objectives and requirements.

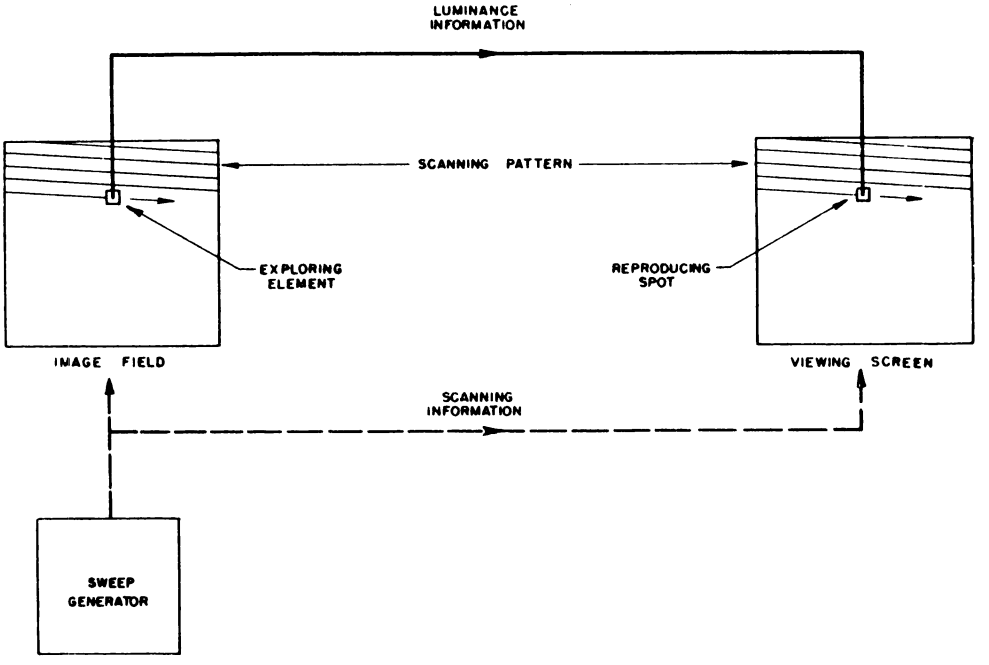
Nature Of Television Signal

Scanning Process

The television signal must contain information in electrical form, from which a picture can be recreated with reasonable fidelity. If we neglect color, a still 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 picture into an electrical signal, that the picture be scanned in some systematic manner so that the large number of luminance values representing a picture are obtained over a period of time. If the scan is sufficiently detailed and rapid,

a satisfactory reproduction of picture detail and motion can be obtained. The basic system consists of a continuous scan in a horizontal line from left to right, starting at the upper left hand corner of the field of view. When the right hand end of a line is reached, the next lower line is explored starting from the left. When the bottom of the field is reached, the process is started again from the top.* The luminance at each scanned point is translated (or coded) into an instantaneous value of signal voltage or current. This scanning process is illustrated in Figure 1 and has been described extensively in the literature.



General Scheme of Television Transmission

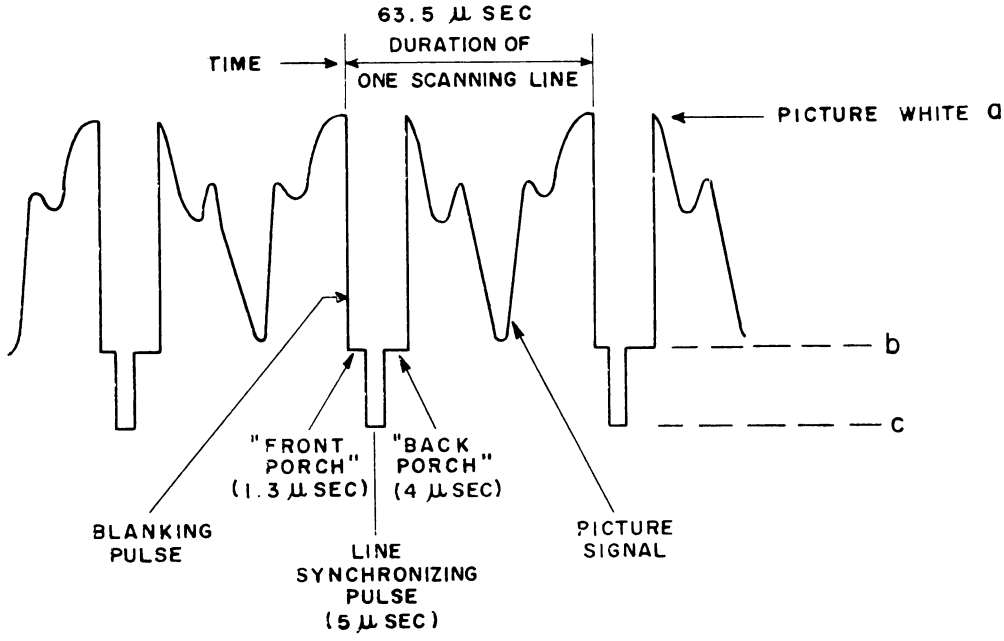
Figure 16-1

For the successful decoding of the signal into a picture at the receiver, it is necessary to transmit a key to the code. In the standard signal, this consists of emitting 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 further motion of the spots, between pulses, is uniform with time in the field of view.

The synchronizing pulses must be distinguishable from the picture signal. This is accomplished by both time and amplitude separ-

*This is simplified. Actually the lines of alternate fields are interlaced.

ation. They are transmitted during "retrace" time. This is the time when the spots are returning from one end of the field of view to the other. During this time no useful picture information can be sent. The picture and synchronizing pulses are also assigned separate amplitude ranges in the total signal. An illustration of this portion of the signal is shown in Figure 2. This shows a trace which can represent either the video signal itself or the envelope of a carrier signal.

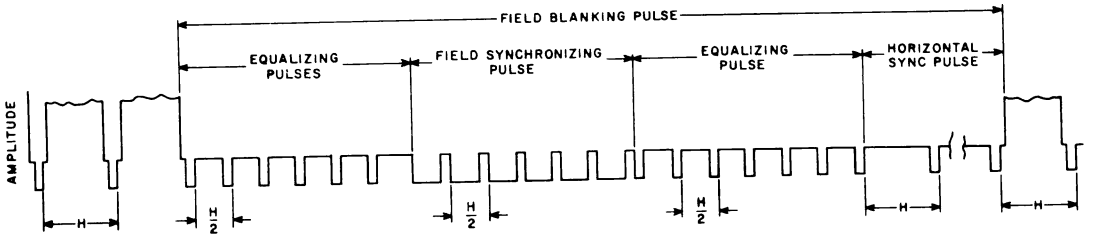


Portions of a Television Signal Showing Line Synchronizing Pulses

Figure 16-2

The picture signal is interrupted during retrace time and replaced by a black signal known as a "blanking pulse". This insures that 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 this region is blacker than black, the sync pulses do not register in the picture.

The line pulses synchronize the individual scanning lines. Similarly it is necessary to synchronize field (vertical) scans. This is done by another pulse in the same amplitude range, as shown in Figure 3. The line and field pulses are identified at the receiver by their durations, which are greatly different. During the field retrace time, the picture is again blanked by a black blanking signal. The equalizing pulses are necessary to insure that both fields making up a frame are properly synchronized.



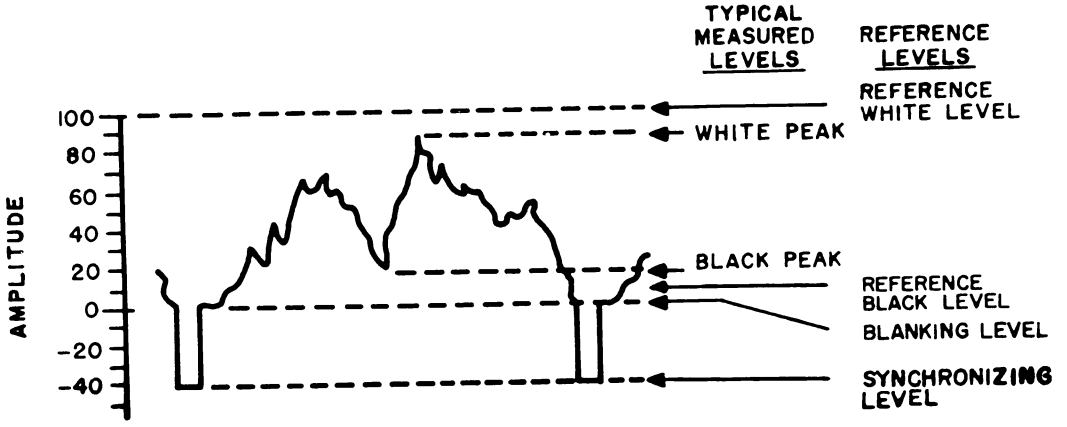
Portion of Television Signal Showing Field Synchronizing Pulses

Figure 16-3

Wave Forms

Figure 3 also shows a portion of the wave forms for both line and field synchronizing pulses. H, the time interval from the start of one line to the start of the next is 63.5 microseconds (15.750 kc line frequency). The picture is scanned vertically at a rate of 60 "fields" per second. Since each complete frame consists of two interlaced fields, the "frame" rate is 30 per second. There are 525 scanning lines per frame of which only about 93% are visible because of loss of time during the field blanking pulse.

The most direct method available for measuring the amplitude and time relationships between the various components of a video signal is to observe the wave forms on an oscilloscope. The method of measurement, and specifications and adjustments of amplitude relationships have been standardized by the IRE. The IRE Standard Scale is shown in Figure 4.



Television Level Measurement

Figure 16-4

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 60 cps field scanning rate, with some energy at lower frequencies. We must, therefore, transmit almost down to dc. How high in frequency must we transmit? The final answer to this question must, of course, come from viewing tests. These tell us that with the optimum viewing conditions and best equipment, a bandwidth of about 4 mc results in very little degradation, and that a greater bandwidth results in an improvement which economically would not be worthwhile. A reduction to 3 mc bandwidth, on the other hand, results in a noticeable degradation.

To appreciate the role which various factors play in this question, it will be instructive to compute the required bandwidth, making certain simplifying assumptions. The standard American black and white picture calls for a frame rate of 30 cps with 525 lines per frame. The ratio of picture width to height (aspect ratio) is 4:3. We assume a scanning spot which is uniformly illuminated over a circular cross-section whose diameter is just equal to the picture height divided by the number of lines per frame.

Vertical Resolution: 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 by a factor of about 70%. Vertical resolution is further reduced to about 93% of its original value because of the loss of lines during blanking time between fields. The resulting number of vertical elements which can be resolved is then:

$$n_v = (525)(0.70)(0.93) = 342$$

Horizontal Resolution: The vertical resolution is controlled by the number of active lines per frame, the size and nature of the scanning spot and the accidental relationships of the scanning pattern

and horizontal lines in the original image. Horizontal resolution, on the other hand, is determined by the highest frequency component which can be resolved along a line. If we make the simplifying assumption 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 that sinusoid (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. Assume, for our purposes, that they are equal.* Since the aspect ratio is 4:3, the desired number of horizontal picture elements would be:

$$n_h = (4/3)(342) = 456$$

Required Bandwidth: A sinusoid which would generate 456 alternate black and white spots would go through 228 cycles along a scanning line. If we allow 16.5% of line scanning time for horizontal blanking, the active (i.e. visible) time for one line will be:

$$t = (0.835)\frac{1}{(30)(525)} = 53 \text{ microseconds}$$

The top transmitted frequency must then be:

$$f_{max} = \frac{228 \text{ cycles per line}}{53 \text{ microseconds per line}} = 4.3 \text{ mc}$$

In spite of all the simplifying assumptions made, this result does not differ substantially from the FCC standard of 4.25 mc.

The significant thing to note in this discussion is that channel bandwidth determines the horizontal resolution. If bandwidth is restricted in any practical case, it is the horizontal detail which suffers.

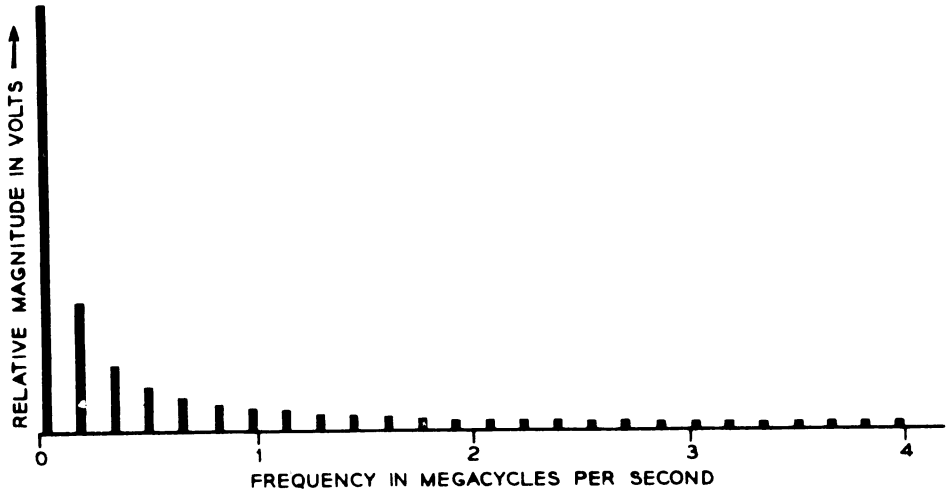
Spectrum

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. In addition, modulation of the line frequency harmonics by the 60 cycle field scan gives rise to 60 cycle 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

---

*In practice the ratio of horizontal to vertical resolution is about 0.94.

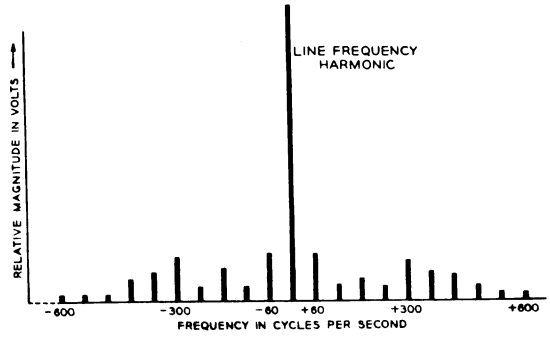
motion in the picture. Each of these frequencies can be considered as a carrier and the effect of motion as adding sidebands around the carrier. The net result is a signal frequency composition similar to that shown in Figures 5 and 6.



Spectrum of Monochrome TV, Showing Every Tenth Line-Frequency Harmonic

Figure 16-5

Figure 5 illustrates the entire 4 mc bandwidth, indicating the levels of line frequency harmonics for a typical signal. Nine-tenths of the harmonics have not been drawn in and the 60 cycle components near zero frequency have been omitted for clarity. A small section of Figure 5 magnified to illustrate the presence of the 60 cycle sidebands that cluster about each line frequency harmonic is shown in Figure 6.



Sidebands Around Each Line-Frequency Harmonic

Figure 16-6

Color Signal

The NTSC* color television system is based on the principle that color may be adequately defined in terms of three characteristics: Luminance, hue, and saturation. Luminance defines intensity or brightness and is the basis on which the present monochrome system operates. Hue defines the color in terms of whether it is red, blue, green, yellow, etc. Saturation defines 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.

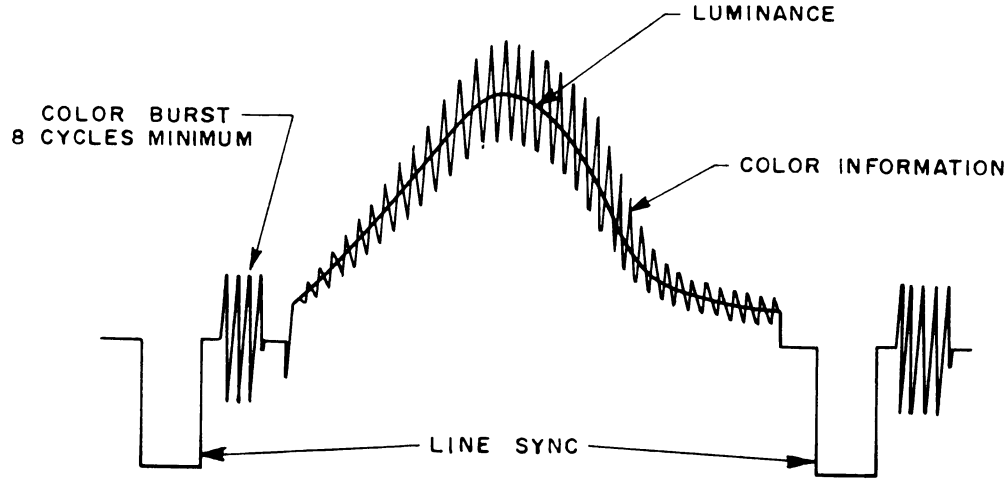
The color signal, therefore, must contain information as to these three characteristics. The color system uses the same type of signal to transmit luminance information as is used in the monochrome or black and white system. To this are added the saturation and hue information which comprise the basic 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 interact with one another and crosstalk occurs.

The saturation and hue information are added to the luminance signal in the form of a new signal called the color sub-carrier. 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.

- -

*National Television Systems Committee

The time or phase relationship of the color sub-carrier to a reference synchronizing signal (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 on Figure 7.



Color Signal Wave Form

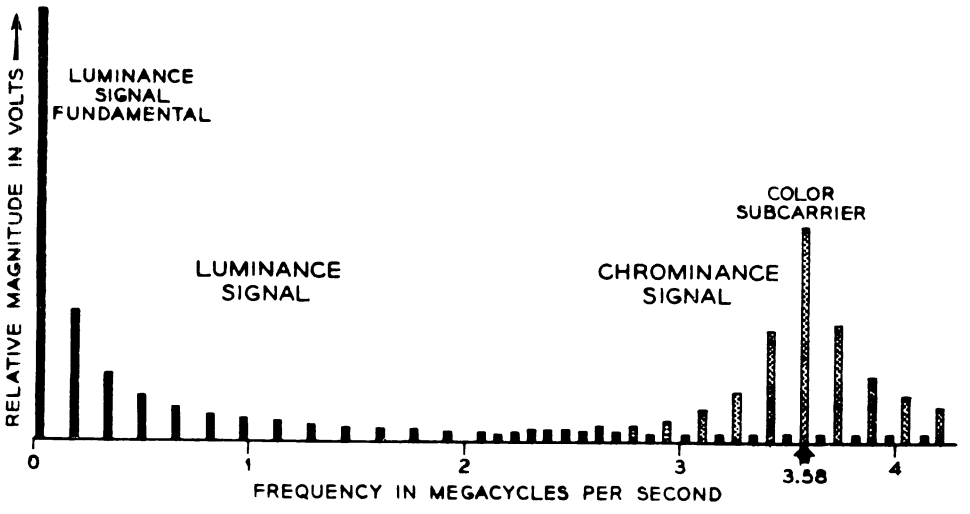
Figure 16-7

Distortions of hue will occur if the 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. The presence of differential phase in a system used to transmit color television results in a change in the hue of the colors.*

The frequency of the color sub-carrier, 3.579545 mc, is chosen to be an odd multiple of half-line frequency (7867 cycles 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 Figure 8. The smaller chrominance components on either side of the sub-carrier are produced by the scanning as in the case for luminance and vary in amplitude and phase in accordance with the hue and saturation information being transmitted.

---

*Such dependence will occur because, for example, the "hot" grid-cathode capacity of the electron tubes is a function of operating point, hence of instantaneous signal voltage.



NTSC Color TV Spectrum

Figure 16-8

TELEVISION TRANSMISSION REQUIREMENTS

The subject of distortion and interference in television pictures is a specialized one. It is a field in which new words have been coined to describe particular phenomena, and where accepted audio terms are sometimes employed to describe visual effects. Hence we look for noise, and ringing, we try to avoid pigeons and glitch,* we do our best to keep the porches flat, and the breezeway open, and bleeding whites merit our serious concern. A complete review of television distortion is quite beyond the scope of this section, of course. What is intended, rather, is to present the more important aspects. We shall attempt to relate cause and effect, and shall indicate what the tolerable limits are for these performance degradations.

Television requirements, like telephone requirements, are based on the results of many subjective tests. Extensive viewing tests have been made in which various amounts of distortion or interference were added to the picture and the result judged using seven preworded comments ranging from "not perceptible" to "unusable". The data for

- -

*These are both descriptive terms for forms of low frequency interference. By "pigeons" is meant bright spots of impulse noise that show up and appear to fly across the picture. "Glitch" is characterized by a narrow horizontal bar which moves through a picture at a slow rate.

all observers were pooled to determine the curve for a median observer and in general the requirement for a particular effect has been set so that the median observer will find the interference or the distortion to be "just perceptible". It should be observed that such a requirement obtains for a complete system. When several systems are connected in tandem the overall requirement must be allocated among the component systems and indeed further subdivided to determine the net contribution for individual amplifiers and filters.

In this section we shall take up picture impairments due to the causes listed below. There is a certain amount of overlapping among some of the items, however, and the decision to refer to a particular effect as band limiting rather than as a transmission deviation, for example, is mostly a matter of judgment.

1. Bandwidth Limitations
2. Transmission Deviations
3. Crosstalk
4. Random Noise
5. Single Frequency Interference
6. Non-linear Effects

We shall confine the discussion which follows to video system effects. The use of carrier facilities introduces other difficulties which will be discussed later.

Bandwidth Requirements

Since the television signal is derived from a scanning process the vertical resolution in the picture is determined by the number of scanning lines used per frame and the horizontal resolution is determined by the bandwidth of the system. To make the horizontal resolution approximately equal to the vertical resolution, a bandwidth of about 4.2 mc is required. An unimpaired test signal or test pattern as reproduced on a television monitor is shown in Figure 9. We shall use this test pattern as our reference undistorted picture and show the effects of several types of impairments upon picture quality. Note that the vertical lines in the pattern of Figure 9 can be resolved right into the central circle.

If a low pass network having a transmission characteristic which is approximately flat to 1 mc and then falls off at the rate of about 6 db per octave is inserted between the picture source and the monitor, the test pattern will be reproduced as shown in Figure 10. The picture is no longer "crisp", the vertical bars cannot be resolved close to the central circle (see the lower wedge) and the words "New York"

Figure 16-9 Unimpaired Signal

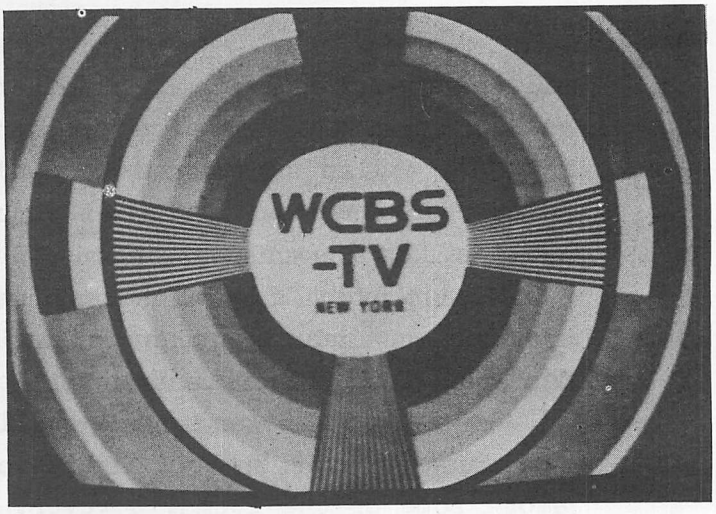


Figure 16-10 1 mc Roll-off

Figure 16-11 Negative Streaking

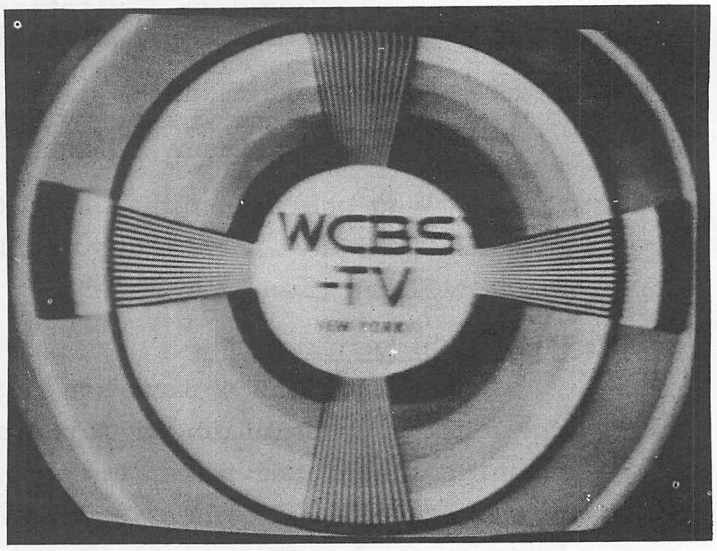


Figure 16-12 Smearing

Figure 16-13 Overshoot

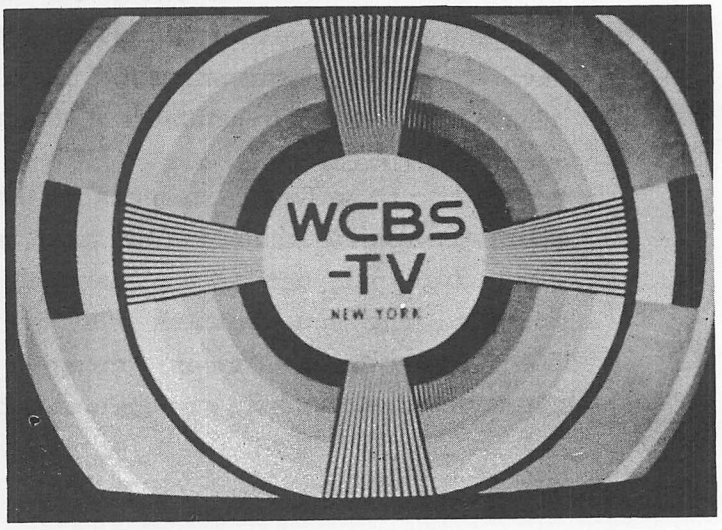


Figure 16-14 4 mc Ringing

are no longer clear. In short, fine detail is not reproduced well. In addition to frequency response, excessive noise and some echo patterns can also mask fine picture detail and result in apparent loss of resolution.

The Bell System objective for television transmission is therefore to provide facilities having flat transmission and delay characteristics between a very few cycles and at least 4.2 mc. The question which naturally arises is, how flat must the gain and delay characteristics be to guarantee acceptable transmission?

Transmission Deviations

Before attempting an answer to the question which has just been raised, it is desirable to examine a few more distorted test patterns. Figures 11 through 14 illustrate some common faults. In order to give a clear qualitative idea of the types of penalties associated with not meeting transmission requirements, large amounts of distortion are shown in these pictures.

1. Streaking and Smearing. Streaking is caused by transmission distortions in the frequency range up to about 200 kc. 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 say 5 kc, are relatively less critical because of the use of electronic circuits called clampers. Clampers effectively reinsert low frequency signal components which were not faithfully transmitted. They permit a 35 db relaxation of gain and phase distortion at 60 cycles but their effectiveness decreases with frequency. All Bell System television networks include clampers.

Streaking and smearing are usually not separate and distinct distortions. A picture which exhibits smear also has streaking. Figure 11 shows the test pattern containing streaking - negative streaking in this case. If the test pattern letters were extended as blacks or grays, the distortion would be described as positive streaking. A badly smeared picture is shown in Figure 12. Figures 11 and 12 are both clamped signals; similar distortions in the vertical direction are attenuated by the clamper.

2. 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.

Figure 13 shows this effect on a typical picture. There is a black outline to the right of white objects and a white outline to the right of black objects.

3. Ringing. Ringing generally results from the transmission of sudden tonal transitions over a system that has a finite pass band with a sharp cut-off at the upper end of the frequency range. It may also result from a marked transmission irregularity at some frequency below cut-off. When a signal containing a sudden transition is applied to such a circuit, damped oscillations or ringing will occur at approximately the frequency of cut-off or other discontinuity, the duration of the ringing depending upon the sharpness or degree of the irregularity. Ringing will be accenuated by a rising gain characteristic preceding the discontinuity. Figure 14 shows the effect of introducing a low pass filter having a sharp cut-off at 4 mc between the signal source and the monitor. The ringing can be seen to the right of the vertical bars in the region where the spacing of the bars corresponds to frequencies approaching 4 mc. It is evident that ringing also causes an apparent loss of resolution.

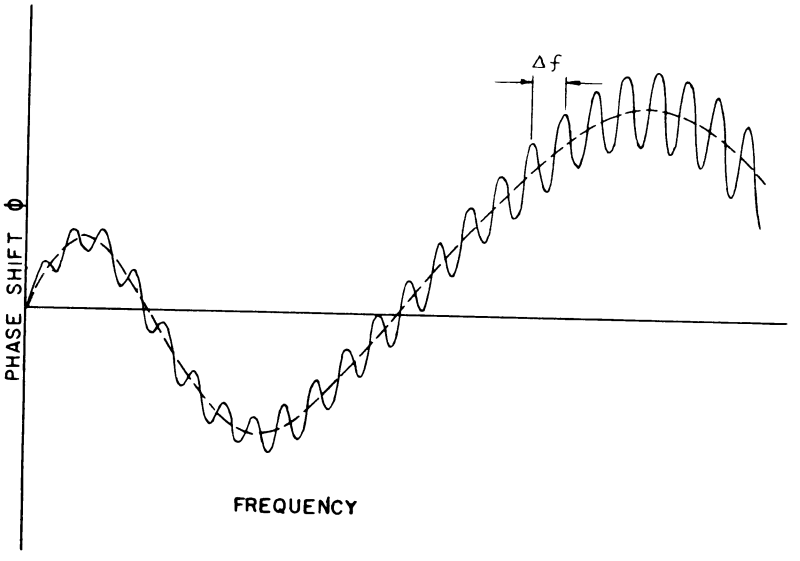
4. Echoes. An echo signal, or ghost, can be defined as a duplicate of the original video signal displaced horizontally from the original signal. Ghosts and echoes are due to impairments in the transmission circuit which cause the signal pulses to reach the viewer at two or more discrete times. The impairment effect of the echo picture not only varies with echo signal strength 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 conditions that give rise to echo signals are usually not continuous throughout the band. Echoes may be either leading or lagging and may be either positive or negative. Figure 23 is an example of a positive echo. (See page 33).

Requirements on Transmission Deviations

The several types of distortion that have been considered all have one thing in common - they are all caused by transmission deviations and can be eliminated by introducing compensating gain and phase equalization.

The requirements placed on the transmission characteristic to hold these picture impairments to tolerable levels are sometimes given in terms of "coarse structure" and "fine structure" deviations.

Figure 15 shows an illustrative steady-state phase curve after the linear component has been subtracted. Widely spaced variations, like that indicated by the dotted line, are known as "coarse structure". Such variations are also described, in a purely descriptive way, 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 540 kc, coarse structure if Δf is much more than 540 kc. The devision is obviously a somewhat arbitrary one.



Steady State Phase Curve Illustrating "Coarse" and "Fine" Structure Transmission Deviations

Figure 16-15

The coarse structure requirements for monochrome television from 7875 cycles (half line frequency) to the upper cut-off are given as

$\pm$ .03 microsecond envelope delay, and
$\pm$ 1.7 db/mc gain slope.

For color transmission this fairly large gain slope cannot be tolerated and the transmission at 3.6 mc should be very nearly the same as at low frequencies.

The fine structure requirements for an overall system for monochrome are tabulated below:

Video Frequency	Phase Distortion	Gain Deviation
60 cycles	± 1 degree	± .1 db
15.75 kc	± 1 degree	± .1 db
500 kc	± 2 degrees	± .2 db
1 mc	± 4 degrees	± .3 db
2 mc	± 8 degrees	± .6 db
3-4 mc	±10 degrees	± .8 db

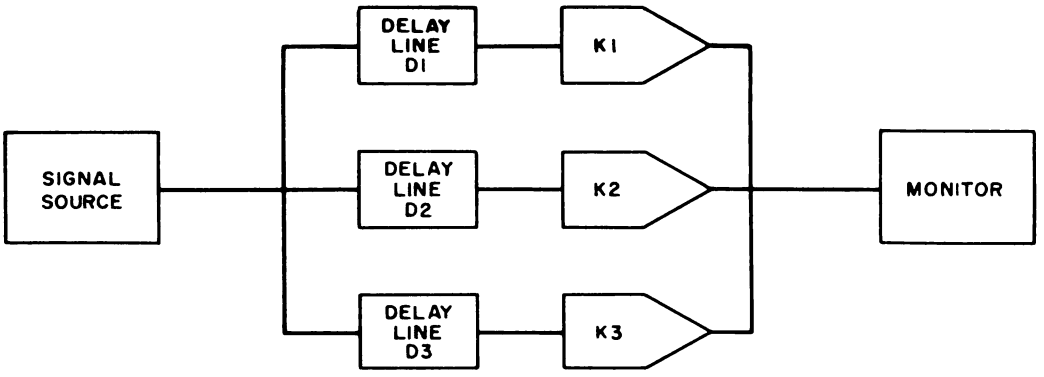
Here again, for color, the requirements between 3 and 4 mc become almost the same as those at 15.75 kc. The above fine structure requirements are for single departures of delay and attenuation. When three or more cycles occur in succession, the requirement must be halved; and where both attenuation and delay occur together the requirement must be divided by $\sqrt{2}$. While these limits serve as a guide, they are not used directly in system design or evaluation. Current practice is to employ the method to be discussed next - the echo rating technique. This is based on the fact that all the transmission deviation effects which have been discussed can be considered as different types of echoes.

The next section will demonstrate the effects of transmission impairments in producing echoes, and how echo delay is related to the periodicity of the variations.

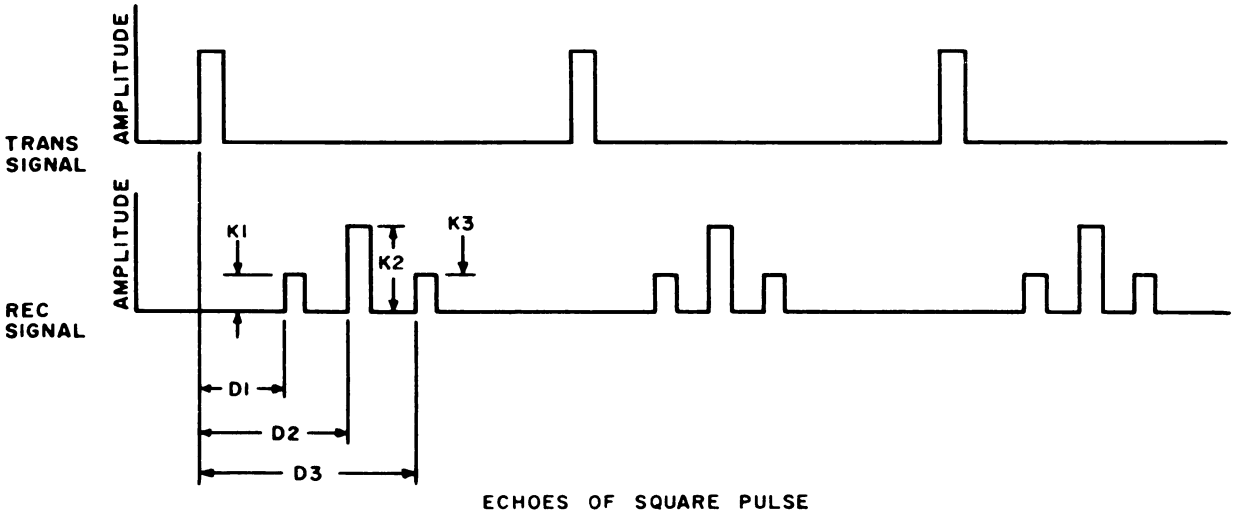
Echo Rating Technique

Before discussing how an echo pattern can be related to the transmission characteristics of a network or amplifier, it is helpful to consider how echoes are generated. They are often thought of as resulting from a discontinuity in a transmission medium or from an impedance mismatch of some sort. The "talker echoes" which were discussed in Chapter 2, for example, arose from improper terminations. The echoes we are concerned with in television are more nearly analogous to "listener echo", or to reverberation effects in acoustics. Such "echoes" can result from multi-path transmission. They can equally well result from transmission over a single path having a non-ideal transmission characteristic. We can set up a paper experiment to show this.

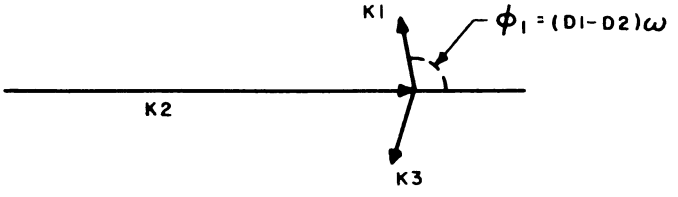
Consider the circuit of Figure 16. The signal is applied to three paths, each having attenuation and delay. As an example, assume that the losses of paths 1 and 3 are equal, so that K1 = K3 = K, and



MULTIPATH TRANSMISSION
(a)



ECHOES OF SQUARE PULSE
(b)



VECTOR DIAGRAM FOR A PARTICULAR FREQUENCY COMPONENT f_1 OF RECEIVED SIGNAL
(c)

Experimental Method of Generating Echoes

Figure 16-16

let K2 = 1. Also, let the delay of path 2 be D and define the delays of paths 1 and 3 so that D1 = D - T and D3 = D + T. The K's and D's thus define the transfer characteristic of the transmission path formed by these three networks in parallel. Thus, if we apply a signal $\varepsilon^{j\omega t}$ at the input, the output voltage can be written:

$$E_{OUT} = K_1\varepsilon^{j\omega[t-(D-T)]} + K_2\varepsilon^{j\omega(t-D)} + K_3\varepsilon^{j\omega[t-(D+T)]} \qquad (16\text{-}1)$$

which, recalling the values of the K's, can be simplified to

$$E_{OUT} = \varepsilon^{j\omega(t-D)}\left[1 + K(\varepsilon^{j\omega T} + \varepsilon^{-j\omega T})\right]$$

$$= \varepsilon^{j\omega(t-D)}\,[1 + 2K\cos\omega T] \qquad (16\text{-}2)$$

Suppose now that the multipath circuit under consideration were replaced by a single equivalent network having a cosinusoidal gain characteristic and zero delay distortion. An observer at the receiving end of such a circuit would be unable to distinguish between the three-path circuit of Figure 16 and a single equivalent circuit having the transmission characteristic of Equation (2). In this simple case, then, we see that echoes can arise from a non-ideal transmission characteristic.

The same phenomenon can be considered in terms of the vector diagram of Figure 16c. Suppose the signal generator is an oscillator which is slowly swept over the transmission band. The subsidary vectors caused by the K1 and K3 paths will rotate about the K2 vector which represents the main path transmission. At any given frequency the phase differences between these vectors will be a function of the delays involved. For the case illustrated (equally spaced leading and lagging echoes of equal amplitude) the result, on the frequency scale, will be analogous to amplitude modulation on the time scale. (Compare, for example, with Figure 19-10.) The resultant vector (total signal) will grow and shrink as we sweep over the transmission band. The greater the delay, the closer together these maxima and minima of transmission vs frequency will be.

We can see this also by referring to Figure 15 and identifying T with $1/\Delta\omega$. T is also, of course, the delay between the received pulses of Figure 16. We see that if T is small, corresponding to an echo close to the signal, the deviation in the gain-frequency characteristic will have a coarse structure. If, for example, we had a coarse

cosinusoidal gain ripple with Δf = 2 mc the echoes would be one half microsecond away from the signal, or about 0.16 inches on a television screen seventeen inches wide (screen with twenty-one inch diagonal) with the standard scanning rates specified previously. A fine structure deviation ripple (high periodicity) with $\Delta\omega$ = 200 kc would produce echoes about 1.6 inches away from the signal.

Similarly, it can be shown that a small sinusoidal ripple in the phase characteristic only results in a pair of echoes, one leading and one lagging the signal, but of opposite polarities.* If the gain characteristic should happen to have a cosinusoidal ripple and the phase characteristic a sinusoidal ripple, both of the same periodicity, the leading echoes may cancel, leaving only a lagging echo. This will occur if the gain ripple amplitude (in nepers) is equal to the phase ripple amplitude (in radians), since the echo amplitude is a direct function of ripple amplitude.

The illustrations we have been discussing are, of course, simple cases of the general problem. Usually the transmission characteristic we are concerned with does not exhibit sinusoidal deviations vs. frequency, but is more complex. These more complex characteristics can usually be analyzed, however, by Fourier methods (a tool which is discussed in Chapter 21). We can thus find, for a given transmission characteristic, the more or less complex pattern of leading and lagging echoes to which it gives rise. How practical such an analysis will be in a given case depends on the complexity of the characteristic under consideration. Often a reasonably accurate approach to the final answer we seek can be found by approximations which will be discussed a little later.

The final answer we seek is a "rating" of the circuit - a sort of figure of merit which will permit us to say whether or not, for example, a certain residual ripple (after equalization) is better or worse than the slope or bulge we tried to equalize out. This figure of merit is the "echo rating" of the circuit. It is found by assuming that the annoying effect of many small echoes can be added on an r.s.s. basis (an assumption which is fairly well substantiated by subjective tests),

- -

*In the previous case of gain ripple only, an analogy to amplitude modulation was pointed out. Similarly, phase ripples on the frequency scale are analogous to angle (phase or frequency) modulation on the time scale. If the deviation were large, we would have to consider additional leading and lagging echoes, analogous to higher order sidebands in high-index FM.

and that this r.s.s. sum can be expressed in terms of a single equivalent echo. In other words, it is assumed that the interfering effect of a pattern of echoes, however complex, can be simulated by the interfering effect of a single, properly displaced echo of appropriate amplitude. What we mean by "properly placed" will be defined in a moment. The ratio of the amplitude of this equivalent echo to the amplitude of the signal, expressed in decibels, is defined as the echo rating of the transmission circuit.

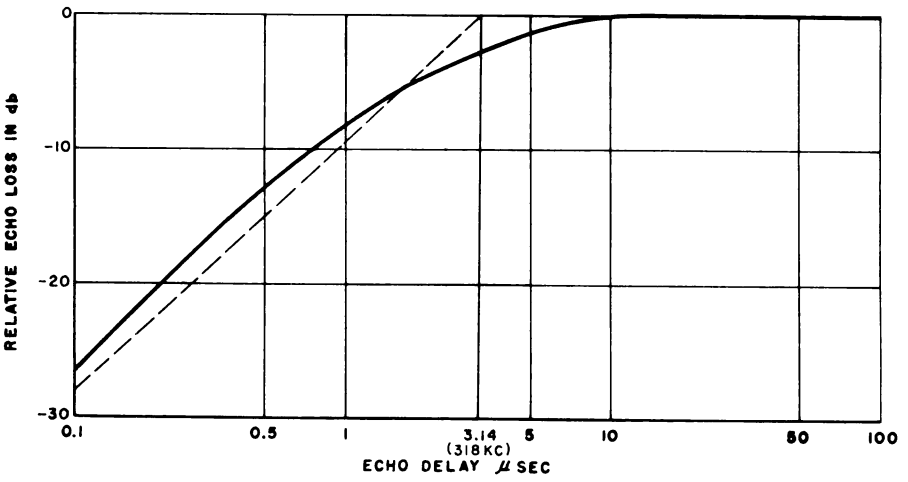
Before we can add the annoying effects of the component small echoes to define this "equivalent echo", we need to know a good deal more than we now do about the annoying effects of echoes. Returning to Figure 16, for a moment, we would expect the viewer's opinion of the picture to change if a), we change T, or b), we make K1 and K3 functions of frequency. We need quantitative information on this point. This can be obtained only from subjective tests.

Time Weighting

Subjective tests indicate that the interfering effect of an echo on a television screen, that is, the annoyance to the viewer, is a function of the spacing between signal and echo. Echoes close in tend to be masked by the signal. Those further out stand more by themselves and tend to be more annoying. The results of a large number of subjective tests are summarized in the time weighting curve of Figure 17. For example, an echo 0.5 microseconds away from the signal is about 12 db less annoying than one of the same amplitude spaced 10 microseconds away. This difference in annoyance associated with time weighting must be taken into consideration in adding echoes or deriving an echo rating. This brings us to the definition of the "properly placed" echo used in echo rating. This equivalent echo is considered to be far out, that is, to lie in the region above 10 microseconds.

Frequency Weighting

So far, we have discussed only cases in which the deviation ripples extend with constant amplitude across the entire frequency band of interest. This is not generally the case. Echoes need not necessarily be miniature duplicates of the signal, but can contain frequency components in different proportions so that the echo of a square pulse might well have rounded corners or a sloping top and so on. Now, we already know that most of the energy of a TV picture is concentrated at low frequencies, and that the high frequency components carry information about fine detail only. It is not surprising, then, to find that a secondary



Echo Time Weighting Function

Figure 16-17

path which transmits the low frequency components is more objectionable than one which transmits only the high frequency components. The former produces the main outlines of the picture in echo form; the latter gives a bit of fuzz which is lost in the noise. Again we can readily deduce the relationships between echoes and transmission deviations in this case by considering the frequency components of the signal at the receiver in Figure 16 as vectors whose phase is a function of the delay. It is clear that if K1 and K3 are relatively large at low frequencies and approach zero transmission at high frequencies, two effects will follow: 1) the echo will consist of the low frequency (high energy) components; and 2) the total transmission characteristic will exhibit relatively large ripples at low frequencies and very little ripple at high frequencies.

Consider, for example, a high periodicity consinusoidal ripple (only a few kc between ripples on the frequency scale) which extends only over a part of the transmitted band. This would correspond, in Figure 16, to a relatively long delay in a secondary path which also included a band pass filter. In order to evaluate such an echo, two factors must be considered - over what fraction of the total band does the deviation extend, and where is it centered? Ignoring the latter question for a moment, the assumption can be made that an echo due to a ripple over part of the frequency band is less disturbing than an echo due to a ripple across the band, in direct proportion to the fraction of the band involved. Thus, an echo due to a cosinusoidal ripple extending

with constant amplitude over, say, 400 kc of a 4 mc band would be less disturbing than an echo due to a ripple across the full 4 mc by a factor of 10 log 4.0/0.4 = 10 db. The factor here is known as "bandwidth advantage" and can be written:

$$\text{Bandwidth Advantage} = 10 \log \frac{B}{\Delta B} , \qquad (16\text{-}3)$$

As mentioned earlier, however, the location of the frequency deviation in the band of interest is of importance. Figure 18 shows the relative annoying effect of ripples as a function of their location in the band. The figure also indicates how the curve must be modified for color transmission, which contains important information around the color carrier frequency at about 3.6 mc. After demodulation, the sidebands of the 3.6 mc carrier appear as low video frequency color components, leading to a more stringent requirement around 3.6 mc for color than for monochrome transmission. Note that, while the addition of color increases the sensitivity to transmission deviations around 3.6 mc by some 15 db, the requirements at low frequencies remain more severe by about 11 db.

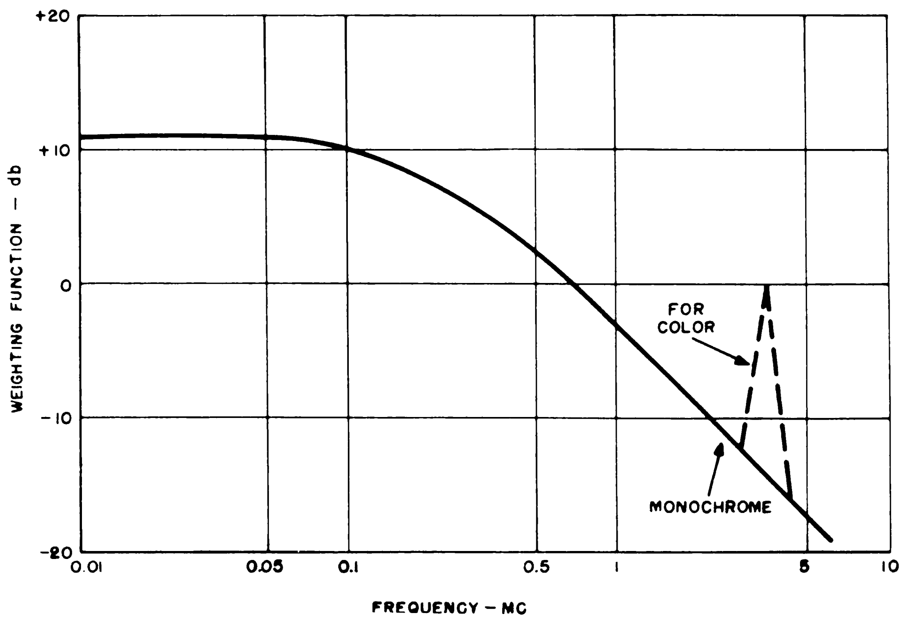
Consider a simple numerical example. We commented above on a cosinusoidal deviation which extended over a band 400 kc wide. Suppose that the center frequency of this region of deviation was about 500 kc. Figure 18 shows the penalty to be about 3 db.* The complete frequency weighting for this example would combine the 10 db advantage for bandwidth and the 3 db penalty for center frequency to give a net advantage of 7 db over an echo resulting from a deviation extending over the entire band. To complete the picture, time weighting would also have to be taken into consideration. We now have all the tools required to consider a general method for estimating, to a first degree of approximation, the echo rating of a transmission system whose gain and phase deviations are known.

Estimating Echo Rating

We have already noted that the problem of determining echo response from information on transmission deviations is theoretically

- -

*Taking the value of this weighting function at the center frequency of the deviation band is reasonably accurate if the weighting does not vary much over the deviation band. A more accurate approximation results if the band is divided into narrow strips, weighting each by the amount given in Figure 18. The results can then be added on a power basis.



Echo Frequency Weighting Function for
Monochrome And Color Television

Figure 16-18

best attacked by application of Fourier Transforms. While this approach results in concise formulation of the problems and results, the actual operations are generally difficult to perform. The integrals can be evaluated for a number of elementary shapes, however. If we can analyze a given, complex deviation shape as the approximate sum of a number of such elementary shapes it may be possible to arrive at an echo rating for the complex deviations by summing up the echo ratings for the simpler shapes on an r.s.s. basis.*

Such a method has been derived. Figure 19 shows a number of deviation shapes whose amplitudes are such that each corresponds to a -60 db echo rating. That is, each of these "echo patterns" is just as annoying as a single echo which perfectly duplicates the signal in terms of spectrum, which lies at least 10 microseconds away from the signal and which is 60 db below the signal.

---

*Strictly speaking, the elementary shapes used should be orthogonal functions.

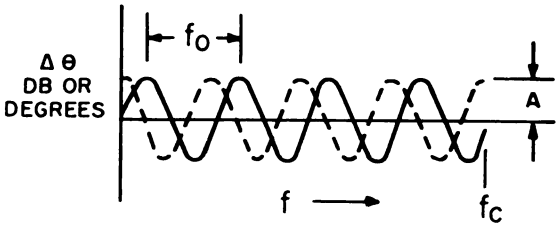
DEVIATION AMPLITUDES ARE TABULATED AT MAGNITUDES CORRESPONDING TO AN ECHO RATING OF −60db FOR MONOCHROME TELEVISION

(a) SINUSOIDAL DEVIATIONS, FINE STRUCTURE,

$f_0 < 318$ KC

A = 0.012 db

A = 0.08 DEGREES



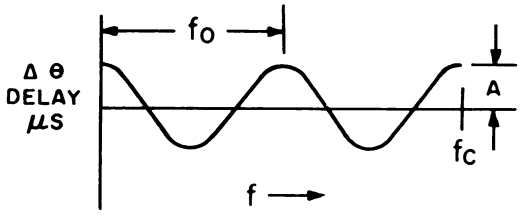
$$\Delta\theta = A \begin{bmatrix} \text{SIN} \\ \text{COS} \end{bmatrix} \omega_0 t$$

(b) SINUSOIDAL DELAY DEVIATIONS, COARSE STRUCTURE

$f_0 > 318$ KC

A = 0.7 Mμs

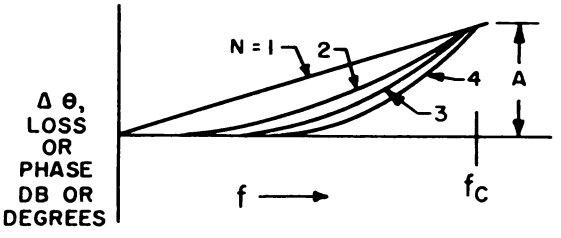
FOR GAIN RIPPLE, USE PART (a)
AND TIME WEIGHTING (SEE PAGES 25 AND 26)



$$\Delta\theta = A \cos \omega_0 t$$

(c) BROAD DEVIATIONS INCREASING WITH FREQUENCY

N =	1	2	3	4
A (db) =	0.5	0.9	1.1	1.3
A (DEG) =	—	5.9	7.2	8.6



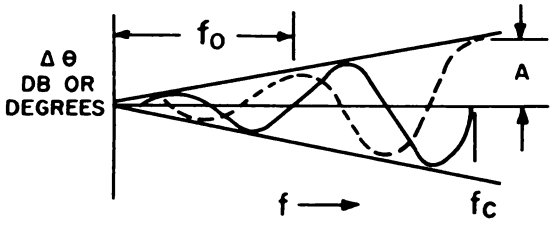
$$\Delta\theta = A \left(f/f_c\right)^N$$

(d) EXPANDING SINUSOIDAL DEVIATIONS,

$f_0 < 540$ KC

A = 0.067 db

A = 0.45 DEGREES



$$\Delta\theta = A \left(f/f_c\right) \begin{bmatrix} \text{SIN} \\ \text{COS} \end{bmatrix} \omega_0 t$$

Echo Rating for Elementary Deviations

Figure 16-19

Consider first the sinusoidal deviations shown in Figure 19a. The gain ripple shown would correspond to well-displaced leading and lagging echoes of the same polarity, each 63 db below the signal. Subjectively, these would add on a power basis to give the annoying effect corresponding to a single echo 60 db below the signal. From a consideration of the vector magnitudes involved, we can easily check the magnitude relationships: thus, in Figure 16, the case illustrated would correspond to K2 = 1, K1 = K3 = .0007. At frequencies where the three signals are in phase, the maximum vector length becomes 1 + .0007 +.0007, which is .012 db greater than the unperturbed value of unity.

Similarly, we can see that the phase ripple shown could be produced by two echoes of the same absolute magnitude but shifted in phase so that K3 = -K1 = .0007. These would add to give a peak phase deviation of .0014 radians or .08°.

The corresponding delay deviation would be a function of f_o. The delay deviation is found by taking the slope of the phase deviation curve. If we keep the peak phase deviation unchanged and double f_o, we halve the slope and therefore get half as much delay deviation.

Consider now the curve of Figure 19b. This applies for close-in echoes produced by coarse structure phase deviations. Such echoes are subject to the time-weighting curve in the region where it has a 6 db slope per octave of f_o. When we double f_o, the time weighting curve tells us we could double the echo amplitude for the same echo rating. From the previous remarks on the relationship between phase and delay, we see that this would leave the delay deviation unchanged - doubling the phase ripple amplitude and doubling f_o leaves the maximum slope of the phase curve unchanged. For echoes this close to the signal, then, the annoying effect is directly related to the magnitude of the delay deviation. Figure 19b is therefore plotted in terms of delay - a practical advantage, since coarse structure delay distortion is a common system problem.

For the purposes of the discussion above, coarse structure is defined to include cases in which the ripple periodicity in the frequency plane is greater than 318 kc. Fine structure includes cases with periodicity less than 318 kc. This choice of 318 kc as the dividing

point between fine and coarse structure has the effect, then, that we automatically include the effect of time weighting when we scale the ripple amplitudes of Figures 19a and 19b. We are, in effect, approximating the time weighting curve of Figure 17 with straight lines, indicated as dashed lines in the figure. For long delays, there is no time weighting; for short delays, the echo rating improves directly with ripple frequency. By scaling Figure 19a in terms of phase and Figure 19 in terms of delay, the effects of the straight line approximation to time weighting are automatically included.*

As stated earlier, these curves show the values of gain or phase (or delay) deviation which will yield a -60 db echo rating. To find the echo rating where the deviation magnitude is different from that given, scale on a 20 log basis. That is, if we have a fine structure sinusoidal gain deviation, with maximum deviation equal to .024 db, the echo rating should be $-60 + 20 \log \frac{.024}{.012} = -54$ db. As we have already indicated, these curves can be used when ripples cover only part of the band if the proper bandwidth advantage and frequency weighting are used. This will be discussed in more detail later.

The next group of curves (Figure 19c) covers broad deviations which increase with frequency, corresponding to linear, square, cube and fourth power shapes. Again the table shows the values of gain or phase deviation at the top frequency which, with the appropriate shape, will yield a -60 db echo rating. Our simple technique does not permit this group of curves to be used when the deviation extends over only part of the band.

The expanding sinusoidal deviations of Figure 19d are governed by the same comments as those covering Figure 19c.

The application of this technique is best demonstrated with an illustrative example. Suppose that we have a fictional pattern of gain and phase deviation which we have broken down into the following components:

---

*The choice of 318 kc as the dividing point between coarse and fine structure is somewhat arbitrary, a delay of 1/318 kc = 3.14 microseconds, corresponding to the 3 db point of the time weighting curve. We might have used a straight line asymptote to the sloping part of the weighting curve. This would give a dividing point of 540 kc and would require that the value of A in Figure 19b be changed to 0.4 millimicroseconds. (Note that delay = frequency x time and that the values of A in Figures 19a and 19b should correspond at the arbitrary crossover frequency. The 318 kc line gives a better approximation in the crossover region. In the straight line portion, where the 540 kc approximation is more exact, the echo rating will be relatively far down and we are less concerned with accuracy.

a. A parabolic gain shape with A = 1.8 db at the top frequency, extending across the 4.25 mc bandwidth.

b. A consinusoidal phase shape extending from 0.3 mc to 3.0 mc. The ripple periodicity is 70 kc. A = 0.12 degrees.

Find the echo rating; that is, find the equivalent well-displaced echo:

From Figure 19c, the gain deviation echo rating is:

$$-60 + 20 \log \frac{1.8}{0.9} = -54 \text{ db}$$

To find the rating for the sinusoidal phase deviation, use Figure 19a.

For a deviation all across the band, the echo rating is

$$-60 + 20 \log \frac{0.12}{0.08} = -56.5 \text{ db}$$

Since the ripple extends over a region of the spectrum where there is a large change in weighting from the lower to the upper frequency (Figure 18), we will divide up the region into sub-bands. For each sub-band we will read off a value corresponding to the center frequency from Figure 18. From this we subtract the bandwidth advantage calculated for each sub-band which gives the net frequency weighting for each sub-band. To obtain the overall net weighting, the values for the sub-bands are added on a power basis.

Sub-bands	Weighting (Fig. 18)	Bandwidth Advantage	Net Weighting
.3 to 5 mc	+ 4 db	13.3 db	-9.3 db
.5 to 1	0	9.3	-9.3
1 to 3	-7	3.2	-10.2

(-9.3) "+" (-9.3) "+" (-10.2) = -4.8 db

This makes the echo rating for the phase deviation

-56.5 - 4.8 = - 61.3 db

Adding the two echo ratings on a power basis, the final rating becomes

(-61.3) "+" (-54) = - 53.3 db

This technique of estimating echo rating can be a powerful tool in designing television systems and is one with which the designer must be familiar. It makes possible the analysis of loss and phase shift

deviations of a network, in terms of the picture distortion produced by an equivalent echo, which can then be compared to the allocated echo rating. Graphical integration procedures have also been developed for handling this problem with the aid of computing machines.

Echo Rating Objective

The Bell System echo rating objective for a 4,000 mile television network, including radio and coaxial toll circuits, video circuits, and all switching facilities is -40 db. As suggested above, this is an all-inclusive performance figure taking into account all effects that can be cured by equalization.

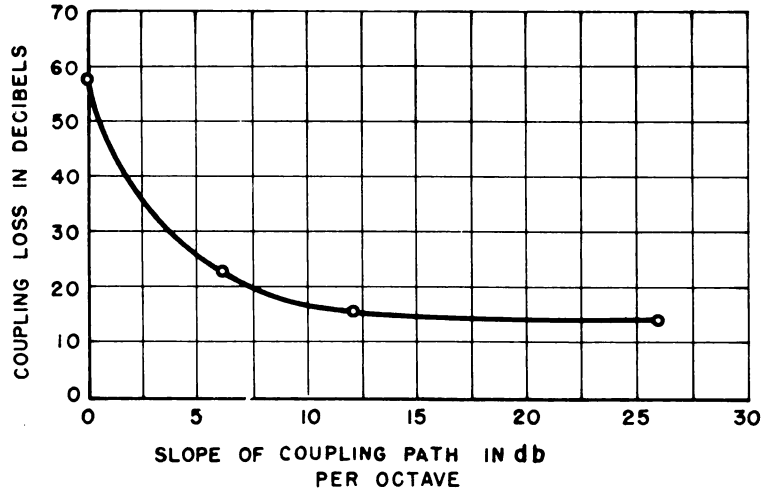
Crosstalk

Video crosstalk becomes an important consideration when two or more video transmission systems operate on adjacent facilities. If coupling between systems is excessive, crosstalk from one system will seriously impair the picture transmitted by the other. To eliminate crosstalk coupling entirely is usually impractical, if not impossible; to reduce it by even modest amounts is sometimes difficult and expensive. The question then arises as to how much crosstalk can be tolerated.

If strong enough, video crosstalk appears as an image of the unwanted signal moving erratically back and forth across the wanted picture. This motion occurs because of the lack of exact synchronism between independent video systems. As the crosstalk image moves across the main picture, it appears to be framed. This frame is formed by the horizontal and vertical synchronizing pulses and is much more noticeable than any feature in the interfering image. Since the side frame of the image extends from top to bottom of the wanted picture, no part of the latter escapes interference. At the threshold of interference, there is no semblance of frame or image; just a slight flicker appears as the frame moves across some sensitive portion of the main picture. When the coupling has a sloping characteristic, rather than flat or undisturbed crosstalk, the differentiated crosstalk image will appear in bas relief.

The crosstalk coupling loss requirement is plotted in Figure 20, in which coupling loss required at 4 mc is plotted against the slope in db per octave of the coupling path. If the coupling path from one video circuit to another is flat vs frequency, the loss required at all frequencies is 58 db. If, however, a high loss is obtained at low frequencies, the 4 mc requirement is eased. For example, if the coupling path loss decreases at 6 db per octave, a loss of 23 db at 4 mc is satisfactory; the corresponding loss at 20 kc will then be 69 db.

The curve of Figure 20 represents lumped rather than distributed coupling. When the coupling is distributed over a long distance, as in adjacent pairs, the crosstalk image will be less clear and probably somewhat less severe requirements would be imposed.



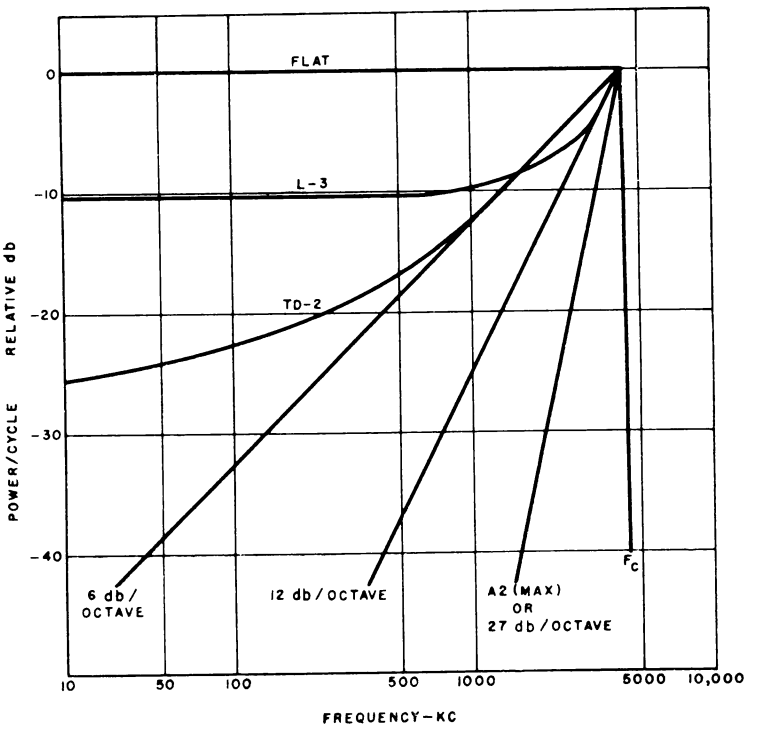
Crosstalk Requirement Coupling Loss at 4 MC

Figure 16-20

Despite the easing of 4 mc coupling loss requirements by as much as 35 to 44 db 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: The magnitude of the incoming signal is decreased and the steepness of the slope of the equalization 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

By random noise we mean fluctuation noise or the type of noise obtained from vacuum tube amplifiers - thermal noise. At the source the spectrum of random noise is usually flat with frequency over a very wide band but communications networks generally contain many selective networks which introduce slope. Figure 21 shows the energy distribution of random noise for several transmission systems. While the effective

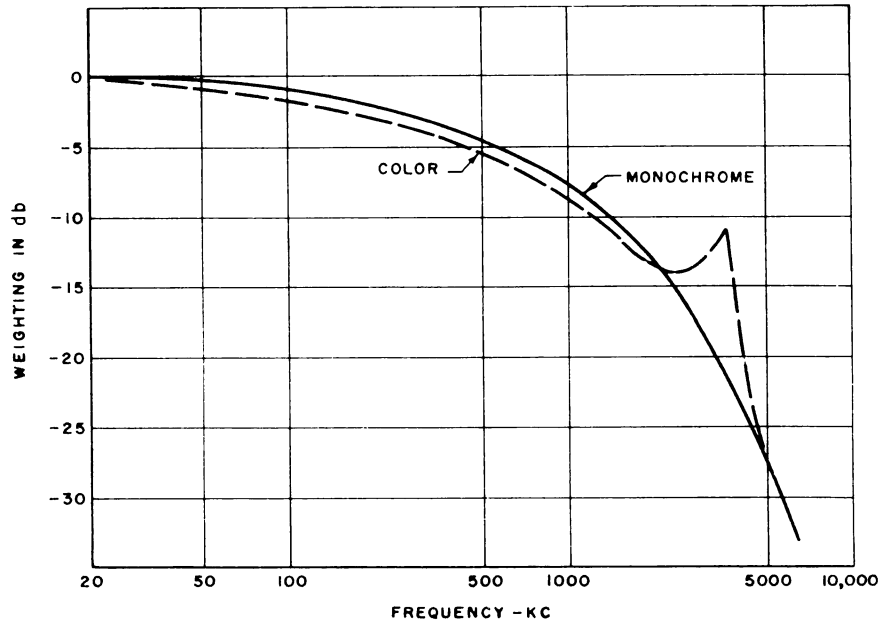


Energy-Frequency Distributions of Random Noise

Figure 16-21

bandwidth of the noise is reduced by these systems the bands are not so narrow as to approach single frequencies, hence there is no correlation from scanning line to scanning line, which would produce bar patterns.

Subjective tests have been made to determine the interfering effect of broad, narrow and mixed bands of random noise distributed throughout the television band. The frequency weighting derived from these judgment tests is shown in Figure 22. This weighting is for use with a simple power summing device such as a power meter. Figure 22 also shows a proposed modified noise weighting for color television. The reason for the peak in the noise weighting curve for color television is that the color receiver demodulates energy in the vicinity of 3.6 mc down to low video frequencies. Hence 3.6 mc noise will appear in the picture as low frequency noise having a relatively coarse pattern in the colored areas of the picture.



Random Noise Weighting for Monochrome and Color Television

Figure 16-22

The general principles derived from these tests have been summarized as follows:

1. Low frequency noise is judged much more interfering than high frequency noise of equal power.
2. A given amount of noise power is judged 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 sum 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.

The weighting curve of Figure 22 may be used to weight the several noise spectra shown on Figure 21. The weighting factors listed in Table 16-1 indicate the extent to which the signal-to-noise ratio may be relaxed for the given noise spectra.

Table 16-1

Energy Distribution of Noise	Weighting Factor in db	
	Monochrome	Color
Flat	9.0	9.5
L-3	12.8	12.3
TD-2	16.9	14.1
A2 (for max. repeater spacing)	23.7	16.3
Uptilted Ndb/octave (6<N<27)	$\frac{72+28N}{8+N}$	$13.7 + \frac{N}{10}$

The overall signal-to-weighted noise requirement as determined from subjective tests and making use of the weighting curve of Figure 22 is:

$$20 \log \frac{E_{video}\ (\text{peak-peak volts})}{E_{\text{weighted noise}}\ (\text{rms volts})} = 54 \text{ db}$$

Here we compare a peak-to-peak voltage with an rms voltage, an unusual procedure. These voltages are the values most readily measured, however, and for the video signal the peak-to-peak voltage is the only one having any meaning.

The 54 db number represents an overall transmission requirement. The requirement on a component video system is considerably more stringent, of course. The current allocation of the noise requirement allows 59 db for all the video systems in an overall television network. If we assume that of the many video systems in the circuit, seven of the video sections are of maximum line length and hence absorb all the requirement, the requirement for such a maximum length section becomes 67.5 db.

Single Frequency Interference

The addition of a single frequency to the television signal superimposes a bar pattern upon the ultimate viewed 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 appear to move. Figure 25 shows a case of bad 1000 cycles interference, Figure 26 shows 311 kc while Figure 27 shows 3.6 mc interference.

Figure 16-23 Unimpaired Signal

Figure 16-24 Positive Echo

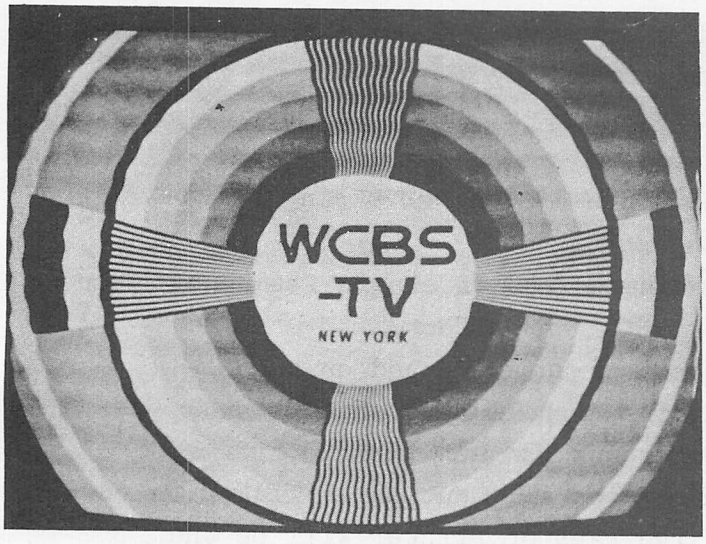


Figure 16-25 1000 Cycle Interference

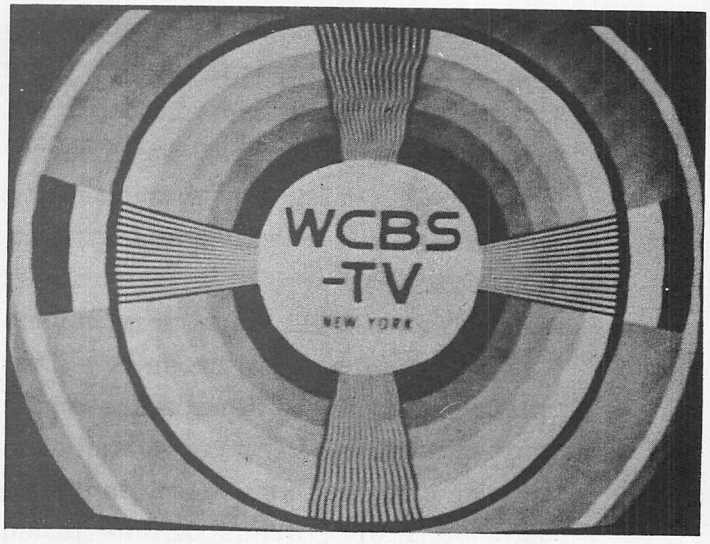


Figure 16-26 311 kc Interference

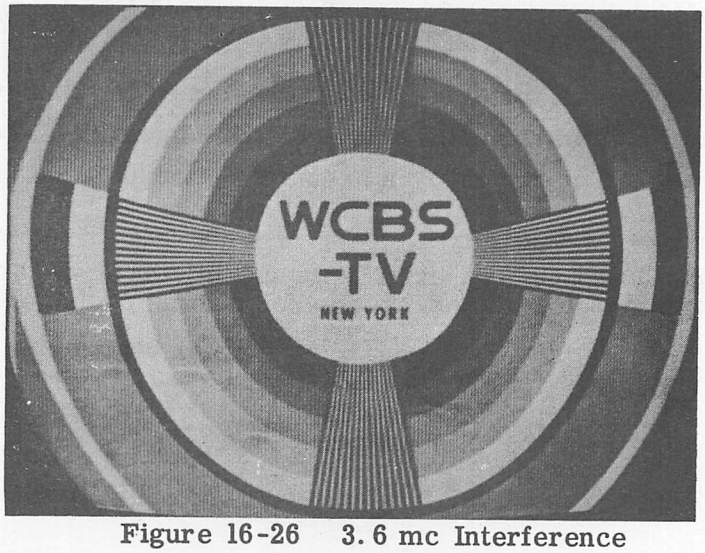


Figure 16-26 3.6 mc Interference
(Note fine vertical bar pattern)

The visibility of a bar pattern, either vertical or horizontal, depends upon the amplitude of the interference and on the angle which the bars subtend at the eye. For a given viewing distance, therefore, a 4 mc pattern is more difficult to see than a 1 mc pattern. A high frequency interference pattern which is synchronous or nearly synchronous with a line scanning component is much more disturbing than a frequency which falls midway between line scanning components.* This fact is made use of in the color signal where the color carrier, 3.579545 mc was chosen to fall 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 were made in connection with the design of the coaxial cable carrier systems to determine a requirement for this type of interference. A complete plot of the results would show a number of maxima and minima, as the interfering frequency approaches or recedes from synchronization with multiples of the line scanning frequency. If we plot only the most stringent requirements, we obtain the curves of Figure 28, which is thus really the envelope of the worst single frequencies in the various portions of the spectrum.

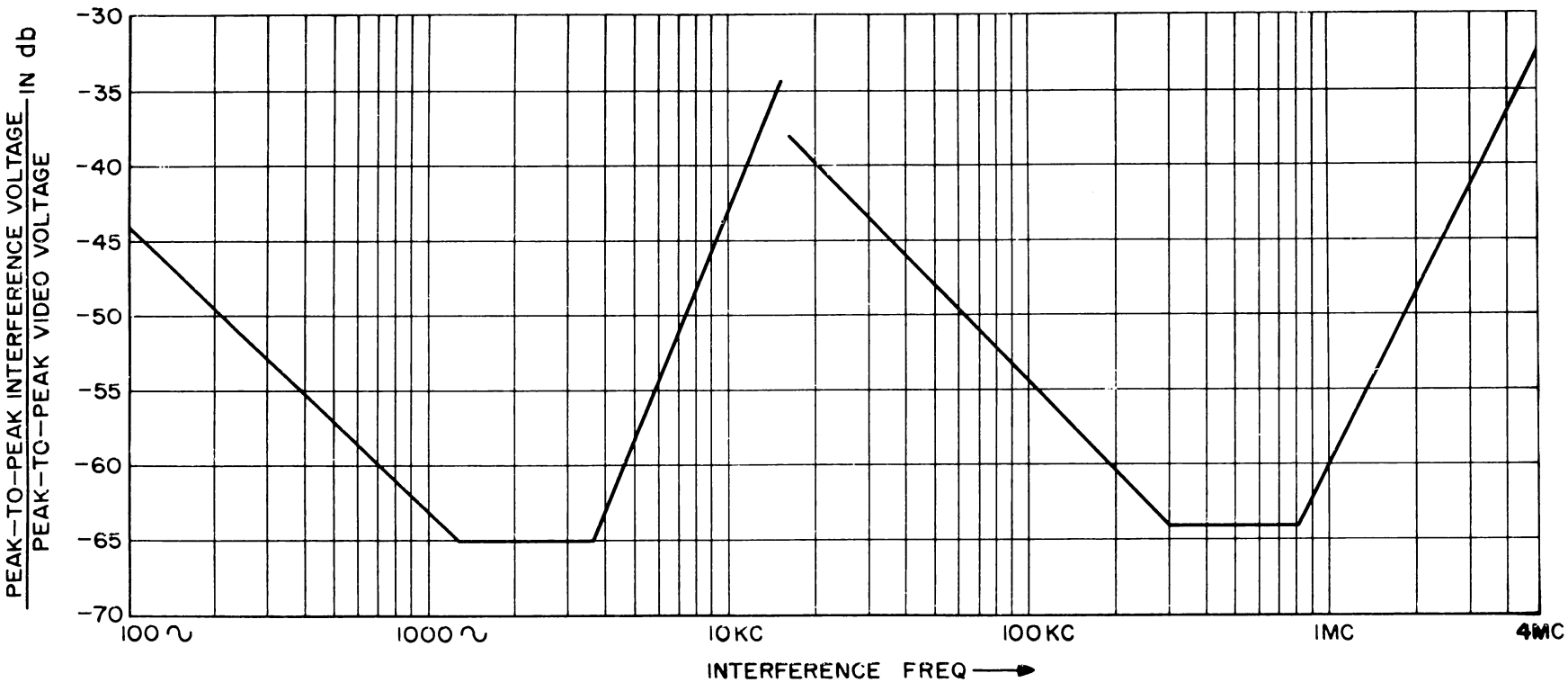
Below 100 cycles, the disturbing effect of single frequency interferences is made more severe by a different effect.

When low frequency interference is superimposed on a television picture - a normal scene containing high lights, 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 as a flicker in some sensitive areas of the picture. The rate of flicker will be the beat frequency between the interfering and the 60 cycle field frequency. This flicker is much more noticeable and disturbing than the brightness distortion casued 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

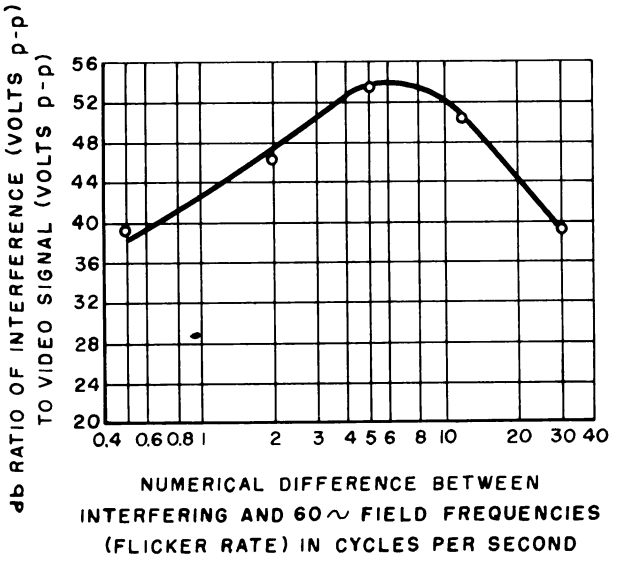
- -

*Ref. 4, p. 210-211 gives an excellent explanation of this effect, as well as the reasons for choosing the location of the color subcarrier.



Tolerable Value for the Worst Single Frequency
Interference in Any Frequency Region

Figure 16-28



Tolerable Limits for Low Frequency Interferences

Figure 16-29

the requirement. This curve has been reproduced in Figure 29, and shows that the most sensitive flicker rate is in the vicinity of five cycles per second.*

Non-Linear Effects

Non-linear distortion is contributed by the system amplifiers; in amplifier design work it is usually evaluated by measuring second and third order modulation. Video amplifiers are generally designed on a balanced basis in order to obtain balance against even order distortion.** The odd order terms, particularly the third, contribute a fundamental term which can be interpreted as a compression term. As we approach the overload point of the amplifier in question, there is no longer a 1:1 relationship between input and output and the output is said

---

*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. Requirements on transmission systems with respect to this effect have been greatly eased by the use of clampers at the TV terminals, but the phenomenom still merits attention in the design of a new system.

**Feedback amplifiers are not used - partly because of the design difficulties associated with the great number of octaves to be transmitted, but mainly to get the better differential phase performance which can be obtained from non-feedback amplifiers. Feedback amplifiers reduce compression effects at the cost of introducing differential phase, and the trade is not an advantageous one.

to be compressed. In television transmission this may result in compression of the synchronizing signal, compression of the picture whites, or both. Requirements on compression for transmission systems are determined by the effect on color, however.

Non-linear effects may be evaluated readily for color transmission by measuring the differential phase and gain performance of the system. These terms have already been defined. It has been established that a critical observer can detect hue changes for a phasing error of 5°. Observers are most critical of flesh tones. They are generally less critical of saturation changes and a saturation change corresponding to about 2 db is not considered to be objectionable. As before, these values obtain for an overall transmission system. When color transmission is satisfactory, monochrome transmission will be better than just adequate as far as compression is concerned. Since any television network may be called upon to transmit color, monochrome compression is therefore not a problem.

Summary of Video Requirements

The various requirements on the overall system for satisfactory transmission of television signals may be summarized as follows:

1. Bandwidth: Approximately 4.2 mc, preferably with gentle roll-off above that frequency. See also transmission deviations, below.

2. Noise: The weighted rms noise shall be 54 db below the peak-to-peak video signal at a flat level point. Figure 22 gives the weighting curves for monochrome and for NTSC color. See also impulse noise, below.

3. Transmission deviations - gain and delay: So much depends on particular characteristics of the deviations that the requirement must be phrased in terms of the echo rating, which must not be worse than-40 db. An idea of the orders of magnitude involved can be obtained from pp. 15 and 16. Except for the different frequency weighting curves (Figure 18) no distinction is made between monochrome and color.

4. Single tone interferences: The requirements are given in terms of bar patterns and flicker - see Figures 28 and 29.

5. Differential gain and phase is change in transmission of color sub-carrier caused by non-linear effects as luminance

varies from 0 to 100 on ERMA scale (see Figure 4). Occurs in video and carrier systems as a result of power-series behaviour of repeaters; in microwave FM systems as a result of transmission deviations (g.v.). The Bell System requirement on differential phase is 5° (this leaves some for broadcaster, who works to a total of 9° or 10°. Even 5° is noticeable, however). For differential gain, the Bell System requirement is 2 db.

6. Crosstalk: A coupling of 58 db is satisfactory if flat with frequency; for non-flat coupling, see Figure 20 which can also be applied to color except for large coupling path slopes.

7. Impulse Noise: The ratio of peak-to-peak signal to peak-to-peak noise shall be at least 14 db. The apparent leniency of this requirement is explained by the short duration of this interference.

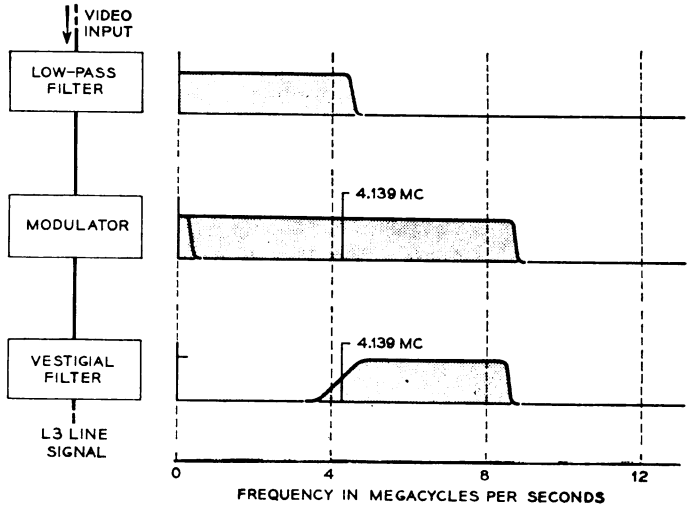
Carrier Transmission of Television Signals

Up to this point, we have considered the phenomona associated with television transmission in terms of the video band. For the longer lines between cities, carrier transmission is used, either over coaxial (using amplitude modulation, and transmitting one sideband plus a vestige of the other) or over radio relay (using frequency modulation). This introduces additional problems. To use many of the subjective requirements stated above, imperfections in the carrier system must be translated into video terms. This is often difficult; in addition, there are other problems peculiar to carrier transmission. Space does not permit of more than a cursory survey of these points.

Vestigial Sideband Transmission

To conserve bandwidth, it would be desirable to be able to transmit single sideband. It can be shown, however, that this would result in intolerable distortion; the practical solution is to use vestigial sideband transmission. This problem, and others associated with amplitude modulation of TV signals, are discussed in Monograph 2090 in connection with the L3 coaxial system, and the following discussion of them is abstracted from that source with only minor changes.

Figure 30 shows the translation (in one step of modulation) of a video signal somewhat more than 4 mc wide to its carrier frequency position in the L3 spectrum. The carrier frequency is 4.139 mc; the upper sideband is nearly 4.2 mc wide, and the lower, vestigial sideband is 500 kc wide.



Vestigial Sideband Modulation

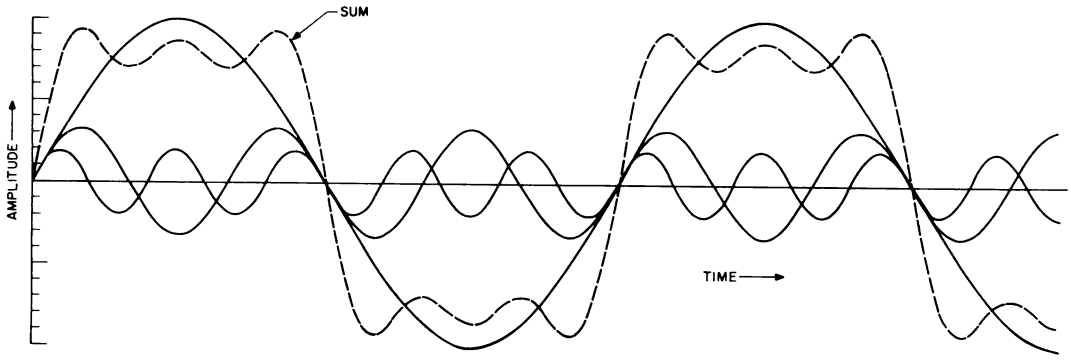
Figure 16-30

The vestigial sideband signal is produced by a band-shaping filter following the modulator. In this filter the lower sideband is suppressed completely except for those frequencies which are within 500 kc of television carrier. From 500 kc below the carrier to 500 kc above the carrier, the transmission characteristic is shaped to achieve a response function which is symmetrical about the carrier. Frequencies more than 500 kc above the carrier are transmitted as in normal single-sideband practice. The factors to be considered in choosing the particular shape of the symmetrical characteristic in the vestigial sideband region ($f_c \pm 500$ kc) are (a) the practical problems of filter and shaping network design and (b) the effect of various shapes in increasing the peak factor of the signal.

It is convenient in a discussion of vestigial sideband transmission to consider the transmission as made up of two components, each symmetrical about carrier frequency, a real or in-phase component and a quadrature component which is a distortion term. The concepts of in-phase and quadrature terms can be illustrated by the following simplified example. Suppose that we are called on to transmit some function of time f(t) such as a repetitive pulse train of period 2π seconds. Such a signal can be shown to consist of a summation of discrete frequencies which are odd harmonics of the repetition rate, having specified amplitudes and phases relative to each other. In general terms,

$$f(t) = \Sigma\, a_n \sin (nt + \theta_n) \qquad (16\text{-}4)$$

The first three terms of such a series, and their sum, are shown in Figure 31. If more terms were included, the approximation to a series of rectangular pulses would become better.



Three Terms of P(t)

Figure 16-31

If we amplitude-modulate a carrier of frequency ω (radians per second) with such a function of time f(t) we can write the resulting double-sideband signal and the transmitted carrier component as

$$E(t) = [1 + f(t)] \cos \omega t \qquad (16\text{-}5)^*$$

Substituting (16-4) in (16-5) we obtain

$$E(t) = [1+\Sigma\, a_n \sin (nt+\theta_n)] \cos \omega t \qquad (16\text{-}6)$$

Since $\sin x \cos y = \frac{1}{2} \sin (x+y) + \frac{1}{2} \sin (x-y)$,

$$E(t) = \underbrace{\cos \omega t}_{\text{carrier}} + \underbrace{\Sigma \frac{a_n}{2} \sin [(\omega+n)t+\theta_n]}_{\text{upper sideband}} - \underbrace{\Sigma \frac{a_n}{2} \sin [(\omega-n)t-\theta_n]}_{\text{lower sideband}} \qquad (16\text{-}7)$$

Now for the sake of simplicity, assume that instead of the shaped vestigial filter characteristic of Figure 30 we have a filter which completely eliminates the lower sideband but transmits the carrier and upper sideband without amplitude or phase distortion. (This would be an impossible filter to build, but recall that we are merely trying to illustrate the concept of in-phase and quadrature components.)

*A more general form of 16-5 would be $E(t)=A_c[1+m\, f(t)] \cos \omega t$ - we have merely normalized for convenience of notation without affecting the generality of the subsequent discussion.

If we eliminate the lower sideband from Equation (7) we have left:

$$E_v(t) = \cos \omega t + \Sigma \frac{a_n}{2} \sin [(\omega+n)\, t + \theta_n] \qquad (16\text{-}8)$$

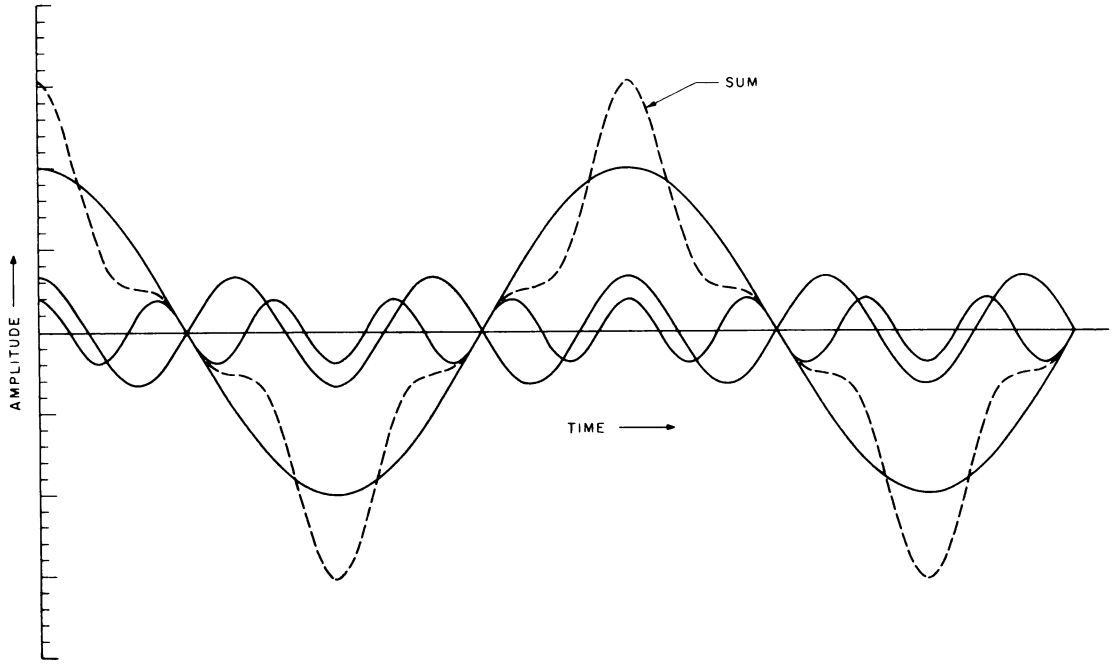
Since $\sin (x+y) = \sin x \cos y + \cos y \sin x$,

$$E_v(t) = \cos \omega t + \Sigma \frac{a_n}{2} \sin (nt+\theta_n) \cos \omega t + \Sigma \frac{a_n}{2} \cos (nt+\theta_n) \sin \omega t \qquad (16\text{-}9)$$

$$= (1+\Sigma \frac{a_n}{2} \sin(nt+\theta_n)) \cos \omega t + \Sigma \frac{a_n}{2} \cos(nt+\theta_n) \sin \omega t \qquad (16\text{-}10)$$

$$= [1+P(t)] \cos \omega t + Q(t) \sin \omega t \qquad (16\text{-}11)$$

Observe that $P(t) = \frac{f(t)}{2}$, so that except for a 6 db factor, P(t) is an undistorted replica of our original modulating function. If it (and the carrier) were the only component of E_v, the original function could be recovered by envelope detection. But Q(t) is also present; it differs from P(t) in that each component has been shifted 90°. It contains cos $[nt+\theta_n]$ where P(t) contains sin $[nt+\theta_n]$, and multiplies sin ωt instead of cos ωt. Hence the name "quadrature component". A plot of the first three terms of Q(t) and their sum is shown in Figure 32.

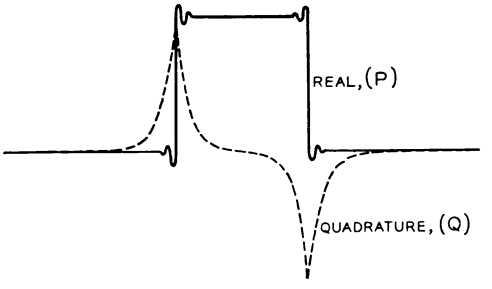


Three Terms of Q(t)

Figure 16-32

Since the components of $E_v(t)$ are in quadrature, the magnitude of $E_v(t)$ is given by the square root of the sum of the squares. The envelope of the wave is, therefore, $\sqrt{[1+P(t)]^2 + Q^2(t)}$.

Quadrature distortion therefore produces an output, if envelope detection is used, which for a rectangular pulse input looks like Figure 33. It might be observed in passing that the effect of Q(t) as a distortion in envelope detection can be decreased by making P(t) and therefore Q(t) small compared to the unmodulated carrier amplitude (which in this case means small compared to unity). This is the approach used in broadcasting TV signals to home sets. Another method is to use a product demodulator instead of an envelope detector.

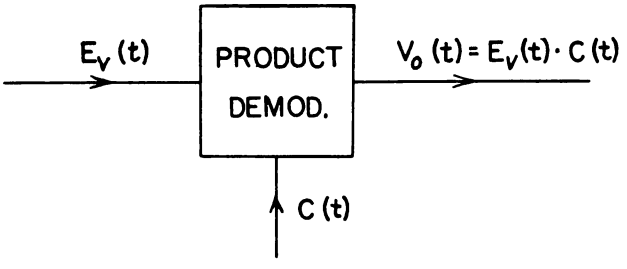


Real and Quadrature Components of Rectangular Pulse in a Vestigial Sideband System

Figure 16-33

Product Demodulation

Figure 34 illustrates the idea of product demodulation; the output of this circuit is the product of the two inputs.



Product Demodulator

Figure 16-34

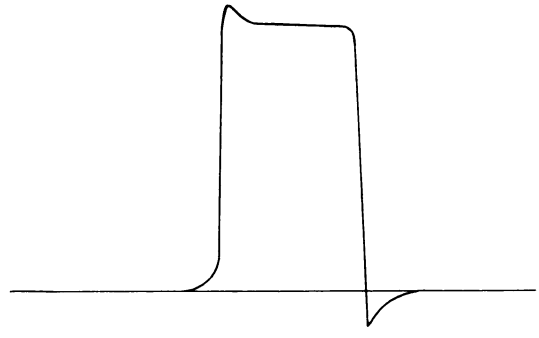
Let $E_v(t) = [1 + P(t)] \cos \omega t + Q(t) \sin \omega t$, and $C(t) = \cos(\omega t + \Phi)$, representing a local carrier supply synchronized with the carrier component $\cos \omega t$ but having a small phase error Φ. Then the output is

$$V_o(t) = E_v(t) \cdot C(t) = \frac{1}{2}[1+P(t)] \cos(2\omega t + \Phi) + \frac{1}{2}[1+P(t)] \cos \Phi + \frac{1}{2} Q(t) \sin(2\omega t+\Phi) - \frac{1}{2} Q(t) \sin \Phi$$

If the carrier at the point of demodulation is at least twice the frequency of the highest frequency component of f(t), the terms containing $2\omega t$ and their sidebands can be eliminated by a low pass filter without affecting the highest frequency components of P(t). The output then becomes

$$V_o(t) = \frac{1}{2}\left([1+P(t)] \cos \Phi + Q(t) \sin \Phi \right)$$

which, if $\Phi = 0$, contains only the wanted P(t) plus a d.c. term, since $\sin 0 = 0$. We see, then, that by providing carrier exactly in phase with the real component of the signal the quadrature component in the output may be suppressed completely. For small but non-zero values of Φ, the output waveform, after filtering out the 2ω components, will have the shape sketched in Figure 35, assuming the input is a repetitive rectangular pulse.



Output of Product Demodulator for Carrier With Phase Error

Figure 16-35

The foregoing discussion has been on the assumption that one sideband is completely suppressed. The partial transmission of one sideband can be thought of as approaching double sideband transmission, which if carried to the extreme would eliminate quadrature distortion. The combination of vestigial sideband transmission, as against true single sideband, and product demodulation, makes for feasible filter and carrier supply solutions. The requirements on the phase of the carrier supply are still severe, however. It has

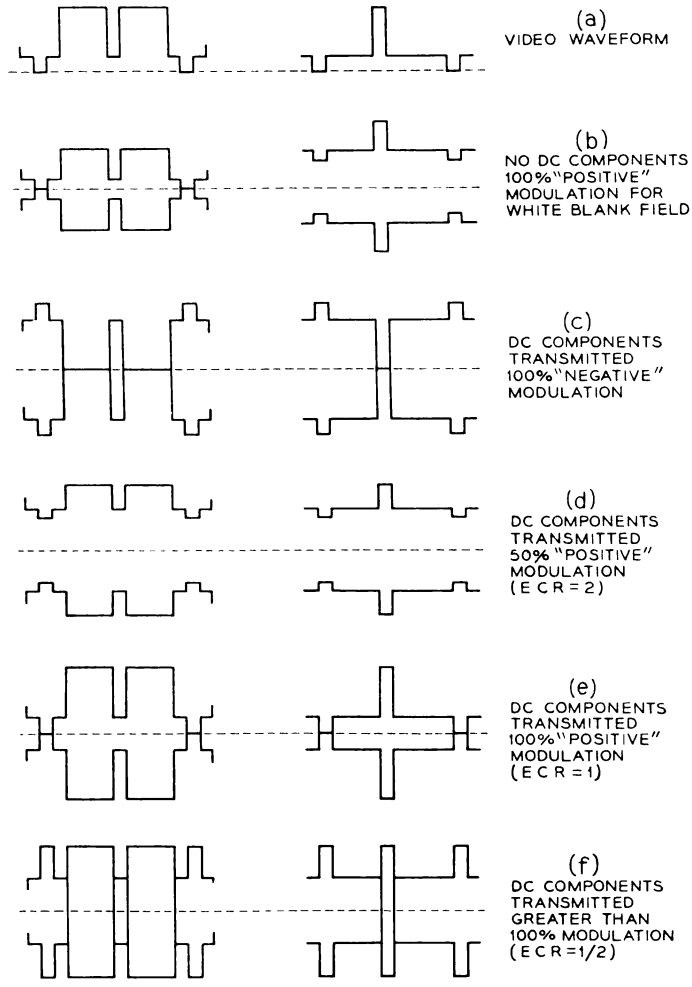
been determined that to suppress the quadrature component resulting from the L3 vestigial band shape to barely perceptible (threshold) values the phase angle of the carrier regenerated at the receiver must be maintained to an accuracy of plus or minus 2.5 degrees. A requirement for one demodulator, when six pairs of terminals contribute to produce quadrature distortion at threshold value, becomes 2.5 degrees divided by the square root of six, or about one degree.

Percent Modulation

Other aspects of the problem of transmitting TV signals in an amplitude modulated system are (1) the transmission of dc components of the video signal; (2) the per cent modulation of the carrier which for convenience is defined in terms of "excess carrier ratio"* and (3) the sign or sense of modulation, that is, whether increasing or decreasing brightness should correspond to increasing signal voltage on the high frequency line. The selection of the optimum method requires an understanding of how the various alternatives would be affected by the system noise and linearity performance and an understanding of representative television viewing tube performance with respect to susceptibility to different types of interference.

Consider, for example, the two video wave forms of Figure 36, one corresponding to a black bar on a white field, the other a white bar on a black field. Figures 36b to f show the corresponding carrier envelopes which would result for various choices of modulation methods. All of these have been drawn with the same maximum peak-to-peak carrier amplitude, on the assumption that overload is limiting the system performance. Since each would have the same magnitude of noise component, the relative signal to noise ratios for the various choices can be estimated by comparing the useful signal magnitudes, given by the distance from sync tip to white. Thus, for example, we see that because the carrier amplitude, in the absence of dc information, will be a function of average video voltage, (b) is inefficient. Here a white bar results in little useful signal for a given envelope magnitude, and a black bar is penalized by the necessity of lowering levels to allow for white bars. Between (c) and (e) there is no choice, on a noise basis, since signal

*A generalized definition of "ECR" is difficult to frame. For our purposes we can define "ECR" (excess carrier ratio) as follows: if we obtain the carrier frequency envelope shown by Figure 36f by starting with waveform (c) and subtracting carrier, then ECR = peak amplitude of carrier during sync pulses, divided by the sync tip to "white" amplitude measured in the carrier envelope. Had we started with Figure 36e instead, the numerator would be the carrier during "white" rather than during sync pulses. The ECR concept becomes meaningless if we do not transmit dc components, thus letting carrier magnitudes be a function of picture content as in Figure 36b.



Alternative Carrier Frequency Envelopes

Figure 16-36

amplitudes are the same - but an examination of modulation products would show an advantage for (e). The 50% case, (d) is obviously poorer than either of these from the signal to noise standpoint. Best of all is (f), where the "folding over" effect of subtracting carrier gives a signal voltage twice that of (c) or (e) for the same total carrier envelope.

The advantages to be obtained by optimizing the carrier signal wave form are substantial, as Table 16-2 illustrates for the L3 coaxial system case. The table shows the signal-to-noise and signal-to-modulation performance which would be obtained for the waveforms of Figure 33, relative to the reference case of ECR = 1/2. Since negative values indicate poorer performance, the ECR = 1/2 case is the best choice from all standpoints.

Table 16-2 Relative Performance of Alternative Television Waveforms

Waveform	Relative Signal-to-Noise Ratio in db	Relative Signal-to-Modulation Ratio (Bar Patterns) in db	
		Group 1*	Group 2*
b no dc	-10.2	-11	-11.3
c neg. mod.	-6	-9.5	-12.5
d ECR = 2	-12	-14	-15.5
e ECR = 1	-6	0	-6
f ECR = 1/2	0	0	0

*Group 1 products are those whose magnitudes are directly proportional to the carrier magnitude. Group 2 products are those whose magnitudes are proportional to the square of the carrier magnitude.

Modulated signals of the forms of b, c, d and e may be detected by rectification, i.e., envelope detection. However, rectification of the waveform (f), produces a spurious envelope wherein video signals which exceed a particular value are inverted. It is necessary to employ homodyne detection, that is, a demodulator driven by a locally generated carrier which is synchronous in phase angle and frequency with the carrier component of the signal wave. As discussed earlier, homodyne detection also makes possible the necessary suppression of the quadrature distortion associated with vestigial sideband transmission.

References

1 - Kenneth Fowler and Harold Lippert, "Television Fundamentals - Theory, Circuits, and Servicing", McGraw-Hill, New York, 1953.

2 - Donald G. Fink, "Television Engineering", McGraw-Hill, New York, 1952.

3 - Fundamentals of Television Transmission: Bell System Practices, Section AB96.100 and 6 Appendices.

4 - John W. Wentworth, "Color Television Engineering", McGraw-Hill, New York, 1955.

5 - S. Doba, Jr., J. W. Rieke, "Clampers in Video Transmission", Bell Telephone System Monograph 1738; Trans. A.I.E.E., Vol. 69, Pt. 1, 1950, pp. 477-487.

6 - H. A. Wheeler, "The Interpretation of Amplitude and Phase Distortion in Terms of Paired Echoes", Proc. I.R.E., June, 1939, pp. 359-385.

7 - P Mertz, "Influence of Echoes on Television Transmission", Bell Telephone System Monograph 2144; Jour. Society of Motion Picture and Television Engineers, Vol. 60, May 1953, pp. 572-596.

8 - P. Mertz and F. Gray, "A Theory of Scanning", BSTJ, Vol. 13 July 1934.

9 - Donald G. Fink, "Television Engineering Handbook", McGraw Hill (1957), Chapters 1 and 2.

10 - A. D. Fowler, "Observer Reaction to Video Crosstalk", Bell Telephone System Monograph 1928; Jour. Society of Motion Picture and Television Engineers, Vol. 57, November, 1951, pp. 416-424.

11 - J. M. Barstow, H. N. Christopher, "Measurement of Random Monochrome Video Interference", Bell Telephone System Monograph 2259; Trans. A.I.E.E., Vol. 73, Pt. 1, January, 1954, pp. 735-741.

12 - A. D. Fowler, "Low-Frequency Interference In Television Pictures" Bell Telephone System Monograph 1904; Proc. I.R.E. Vol. 39, October, 1951, pp. 1332-1336.

Chapter 17

INTRODUCTION TO MICROWAVE SYSTEMS

The building blocks of long and short haul microwave radio systems are described in order to introduce the general subject of FM transmission as it applies to these systems. AM wire and FM radio systems are compared, and some of the problems of FM transmission which will be covered in subsequent chapters are briefly discussed.

Introduction

The purpose of the following chapters is to point out some of the important factors which influence the design, installation and application of microwave systems, and to 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 will be discussed. Simplifying assumptions will be made in some cases in order to concentrate attention on the fundamental problem involved.

The present chapter discusses microwave systems in a general way in order to point out the nature of the problems to be dealt with. Subsequent chapters deal with specific background topics such as radio propagation, frequency modulation theory, distortion mechanisms, and radio channel allocation. The final chapter describes the simplified design of a system in order to illustrate the methods used.

Over-all Block Diagram

The primary blocks that form a microwave system such as TD-2 and TH are shown in Figure 1. (See end page of this chapter.) The telephone multiplex terminal at the left hand side stacks the individual 4 kc telephone channels to form an AM carrier signal. This signal and the television signal are commonly referred to as baseband signals. The path of the signal through the system can be seen by considering Channel 1 W-E (West-to-East). The output of the telephone multiplex terminal feeds the FM transmitting terminal. Here the baseband signal is translated into a frequency modulated wave with carrier typically of the order of 70 megacycles and having lower and upper sideband components extending from perhaps 60 to 80 megacycles. In the radio transmitter this 60 to 80 mc band of signals is modulated up to the proper microwave frequency and amplified. Waveguide then carries the microwave signal from the radio transmitter

to the antenna where it is radiated. After propagation over the radio path, the attenuated signal is intercepted by the receiving antenna and is modulated down to the 70 megacycle region by the radio receiver. In the repeater station the receiver output is connected to the radio transmitter for retransmission to the next relay station. The figure shows only one repeater station whereas in the usual case there are many between terminal stations. At the terminal station the receiver output goes to an FM receiving terminal where it is translated into the original baseband AM signal for connection to the telephone multiplex terminal.

The Channel 1 W-E, which has just been traced through the system, is an example of a one-way radio channel. In the figure, Channels 1 W-E and 2 W-E are shown completely, and connecting arrangements for four additional channels are shown dotted. For telephone transmission, radio channels must always be used in pairs, one West-to-East and the other East-to-West. In the lower half of the figure, equipment is shown to transmit six channels in the E-W direction.

Television, on the other hand, is a one-way service. In Figure 1, Channel 2 is shown in television service. Channel 2 W-E carries one program, and Channel 2 E-W carries another in the opposite direction.

FM Transmitting Terminal

Figure 2 shows a block diagram of an FM transmitting terminal, such as found in the TD-2 system.

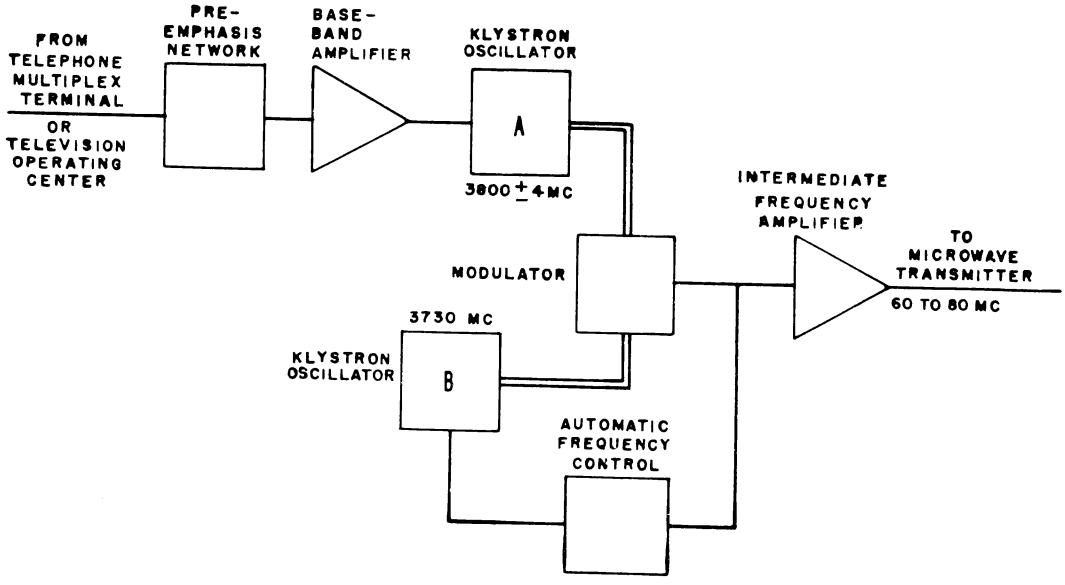
The functions of the various components are discussed in the following paragraphs; subsequent sections cover receiving terminals and the repeaters.

The pre-emphasis network emphasizes high baseband frequencies relative to low, in order to gain certain system performance advantages. This is discussed in Chapter 19.

The baseband amplifier amplifies the signal and applies it to the repeller of klystron A.

Klystron A is an oscillator which is frequency modulated by the signal applied to its repeller. This klystron is operated in such a way that there is a very closely linear relationship between the amplitude of the signal voltage on the repeller and the frequency of oscillation.

Klystron B provides a beating oscillator signal which is 70 megacycles above (or below) the center of unmodulated frequency of klystron A.



FM Transmitting Terminal

Figure 17-2

The modulator is a crystal diode mixer which combines the signals from the klystrons. Its output is the difference product of the inputs. At the output of the modulator, then, there is a frequency modulated signal with a center frequency of 70 megacycles.

The automatic frequency control acts to maintain the difference in the average frequency of the two klystrons equal to 70 megacycles.

The intermediate frequency amplifier raises the level of the signal for connection to the microwave radio transmitter.

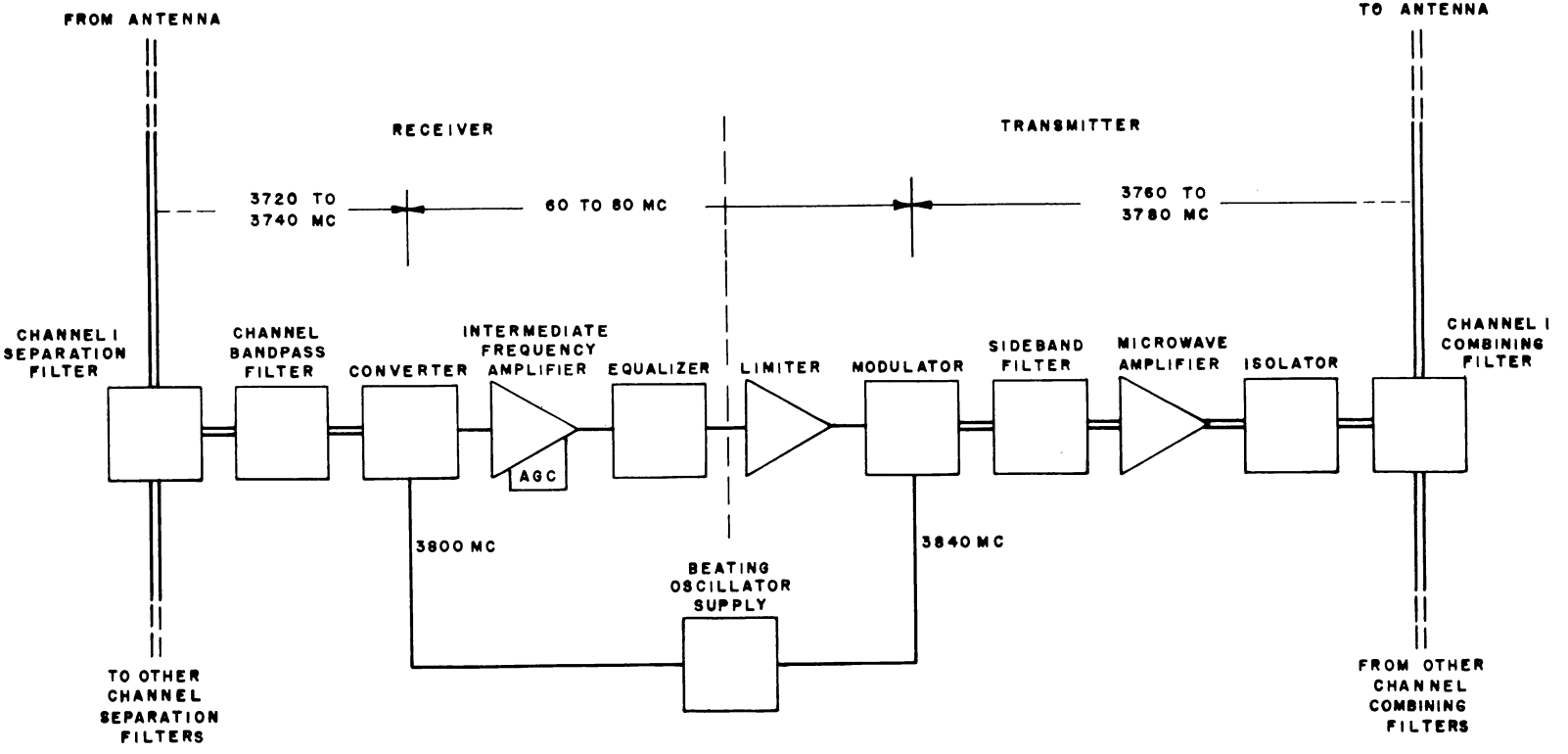
Radio Relay Repeater

The components which typically are employed in a microwave repeater station to handle a one-way radio channel are blocked out in Figure 3. In a terminal station the receiver and transmitter are not connected back-to-back as shown here, but connect to the FM terminals.

The channel separation filter at the left side of the figure separates off the Channel 1 signal and allows all other signals to continue down the waveguide.

The channel bandpass filter has a high loss to all signals except those falling in Channel 1, and hence attenuates unwanted signals that the imperfect operation of the channel separation filter allows to appear at its input.

The converter, which is similar to the modulator in the FM transmitting terminal, makes use of crystal diodes to produce the difference frequencies between the incoming signal and a beat oscillator.



Block Diagram of Typical Microwave Repeater

Figure 17-3

The difference product, which falls in the intermediate frequency band between 60 and 80 megacycles, is fed to the IF amplifier which follows.

The intermediate frequency amplifier provides part of the gain in the repeater. This amplifier has an automatic gain control which compensates for relatively slow time variations in the loss of the radio path and holds the amplitude of the signal at amplifier output essentially constant.

The final component of the microwave receiver is the fixed equalizer. The importance of obtaining flat transmission and a constant delay characteristic will be shown in Chapter 22. The fixed equalizer is used to correct for transmission deviations arising in the various repeater components, the most important being parabolic delay distortion to which the microwave filters and intermediate frequency amplifier contribute about equally.

The next component shown is the limiter, which calls for a bit of explanation. When an FM signal encounters certain transmission deviations, a part of the frequency (or phase) modulation is converted into amplitude modulation. Since our equalization is never perfect, this effect is always present. Furthermore, the transmission phase of microwave amplifiers (e.g., travelling wave tubes) is a function of signal amplitude. If the spurious AM engendered by transmission deviations were permitted to reach the microwave amplifier, spurious phase modulation (frequency modulation) would therefore result. The information being transmitted as phase or frequency modulation would thus suffer distortion. The function of the limiter is to hold the amplitude of the signal constant on a cycle by cycle basis, thus stripping off the spurious AM and avoiding AM to PM conversion in the microwave amplifier. In some systems (TH is an example) limiters are used at every repeater; in other (TD-2) they are not - not because the distortions referred to are non-existent, but because they are swamped out by other distortions, so that using limiters would not give enough improvement to be worthwhile.

The modulator in the radio transmitter is quite like the modulator in the FM transmitting terminal and the converter in the radio receiver. Instead of beating two microwave signals to get an IF signal, however, it beats the IF signal from the limiter and a microwave carrier to get a microwave signal. In the process, the frequency of the transmitter output is shifted with respect to that of

the receiver input in order to minimize the effects of unwanted coupling from the transmitting antenna into the receiving antenna.

The IF signal is a double sideband signal with a sideband above (70 to 80 mc) and a sideband below (60 to 70 mc) the 70 mc carrier frequency. The output of the modulator is a pair of double sideband signals, one centered on a frequency 70 mc above the microwave carrier and the other centered on a frequency 70 mc below the microwave carrier. The waveguide sideband filter on the output of the modulator allows only one of the double sideband signals to pass, the other being attenuated in the filter.

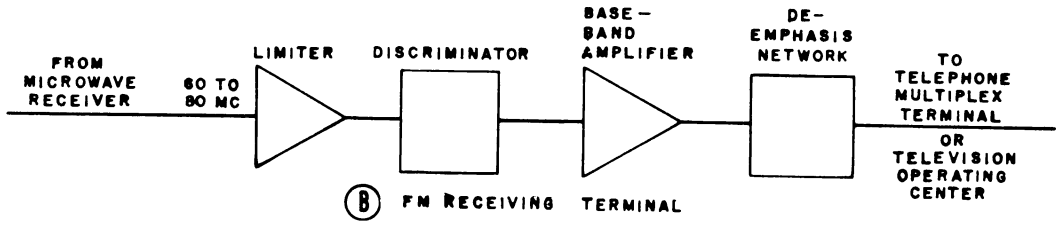
The microwave amplifier boosts the power of the microwave signal from the modulator in preparation for its radiation from the antenna. The power output of the microwave amplifier may typically be one-half to five watts.

The isolator which follows the microwave amplifier is a newly developed component which has proved very useful in overcoming a problem which plagued earlier microwave systems. It is not possible to achieve a perfect power match from a waveguide to an antenna to free space. In spite of all precautions, part of the energy from the transmitter will be reflected back toward the microwave amplifier. This amplifier is not normally perfectly matched to the waveguide. In the absence of an isolator, a second reflection occurs and an echo of the original signal is created. As was pointed out in Chapter 16, very little echo can be tolerated in systems transmitting television. The isolator, although a passive component, is essentially a one-way transmission device. The output of the microwave amplifier is transmitted practically undiminished, but any energy arriving at the isolator from the opposite direction is almost completely absorbed in the isolator and hence is not permitted to be reflected back toward the antenna.

The channel combining filter is identical with the channel separation filter at the receiving input. It provides low loss transmission between its side and top connections for frequencies in Channel 1, and between bottom and top connections for signals in all other channels.

FM Receiving Terminal

A block diagram of an FM receiving terminal is shown in Figure 4.



FM Receiving Terminal

Figure 17-4

The limiter serves a purpose similar to that in the radio transmitter. The discriminators used in FM systems are normally sensitive in some degree to amplitude modulation. If the limiter were omitted, the input to the discriminator would be amplitude modulated by both noise and the baseband signal, resulting in extraneous noise and distortion products at the discriminator output.

The discriminator converts the intermediate frequency FM signal into a baseband AM signal.

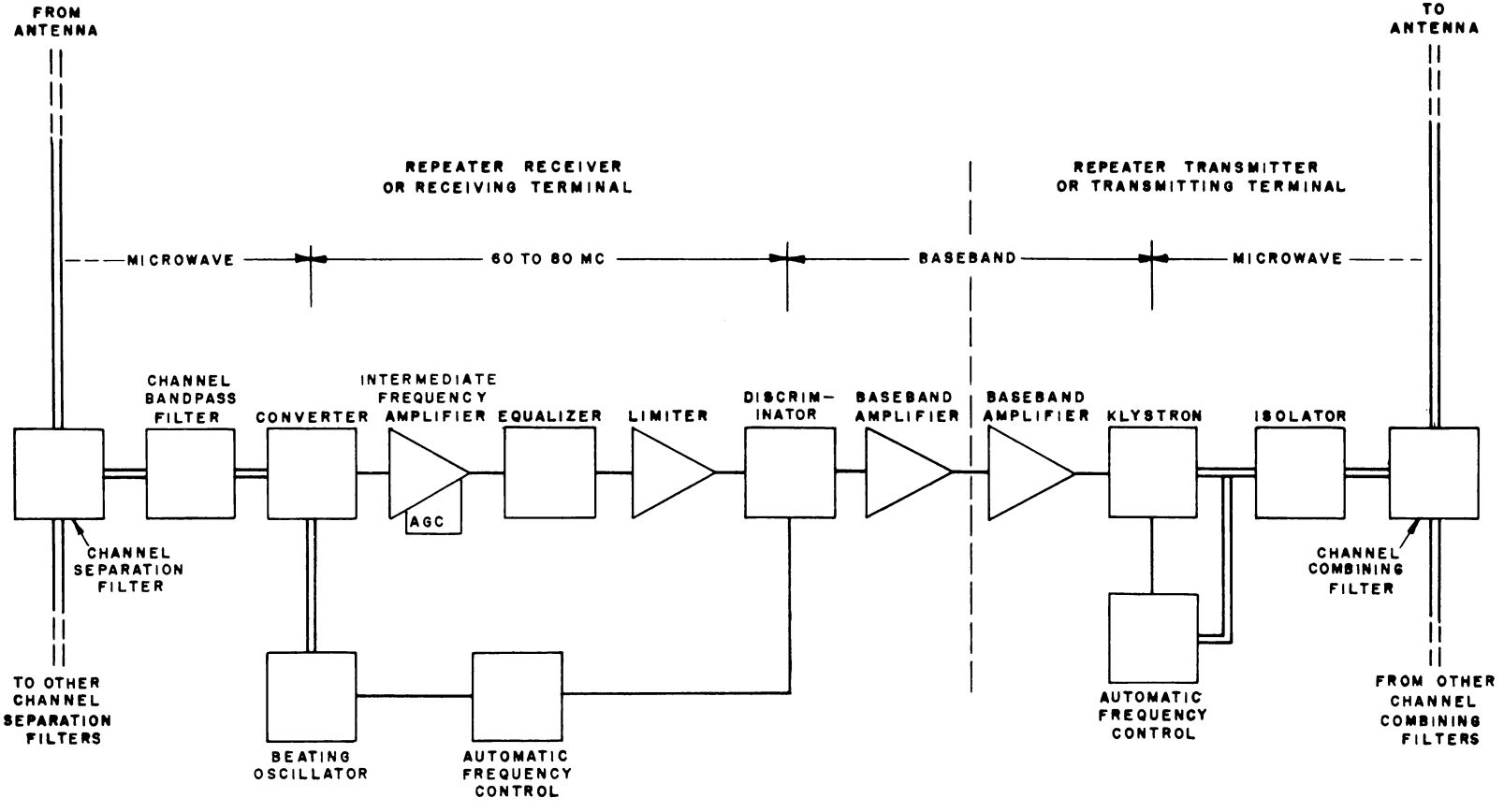
The baseband amplifier raises the level of the signal output of the discriminator.

The de-emphasis network has a transmission inverse to that of the pre-emphasis network. It restores the spectrum of the baseband signal to its original condition.

Baseband Repeater System

As previously pointed out, the description in the paragraphs above applies to such systems as TD-2 and TH. These are systems in which the microwave receiver and transmitter are interconnected at the intermediate frequency, and the baseband signal is recovered only at a terminal station.

There are other microwave systems which are based on a somewhat different transmission plan. Certain short haul, light route, systems such as TJ make use of a so-called baseband repeater. In these systems each repeater contains a rudimentary FM terminal, and the baseband signal is the point of interconnection between receiver and transmitter. The block diagram of a typical baseband repeater system is shown in Figure 5. The components to the left of the dashed division line can be used either as the receiver in a repeater station or as the receiving terminal in a terminal station. Likewise the components on the



Block Diagram of Typical Baseband Repeater System

Figure 17-5

right side of the dashed line serve the dual roles of repeater transmitter or transmitting terminal.

Little comment is needed on the receiver since most of the component blocks have already been described. The beating oscillator is usually a klystron and its frequency must be stabilized by an automatic frequency control.

The transmitter employs a repeller-modulated klystron as the microwave frequency source. A microwave amplifier is avoided by choosing a klystron that can be operated with a reasonably high output power - typically one-tenth watt to one watt.

The advantages and disadvantages of this type of system will be discussed in later chapters.

Comparison of AM Wire Systems and FM Radio Systems

It seems appropriate to pause at this point to comment on certain similarities and differences between FM microwave systems and AM wire systems. Some of the concepts discussed in previous chapters are directly applicable to the FM microwave system and others apply in a modified form. First let us answer a question that very likely is already concerning some readers.

Why Use FM?

The radio transmission of broadband carrier telephone or television is done in the microwave frequency range. It would be very difficult to obtain the allocation of channels several megacycles wide at lower frequencies - the radio spectrum is already much too crowded. In addition, below 100 mc, sky wave propogation (see Page 18-1) makes a radio relay system of the type under discussion impossible. Lastly, higher frequencies have the advantage that antenna gains are greater and the loss from transmitter to receiver can be reduced.

The primary reason that FM transmission is used in microwave relay systems is that linear amplifiers with adequate gain and power output are not available at these frequencies. In AM wire systems great pains are taken to achieve amplifiers having very low distortion - for example, by designing for the maximum possible feedback. The microwave amplifiers used in radio systems, on the contrary, are by no means distortion-free, and, in fact, it is normal practice to operate them under conditions of considerable overload. The large amount of intermodulation which would result from the transmission of a broadband AM signal through these amplifiers would be intolerable. The FM signal, however, is insensitive to this type of non-linear distortion. It car thus be transmitted through overloaded amplifiers which have pure compression with no penalty.

Reduction in noise is sometimes an additional advantage of the use of FM. This noise advantage is obtained only when the peak frequency deviation is approximately equal to, or larger than, the baseband bandwidth. It is not realized in a system such as TH where the baseband bandwidth is 10 mc, but the peak frequency deviation is only 4 mc.

Thermal Noise

In both AM and FM systems, thermal noise sets the minimum level to which signals can be allowed to fall. In the AM case the critical point is the repeater amplifier input, while in the FM case it is the receiver converter. The converter loss, noise in the converter crystal diodes, and the noise of the first tube in the intermediate frequency amplifier result in a noise figure of around 10 to 12 db for the receiver converter. An overall or effective noise figure for the repeater is sometimes a convenient parameter to use and is found by adding the losses of the receiving antenna, input waveguide run, and input filters to the receiver converter noise figure. This results in an effective noise figure of the order of 15 db for a typical repeater.

The method of calculation of noise in FM systems will be covered in Chapter 20.

Modulation Noise

As was pointed out in earlier chapters, modulation noise arising in AM systems sets a "ceiling" on the maximum level of signal transmission. The amplifiers of an AM system are the source of modulation noise. The vacuum tubes, and to a lesser extent the transformers and inductors, are non-linear. To minimize modulation noise we make the peak signal current of the output tube as small a fraction of the total plate current as possible, and employ as much feedback as can be attained. Having built an amplifier that is as linear as we can make it, we set the signal level low enough that the modulation noise generated is tolerable.

The FM case is markedly different. The FM signal is unaffected by the non-linear transmission which causes the modulation noise in AM systems. In FM systems a different mechanism operates, and modulation noise arises because the transmission gain and delay of the system is not the same at all frequencies within the transmission band. The passage of an FM signal through a path with typical transmission deviations will result in the creation of distortion products

unless the transmission deviations are equalized ahead of the discriminator. These gain and delay deviations produce only a baseband equalization problem in the AM case. In an FM system they produce non-linear distortion in addition to requiring baseband equalization.

Whereas modulation noise is a function of signal level in an AM system, it is a function of the amplitude of frequency deviation in the FM system. The frequency deviation of an FM system is analogous to the signal level in an AM system in several ways, and these will be pointed out in the following chapters.

Repeater Spacing

Microwave systems differ from AM wire systems in that the average repeater spacing does not vary greatly from one system to another. Another difference is that in microwave systems, there may be considerable variation in length from one repeater section to another within a system. This is due to the importance of geographic considerations and because, unlike the AM wire case, repeater spacing is not one of the critical parameters in the design of a microwave system.

The distance between microwave repeater stations is determined primarily by two restrictions. First, the microwave radio systems considered in these chapters require that a clear line of sight be established between the antennas, so that as the repeater spacing is increased the tower heights must be increased. As a result, tower economics and geography play an important role in setting repeater spacing. The second restriction is fading. As the repeater spacing is increased the problems of fading becomes more and more acute.

Contrary to the AM wire case, there is little to be gained by using repeater spacings less than the restrictions above allow. Cable loss in db varies directly with distance, and AM system performance can be improved greatly by a small decrease in repeater spacing. If, for example, twenty miles of cable has 60 db loss at the top of the desired band, ten miles will have only 30 db loss. The loss of a radio path, on the other hand, varies directly as the square of the distance. Halving the spacing of a typical radio relay system would only reduce the repeater-to-repeater loss 6 db. Thus, system performance is gained only very slowly as spacing is reduced.*

It follows that the problem of choosing repeater spacings in a radio system is not as clear cut as it was in the AM wire case. There a definite solution could be found in terms of repeater performance and

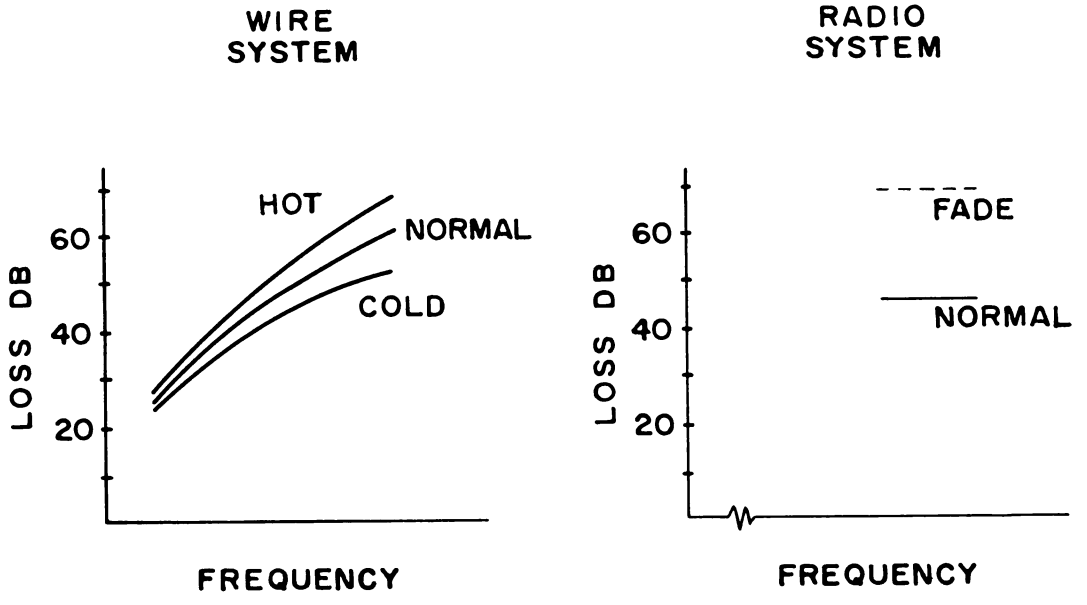
- -

*Repeater-to-repeater loss and system performance obviously can be varied by antenna choice. This is discussed in Chapter 18.

system requirements; here we have a fuzzy problem involving tower economics, geography, and fading considerations. Consideration of these factors results in an average repeater spacing of 30 miles in TD-2 and TH, and from 15 to 25 miles in TJ.

Equalization - Regulation

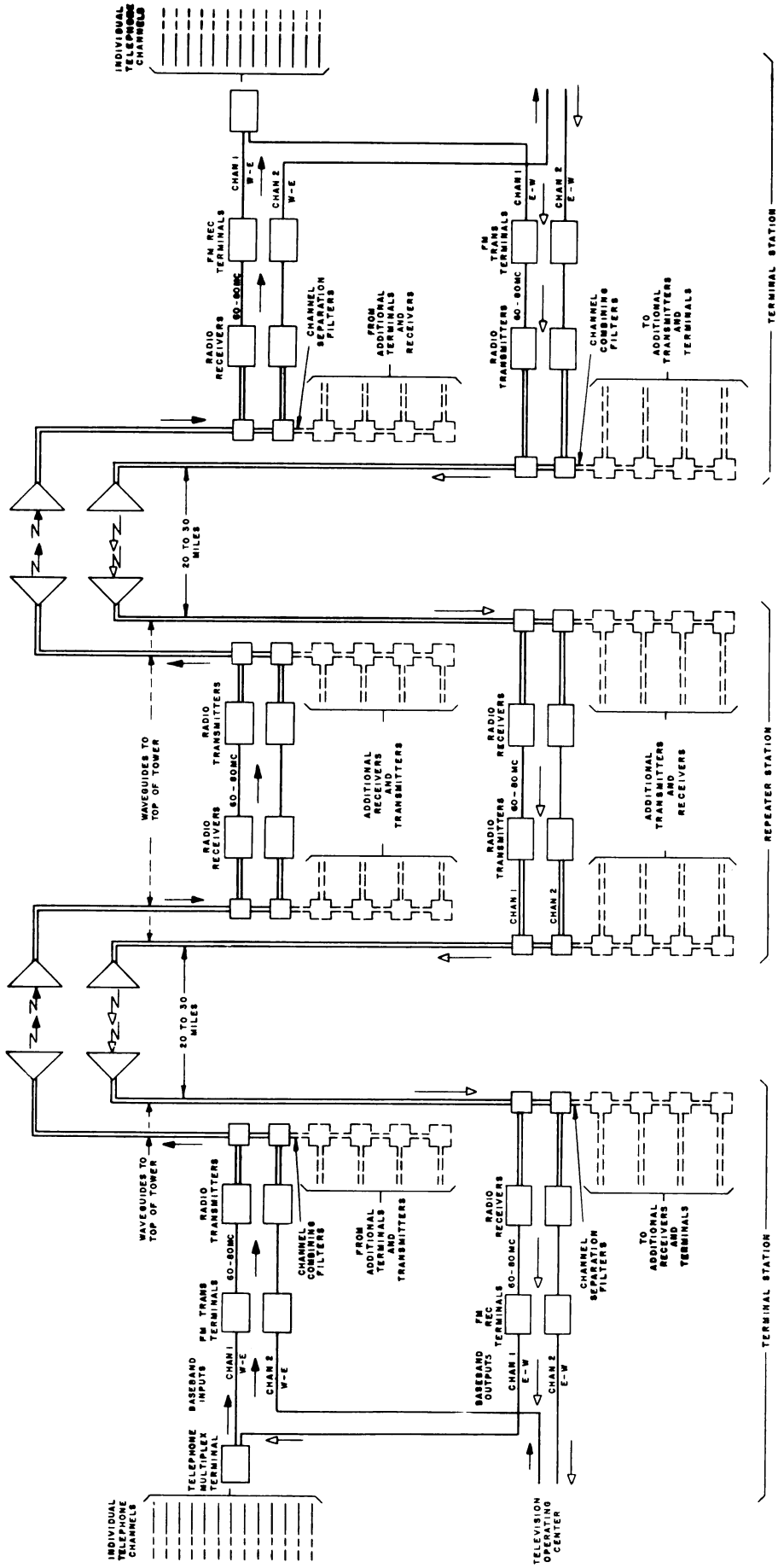
A problem of some magnitude in broadband wire systems is the variation in loss of the cable with respect to frequency. Equalization must be provided in order that the repeater closely match the loss characteristic of the cable. Conversely, the loss of the radio path is essentially constant for all frequencies within a radio channel. Thus the radio system requires equalization only to correct the deficiencies of the repeater equipment itself. Figure 6 compares the loss characteristics of the wire and radio paths.



Typical Repeater Section Losses

Figure 17-6

Both wire and radio media are subject to variation with time. Cable loss varies with temperature, and broadband AM systems normally have automatic means of regulating repeater gain to match the changing cable loss. Radio path loss varies in accordance with atmospheric conditions. These fades may be quite large - 20, 30, or 40 db, and will be different functions of frequency at different times. The effect of a fade within a single radio channel is often relatively flat.



Block Diagram of Typical Microwave Relay System

Figure 17-1

Chapter 18

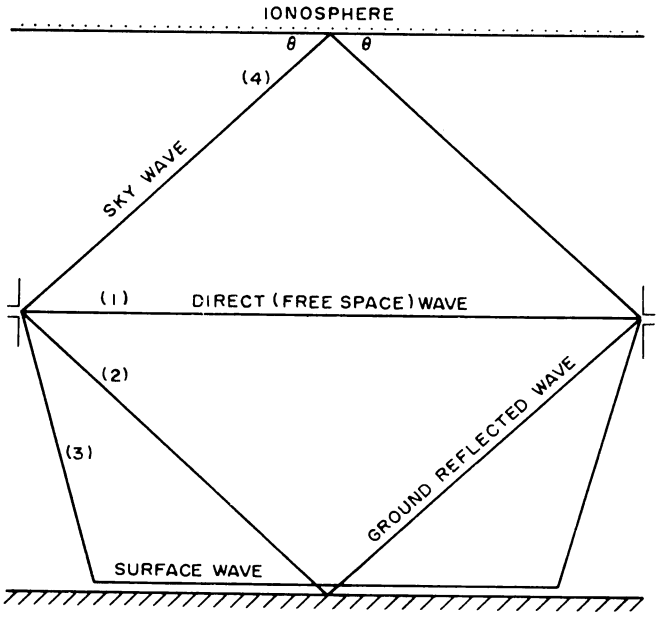
RADIO PROPAGATION

The nature of the electromagnetic wave radiated from an antenna is qualitatively discussed. Expressions are then given from which the antenna gain and free space path loss can be computed. Various types of microwave radio antenna systems and their characteristics are described. The chapter concludes with a discussion of path clearances, fading, and absorption phenomena at microwave frequencies.

Introduction

Some knowledge of radio propagation and antennas is essential to an understanding of transmission in microwave radio systems. The present chapter is meant to provide the student with certain basic concepts concerning this subject.

The normal propagation paths which exist between two antennas are illustrated in Figure 1. The direct or free-space wave is shown as line 1, and the wave reflected from the ground is line 2. Line 3 indicates a surface wave which consists of the electric and magnetic fields associated with the currents induced in the ground. Its magnitude depends



Transmission Paths Between Two Antennas

Figure 18-1

on the constants of the ground and the type of polarization used. The sum of these three, taking into account 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.

Line 4 indicates the type of transmission path between two antennas that is called the sky wave. This path depends on the presence of the ionosphere, an ionized layer about the earth that reflects back some of the energy that normally would be lost in outer space.

All of the possible paths shown in Figure 1 exist in any radio propagation problem, but some are negligible in certain frequency ranges. At frequencies less than about 1500 kc the surface wave provides the primary coverage and the sky wave helps to extend this coverage at night when the absorption of the atmosphere is at a minimum. At frequencies above about 30-50 megacycles the free space and ground reflected wave 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 ordinarily a source of occasional long distance interference rather than a reliable signal for communication purposes.

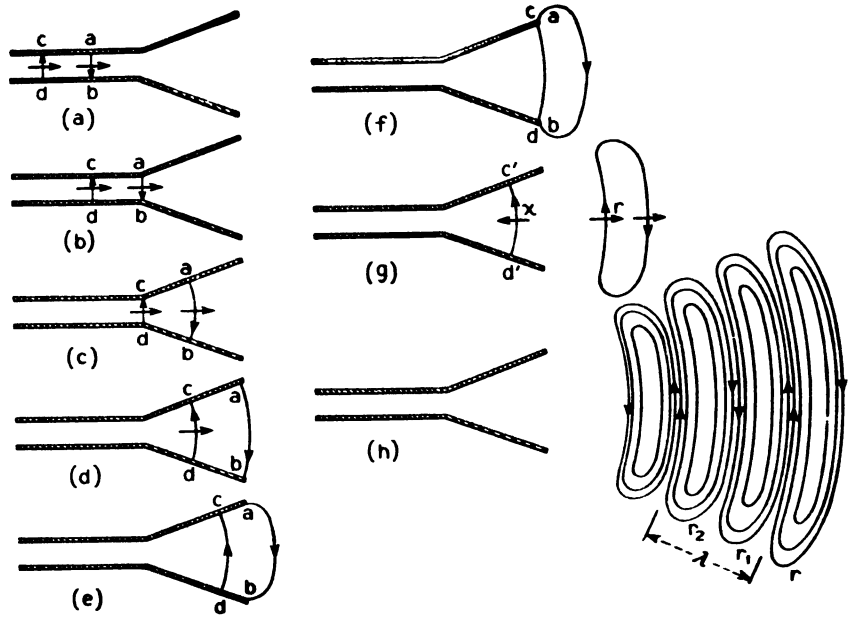
At frequencies of the order of thousands of megacycles the free space wave is usually controlling on good optical paths, although reflected waves are contributors to fading phenomena. Since our interest is in microwave systems, the surface and sky waves can be neglected and attention focused only on those phenomena that affect the direct and reflected waves. Free space transmission will be discussed first, then antenna properties and types, and finally deviations from free space transmission.

Free Space Transmission

Straight-line transmission through a vacuum or an ideal atmosphere, with no absorption or reflection of energy by near-by objects, is referred to as "free space" transmission. The microwave line-of-sight path fits this description quite well if we overlook atmospheric anomalies.

In a simple qualitative way, the energy radiated by an antenna may be thought of as consisting of packets of electric and magnetic lines of force. The electric lines of force (the "E" component of the electromagnetic wave) are a measure of the direction and magnitude of the force which would be exerted on a unit positive charge at any point in the

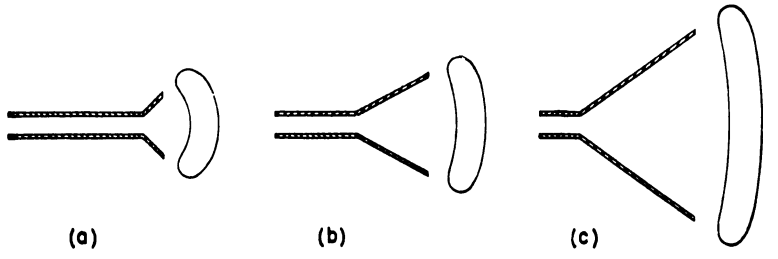
field; in an analogous way, the magnetic lines of force (H) give vectorial information on the force which would be exerted on a northseeking magnetic pole at any given point. The E and H components in the radiated wave front are everywhere in phase and at right angles to each other and to the direction of propagation. Figure 2 illustrates, in over-simplified and idealized form, the radiation from a flared two-wire transmission line, for example. Here only electric lines of force are shown; in the successive illustrations (a) to (g) we see two lines 180° apart as they progress down the transmission line, until at the discontinuity at the end some of the energy is radiated (r) and some reflected (x). Between the two lines of force shown, and on either side of them, there are of course innumerable other lines of force, and there are magnetic components of the field at right angles to the electric lines, into and out of the paper. Some of the other electric lines of force are shown in the final drawing (h), illustrating the successive packets of lines r_1, r_2, r_3.



Successive Epochs in a Highly Idealized Representation of Radiation from the Flared End of a Transmission Line

Figure 18-2

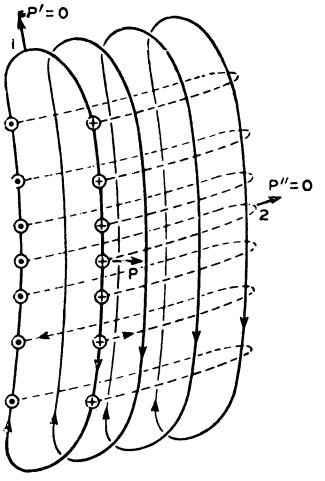
Several points brought out by this representation are to be noted. In the first place, observe the dispersion of the energy as it is propagated out from the antenna in (h). Since this dispersion will be horizontal as well as vertical, the energy per unit area of wave front will decrease as the square of the distance from the antenna. If we are trying to transmit power to another antenna at some distance away, this dispersion militates against energy collection at the receiving antenna. One remedy is to make the receiving antenna very large, of course, and so intercept the energy in spite of the fact that it is dispersed -- or at least collect more of it than we could with a smaller antenna. Another remedy, however, is to make the transmitting antenna large -- for, as Figure 3 illustrates (again in an over simplified way) this will reduce the dispersion. Here we have one example of the reciprocal nature of radiation and reception -- a good radiator is a good absorber, in the same way and to the same degree.



Illustrating How Radiating Systems Of Large Aperture May Give Rise to Wave Fronts of Large Radius of Curvature And Hence Lead to Increased Directivity

Figure 18-3

Another point to note is the fact that the wave is polarized. Consider, for example, the wave packet shown in Figure 4, which would be radiated by a properly arranged array of dipoles like the one of Figure 2, or by a horn. Here the solid lines represent the electric lines of force, and the dotted are the magnetic lines of force. We would describe this wave as vertically polarized, using the direction of the E vectors or lines of force as our reference, and meaning that a unit charge would move vertically if released in the middle of this field as shown by the arrows on the E lines. Another way of thinking of the polarization is to notice, by comparing this diagram with Figure 2 and remembering the



Highly Idealized Representation of a Wave-packet Radiated by a Typical Microwave Source. One Half of the Total Packet is Assumed to be Cut Away.

Figure 18-4

identical way in which receiving and transmitting antennas behave, that a simple dipole would collect energy most efficiently from this wave if the dipole were vertical, least efficiently if horizontal.

The energy contained in any unit volume of the wave, which could be collected by an absorber of unit area placed perpendicular to the direction of propagation, is given by the product (strictly speaking, the vector cross product) of the E and H components -- the so-called Poynting vector. The "P" vectors represent Poynting vectors; it is apparent that the value of the Poynting vector is large and nearly constant over the middle of the wave front, and diminishes to almost zero where the H vector approaches zero (P' at 1) or the E vector reaches its minimum (P" at 2), though the other vector in each case is still large.

Received Power for Field E and Antenna Area A

These concepts may be illustrated and made more quantitative by the following example - which is not, however, necessary for the main line of our discussion of microwave systems. Suppose we investigate the power received by an antenna having some arbitrary area. To do this it is convenient to start with the Poynting vector

$$\vec{E} \times \vec{H}$$

which represents the power flowing in an electromagnetic wave. The units of E and H are volts per meter and amperes per meter, respectively. The analogy to the more usual lumped parameter circuit cases is evident.

We now integrate the Poynting vector over an area.

$$P = \oint (\vec{E} \times \vec{H}) \cdot dA \qquad (18\text{-}1)$$

This is the power flowing through the area. If we consider the receiving location to be at a distance from the transmitting antenna, the wave will be plane and uniform over the area in question. If the area of the receiving antenna is A, perpendicular to the direction of travel of the wave, then (18-1) simplifies to

$$P = E\,H\,A \qquad (18\text{-}2)$$

There is however a fixed relationship between the magnitudes of the E and H vectors.

$$E = \sqrt{\frac{\mu}{\epsilon}}\, H \qquad (18\text{-}3)$$

$$= 120\pi\, H \text{ in free space} \qquad (18\text{-}3a)$$

Here μ is the permeability and ϵ is the dielectric constant. The ratio of E to H, 120π, is called the intrinsic impedance of space.

Combining (18-2) and (18-3a), the received power P_R for a receiving antenna of area A is

$$P_R = \frac{E^2 A}{120\pi} \qquad (18\text{-}4)$$

Power Loss - Isotropic Transmitting Antenna to Receiving Antenna

Let us now consider the power impinging on a receiving antenna when a given amount of power P_T is being radiated from a transmitting antenna. From what has been said earlier, it is clear that the field strength at the receiving antenna will be a function of the area of the transmitting antenna - that is, a function of the extent to which the transmitting antenna focuses the radiated energy in the direction of the receiving antenna. A convenient reference case to consider is the highly artificial one of an "isotropic" transmitting antenna - that is, one which radiates with equal efficiency in all directions.* Any physical

*Some texts use the half-wave dipole as the reference antenna and consider its gain as unity or 0 db. With respect to the isotropic radiator, the half-wave dipole has a gain of 2.15 db.

antenna that we shall be interested in will be more efficient in delivering power to the receiving antenna - i.e., will have "gain" relative to the reference case of the isotropic antenna. (Physically, in terms of the sort of simplified pictures of radio propagation that we have been considering, isotropic antennas are difficult to think about. An isotropic transmitting antenna would be a single point whose potential varied positive and negative at the frequency of the radiated wave, so the question of the polarization of the wave becomes somewhat paradoxical. Equally paradoxical is the problem of receiving energy with an isotropic antenna having no area. These problems are best avoided by saying firmly that isotropic antennas radiate equally in all directions, have an effective area of $.08\ \lambda^2$, and have no other characteristics or existence.)

Imagine a sphere of radius d centered on an isotropic antenna which is radiating a power P_T. All the radiated power will pass through the area of the sphere's surface, which is $4\pi d^2$. The power passing through an area A will be, therefore,

$$P_R = \frac{A}{4\pi d^2} P_T \qquad (18\text{-}5)$$

It is convenient at this point to start distinguishing areas, since we shall soon be concerned with the area of a physical transmitting antenna. Calling the area of the receiving antenna A_2, then, the power loss from isotropic transmitter to the receiving antenna is

$$\frac{P_T}{P_R} = \frac{4\pi d^2}{A_2} \qquad (18\text{-}6)$$

Antenna Gain

We still need an expression for the "gain" of the transmitting antenna - the amount by which focused transmission is more efficient than transmission which is equally good in all directions - before we can compute the loss in power between two physical antennas. The derivation of the expression for antenna gain would require the presentation of considerable background material and will not be attempted here. From what has been said, it is clear that what we seek is the ratio of two powers - one, the power intercepted by a given receiving antenna when some power P_T is radiated in a nearly plane wave by a transmitting antenna of area A_1, - the other, the received power when P_T is radiated by an isotropic antenna at the same location. The ratio of these two powers we will define as the "gain" of an antenna of area A_1; it is found to be:

$$\text{Transmitting antenna gain} = \frac{4\pi A_1}{\lambda^2} \qquad (18\text{-}7)$$

Here λ is the wavelength in units consistent with the units in which A_1 is measured, so that the term $\frac{A_1}{\lambda^2}$ is numerically the area of the transmitting antenna in square wavelengths.

In actual antennas there are losses which result in the gain being less than theoretical. This discrepancy between actual and theoretical gains is usually 2 or 3 db. It is often taken into account by assigning to the antenna an "equivalent area" which is less than its physical area.

Loss Between Two Antennas

Collecting the results so far, we can write, for the ratio of power received by an antenna of area A_2 from an antenna of area A_1, a loss expression obtained by combining 18-6 and 18-7:

$$\frac{P_T}{P_R} = \frac{4\pi d^2}{A_2} \cdot \frac{\lambda^2}{4\pi A_1} = \frac{d^2\lambda^2}{A_1 A_2} \qquad (18\text{-}8)$$

Two Illustrative Examples

In the next section we shall define certain quantities (receiving antenna gain, free space path loss) which are commonly used in microwave system discussions. Before we do so, it is worthwhile to point out that we can solve some propagation problems with only the concepts developed thus far. This approach has the merit of emphasizing the physical realities of the problem, and the way in which the various parameters are related to each other. The electric field intensity, E, is the focal point of interest. For our purposes here we can summarize the results of Page 18-6 as follows:

(a) Power density at any point is proportional to the square of field intensity

$$P \text{ density} \sim E^2 \qquad (18\text{-}9)$$

(b) The received power is proportional to the square of the field intensity and the receiving antenna area

$$P_R \sim E^2 A_2 \qquad (18\text{-}10)$$

Lastly, we can write equation 18-8

$$P_R = \left[\frac{A_1}{\lambda^2}\right] \cdot \frac{1}{d^2} A_2 P_T \qquad (18\text{-}11)$$

where $[\frac{A_1}{\lambda^2}]$ is the area of the transmitting antenna in square wavelengths and d^2 and A can have any length^2 units, as long as they are the same. After this preliminary work, we are now ready to look at a couple of simple examples.

A. Two antennas face each other on towers twenty miles apart. The overall loss from antenna 1 input to antenna 2 output is 60 db at 5000 mc. What is the loss at 10,000 mc?

When the frequency is doubled, the solid angle of radiation is halved - the concentration of energy is increased four-fold. Another way of stating this is to say that the area in square wavelengths of the transmitting antenna is quadrupled by the doubling of the frequency. Hence, since distance remains the same, E^2 will be four times as great as previously for the same transmitted power. Receiving antenna area remains the same so the received power is up by a factor of four. The loss at 10,000 mc is 54 db.

B. Two antennas of four-foot diameter face each other on towers twenty miles apart. The overall loss is 60 db at 5000 mc. What would the loss be if the diameter of the antennas was made 8 feet?

Quadrupling transmitting antenna area gives four times the original E^2 and thus four times the power density at the receiving antenna. Quadrupling the receiving antenna area results in a second factor of four. Loss is hence decreased by 12 db.

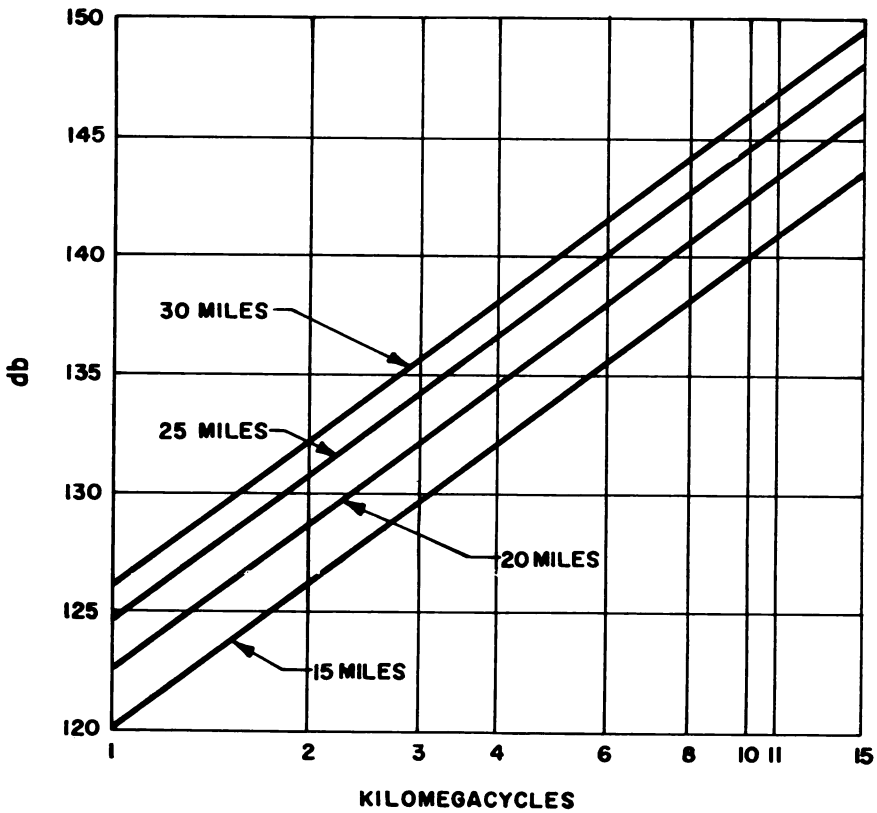
Free Space Path Loss

In more complex problems, it is adventageous to be able to consider the antenna gains and the path loss independently. The path loss between isotropic radiators can be determined from the same derivation in which the antenna gain is determined or can be obtained by mathematical manipulation of Equation 18-8. We start by recalling the reciprocal nature of antenna performance with respect to radiation and absorption, and set up the definition that "antenna gain" is the same whether we speak of transmitting or receiving. By this definition

$$\text{Receiving Antenna Gain} = \frac{4\pi A_2}{\lambda^2}$$

Equation 18-8 can then be rewritten to include the transmitting antenna and receiving antenna gains as:

$$\underset{\text{Overall Loss}}{\frac{P_T}{P_R}} = \underbrace{\left(\frac{4\pi d}{\lambda}\right)^2}_{\text{Free Space Path Loss}} \div \underbrace{\left(\frac{4\pi A_1}{\lambda^2}\right)}_{\text{Trans Ant Gain}} \cdot \underbrace{\left(\frac{4\pi A_2}{\lambda^2}\right)}_{\text{Rec Ant Gain}} \qquad (18\text{-}12)$$



Free Space Path Loss

Figure 18-5

The part of the equation not included in transmitting and receiving antenna gains is the "free space path loss". It will be noted that, since both the transmitting and receiving antenna gains are given as gain relative to an isotropic radiator, the term $[\frac{4\pi d}{\lambda}]^2$ is the free space path loss between isotropic radiators and is a function of the distance between antennas in wavelengths.

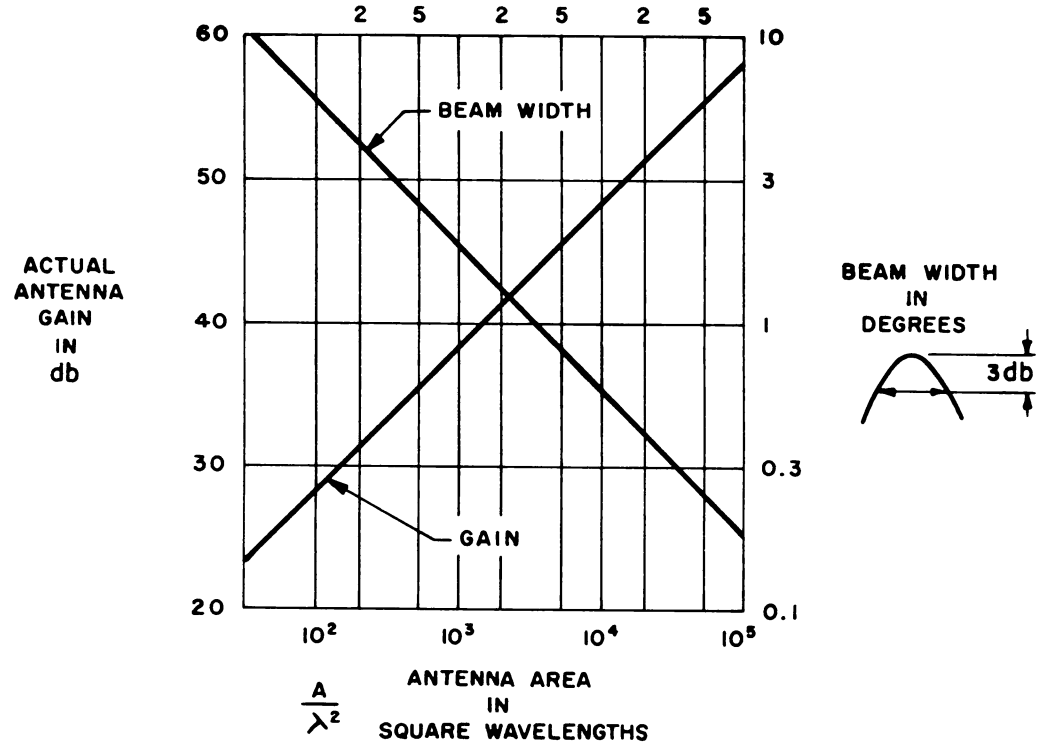
In db we have

$$\text{Loss} = 20 \log \frac{4\pi d}{\lambda} - 10 \log \frac{4\pi A_1}{\lambda^2} - 10 \log \frac{4\pi A_2}{\lambda^2} \qquad (18\text{-}13)$$

The free space path loss, $20 \log \frac{4\pi d}{\lambda}$, is plotted versus frequency in Figure 5, with "d" in miles as a parameter.

Antenna Characteristics

There are a number of antenna characteristics which are of primary importance in a microwave system.



ANTENNA	4000MC	6000MC	11000MC
	$\frac{A}{\lambda^2}$		
8 FT PARABOLOID	8×10^2	2×10^3	6×10^3
4 FT PARABOLOID	2×10^2	5×10^2	1.5×10^3
8 FT SQUARE	1×10^3	2.5×10^3	8×10^3

Approximate Antenna Gain And Beam Width

(Abcissa is actual antenna area and "actual antenna gain" is taken to be 3 db below theoretical)

Figure 18-6

The first of these is antenna gain. An antenna has gain because it concentrates the radiated power in a narrow beam rather than sending it uniformly in all directions as an isotropic antenna does. Since it reduces the net path loss, high antenna gain is obviously desirable. Antenna gain is increased by increasing the antenna area.

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 must decrease as the antenna gain is increased. Antennas used in microwave systems ordinarily have beam widths of the

order of one degree (see Figure 6). A narrow beam is desirable in order to minimize interference from outside sources and adjacent antennas. Too narrow a beam, however, imposes severe mechanical stability requirements and leads to problems in antenna lineup and fading.

There are several antenna characteristics which are important in evaluating the interference to be expected between adjacent transmitting and receiving antennas. One property is the front-to-back ratio. This is defined as the ratio of the power received from (or transmitted by) the front side of the antenna to the power received from (or transmitted to) the back side, and is usually expressed in db. This ratio can generally be read directly from the radiation pattern of the antenna. Two front-to-back ratios may be given for an antenna. The ideal front-to-back ratio is the ratio that would exist if the antenna were isolated in free space. The effective front-to-back ratio is the ratio that would be measured in a typical antenna installation, and may be 20 to 30 db below the ideal ratio because of relections from the foreground or from objects in or near the main beam of the antenna. The front-to-back ratios for the antennas described in the next section are effective ratios. One use of the front-to-back ratio in systems analysis is in computing the interfering effect between a transmitting antenna on one tower and a receiving antenna on the preceding tower.*

Side-to-side coupling expressed the fraction of transmitted power that is received by a second antenna located along side the transmitting antenna. It is generally expressed in db. The usual practice is to give the effective, rather than ideal, side-to-side coupling for the particular types of antennas, as measured for specific side-to-side orientations of these antennas.

The back-to-back coupling expresses, in db, the fraction of the transmitted power received by a second antenna located to the rear of the transmitting one, on the same tower. The value of back-to-back coupling quoted for an antenna is normally the effective coupling as measured in a typical antenna installation. Both side-to-side and back-to-back couplings are useful in computing the interfering effects between transmitting and receiving antennas located on the same tower.

Some antennas must transmit both vertically and horizontally polarized waves and their cross-polarization discrimination is important.

There should be a good impedance match between the antenna and radio transmitter in order that reflections do not distort the

- -

*This problem is discussed in more detail in Chapter 23.

transmitted signals. The use of isolators, waveguide devices which pass the signal in the desired direction but block the reflections, has reduced somewhat the importance of maintaining a good impedance match. Such a match would be difficult to obtain since the transmitter has a poor output impedance.

The size and weight of the antenna system are important. The antenna tower and associated antenna mounting arrangements are sometimes an appreciable portion of the cost of a microwave system. When this is true a balance must be made between the cost saving possible through the use of a lightweight antenna on a lightweight, inexpensive tower and the improved transmission performance which may result from using a large and heavier antenna with its correspondingly more rugged and expensive tower.

In choosing an antenna the difficulty of construction and maintenance is another consideration. One must consider the mechanical tolerances which must be attained in production and maintained in the field under conditions of ice and wind loading. A common rule of thumb for mechanical tolerances on reflecting surfaces is that dimensions should be held within $\frac{\lambda}{16}$. Since at 11,000 mc this is 1/16 inch, it is readily seen that the construction of large antennas is not simple.

The choice of antenna for any particular system is a result of a careful weighting of the factors noted above to produce the most efficient arrangement within the cost framework dictated by system economics.

Typical Microwave Antennas

Paraboloid Antenna:

This is an inexpensive and simple high gain antenna. It consists of a paraboloid dish reflector supplied with microwave energy by a small feed horn located at the focus of the paraboloid. Disadvantages include poor impedance match, narrow bandwidth (about 10%), and poor suppression of radiation from the rear. It is widely used in spite of these shortcomings. For example, an 8-foot dish is often used on lightly loaded TD-2 routes (i.e., routes using only a few of the available channels) where cost reduction is essential. This antenna has 37 db gain, a beam width of 2.4 degrees, and a front-to-back ratio of 46 db at 4000 mc.

Paraboloid Periscope Antenna:

It is desirable to place the paraboloid dish as near as possible to the transmitter. The objectives here are to reduce the losses in the waveguide run feeding the antenna and to reduce the delay of the echo caused by the poor impedance of the transmitter. In general, the dish is mounted near the ground and directed upwards, and a reflector mounted on

a tower above is used to direct the beam towards the next repeater station. Such an arrangement is used in TJ. Here the basic antenna is a 5-foot parabolic dish mounted on the roof of the equipment building and providing a normal gain of 42.1 db. Two standard reflectors are available: a 6-foot by 8-foot plane reflector and an 8-foot by 12-foot sheet that can be used either as a plane or curved reflector. By suitably spacing the antenna and reflector, the combined antenna system can have as much as 3.5 db gain over the parabolic antenna alone, as shown in Figure 7.* Under the condition of optimum spacing (i.e., for maximum antenna system gain) the beam width is 0.6 degree with the 8 x 12-foot reflector and 0.8 degree with the 6 x 8-foot reflector. The disadvantages of this arrangement are that it places more stringent requirements on the tower stability (± 1/4 degree for sway and ± 1/2 degree for torsion in TJ), makes antenna alignment more difficult, and results in greater interference between adjacent antennas.

Horn Reflector Antenna:

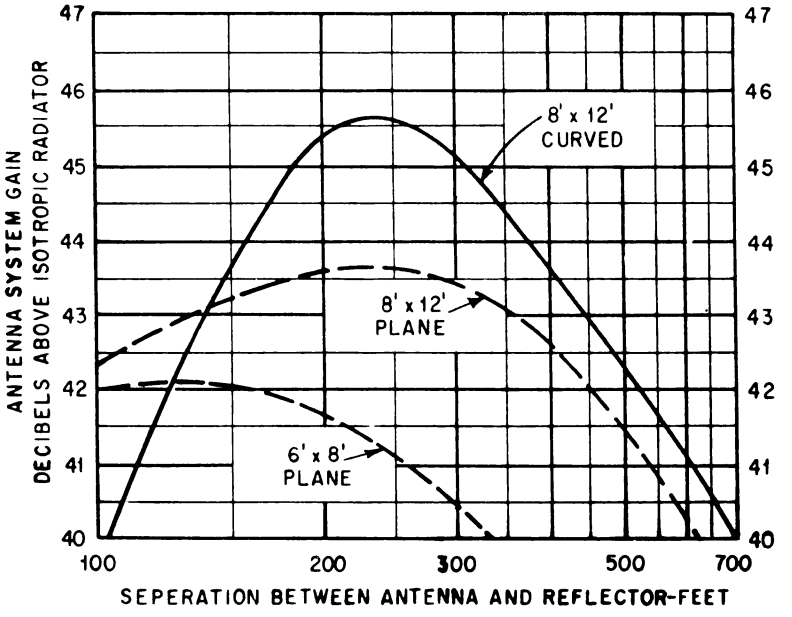
In this antenna a vertical horn is used to illuminate a section of a parabolic surface. Because of the design and size of the horn the impedance of this antenna is very good, the return loss being between 50 and 60 db. It is a broadband antenna and can be used with both vertical and horizontal polarization in the 4000 mc, 6000 mc, and 11,000 mc bands. Nominal characteristics at 6000 mc are 43 db gain and 1.5 degrees beam width. Due to its shielded construction, it radiates very little power to the rear, resulting in a nominal 70 db front-to-back ratio. Measurements made on a large number of antenna installations have shown the average side-to-side coupling to be of the order of 97 db. Back-to-back coupling between these antennas may run as high as 120 db, but the value measured at a particular installation may be 10 to 20 db below this value because of additional coupling caused by leakage of energy at the joints in the waveguide run feeding the antennas. A disadvantage is its bulk (large surface area and about 1 ton weight) and difficulty in mounting. Construction is somewhat expensive.

Delay Lens Antenna:

The delay lens antenna of the TD-2 system uses thin metallic strips supported in Styrofoam as a microwave lens which gives 39 db gain and a beam width of 2 degrees at 4000 mc. A short rectangular horn is used to feed the delay lens. This antenna has the advantages of the horn reflector in respect to good impedance match and suppression of rear

- -

*An analysis of this type of antenna system can be found in references 10 and 11 at the end of this chapter.



TJ Antenna System Gain
(Gain of 5-foot Parabola Alone = 42.1 db)

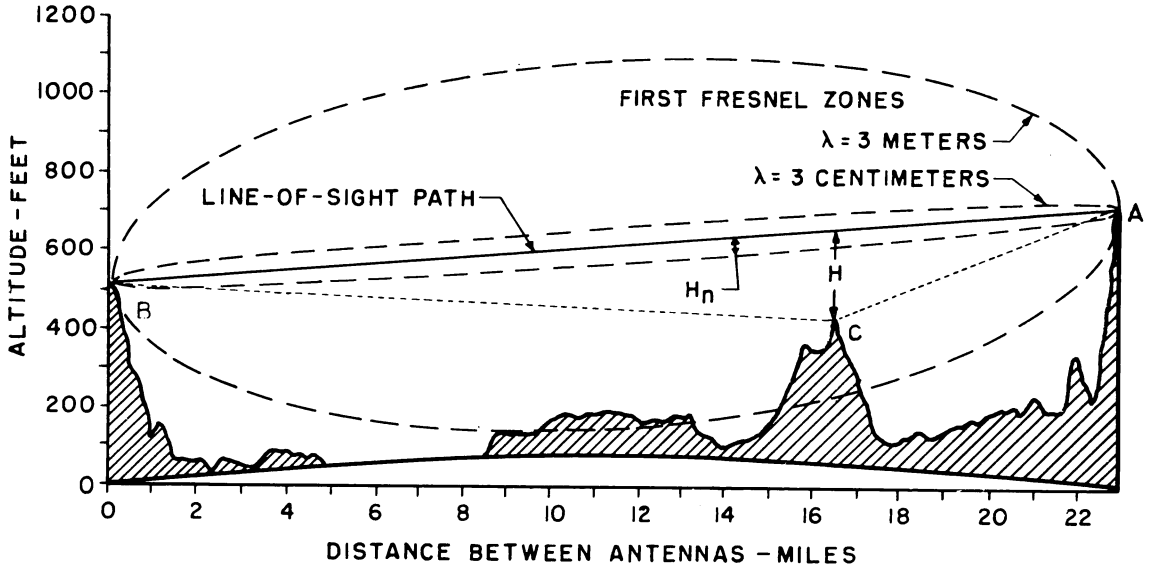
Figure 18-7

radiation. The chief disadvantage is the restricted bandwidth due to the fact that the refraction produced by the lens is a function of wavelength. Its construction is such that it transmits only a single polarization.

Antenna Heights and Path Clearance

Up to this point the present chapter has considered only a free space transmission path, and has not taken into account the effects of the presence of the earth and the non-uniformity of the atmosphere. In order to consider actual operating conditions, we must distinguish between normal and abnormal transmission.

Under normal transmission conditions, i.e., for a large percentage of the time, the path loss of a typical microwave link can be made to closely equal 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 conditions from deviating from the free space value, but also to prevent severe fading problems during periods of abnormal transmission.



Typical Profile Plot, Showing First Fresnel Zones for 100 mc and 10 kmc

Figure 18-8

The importance of adequate clearance can be seen by considering Figure 8, which shows the profile of the path between two antenna sites. For the antenna heights shown, the distance H represents the closest proximity between the line-of-sight path AB and the intervening terrain. Path ACB represents a secondary transmission path caused by the reflection of energy from the projection. If there were no phase reversal at the point of reflection, the energy received over the two paths would cancel whenever AB and ACB differed by one-half wave length, or any odd multiple of a half-wavelength. When the grazing angle of the secondary wave is small, however, which is the usual situation when the obstacle is far from the antenna, a phase reversal will normally occur at the point of reflection. Therefore, whenever AB and ACB differ by one-half wavelength, or any odd multiple of a half-wavelength, the energies of the received signals will actually add, rather than cancel. Conversely, if the two paths differ by a wavelength, or by any whole number of wavelengths (or even number of half-wavelengths) the signals from the two paths will tend to cancel. Evidently, the amount of clearance between the line of sight path and the obstruction must be chosen to minimize the effect of secondary paths.

The amount of clearance is generally described in terms of Fresnel zones. All points from which a wave could be reflected with a delay of one-half wavelength form the boundary of what is defined as the first Fresnel zone. Similarly, the boundary of the n-th 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 n-th Fresnel zone is given by

$$H_n = \sqrt{\frac{n\lambda d_1 (d-d_1)}{d}}$$

where λ is the wavelength. The boundaries of the first Fresnel zones for λ = 3 meters (100 mc) and λ = 3 centimeters (10 kmc) are shown in Figure 8.

Measurements have shown that to achieve a normal transmission loss approximately equal to the free space loss, the line-of-sight path should pass over all obstacles with a clearance of at least 0.6 the first Fresnel zone distance and preferable by an amount equal to the first Fresnel zone. The procedure is to obtain a profile plot of the terrain between the proposed antenna sites and determine the worst obstacle in the path, such as the ridge shown in Figure 8. This obstacle can then be used as a "leverage point" from which the most suitable antenna heights at each location can be chosen to provide the proper Fresnel zone clearance.

So far, nothing has been said about refraction, or the effects of refraction. Refraction refers to the bending of a wavefront and is caused by a change in the velocity of one part of the wave-front with respect to another as the front passes obliquely from one medium to another. Refraction occurs in air when, for example, the wave-front impinges obliquely on two layers of air having different densities. Much of the abnormal transmission phenomena observed at microwave frequencies is believed to be caused by refraction effects.

Refraction will frequently be an important part of normal transmission. Consider, for example, a microwave signal which we attempt to transmit parallel to a tangent to the earth drawn at the transmitting antenna. As the signal progresses, it will, of course, get further and further from the earth as the earth bends away from it. However, in traveling such a path the signal normally encounters air of decreasing density. Since the top of the wave-front reaches the lighter air first,

this portion of the wave will increase its speed relative to the bottom portion, with the result that the wave tends to bend back toward the earth. In other words, the refraction of normal air causes a microwave signal to curve towards the earth, or, in effect, to make the earth appear to flatten out. Frequently, a radius of curvature of 4/3 times that of the actual earth's radius is used to account for the refraction effect. Special profile paper for path plotting is available on which the earth's radius is assumed to be 4/3 the actual radius. Such paper permits plotting the signal path as a straight line and is convenient to use when refraction along the transmission path must be taken into consideration.

The determination of suitable antenna heights for a particular path is frequently a difficult and rather non-exact job. Some general rules have been formulated. For example, in many cases involving flat terrain it is recommended that the Fresnel zone clearance be determined by using an effective 4/3 earth radius, and that the antenna heights be selected to provide little or no excess clearance over that required for the first Fresnel zone. Experience shows that excess clearance for such terrain will frequently increase the fading problem described in the next section. Where the terrain is rough, and only a single major obstruction is present, such as a sharp ridge, the effect of excess clearance is not as significant. Such paths over rough terrain are generally engineered using the earth's true radius, and refraction effects are ignored. In many cases, however, general rules such as these cannot be applied because of peculiar terrain conditions or other factors. It is frequently necessary, therefore, to make actual transmission measurements over the proposed path, using various antenna heights at each repeater location to determine the optimum heights.

Fading

During abnormal conditions, the path loss may differ considerably from the normal. Although the path loss may decrease under abnormal conditions, the more usual case is for increases of 10, 20, 30 or more db to occur for short periods. The margins that must be provided against fading are important in determining system parameters.

The factors involved in fading phenomena are many and complex. In general, the number of fades per unit time increases as both the distance between antennas and the transmitting frequency are increased. However, measurements have shown that the duration of a fade tends to decrease with both distance and frequency. As a result, the percent

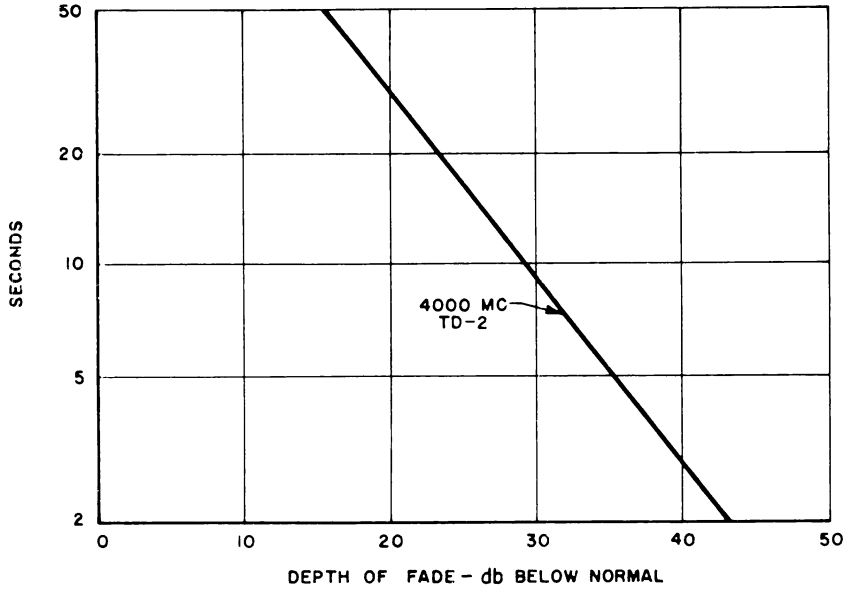
time that a system is experiencing a particular depth of fade tends to be independent of repeater spacing and frequency. Exact explanations for all fading phenomena are not yet available, but some general characteristics can be stated. It is known that during the daytime when the lower atmosphere is thoroughly mixed by rising convection currents and winds, the signals on line of signt paths are normally steady and are near predicted free space values. Also during the winter months, when the humidity content of the atmosphere is low, signal variations are usually small. However, on clear summer nights with little or no wind, non-uniform distributions of temperature and humidity can create steep dielectric constant gradients in the lower atmosphere, thus causing anamalous propagation and fading.

The most common type of fading is that of multiple path transmission. Two, three, and sometimes more signal components will arrive at various angles in the vertical plane, usually above the line of sight. Wave interference among these components produces fading whose severity depends upon the relative amplitudes and delays of the components. In these cases different microwave frequencies fade differently and the signals received on two vertically spaced antennas fade differently. Either space or frequency diversity is useful in combating this type of fading.

Abnormal variations in the dielectric constant of the atmosphere are another cause of fading. Normally the dielectric constant of the atmosphere decreases with height above ground so that the ray path usually has a curvature in the direction of the earth's curvature. Occasionally the dielectric constant increases with height and this results in a situation where the wave is bent away from the earth and passes above the receiving antenna.

Another rarer type of fading is observed when a reflecting layer is situated above the transmission path. The signal then suffers interference between the reflected wave and the energy from the direct path. Ground or water reflections also occasionally play a part in fading when under certain conditions of atmospheric refraction even though under normal conditions the geometry of the path does not permit such reflection.

Atmospheric focusing or trapping is another possible cause of fading. In this case, due to changes in the dielectric constant of the atmosphere, the ray may be transmitted in a kind of waveguide or duct formed by the earth and a reflecting layer or by two elevated reflecting layers. This phenomenon occurs only very seldom.



Median Duration of Fast Fading

Figure 18-9

In assessing the effect of fading, the depth of the fade, the frequency of occurrence and the duration are all important. The depth of the fade can be any amount, but fortunately the deeper the fade the less frequently it occurs and the shorter its duration when it does occur. Figure 9 shows the average duration of fading for various depths of fade on a 4000 mc system where the average repeater spacing is 30 to 35 miles. It is seen that for this spacing the average duration of a 20 db fade is about 30 seconds and the average duration of a 40 db fade about 3 seconds.*

- -

*Recent 4000 mc data shows that for an average repeater spacing of 25 miles, the median duration of a 40 db fade is about 5 seconds. At a 45 mile average repeater spacing the median duration of a 40 db fade is approximately 2 seconds. The fading characteristics plotted for these repeater spacings tend to parallel the curve in Figure 9, at least for fades between 30 and 45 db. Data has also been collected on the duration of particular depths of fades, and it is found that a plot of number of fades versus the logarithm of their duration forms a normal distribution. This data shows, for example, that for a 25 mile repeater spacing, 99% of the 40 db fades will have a duration of about 41 seconds or less; for a 45 mile spacing, 99% of the 40 db fades have a duration of about 21 seconds or less.

A less common but potentially more serious type of fading that may last several hours or more is encountered in long paths. Accurate experimental data is not yet available on such conditions. It may be caused by either reflections from an elevated layer or by an increase in dielectric constant with height.

Fading is very difficult to analyze since it is a result of many varying factors. The only effective method we have for making predictions is a measurement program over a long period of time under varying atmospheric conditions and types of path. The margins which are included in a system for the effects of fading are based on such a propagation study, plus a decision on the frequency and duration of the intervals during which sub-standard system performance can be tolerated.

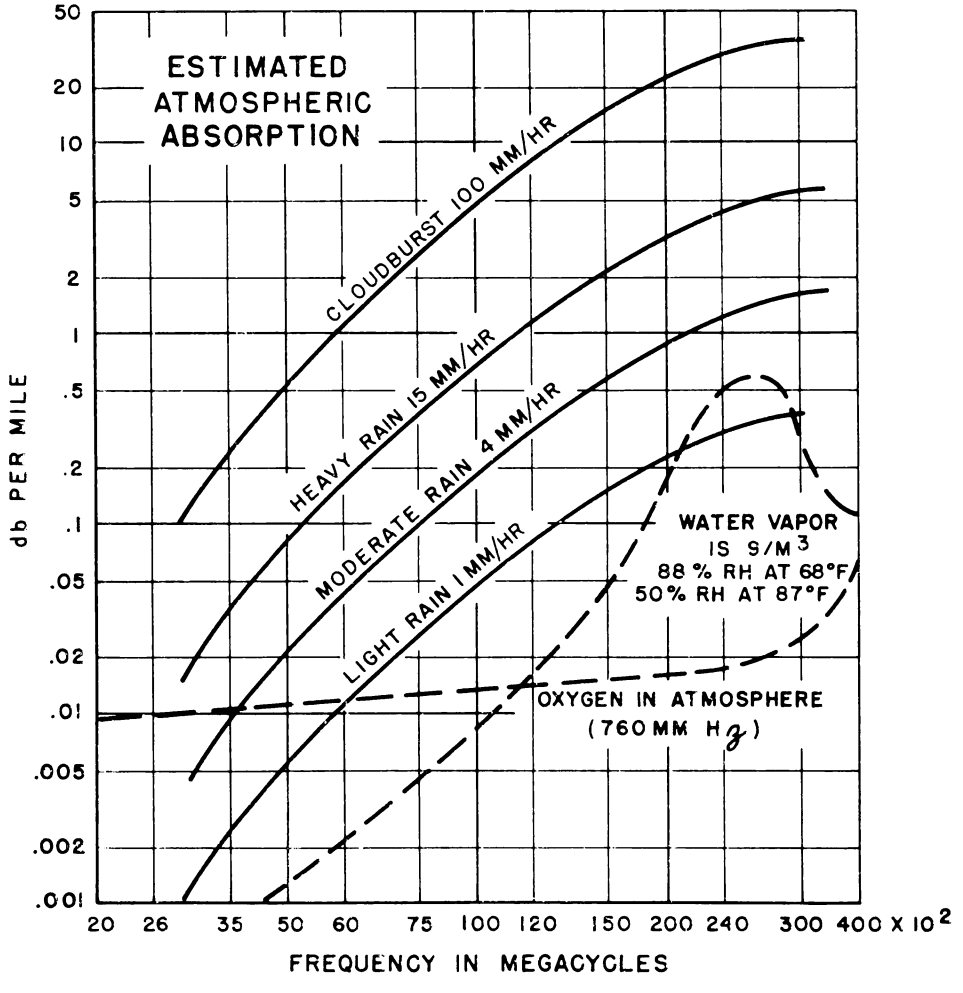


Figure 18-10

Absorption

In addition to the fading discussed above, rainfall and water vapor produce very pronounced 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 indicate that they should be found throughout the electromagnetic spectrum. The first absorption band due to water vapor occurs at about 22,000 megacycles and the first absorption band due to the oxygen in the atmosphere occurs at about 60,000 megacycles.

The adverse effect of rain on microwave radio propagation can be seen in the region of 4 to 6 kmc but is rather small relative to the losses introduced by various fading phenomena. For frequencies approaching 85 kmc and higher, 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 rain drops. Figure 10 indicates the estimated atmospheric absorption for various conditions of rainfall. From this figure it is evident that absorption must be considered in any system at frequencies of 10,000 mc or above and perhaps in lower frequency systems in areas where heavy rains occur frequently. Due to the difficulty associated with control of conditions, it is very difficult to accumulate good data on rain absorption, but experiments are continuing in this regard.

Bibliography

1 - Bullington, Radio Propagation at Frequencies above 30 Megacycles, BTL Monograph 1489.
2 - Crawford and Jakes, Selective Fading of Microwaves, BTL Monograph 1957.
3 - Schelkunoff, "Electromagnetic Waves", Chapter IX.
4 - H. T. Friis, Microwave Repeater Research, BTL Monograph B-1565, pages 19-28.
5 - W. E. Kock, Metallic Delay Lens, BTL Monograph 1519.
6 - Terman, Radio Engineers' Handbook, McGraw-Hill Book Company, 1943, pages 770-871.
7 - Southworth, George C., Principles and Applications of Waveguide Transmission, D. Van Nostrand and Co., 1950.
8 - Schelkunoff and Friis, Antenna Theory and Practice, John Wiley and Son, 1952.

9 - W. C. Jakes, Jr., Theoretical Study of an Antenna-Reflector Problem, Proc. IRE, vol. 41, pp. 272-274; February 1953.

10 - R. E. Greenquist and A. J. Orlando, Analysis of Passive Reflector Antenna Systems, Proc. IRE, vol 42, pp. 1173-1178; July 1954.

11 - "Siting of Fixed Radio Stations", Bell System Practice, R100.010.

12 - Statistical Study of Selective Fading of Super-High Frequency Radio Signals - R. L. Kaylor, BSTJ, vol. 32, pp. 1187-1202 1953 or Monograph 2186.

Chapter 19

PROPERTIES OF THE FREQUENCY MODULATED SIGNAL

Phase and frequency modulation are defined, and the similarities and differences between these two forms of angle modulation are discussed. The expression for the FM or PM signal is analyzed to determine the spectrum when the modulating signal consists of either one or two sinusoids. The problem of extending this analysis to cover more complex modulating signals is then considered .. Other topics include pre-emphasis, the bandwidth required for transmission, the effect of non-linear input-output characteristics, and limiters.

Introduction

The material presented in this chapter is essentially a review of certain aspects of modulation theory which are necessary as background for those chapters on FM systems analysis which follow. For a more detailed discussion of particular points the reader is referred to any standard text* on modulation theory.

Comparison of Amplitude Modulation and Angle Modulation

Any sinusoidal carrier may be subjected to two distinctly different types of modulation. These are amplitude modulation and angle modulation, both of which may be defined with reference to Equation (19-1).

$$M(t) = A \cos [\omega_c t + \varphi] \qquad (19\text{-}1)$$

Here A is the amplitude of the sinusoidal carrier and $[\omega_c t+\varphi]$ is the angle. The carrier frequency is ω_c radians/sec. If the coefficient A is by some means varied with time, amplitude modulation is obtained. If, instead, φ is varied with time the result is angle modulation. The general angle modulated wave might then be expressed as follows.

$$M(t) = A_c \cos [\omega_c t + \varphi(t)] \qquad (19\text{-}2)$$

where $M(t)$ = angle modulated carrier

A_c = a constant

ω_c = carrier frequency in radians/second

$\varphi(t)$ = angle modulation in radians

- -

*See reference at the end of this chapter.

If angle modulation is used to transmit information it is necessary that $\varphi(t)$ be a prescribed function of the modulating signal to be transmitted. For example, if V(t) is the modulating wave or signal to be transmitted, the angle modulation $\varphi(t)$ can be expressed mathematically as

$$\varphi(t) = f[V(t)] \qquad (19\text{-}3)$$

Many varieties of angle modulation are possible depending on the selection of the functional relationship. Two of these are important enough to have the individual names of phase modulation and frequency modulation.

Phase Modulation and Frequency Modulation

The difference between phase and frequency modulation can be readily understood by first defining four terms, as follows:

The instantaneous phase and instantaneous phase deviation are, with reference to Equation (19-2),

$$\text{Instantaneous phase} = \omega_c t + \varphi(t) \text{ radians} \qquad (19\text{-}4)$$

$$\text{Instantaneous phase deviation} = \varphi(t) \text{ radians} \qquad (19\text{-}5)$$

The instantaneous frequency of an angle modulated carrier is defined as the first time derivative of the instantaneous phase. In terms of Equation (19-2) the instantaneous frequency and the instantaneous frequency deviation are

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

$$= \omega_c + \varphi'(t) \text{ radians/second} \qquad (19\text{-}6)$$

$$\text{Instantaneous frequency deviation} = \varphi'(t) \text{ radians/second} \qquad (19\text{-}7)$$

From these definitions the difference between phase modulation (PM) and frequency modulation (FM) is easily defined. Phase modulation is angle modulation in which the instantaneous phase deviation, $\varphi(t)$, is proportional to the modulating signal V(t). Similarly, frequency modulation is angle modulation in which the instantaneous frequency deviation $\varphi'(t)$ is proportional to the modulating signal V(t). Mathematically these statements become,

for phase modulation

$$\varphi(t) = k\ V(t), \tag{19-8}$$

and for frequency modulation

$$\varphi'(t) = k_1\ V(t) \tag{19-9}$$

from which

$$\varphi(t) = k_1 \int^{t} V(t)dt, \tag{19-10}$$

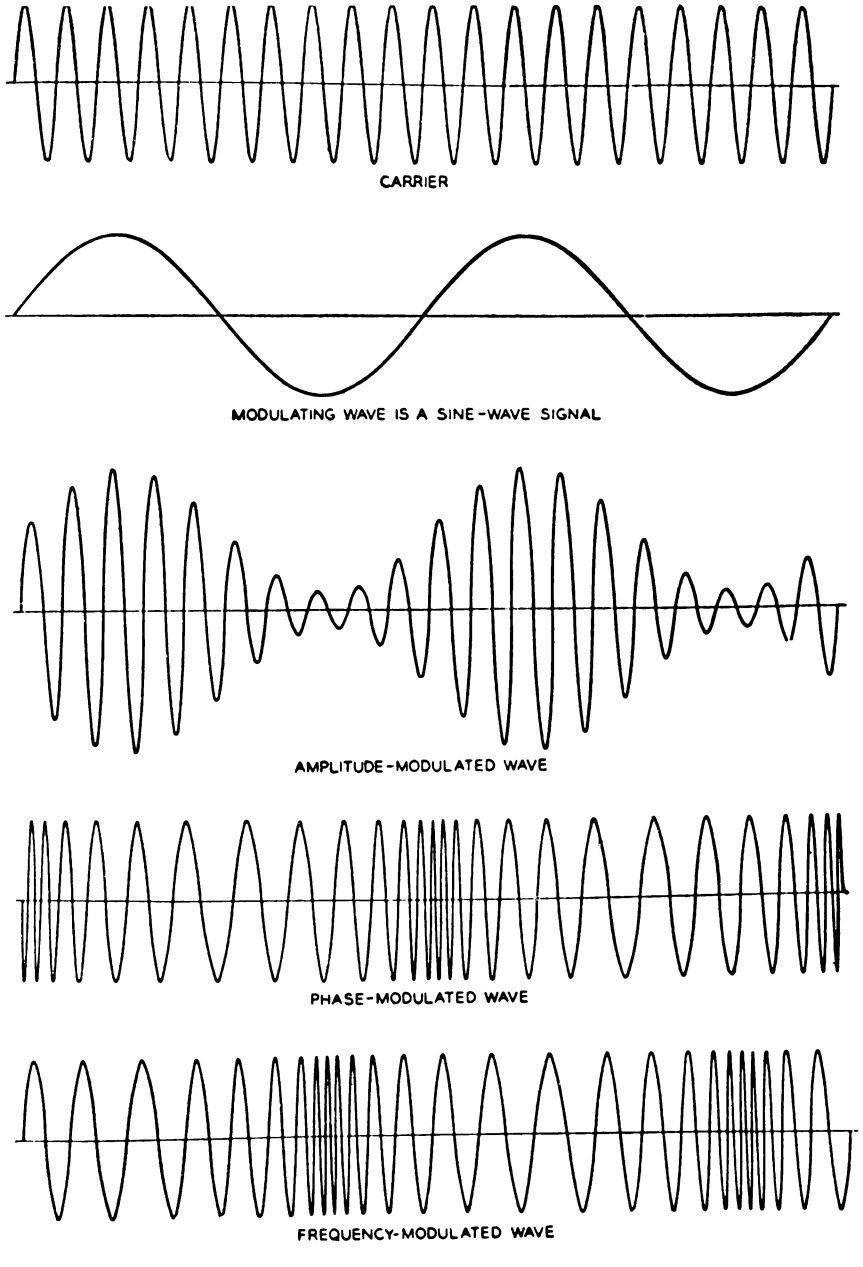
where k and k_1 are constants.

These results are summarized in Table 19-1. This table also illustrates phase modulated and frequency modulated waves which occur when the modulating wave is a single sinusoid.

Table 19-1

Type of Modulation	Modulating Signal	Angle Modulated Carrier
(a) Phase	$V(t)$	$M(t) = A_c \cos[\omega_c t + k\ V(t)]$
(b) Frequency	$V(t)$	$M(t) = A_c \cos[\omega_c t + k_1 \int^{t} V(t)dt]$
(c) Phase	$A_v \cos \omega_v t$	$M(t) = A_c \cos[\omega_c t + kA_v \cos \omega_v t]$
(d) Frequency	$-A_v \sin \omega_v t$	$M(t) = A_c \cos[\omega_c t + \frac{k_1 A_v}{\omega_v} \cos \omega_v t]$
(e) Frequency	$A_v \cos \omega_v t$	$M(t) = A_c \cos[\omega_c t + \frac{k_1 A_v}{\omega_v} \sin \omega_v t]$

Figure 1 illustrates amplitude, phase, and frequency modulation of a carrier, by a signal which consists of a single sinusoid. From these illustrations it should be clear that it is impossible to tell whether a particular angle modulated wave is phase modulated or frequency modulated unless the modulating signal is also known. For example, one cannot look at Equation (19-2) and tell whether it represents an FM or a PM wave. It could be either. A knowledge of the modulating signal, however, will permit the correct identification. If $\varphi(t) = k\ V(t)$, it is phase modulation and if $\varphi'(t) = k_1\ V(t)$, it is frequency modulated.



Amplitude, Phase, and Frequency Modulation of a Sine-Wave Carrier by a Sine-Wave Signal

Figure 19-1

Comparison of (c) and (d) in Table 19-1 shows that both FM and PM waves which are sinusoidally modulated have the form

$$M(t) = A_c \cos [\omega_c t + X \cos \omega_v t] \qquad (19\text{-}11)$$

where

$$X = k A_v \qquad \text{for PM} \qquad (19\text{-}12)$$

$$= \frac{k_1 A_v}{\omega_v} \qquad \text{for FM} \qquad (19\text{-}13)$$

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, and for FM it is inversely proportional to the frequency of the modulating wave. One often hears the terms high index and low index of modulation. It is difficult to define a sharp division. The term low index would normally be used when the peak phase deviation is less than one radian. In a later section the effect of the index of modulation on the frequency spectrum of the modulated wave will be considered.

Phase and Frequency Modulators and Demodulators

A phase modulator, which we may refer to as a PM modulator, is a device which varies the phase of a carrier so that the instantaneous phase deviation is proportional to the modulating wave. On the other hand an FM modulator, often referred to as an FM deviator, produces an instantaneous phase deviation proportional to the integral of the modulating wave. This suggests the following possibility. If a modulating wave V(t) is differentiated before being applied to an FM modulator, the instantaneous phase deviation will be proportional to the integral of V'(t), or in other words proportional to V(t). Thus an FM modulator that is preceeded by a differentiator actually produces an instantaneous phase deviation proportional to the modulating wave and is therefore equivalent to a PM modulator.

Other equivalences are also possible. For example, a PM demodulator is equivalent to an FM demodulator, commonly called an FM discriminator, followed by an integrator. Several equivalences are listed below and illustrated in Figure 2.

PM Modulator = Differentiator + FM Modulator
PM Demodulator = FM Demodulator + Integrator
FM Modulator = Integrator + PM Modulator
FM Demodulator = PM Demodulator + Differentiator

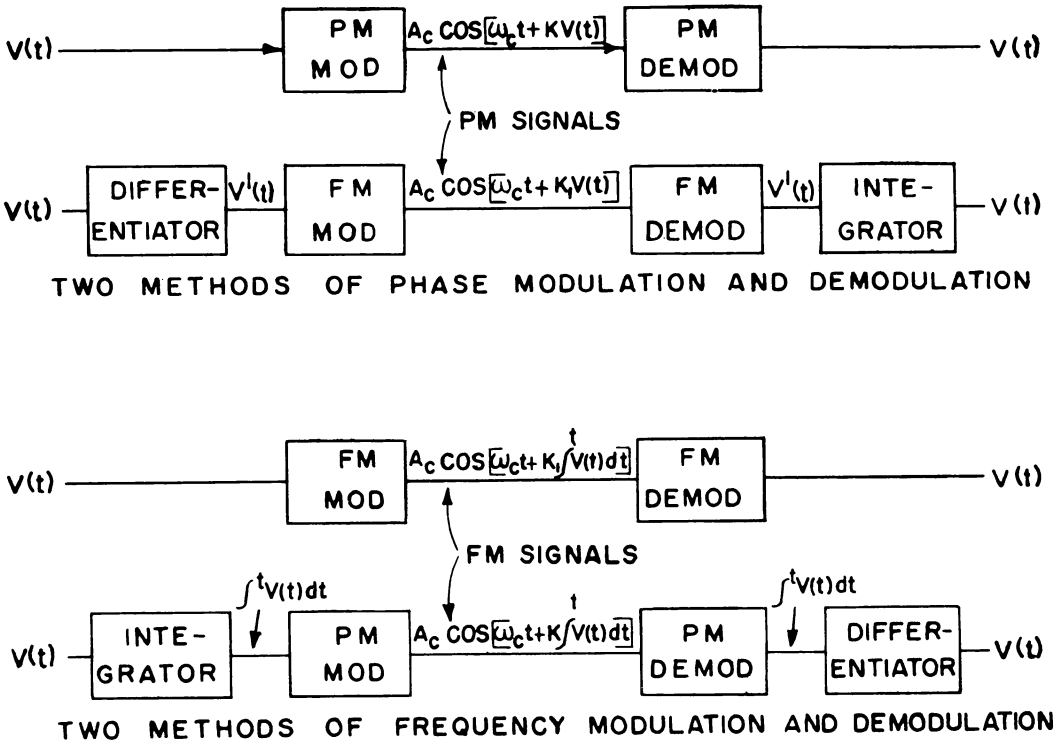


Figure 19-2

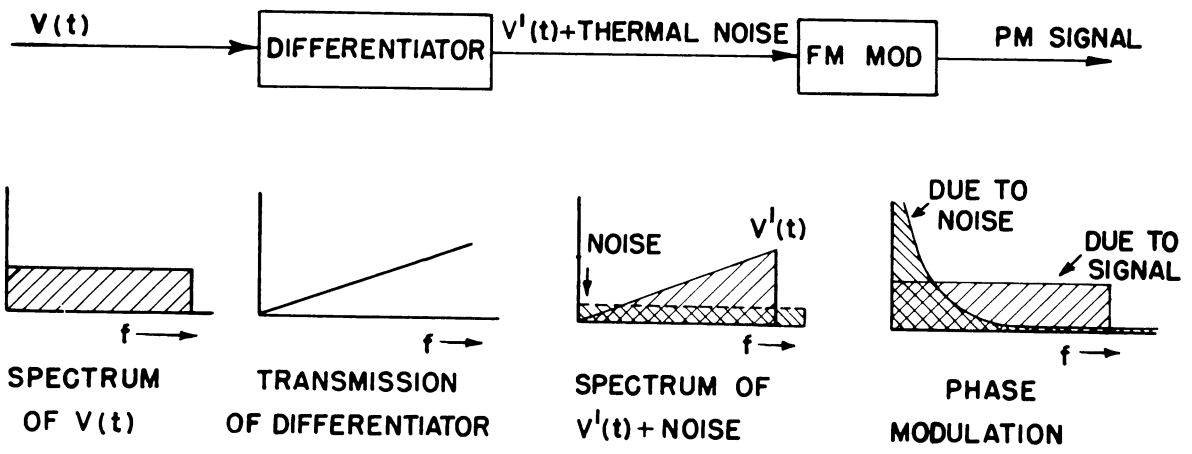
Pre-emphasis: Solution to a Thermal Noise Problem

If we examine the use of a differentiator and an FM modulator to obtain phase modulation a little more closely another problem becomes obvious. This is illustrated in Figure 3. Here we have assumed that V(t) has a flat frequency spectrum*. Thus we may assume that it consists of many sinusoidal components all of equal amplitude and of the form $A_v \cos \omega_v t$. At the output of the differentiator each of these sinusoids would have the form $-A_v \omega_v \sin \omega_v t$. Therefore, the amplitude of any particular component would be proportional to its frequency. The output spectrum corresponding to a flat input spectrum would be triangular as shown in Figure 3.

We can now consider the effect of thermal noise which exists in the circuit at the output of the differentiator. This will be flat

---

*The flat spectrum assumed here for illustrative purposes might, for example, be the base-band signal from a multi-channel telephone terminal.



Noise Problem When FM Modulator And Differentiator Are Used To Obtain Phase Modulation (Vertical Scales Arithemetic)

Figure 19-3

with frequency as shown. At low frequencies it is obvious that the noise level will exceed the signal level. It is therefore impractical to provide phase modulation by this method at the extremely low frequencies. We can, however, change the differentiator characteristic so that we obtain phase modulation at the higher frequencies and frequency modulation at lower frequencies. This is generally spoken of as "pre-emphasis".

Pre-emphasis is provided by passing the modulating wave through a network which shapes the frequency spectrum before the wave is applied to the modulator. Thus, differentiation is actually a form of pre-emphasis. In this case the pre-emphasis would shape a flat spectrum so that it would become triangular.

Other pre-emphasis shapes are possible and are often used to improve the performance of a system. Some reasons for pre-emphasis will be explained in later chapters. For now we shall merely look at a particular pre-emphasis shape which would avoid the noise problem just examined. This is illustrated in Figure 4 which should be self-explanatory when compared with Figure 3. Phase modulation is achieved at the higher baseband frequencies and frequency modulation at the lower frequencies.

It may interest some readers to know that broadcast FM uses a pre-emphasized signal similar to that illustrated above. De-emphasis is provided in the individual home receiviers. It should be noted that any system which is neither pure FM or pure PM is usually

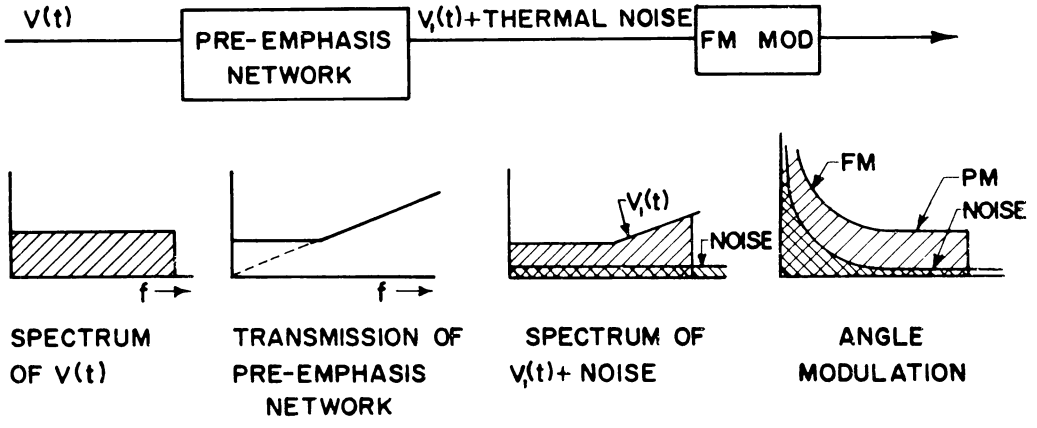


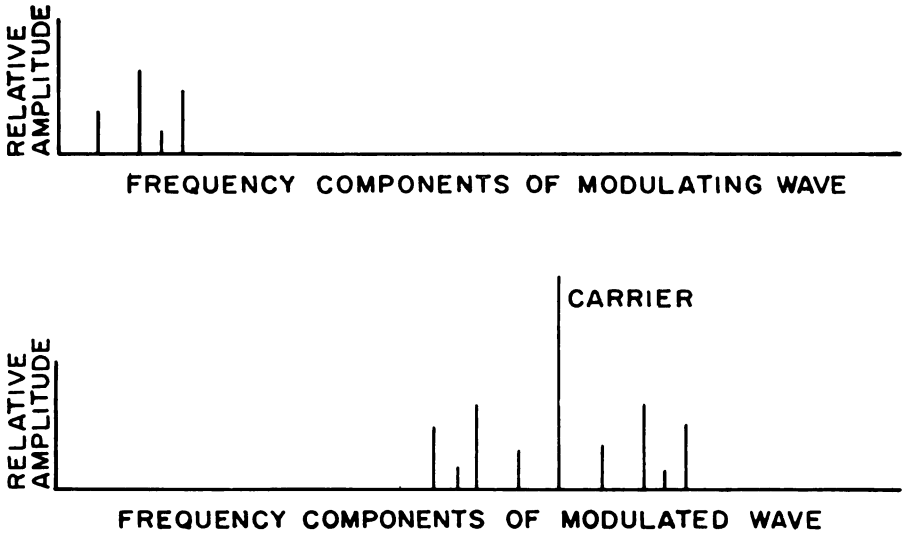
Illustration of Pre-emphasis With FM Modulator

Figure 19-4

referred to as an FM system. In fact it is rather common practice to use the term frequency modulation to denote any form of angle modulation.

Introduction to Frequency Analysis of FM and PM Waves

In the case of amplitude modulation it is easy to demonstrate that the frequency components of the modulated wave 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 wave 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 Figure 5. For every component at a frequency f_v in the modulating wave there are two components in the modulated wave; one at a frequency f_c+f_v and one at a frequency f_c-f_v, 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. This makes amplitude modulation easy to deal with. For example, the bandwidth required to transmit a double sideband AM wave is easily determined. If the highest frequency component in the modulating wave is f_h, the modulated wave is restricted to



Frequency Spectrum of an Amplitude Modulated Wave

Figure 19-5

frequency range which extends from f_c-f_h to f_c+f_h and the required bandwidth is 2 f_h centered at a frequency f_c.

In the case of frequency modulation the frequency components of the modulated wave are much more complexly related to the components in the modulating wave. In a strict mathematical sense a single modulating tone produces an infinity of sideband tones, although many are negligibly small. This in itself complicates the frequency spectrum of an FM wave. In addition the sideband components produced by any single-frequency component in the modulating wave depends on all the rest of the frequency components in the modulating wave. Hence, superposition does not apply.

One might ask if it is really advantageous to deal with the frequency components of an FM wave 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 effect only those frequency components of the signal which are at that frequency. Consider the simple problem of required bandwidth. It is obvious that the required

bandwidth depends on the location of all of the important frequency components in the wave. So in spite of the difficulty, some knowledge of the frequency components of an FM signal is essential.

Phase and Frequency Modulation by a One Tone Signal

The frequency analysis of the FM or PM wave will now be considered for the case where the modulating signal is a single sinusoid:

$$M(t) = A_c \cos [\omega_c t + X \cos \omega_v t] \tag{19-14}$$

As this equation now stands the separate frequency components are not obvious. Fortunately, Bessel Function identities are available which may be applied directly to the problem at hand. Several useful Bessel Function identities are given below.

$$\sin (C+X \sin V) = \sum_{n=-\infty}^{\infty} J_n(X) \sin (C+nV) \tag{19-15}$$

$$\cos (C+X \sin V) = \sum_{n=-\infty}^{\infty} J_n(X) \cos (C+nV) \tag{19-16}$$

$$\sin (C+X \cos V) = \sum_{n=-\infty}^{\infty} J_n(X) \sin (C+nV + \tfrac{n\pi}{2}) \tag{19-17}$$

$$\cos (C+X \cos V) = \sum_{n=-\infty}^{\infty} J_n(X) \cos (C+nV + \tfrac{n\pi}{2}) \tag{19-18}$$

Here $J_n(X)$ is the Bessel function of the first kind of nth order and of argument X. Values of $J_n(X)$ may be obtained in References 2 and 3. Note that the argument X is the index of modulation.

The identity given by Equation (19-18) can be applied to the signal of Equation (19-14) to give

$$M(t) = A_c \sum_{n=-\infty}^{\infty} J_n(X) \cos [\omega_c t + n\omega_v t + \tfrac{n\pi}{2}] \tag{19-19}$$

The first few terms may then be written as

$$\begin{aligned} M(t) = A_c \, [& J_o(X) \cos \omega_c t \\ & + J_1(X) \cos [(\omega_c+\omega_v)t + \tfrac{\pi}{2}] + J_{-1}(X) \cos [(\omega_c-\omega_v)t - \tfrac{\pi}{2}] \\ & + J_2(X) \cos [(\omega_c+2\omega_v)t + \tfrac{2\pi}{2}] + J_{-2}(X) \cos [(\omega_c-2\omega_v)t - \tfrac{2\pi}{2}] \\ & + \ldots\ldots] \end{aligned} \tag{19-20}$$

If we make use of the identity

$$J_{-n}(X) = (-1)^n J_n(X) \tag{19-21}$$

we get

$$\begin{aligned} M(t) = A_c \, [& J_o(X) \cos \omega_c t \\ & + J_1(X) \cos [(\omega_c+\omega_v)t + \tfrac{\pi}{2}] + J_1(X) \cos [(\omega_c-\omega_v)t + \tfrac{\pi}{2}] \\ & - J_2(X) \cos [(\omega_c+2\omega_v)t] - J_2(X) \cos [(\omega_c-2\omega_v)t] \\ & + \ldots\ldots] \end{aligned} \tag{19-22}$$

Equation 19-22 shows that the single sinusiodal modulating wave has produced 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 relative magnitudes of the various sidebands are determined by the coefficients $J_1(X)$, $J_2(X)$, etc. As Table 19-2 and Figure 6 show, the higher order sidebands rapidly become unimportant as the index of modulation, X, becomes less than unity. (For larger values of X, the value of $J_n(X)$ starts to decrease rapidly as soon as n = x.)

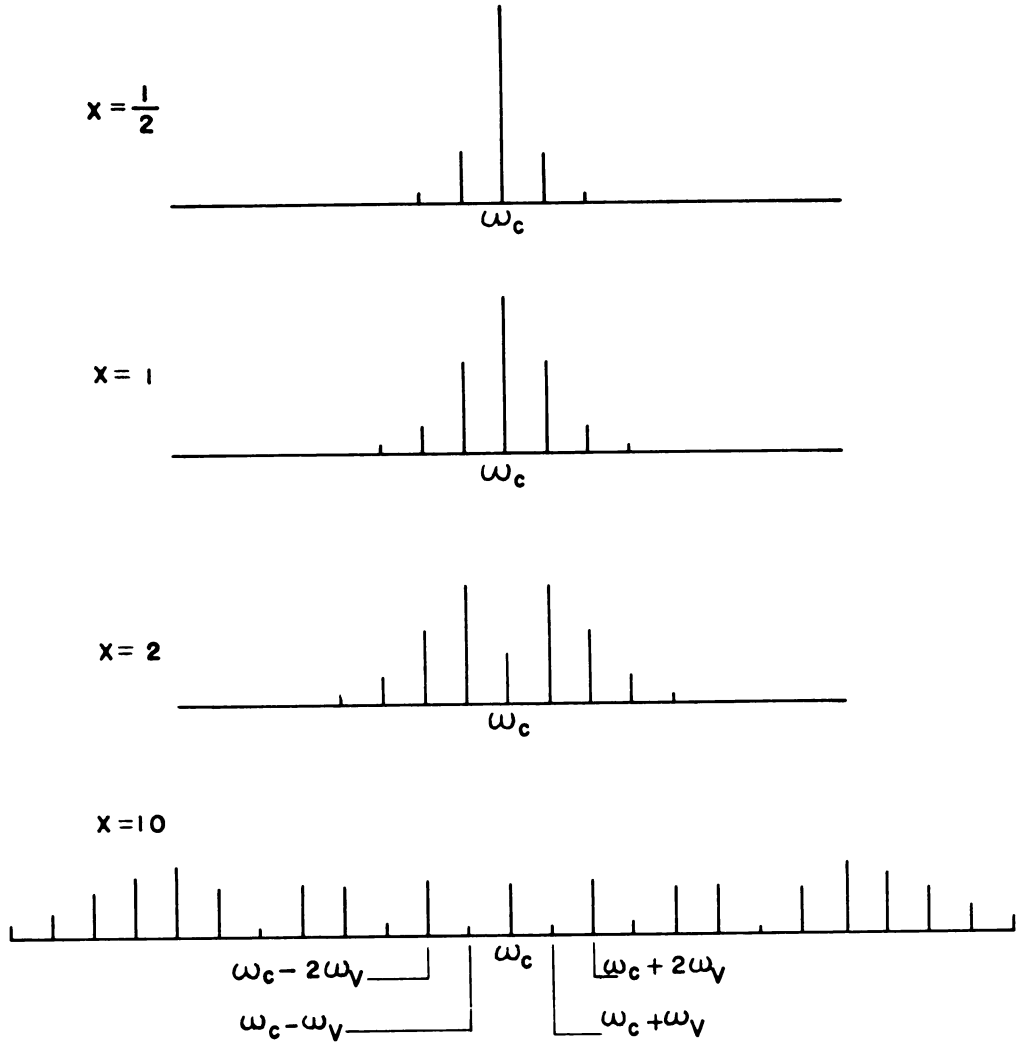
TABLE 19-2

	X = 1/2	X = 1	X = 2	X = 10
$J_O(X)$	0.938	0.765	0.224	-0.246
$J_1(X)$	0.242	0.440	0.577	0.043
$J_2(X)$	0.031	0.115	0.353	0.255
$J_3(X)$	0.003	0.020	0.129	0.058
$J_4(X)$	0.000	0.002	0.034	-0.220

Phase and Frequency Modulation by a Two Tone Signal

In FM and PM the calculation of the sideband components gets increasingly difficult as the number of modulating tones is increased. For example, consider the case of modulation by two sinusoidal waves.

$$M(t) = A_c \cos[\omega_c t + X_1 \cos \omega_v t + X_2 \cos \omega_w t] \tag{19-23}$$



Spectrum of A_c Cos ($\omega_c t$ + X cos $\omega_v t$)
for Various Values of X

Figure 19-6

By the trigonometric identity

$$\cos (A+B) = \cos A \cos B - \sin A \sin B \qquad (19\text{-}24)$$

this can be written as

$$M(t) = A_c \cos[\frac{\omega_c t}{2} + X_1 \cos \omega_v t] \cos[\frac{\omega_c t}{2} + X_2 \cos \omega_w t] \qquad (19\text{-}25)$$
$$- \sin[\frac{\omega_c t}{2} + X_1 \cos \omega_v t] \sin[\frac{\omega_c t}{2} + X_2 \cos \omega_w t]$$

The identities of Equations (19-15) to (19-18) can then be applied.

$$M(t) = A_c \Bigg\{ \sum_{n=-\infty}^{\infty} J_n(X_1) \cos\left[\frac{\omega_c t}{2} + n\omega_v t + \frac{n\pi}{2}\right] \sum_{m=-\infty}^{\infty} J_m(X_2) \cos\left[\frac{\omega_c t}{2} + m\omega_v t + \frac{m\pi}{2}\right] - \sum_{n=-\infty}^{\infty} J_n(X_1) \sin\left[\frac{\omega_c t}{2} + n\omega_v t + \frac{n\pi}{2}\right] \cdot \sum_{m=-\infty}^{\infty} J_m(X_2) \sin\left[\frac{\omega_c t}{2} + m\omega_w t + \frac{m\pi}{2}\right] \Bigg\} \tag{19-26}$$

$$M(t) = A_c \sum_{n=-\infty}^{\infty} \sum_{m=-\infty}^{\infty} J_n(X_1) J_m(X_2) \Bigg\{ \cos\left[\frac{\omega_c t}{2} + n\omega_v t + \frac{n\pi}{2}\right] \cos\left[\frac{\omega_c t}{2} + m\omega_v t + \frac{m\pi}{2}\right] - \sin\left[\frac{\omega_c t}{2} + n\omega_v t + \frac{n\pi}{2}\right] \sin\left[\frac{\omega_c t}{2} + m\omega_v t + \frac{m\pi}{2}\right] \Bigg\} \tag{19-27}$$

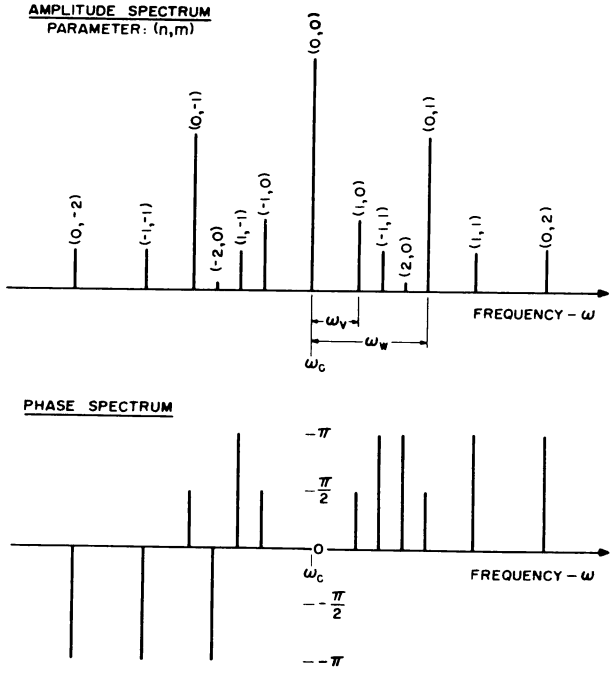
By means of Equation (19-24) further reduction can be obtained.

$$M(t) = A_c \sum_{n=-\infty}^{\infty} \sum_{m=-\infty}^{\infty} J_n(X_1)\, J_m(X_2) \cos\left[(\omega_c + n\omega_v + m\omega_w)t + \frac{(n+m)\pi}{2}\right] \tag{19-28}$$

This then is the desired result. It indicates that not only will there be sideband components displaced from the carrier by all possible multiples of the individual modulating frequencies, but also there will be components displaced by all possible sums and differences of multiples of the modulating frequencies. The sideband components in Equation (19-28) can be split into three types: a) the frequencies which would have been present (as in Figure 6) if only $X_1 \cos \omega_v t$ had been applied as a modulating signal; b) the frequencies which would have been present if only $X_2 \cos \omega_w t$ had been applied; c) all the possible sum and difference components of the form $(\omega_c \pm n\omega_v \pm m\,\omega_w)$.

Figure 7 shows the amplitude and relative phase spectra of the zero, first, and second order components* obtained from Equation (19-28) 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 Equation (19-28) is any component for which $|m| + |n| = 2$.



Amplitude and Relative Phase Spectra
Of Equation 19-28 for

$X_1 = 1/2$ $\qquad |m| + |n| \leq 2$

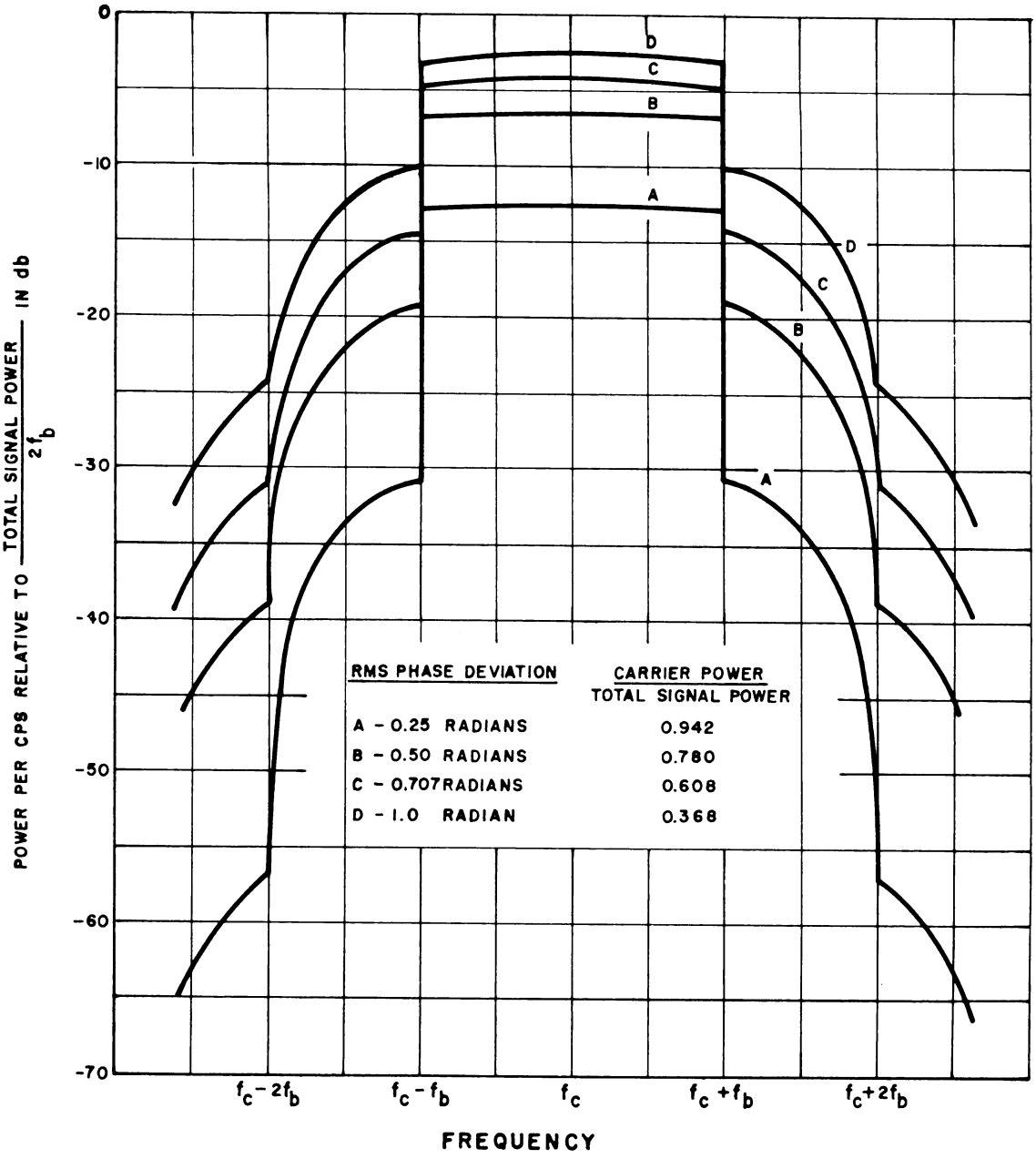
$X_2 = 1$

Figure 19-7

Phase Modulation by a Band of Random Noise

When the modulating wave consists of more than two frequencies the previously described procedures are of little help. In many such cases it may be practically impossible to obtain an accurate knowledge of the actual sideband spectrum. It is obvious, however, that it would contain a great many individual components. One might therefore attempt to find the envelope shape for all these individual components. Such an approach can be useful in particular cases. We shall illustrate the result obtained by one such approach; that of pure phase modulation by a flat band of band-limited random noise.

If the modulating signal is assumed to have a flat frequency spectrum for 0 cps to f_b cps, in which all components are assumed to have random phase, then the sideband spectrum can be statistically determined. The result obtained is a function of the rms phase deviation and is shown in Figure 8 for several values of the rms phase deviation.



Sideband Spectra of Waves Which Have Been Phase Modulated by a Baseband Signal Consisting of a Flat Band of Random Noise Which Extends From 0 to f_b CPS

Figure 19-8

A qualitative understanding of the shape of the spectra in Figure 8 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 $f_c \pm f_b$. Since the spectrum of the modulating signal has been assumed flat, the spectrum of the first order sideband components will also be flat vs. 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 Equation (19-28).

Second order sideband components, which fall within the band bounded by $f_c \pm 2f_b$, arise from combinations involving the carrier frequency and any second order combination of baseband frequencies such as "A+B", "A-B", or "2A". 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 $f_c \pm 2f_b$.

In a similar manner, third order sideband components, which fall in the region bounded by $f_c \pm 3f_b$, 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 $f_c \pm f_b$ portion of the spectrum and drops to zero at $f_c \pm 3f_b$.

The result of power addition of the higher order components to the first order sidebands accounts for the curvature in the spectrum between $f_c - f_b$ and $f_c + f_b$ in Figure 8. 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 to a non-linear device is increased. The same effect accounts for the relatively slow fall-off of higher order sidebands shown by Curve D, as against the rapid fall-off of Curve A.

Exponential Notation and Vector Representation

In certain instances the use of exponential notation for periodic functions is easier than the trigonometric notation which has been used thus far in this chapter. A particularly useful application is in the

vector representation of AM and PM waves as an aid in understanding the respective modulation processes. This will be considered here.

The sinusoidal carrier $\cos \omega_c t$ can also be written as

$$\left[\begin{array}{c}\text{Real}\\ \text{part}\\ \text{of}\end{array}\right] \varepsilon^{j\omega_c t}$$

since

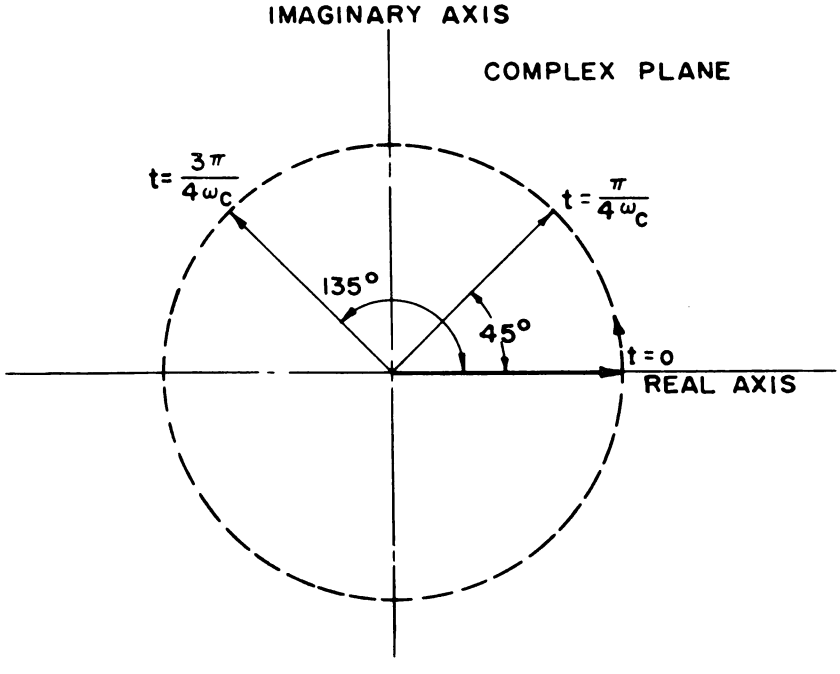
$$\varepsilon^{j\omega_c t} = \cos \omega_c t + j \sin \omega_c t. \tag{19-29}$$

The exponential $\varepsilon^{j\omega_c t}$ is now a rotating vector of unit length in the complex plane and its real part is merely its projection on the real axis. This vector is shown for several values of time in Figure 9.

An amplitude modulated wave with 100% modulation will now be examined.

$$M(t) = (1 + \cos \omega_v t) \cos \omega_c t \tag{19-30}$$

$$= \cos \omega_c t + \frac{1}{2} \cos(\omega_c+\omega_v)t + \frac{1}{2} \cos(\omega_c-\omega_v)t$$



Vector Diagram of $\varepsilon^{j\omega_c t}$ For Various Times

Figure 19-9

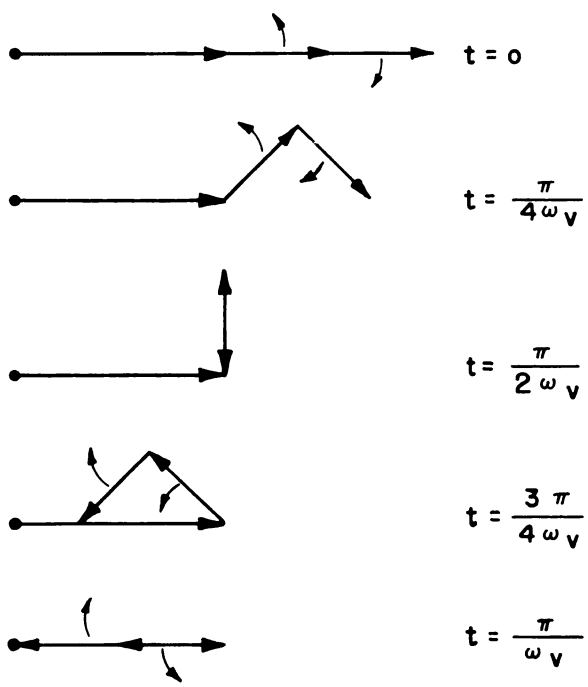
This may now be written in exponential notation.

$$M(t) = \left[\begin{matrix}\text{Real}\\ \text{part}\\ \text{of}\end{matrix}\right] \left[\varepsilon^{j\omega_c t} + \frac{1}{2}\,\varepsilon^{j(\omega_c+\omega_v)t} + \frac{1}{2}\,\varepsilon^{j(\omega_c-\omega_v)t}\right]$$

$$= \left[\begin{matrix}\text{Real}\\ \text{part}\\ \text{of}\end{matrix}\right] \varepsilon^{j\omega_c t}\left[1 + \frac{1}{2}\,\varepsilon^{j\omega_v t} + \frac{1}{2}\,\varepsilon^{-j\omega_v t}\right] \qquad (19\text{-}31)$$

In this form the carrier vector is multiplied by the sum of a stationary vector and two rotating vectors of equal size which rotate in opposite directions. As may be seen in Figure 10 the sum of these three vectors is always real and, consequently, acts only to modify the length of the rotating carrier vector. This produces amplitude modulation as expected.

For the purpose of comparison the frequency modulated wave of Equation (19-22) will be similarly represented. If the index of modulation is taken as 1/2, the second and higher order sidebands will be small enough that they may be neglected. The constant multiplier A_c will be disregarded.



VECTOR DIAGRAM OF $\left[1 + \frac{1}{2}\,\varepsilon^{j\omega_v t} + \frac{1}{2}\,\varepsilon^{-j\omega_v t}\right]$

Amplitude Modulation

Figure 19-10

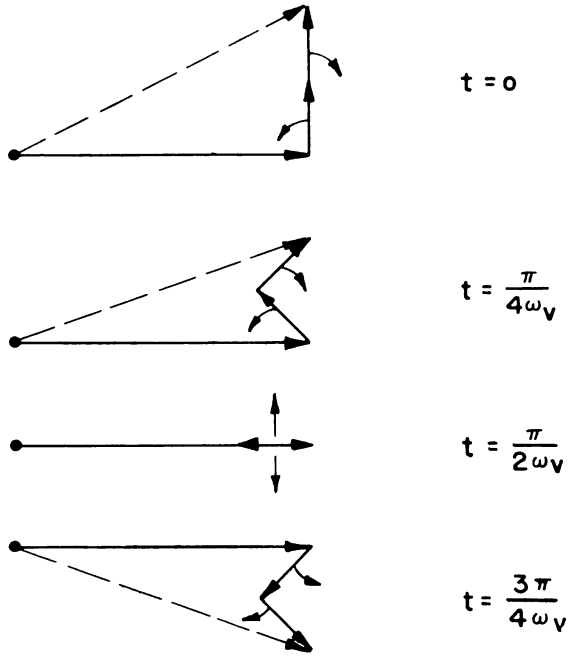
$$M(t) = \left[\begin{matrix}\text{Real}\\ \text{part}\\ \text{of}\end{matrix}\right]\left[J_o(1/2)\,\varepsilon^{j\omega_c t} + J_1(1/2)\,\varepsilon^{j[(\omega_c+\omega_v)t+\frac{\pi}{2}]} + J_1(1/2)\,\varepsilon^{j[(\omega_c-\omega_v)t+\frac{\pi}{2}]}\right] \tag{19-32}$$

$$M(t) = \left[\begin{matrix}\text{Real}\\ \text{Part}\\ \text{of}\end{matrix}\right]\varepsilon^{j\omega_c t}\left[J_o(1/2)+J_1(1/2)\varepsilon^{j(\omega_v t+\frac{\pi}{2})}+J_1(1/2)\varepsilon^{-j(\omega_v t-\frac{\pi}{2})}\right] \tag{19-33}$$

The multiplying vector, after the constants are evaluated, becomes

$$\left[.938 + .242\,\varepsilon^{j(\omega_v t+\frac{\pi}{2})} + .242\,\varepsilon^{-j(\omega_v t-\frac{\pi}{2})}\right]$$

This vector is plotted for several values of time in Figure 11 and it may be seen that it has an essentially constant amplitude but a variable phase. This then corresponds to phase modulation of the carrier. If all of the higher order sidebands had been retained, the multiplying vector would have had a constant amplitude of unity.



VECTOR DIAGRAM OF

$$\left[.938 + .242\,\varepsilon^{j(\omega_v t+\frac{\pi}{2})} + .242\,\varepsilon^{-j(\omega_v t-\frac{\pi}{2})}\right]$$

PHASE MODULATION INDEX OF MODULATION = $\frac{1}{2}$

Figure 19-11

Several interesting conclusions may be observed from this comparison. A low index PM wave and an AM wave are similar in the sense that they both contain the carrier and the same sideband frequency components. The important difference is in the phase of the sideband components. It may therefore be expected that in the transmission of an FM or 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. This will be considered in Chapter 22.

Average Power of an FM or PM Wave

The average power of an FM or 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 as follows by assuming a voltage of the form of Equation 19-2.

$$E(t) = A_c \cos [\omega_c t + \varphi(t)] \tag{19-34}$$

The instantaneous power in a resistance R becomes

$$P(t) = \frac{E^2(t)}{R} = \frac{A_c^2}{R} \cos^2 \left[\omega_c t + \varphi(t)\right]$$

$$= \frac{A_c^2}{R} \left[\frac{1}{2} + \frac{1}{2} \cos[2\omega_c t + 2\varphi(t)]\right] \tag{19-35}$$

The average power is given by the zero frequency terms in the expression above since non-zero frequency terms have an average value of zero. From the previous analysis of FM and PM waves one would expect the second term in Equation (19-35) to consist of a large number of sinusoidal sideband components about a carrier frequency of $2\omega_c$ rad/sec. If we neglect the remote possibility of one of these sidebands falling exactly at zero frequency the average power becomes

$$P_{(average)} = \frac{A_c^2}{2R} \tag{19-36}$$

This, of course, is the same as the average power which would have been present in the absence of modulation.

Bandwidth Required for FM Waves

Examination of the Bessel Function coefficients which occur in the expansion of a sinusoidally modulated FM or PM wave show that for the very low index case, i.e., a peak frequency deviation* much less than the modulating frequency, it is necessary to transmit only the first order sidebands. In the case of modulation by a complex signal of many frequencies it therefore 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 sidebands.

With a high index of modulation it is necessary to transmit several of the higher order sidebands. Again, an examination of the Bessel Function coefficients give an indication of the required bandwidth. From such an examination one may conclude that at least all of the sideband components which differ from the carrier by less than the peak frequency deviation are likely to be important. For the high index case, then, the minimum bandwidth should be at least twice the peak frequency deviation.

A general rule-of-thumb which is attributed to John R. Carson states that the minimum bandwidth required for the transmission of an FM or PM signal is equal to the sum of the peak-to-peak frequency deviation and twice the highest modulating frequency to be transmitted. This 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 actual bandwidth required is to some extent a function of the modulating signal and the quality of transmission desired.

Effect of a Nonlinear Input-Output Characteristic on an FM Wave

Some transmission devices such as electron tubes have nonlinear input-output characteristics which are a source of distortion to an amplitude modulated signal. This was discussed in Chapter 5. For this purpose electron tube nonlinearity was expressed by a power series

$$i_p = a_o + a_1 e_g + a_2 e_g^2 + a_3 e_g^3. \qquad (19\text{-}37)$$

The effect of this same characteristic on an FM signal will now be considered. The FM signal will be taken as

$$e_g = A_c \cos [\omega_c t + \varphi(t)]. \qquad (19\text{-}38)$$

---

*If the peak frequency deviation is less than the frequency of the modulating tone the peak phase deviation is less than one radian.

Substitution in Equation (19-37) then gives

$$i_p = a_o + a_1 A_c \cos [\omega_c t + \varphi(t)]$$

$$+ a_2 A_c^2 \cos^2[\omega_c t+\varphi(t)] + a_3 A_c^3 \cos^3 [\omega_c t + \varphi(t)] \quad (19\text{-}39)$$

The terms may be expanded and collected.

$$i_p = (a_o+\tfrac{1}{2}\ a_2 A_c^2) + (a_1 A_c + \tfrac{3}{4}\ a_3 A_c^3) \cos [\omega_c t + \varphi(t)]$$

$$+ \tfrac{1}{2}\ a_2 A_c^2 \cos [2\omega_c t+2\varphi(t)] + \tfrac{1}{4}\ a_3 A_c^3 \cos [3\omega_c t+3\varphi(t)] \quad (19\text{-}40)$$

The output wave consists of a d-c term and three FM waves centered respectively at the three frequencies ω_c, $2\omega_c$, and $3\omega_c$. If we assume for the moment that a filter can be used to extract the FM wave centered at ω_c we have as the output

$$\text{output} = (a_1\ A_c + \tfrac{3}{4}\ a_3 A_c^3) \cos [\omega_c t + \varphi(t)]. \quad (19\text{-}41)$$

The nonlinear characteristic has done nothing more than modify the gain. This is an important difference between AM and FM and is one of the reasons why FM is used in the microwave systems where nonlinear operation of electron tube amplifiers has thus far been unavoidable at the desired output levels.

We shall now examine the restriction which is necessary in order to achieve the desired output above. It is necessary to separate the FM wave centered at ω_c from the one centered at $2\omega_c$. We shall make use of Carson's rule-of-thumb. If we denote the peak frequency deviation by ΔF and the baseband width by W cps, we find that the FM sidebands of appreciable power about f_c, the carrier frequency in cps, extend upward to a frequency of $(f_c + \Delta F + W)$ cps. Similarly, the sidebands about the carrier at $2f_c$ extend downward to a frequency of $(2f_c - 2\Delta F - W)$. The $2\Delta F$ is required here because the index of modulation of the FM wave at $2f_c$ is twice as great as the index of modulation of the wave at f_c. Thus the frequency deviation will also be twice as great. If we allow the two sidebands to meet but not overlap we get the following restriction.

$$2f_c - 2\Delta F - W \geq f_c + \Delta F + W \quad (19\text{-}42)$$

from which

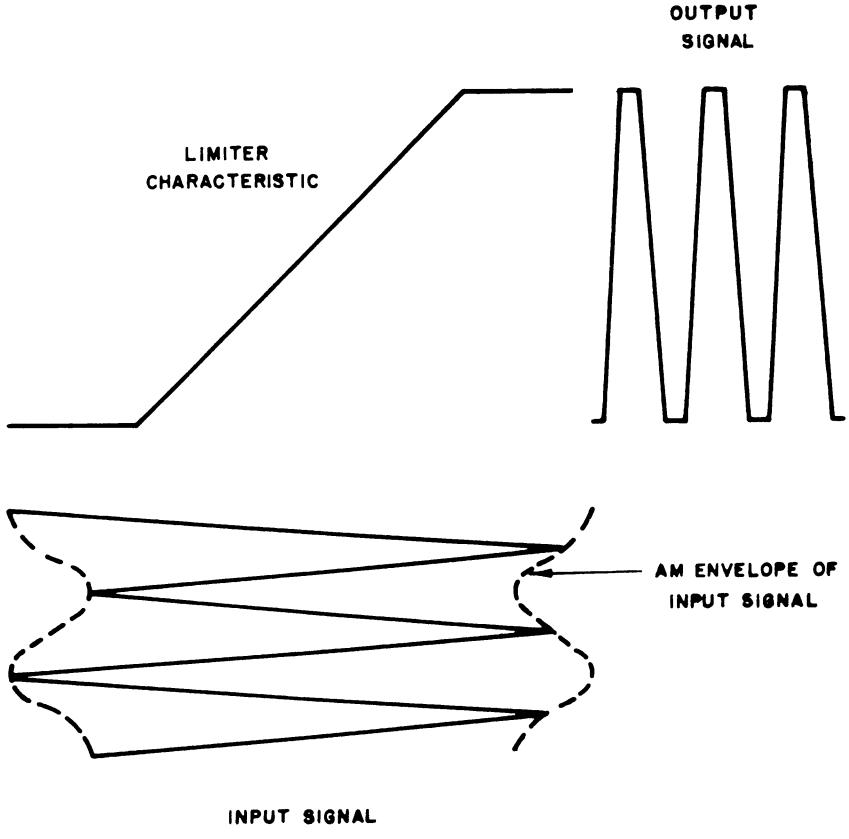
$$f \geq 3\Delta F + 2W \qquad (19\text{-}43)$$

If this restriction is met it is possible to recover the fundamental FM wave without distortion.

Limiters

Some devices which are used in FM systems produce distortion if the frequency modulated wave is also amplitude modulated. Traveling-wave tubes and some types of FM demodulators are examples. They make it necessary to either prevent amplitude modulation or provide methods for suppressing it. Amplitude modulation can be caused by imperfect FM modulators or by transmission deviations which may convert FM sidebands into AM sidebands. It is therefore not easily prevented.

Limiters are devices which ideally clip the peaks of an FM wave at a predetermined level. In this manner most of the undesired amplitude modulation of the wave may be removed. The characteristic of an ideal limiter is shown in Figure 12. The output signal has flat tops



Ideal Limiter Characteristic

Figure 19-12

which indicate that the output of the limiter consists of a fundamental and many harmonics. The harmonics may be removed with a filter subject to the same restriction which was discussed in the previous section. The fundamental output which remains will have the same frequency modulation as the input wave, but the amplitude modulation of the wave may be greatly reduced. It is, therefore, possible to use limiters in an FM system to suppress amplitude modulation before the signal is applied to AM sensitive devices. This can cause a large reduction in the distortion which would otherwise be produced in these devices.

A Few Words of Caution

In the last two sections we have seen that inadvertent nonlinearity does not produce distortion of the FM wave and that particular types of nonlinearity can actually reduce the distortion. There are several things which should be realized in connection with this, however. In the first place, an ideal nonlinear circuit has been assumed in both of the cases discussed in the last two sections. In practice the nonlinear components are often surrounded by inductors and capacitors. Steady state measurements in such circuits have shown that the output phase is often a function of input level. This phenomenon is often referred to as level-to-phase or gain-to-phase conversion. It is, therefore, possible that phase distortion could be produced by passing an amplitude modulated FM wave through such a circuit. Since this problem has not been resolved it is important to use caution when dealing with such nonlinear devices in FM systems.

In addition, it should be emphasized that an FM system with limiters is inherently highly nonlinear and many familiar concepts based on the principle of superposition have to be abandoned. As an illustration, consider the effect of a phase deviation in the transmission characteristic of a network used in an FM system. It will be helpful to refer to Figure 11, which shows the vector diagram of a low-index signal for the case of a carrier modulated by a single sinusoid. It should be evident that a phase distortion in the transmission path can shift one sideband component (vector) relative to the other. When this occurs, each sideband vector can be separated into two components in such a way as to form one pair of vector components which corresponds to the phase modulation of the carrier and a second pair which represents an unwanted amplitude modulation of the carrier arising from the phase distortion. The AM can be removed by a limiter, so that only the FM output will appear at baseband after demodulation. However, since the FM modulation is now represented by components of the original sideband vectors instead of the whole vectors, the baseband output will be an attenuated replica

of the original signal. It follows, then, that the effect of a phase distortion has been to produce amplitude distortion at baseband, and that the only equalizer which can be used after the limiter to remove the effects of the phase distortion is a gain equalizer. 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 any measurements for equalization purposes which are made through limiters by ordinary sweep-frequency techniques are useless. One solution to this problem might be the use of a low-index FM test signal for equalization measurements in systems using limiters, but as yet such a test set has not been developed.

References

1 - Modulation Theory - H. S. Black-D. Van Nostrand Company, Inc.
2 - Frequency Analysis, Modulation and Noise - Stanford Goldman - Mc Graw-Hill Book Company, Inc.
3 - Tables of Functions - Eugene Jahnke and Fritz Emde - Dover Publications.

Chapter 20

RANDOM NOISE IN FM AND PM SYSTEMS

The unwanted amplitude and phase modulation produced by an interfering sinusoid is analyzed for a low index FM or PM system. The principles of the analysis are then extended to determine the modulation caused by a flat band of random noise. Numerical examples illustrate the means of applying the results to the computation of noise in a low index FM or PM system. The problem of analyzing the effect of the interference in a high index system is briefly discussed. Other topics include the comparison of noise in FM, PM, and AM 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. In previous chapters the emphasis was on random noise in voice frequency and amplitude modulation systems. In this chapter the effect of random noise in phase and frequency modulated systems will be considered.

If an unmodulated carrier wave is combined with a band of random noise, the resultant wave is equivalent to a carrier wave which has been both amplitude and phase modulated by random noise. If, then, the resultant wave is demodulated by either an ideal amplitude detector or an ideal phase detector, a random noise output is to be expected. Since phase modulation and frequency modulation are so closely related it is obvious that an FM demodulator would also have a random noise output. However, the output is not the same in an FM system as it is in a PM system. As we shall demonstrate in this chapter, the noise voltage at the output of a 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.

Emphasis throughout will be placed on the way in which noise produces unwanted phase and frequency modulation. We shall first consider the unwanted amplitude and phase modulation of a carrier which is produced by an interfering sinusoidal signal, such as a spurious tone in the transmission band. For a sinusoidal interference, the frequency modulation can easily be deduced from the phase modulation. Having developed the necessary equations for this simple case of a single-tone interference (which is, of itself, an important problem in radio relay

systems), we shall then take up the random noise case. This approach is justified by the fact that random noise can be considered to be the sum of a very large number of equally spaced and randomly phased interfering sinusoids of equal amplitudes.

The reader should note that in most of the following sections the carrier to which the interfering sinusoid or noise is added is assumed to be unmodulated. The results, however, are applicable for noise or interfering signals which are added to a low index FM wave since most of the power is then in the carrier component. The problem of high index systems is treated briefly in a later section.

Amplitude and Phase Modulation of a Sinusoidal Carrier by an Interfering Sinusoidal Signal

In this section we shall discuss the amplitude and phase modulation of a sinusoidal carrier which is produced by an interfering sinusoidal signal. As a start we shall write the combined signal as

$$M(t) = \underbrace{A_c \cos \omega_c t}_{\text{carrier}} + \underbrace{A_n \cos \left[(\omega_c+\omega_n)t + \theta_n\right]}_{\text{interfering sinusoid}} \qquad (20\text{-}1)$$

where,

A_c = carrier amplitude in volts

ω_c = carrier frequency in radians/sec

A_n = amplitude of interfering sinusoid in volts

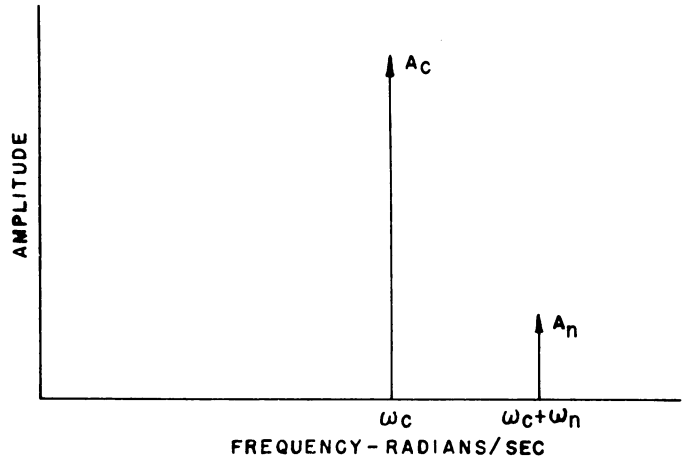
$(\omega_c+\omega_n)$ = frequency of interfering sinusoid in radians/sec

θ_n = phase angle of interfering sinusoid

The frequency spectrum of this combined signal is shown in Figure 1. It is not immediately obvious from either Equation (20-1) or Figure 1 that the combined signal is equivalent to a carrier with frequency ω_c which has been simultaneously amplitude and phase modulated at a radian frequency ω_n. This can be made apparent, however, after some trigonometric manipulation, as outlined in the appendix.

If $A_n \ll A_c$ we obtain the approximate results,

$$M(t) \doteq A_c \left[1 + \frac{A_n}{A_c} \cos(\omega_n t+\theta_n)\right] \cos\left[\omega_c t + \frac{A_n}{A_c} \sin(\omega_n t+\theta_n)\right], \qquad (20\text{-}2)$$



Spectrum: Carrier Plus Interfering Sinusoid

Figure 20-1

which shows that the carrier has been simultaneously amplitude and phase modulated at the difference frequency between the sinusoidal component and the carrier. The peak phase deviation in radians is given by the ratio of the amplitude of the unwanted sinusoid to the carrier amplitude.

$$\underline{\text{Peak phase deviation}} = \frac{A_n}{A_c} \text{ radians} \qquad (20\text{-}3)$$

Since the phase modulation is sinusoidal the rms phase deviation is equal to the peak phase deviation divided by $\sqrt{2}$. Hence,

$$\underline{\text{rms phase deviation}} = \frac{A_n}{\sqrt{2}\, A_c} \text{ radians} \qquad (20\text{-}4)$$

We could have written the rms phase deviation directly (without first considering the peak phase deviation) as a_n/A_c radians, where a_n is the rms voltage of the interference. This will be done later when we consider random noise, since for random noise the rms voltage is readily defined, whereas the peak voltage is not.

Another point which will be of importance in the random noise case is that ω_n can be either positive or negative depending on whether the interference frequency is above or below the carrier frequency. Thus a noise component at a frequency of either $\omega_c+\omega_n$ or $\omega_c-\omega_n$ will produce a baseband output at the same frequency ω_n. When two such noise components are simultaneously present (as they usually are in the cases we shall consider) they will add on a power basis, since they arise from uncorrelated voltages.

A Vector Representation of This Same Problem

Equation (20-1) can be written in exponential notation as follows:

$$M(t) = \left[\begin{matrix}\text{Real}\\ \text{Part}\\ \text{of}\end{matrix}\right] \left[A_c \varepsilon^{j\omega_c t} + A_n \varepsilon^{j[(\omega_c+\omega_n)t + \theta_n]} \right] \qquad (20\text{-}5)$$

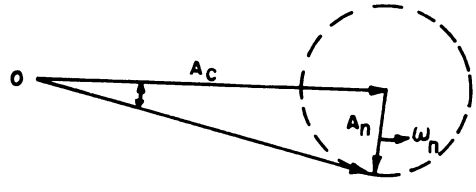
which can be written as

$$M(t) = \begin{matrix}\text{Real}\\ \text{Part}\\ \text{of}\end{matrix} \left(A_c \varepsilon^{j\omega_c t} \left[1 + \frac{A_n}{A_c} \varepsilon^{j(\omega_n t + \theta_n)} \right] \right) \qquad (20\text{-}6)$$

The multiplying vector,

$$1 + \frac{A_n}{A_c} \varepsilon^{j(\omega_n t + \theta_n)},$$

is shown in Figure 2. It is obvious that this vector varies in both amplitude and phase as a function of $\omega_n t$. When $A_n \ll A_c$ the peak phase deviation is approximately equal to A_n/A_c radians which is the same result obtained in the previous section.



VECTORS ARE SHOWN IN POSITION OF PEAK PHASE DEVIATION.
PEAK PHASE DEVIATION IS APPROXIMATELY EQUAL TO $\frac{A_n}{A_c}$ RADIANS.

Carrier Vector Plus Interfering Sinusoid

Figure 20-2

Frequency Modulation of a Sinusoidal Carrier by an Interfering Sinusoidal Signal

We shall now determine the unwanted frequency modulation produced by the sinusoidal component which was considered in the last two sections. This can easily be obtained from a knowledge of the phase modulation since the instantaneous frequency deviation is defined as the derivative of the instantaneous phase deviation. If we assume that the carrier is much larger than the sinusoidal component we can start with the instantaneous phase deviation given in Equation (20-2).

$$\text{Instantaneous phase deviation} = \frac{A_n}{A_c} \sin(\omega_n t + \theta_n) \text{ rad} \qquad (20\text{-}7)$$

Taking the derivative gives

$$\text{Instantaneous frequency deviation} = \frac{A_n \omega_n}{A_c} \cos(\omega_n t + \theta_n) \frac{\text{rad}}{\text{sec}} \qquad (20\text{-}8)$$

The peak frequency deviation in radians per second is given by

$$\underline{\text{Peak frequency deviation}} = \frac{A_n \omega_n}{A_c} \text{ radians/sec} = \frac{A_n}{A_c} f_n \text{ cps} \qquad (20\text{-}9)$$

The rms frequency deviation for the case of an interfering sinusoid is given by the peak frequency deviation divided by $\sqrt{2}$. Thus,

$$\underline{\text{rms frequency deviation}} = \frac{A_n \omega_n}{\sqrt{2} A_c} \text{ radians/sec} = \frac{A_n}{\sqrt{2} A_c} f_n \text{ cps} \qquad (20\text{-}10)$$

It should be observed that the peak frequency deviation is a function of the difference frequency ω_n. Consequently, sinusoidal components which are well displaced from the carrier frequency produce larger frequency deviations than sinusoidal components close to the carrier frequency. In consequence, we shall see that in the random noise case the rms frequency deviation is not as easily defined as the rms phase deviation was. For the rms phase deviation, it was sufficient to know the ratio of a_n, the rms interference voltage, to the peak carrier voltage. For the rms frequency deviation, it is necessary to take into account the spectrum of the interference.

Illustrative Example 1

In order to illustrate the concepts which we have described in the preceding sections we shall consider the following example of a single interfering sinusoid.

An FM signal with a 70 mc carrier is frequency modulated with a 7.75 mc baseband sine wave. The resulting peak frequency deviation is 4 mc. A 62 mc sinusoidal interference is added to the FM wave. If the power of the interfering tone is 40 db below that of the FM wave, what is the signal-to-interference ratio at the output of the system?

We may note that the index of modulation is small and that, consequently, most of the power in the FM wave is in the carrier. We can assume with very little error that it is all there when we are determining the frequency deviation produced by the interference. Since

the interference power is 40 db below the carrier power we have, for the corresponding peak voltages,

$$\frac{A_n}{A_c} = .01 \qquad (20\text{-}11)$$

The peak phase deviation produced by the interference is therefore .01 radians. The peak frequency deviation due to the interference is equal to the product of the peak phase deviation produced by the interference and the difference frequency f_n between the interference and the carrier. (Equation 20-9) Since in this case f_n is 8 mc, the peak frequency deviation is .08 mc. This may be compared with the peak frequency deviation due to the transmitted signal to determine the signal-to-interference ratio, S/N:

$$\frac{S}{N} = 20 \log \frac{4}{.08} = 34 \text{ db}$$

RMS Phase and Frequency Deviations: Carrier Plus Random Noise

Let us now consider what happens when a band of random noise is added to a sinusoidal carrier. We shall consider the effects of this interference in PM and FM systems. We shall discuss a) the total noise in the baseband (this is of interest if we are using a microwave signal to transmit TV, for example) and b) the noise in particular baseband slots (for example, the noisiest channel when we transmit a telephone multiplex signal).

First we note that for this purpose the random noise can be assumed to consist of an extremely large number of equally spaced sinusoidal components each with equal amplitude and arbitrary phase. It is convenient to deal with noise problems on a "per cycle" basis; thus, for example, we can replace a 3000 cycle band of noise by 3000 uniformly spaced sinusoids each having the same power as a one cycle band of noise.

If the single sinusoidal component of Equation (20-1) is replaced by a summation of N sinusoidal components we get

$$M(t) = A_c \cos \omega_c t + \sum_{n=1}^{N} A_n \cos \left[(\omega_c + \omega_n)t + \theta_n \right] \qquad (20\text{-}12)$$

A derivation which is practically identical with the one given in the appendix for the case of a single interfering sinusoid leads to the following result:

$$M(t) = A_s(t) \cos \left[\omega_c t + \varphi_s(t) \right] \qquad (20\text{-}13)$$

where, when the noise power is much less than the carrier power,

$$A_s(t) \doteq A_c + \sum_{n=1}^{N} A_n \cos [\omega_n t + \theta_n] \qquad (20\text{-}14)$$

and

$$\varphi_s(t) \doteq \frac{1}{A_c} \sum_{n=1}^{N} A_n \sin [\omega_n t + \theta_n] \text{radians} \qquad (20\text{-}15)$$

Comparing (20-15) with the instantaneous phase deviation given in (20-2) for the case of a single interference component, which was

$$\frac{A_n}{A_c} \sin (\omega_n t + \theta_n)$$

we see that the principle of superposition holds, subject of course, to the restriction imposed by the original assumption on which both equations are based, that the carrier power is much greater than the total interference power. By this we mean that the phase modulation produced by a band of random noise is equal to the summation of the phase modulation components which would have been produced by each noise component separately. In other words, the phase modulation produced by a particular noise component or group of components does not depend appreciably on the other noise components which are present, because each component (and their total) is so small compared to the carrier.

The summation shown in (20-15), which is

$$\sum_{n=1}^{N} A_n \sin [\omega_n t + \theta_n]$$

represents a function of time which has an rms value equal to the square root of the sum of the squares of the rms values of the individual sinusoids of which it is composed. Defining this total rms voltage as a_N, we can write, for the total rms phase deviation*

---

*At first glance there might seem to be some restrictions attached to this statement since in the noise voltage all components are assumed to have distinct frequencies, whereas, in the phase modulation there is the possibility of having two components at any given frequency. This is because two noise components, respectively above and below the carrier by the same frequency difference, produce phase modulation at the same frequency. We know that sinusoidal voltage components with different frequencies add on a power, or root-sum-square (rss), basis. This type of addition is not necessarily justified where more than one component can fall at a particular frequency. However, it can be shown to be justified when the two components which can fall at any given frequency have phase angles which are random and unrelated.

$$\underline{\text{total rms phase deviation}} \doteq \frac{a_N}{A_c} \text{ radians}$$

The value of a_N is readily computed for a band of flat random noise (white noise). In a practical problem we will usually know the bandwidth of interest in cycles per second, and the repeater noise figure (NF) in db. Thermal noise power can be expressed as being -174 dbm per cycle; if we are interested in a band of f_1 cps on either side of the carrier, the total noise power in dbm is

$$[-174 + NF + 10 \log 2 f_1] \text{ dbm}$$

where the factor of 2 represents the fact that we must consider the noise components on both sides of the carrier. a_N is, then, the rms voltage which would produce this power in whatever load impedance we are discussing. Usually we can ignore the particular impedance involved, and compare the noise power with the carrier power without reducing either to voltage units.

If we let a_n represent the rms noise voltage per cycle (thermal noise increased by the noise figure), we can write

$$a_N = a_n \sqrt{2 f_1}$$

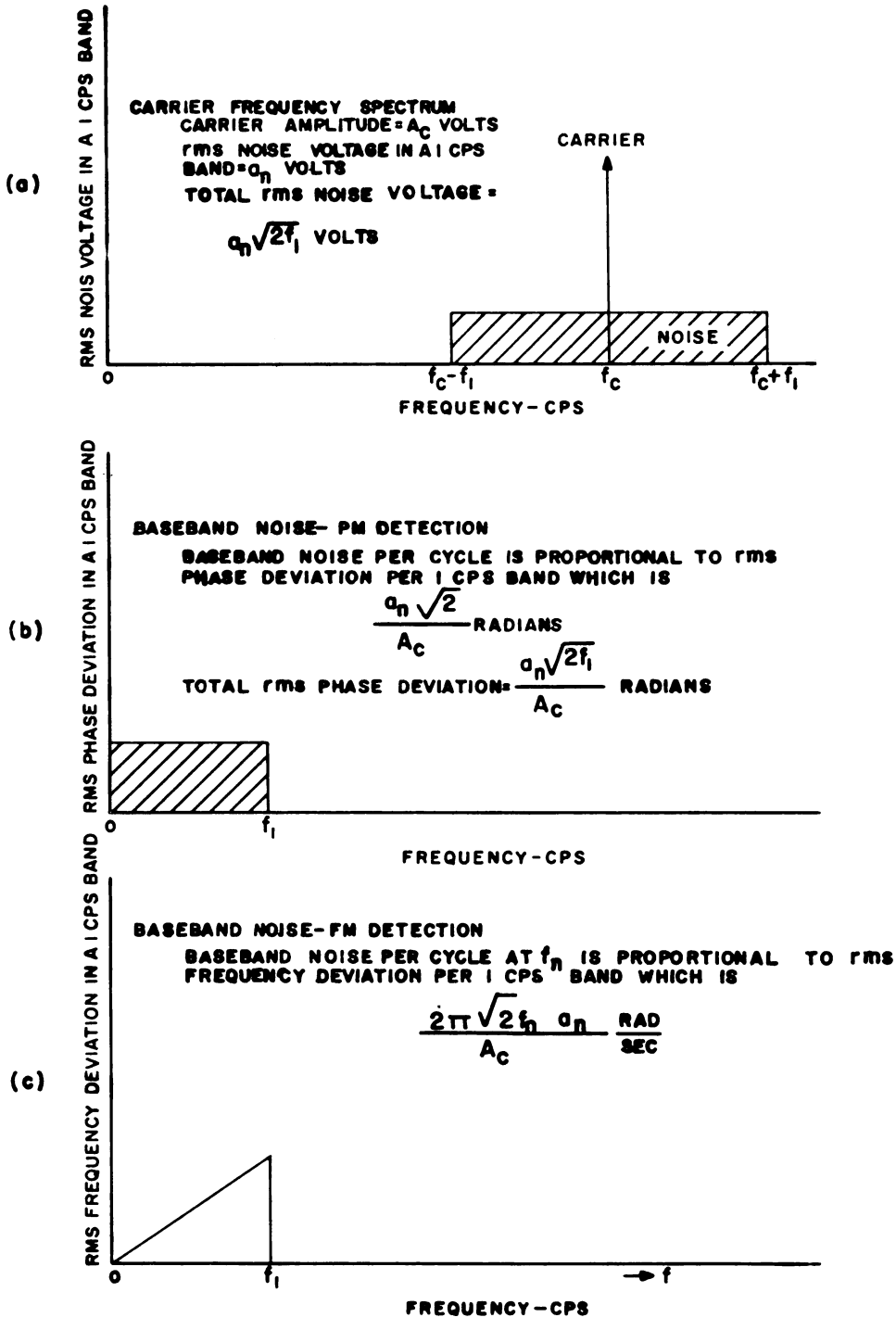
PM System Noise

If the random noise is flat vs. frequency over the transmitted band from $f_c - f_1$ to $f_c + f_1$, as shown in Figure 3(a), the noise per cycle in the baseband output of a phase modulation system will similarly be flat vs. frequency. This can readily be seen by decomposing the interference into component narrow bandwidths or sinusoids, a procedure which is valid as long as the principle of superposition holds i.e., as long as carrier power is much greater than noise power. Specifically, the rms phase deviation produced by two one cycle bands of noise, one ω_n radians per second above and the other ω_n radians per second below the carrier will be

$$\frac{a_n \sqrt{2}}{A_c} \text{ radians}$$

This formula is given as part of the second diagram of Figure 3.

If we were transmitting telephone multiplex over such a system, all channels would be equally noisy. In order to find the noise in dba at zero transmission level, we would, of course, need some relationship between rms phase deviation in the radio system and power at the zero level point. This relationship could be stated in a number of ways; we shall consider the similar problem in the FM case below in more detail.



Addition of a Flat Band of Noise to a Carrier: Resultant Baseband Noise in PM and FM Systems

Figure 20-3

FM System Noise

Figure 3(c) shows the baseband noise output of an FM system under the same assumptions as in Figure 3(a) - i.e., flat noise from f_c-f_1 to f_c+f_1, and carrier much greater than noise power. In this case the baseband noise is proportional to the rms frequency deviation rather than to the rms phase deviation. The instantaneous frequency deviation is obtained by differentiation of the instantaneous phase deviation. This is equivalent to multiplying the amplitude of each frequency component in the phase deviation by ω, as is evident from Equation (20-8). Hence, the rms frequency deviation per cycle of bandwidth varies directly with the frequency separation between the carrier and the one cycle band being considered. Specifically, the rms frequency deviation produced by two one cycle bands of noise, one ω_n radians per second above and the other ω_n radians per second below the carrier will be

$$\frac{\omega_n a_n \sqrt{2}}{A_c} \quad \text{radians/sec.}$$

where a_n and A_c have the same definitions as before. This formula is given as part of the third diagram of Figure 3, with $2\pi f_n$ written for ω_n.

The total noise in the entire baseband spectrum will be a direct function of the total rms frequency deviation. The total rms frequency deviation may be obtained by integrating the density of the mean square frequency deviation from 0 to f_1 cps and then taking the square root. This is analogous to summing power over a band. If we replace ω by $2\pi f$ the calculation becomes

$$\text{total rms frequency deviation} = \sqrt{\int_0^{f_1} \left(\frac{2\pi f \sqrt{2}\, a_n}{A_c}\right)^2 df}$$

$$= \frac{2\pi\sqrt{2}\, a_n}{A_c} \sqrt{\frac{f_1^3}{3}} \quad \frac{\text{radians}}{\text{sec}}$$

$$= \frac{\sqrt{2}\, a_n}{A_c} \sqrt{\frac{f_1^3}{3}} \quad \text{cps} \qquad (20\text{-}16)$$

Similarly, we can find the total rms frequency deviation for a portion of the spectrum by changing the limits of integration in the above expressions. A case of particular interest (corresponding to the top

channel in a multi-channel telephone multiplex signal) is that in which the bandwidth under consideration is small compared to its separation from the carrier. For example, we might be interested in a 3 kc band around a baseband frequency of 4 mc. For such a δf band at baseband frequency f_1, where $f_1 \gg \delta f$, the rms frequency deviation is

$$\text{rms freq. dev. for } \delta f \text{ band at } f_1 = \frac{2\pi\, a_n\sqrt{2\delta f}\, f_1}{A_c} \text{ radians/sec.}$$

$$= \frac{a_n\sqrt{2\delta f}\, f_1}{A_c} \text{ cps}$$

Another way of expressing this is to say that the rms phase deviation is given by the ratio of the total rms noise voltage ($a_n\sqrt{2\delta f}$) to the peak carrier voltage (A_c). The corresponding vector diagram is analogous to Figure 2. The rms frequency deviation is obtained by multiplying this ratio by ω_1 or f_1 (radians/sec or cps).

Noise at Zero Transmission Level

We have developed a formula for the rms frequency deviation for the noise in a narrow slot of the spectrum, such as a telephone channel. If we are discussing an FM system carrying a telephone load only, and know the peak frequency deviation ΔF for which the system is designed, we can find the noise at zero level.

First we establish a relationship between the baseband signal and the peak frequency deviation. We do this by recalling that load carrying capacity requirements on amplitude modulated telephone systems are phrased in the following terms: if the system is not to be overloaded more than 1% of the busy hour, then it must be capable of carrying a sine-wave signal at a carrier frequency zero db transmission level point of P_s dbm.* We can assume in an FM system, therefore, that the peak frequency deviation ΔF will correspond to the peaks of a sinusoidal baseband signal of P_s dbm, the maximum signal that we engineer telephone systems to transmit.

We can express these relations in rms terms. The rms value of the frequency deviation is $\Delta F/\sqrt{2}$ if the peak value is ΔF and the variation is sinusoidal. This is the rms deviation corresponding to a power at a 0 db TLP of P_s dbm. Let us define this rms frequency deviation of $\Delta F/\sqrt{2}$ as $\overline{F}$.

- -

*Such that the peaks of the sine wave reach a voltage that the telephone signals (during 1% of the busy hour) reach 0.1% or 0.01% of the time.

Clearly any rms frequency deviation equal to $\overline{F}$ will produce P_s dbm of power at the zero level. Consequently the baseband power in dbm at the zero level point produced by any other rms frequency deviation $\overline{f}$ could be expressed as follows:

$$\text{Baseband power} = \left(P_s + 20 \log \frac{\overline{f}}{\overline{F}}\right) \quad \text{dbm} \qquad (20\text{-}17)$$

In order to get the noise at zero level produced by random noise at the repeater input we need only to get the rms frequency deviation due to the noise and then use the expression above.

Illustrative Example 2

Find the noise (in dba at the -9 db transmission level) in the noisiest channel at the output of an FM system due to thermal noise at the repeater inputs. The system constants can be taken as follows:

Baseband signal - 1000 telephone channels where each channel is 3 kc wide but spaced every 4 kc to allow for separation filters. The baseband signal will be assumed to extend from zero to 4 mc.

Repeaters - The system consists of 100 repeaters in tandem where each repeater has an input power of -30 dbm and a noise figure of 12 db. The bandwidth at the repeater input will be taken as 20 mc.

Peak frequency deviation = 4 mc./s.

Solution*

From the material of Chapter 12, P_s for this system is 24.5 dbm. This produces a peak frequency deviation of 4 mc, which (since P_s is a sinusoid) is an rms frequency deviation of $4/\sqrt{2}$ or $2\sqrt{2}$ mc.

The rms frequency deviation caused by single repeater noise in a 3 kc band whose baseband frequency location is 4 mc (this is the top, noisiest, channel) can be found from the formula

---

*We should first justify using, in this case, the methods derived earlier. These assumed that the noise was small compared to the unmodulated carrier; is this assumption valid here? Let us find the total noise at the repeater input and compare this with the carrier power. The thermal noise in a 3 kc band is -139 dbm. This is -174 dbm per cycle or -101 dbm in a 20 mc band. If we increase this by the 12 db noise figure, the total noise is -89 dbm. With only small error in the final result we can assume that we have a low index system with a carrier power of -30 dbm. Thus, the carrier to noise ratio is 59 db and the approximate results of the preceeding sections may be used.

$$\text{rms freq. dev. for } \Delta f \text{ band at } f_1 = \bar{f} = \frac{a_n \sqrt{2\delta f}\, f_1}{A_c} \text{ cps}$$

Recalling the significance of these terms, we observe that at repeater input $a_n\sqrt{2\delta f}$ corresponds to the rms noise voltage for thermal noise power in a 6 kc band (-136 dbm) increased by the 12 db noise figure. Thus $a_n\sqrt{2\delta f}$ corresponds to -124 dbm. A_c corresponds to the peak voltage of a sine wave (the carrier) whose power is -30 dbm; the peak to rms relationship is 3 db, so the power corresponding to the peak voltage is -27 dbm.

The ratio $a_n\sqrt{2\delta f}\, f_1/A_c$ therefore corresponds to -97 db. Since we are comparing voltages, db ~ 20 log • ratio, and we find $-97 = 20 \log \sqrt{2}\, 10^{-5}$, whence

$$\bar{f} = \sqrt{2}\, 10^{-5} \cdot f_1$$

$$= \sqrt{2}\, 10^{-5}\, 4 \text{ mc/s}$$

From the relationship

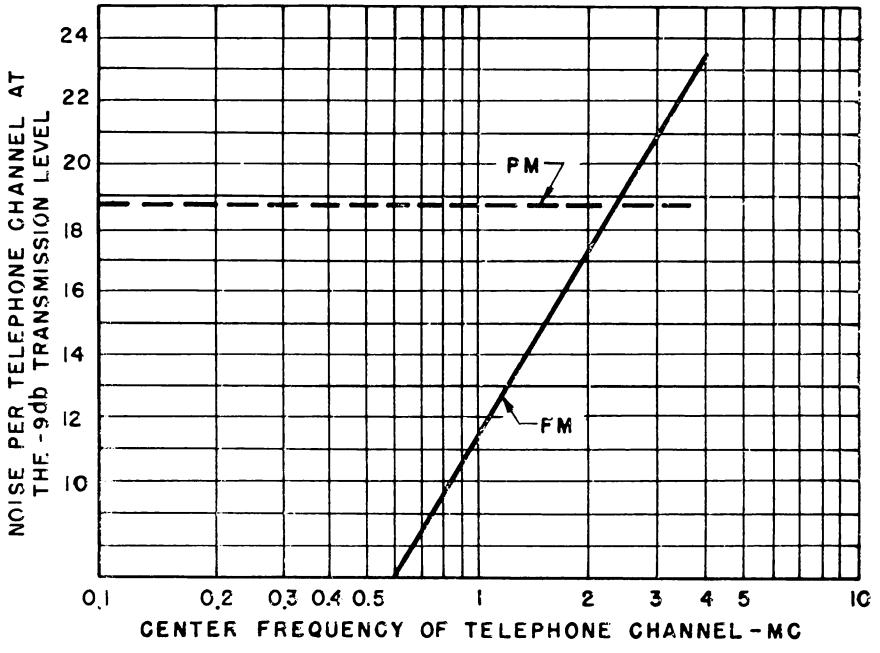
$$\text{Baseband power} = \left(P_s + 20 \log \frac{\bar{f}}{F}\right) \text{ dbm}$$

we find the baseband power at zero TLP to be -69.5 dbm. Converting to dba at the -9 point (assuming 0 dbm = 82 dba) and adding 20 db to allow for the fact that we have 100 repeaters (power addition) the final result is 23.5 dba at -9 db TLP.

The reader should note that this is the noise in the top channel and that the noise has a 6 db per octave slope. The noise in telephone channels located at lower baseband frequency positions will be less. For example, in this problem at a frequency equal to one-half the top frequency the noise in a telephone channel would be 17.5 dbm. The channel noise vs. frequency of channel is shown by solid line in Figure 4. The dash line will be explained in the next section.

Comparison of FM and PM System Noise

The illustrative problem of the previous section 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 requirements, the lower frequency channels have unnecessary margin. This is not very efficient.



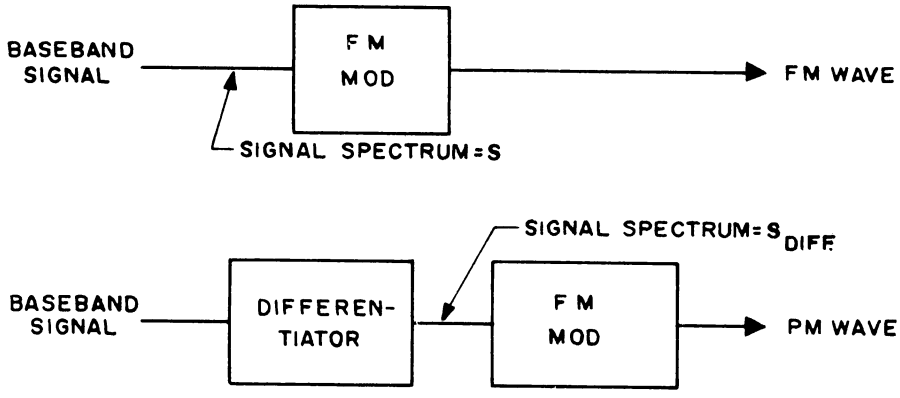
Noise Per Channel At the Output of the FM System Considered in Illustrative Example 2. The dash Line Shows The Noise Per Channel in a PM System With the Same rms Frequency Deviation.

Figure 20-4

In a phase modulated system the noise is the same in all the telephone channels, since the phase modulation due to the signal and the phase modulation due to the random noise are both flat with frequency. We shall now determine how the noise in a PM system compares with the noise in the top channel of an FM system, under the condition that the rms frequency deviation caused by the transmitted signal is the same for the two systems.

An easy way to approach this problem is to compare the output of an FM modulator with the output of a PM modulator where the PM modulator consists of a differentiator and an FM modulator. This is shown in Figure 5. We shall assume that the power spectrum of the baseband signal is flat from 0 cps to f_1 cps. Thus in the first case, where an FM wave is produced, the power spectrum S of the signal applied to the FM modulator can be written as

$$S = a^2 \text{ watts/cps } 0 < f < f_1 \text{ cps}$$
$$= 0 \text{ elsewhere} \qquad (20\text{-}18)$$



FM And PM Modulators

Figure 20-5

The total power would be

$$\text{Power} = a^2 f_1 \text{ watts} \tag{20-19}$$

In the second case where a PM wave is produced, the power spectrum S_{Diff} of the signal applied to the FM modulator would be parabolic. This is because the voltage spectrum at the output of the differentiator is proportional to the frequency and the power spectrum is proportional to the square of the voltage spectrum (i.e., $S_{Diff} = kf^2$, where k is a constant). If, for the moment, we let the power spectrum in the top channel at differentiator output be equal to that in the FM case, a^2, it follows that $k = a^2/f_1^2$ and the expression for S_{Diff} becomes

$$\begin{aligned} S_{Diff.} &= \frac{a^2}{f_1^2} f^2 \text{ watts/cps } 0 < f < f_1 \text{ cps} \\ &= 0 \text{ elsewhere} \end{aligned} \tag{20-20}$$

From this, the total power would be

$$\text{Power (Diff.)} = \int_0^{f_1} \frac{a^2}{f_1^2} f^2 \, df = \frac{1}{3} a^2 f_1 \text{ watts} \tag{20-21}$$

We are now in a position to make a comparison. We have deliberately set the signal levels in the top channels equal (at FM modulator input). The

frequency (or phase) deviation caused by the top channel signal will therefore be the same in the two systems, and the top channel S/N ratios in the radio links of the two systems will be identical. At final baseband output, the zero TLP noise in the top channels of the two systems must therefore be the same.

However, we may note that the total power applied to the FM modulator which is producing phase modulation is only one-third as great as that applied to the other modulator. Therefore, the rms frequency deviation will be less by a factor of $\sqrt{3}$. In the PM system this permits us to raise the signal level by 20 log $\sqrt{3}$ or 4.8 db and make the rms frequency deviation the same in each case. We therefore find that for the same rms frequency deviation there is a 4.8 db signal-to-noise advantage for pure PM over pure FM. In other words, for the same rms frequency deviation the random noise in each of the channels at the output of PM system is 4.8 db below the noise in the worst channel at the output of an FM system. This advantage is shown in Figure 4 for the illustrative problem of the previous section.

The 4.8 db advantage derived above presents the maximum advantage which can be obtained by pre-emphasis in front of FM terminals. In practice, as illustrated in Chapter 19, it is not possible to provide differentiation of the telephone signal clear down to zero frequency without getting into noise difficulties at the output of the differentiator. For this reason the advantage obtained by pre-emphasis is usually between 3 and 4 db.

FM Advantage

By using frequency modulation it is possible to get better signal-to-noise performance than would be obtained in an AM system with the same transmitted power. In order 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 the signal-to-noise performance which is obtained by using wider bandwidths is sometimes referred to as the FM advantage. We shall now examine this quantitatively.

A single sideband AM system with suppressed carrier will be compared with an FM system. The peak power will be assumed to be the same for both systems. We will examine the AM system first.

Let

P_S = power in dbm, at zero transmission level, of the largest sine wave that the system is designed to transmit.

N = noise in dbm in a 3 kc band at some low level point - say repeater input.

P_R = power in dbm of the largest sine wave that the system is designed to transmit, at the low level point in the system. (Hence, $P_S - P_R$ is the gain in the system between the low level point and zero transmission level.)

The noise in a 3 kc telephone channel at zero level can then be written as

$$\text{Noise}_{AM} = P_S - P_R + N \text{ dbm} \tag{20-22}$$

In an FM system we must define two additional quantities. These will be the peak frequency deviation, and the center frequency of the top telephone channel. The noise will be the highest in this channel.

ΔF = peak frequency deviation - cps

f_n = center frequency of top channel - cps

We observe that the application of P_s dbm at 0 TLP will produce a peak frequency deviation of ΔF.

The rms phase modulation produced by the noise in two 3 kc bands respectively above and below the carrier is given by the ratio of the rms noise voltage (due to both bands) to the peak carrier voltage. At the low level point the total noise power is, therefore, N + 3 dbm. Furthermore, the carrier power at this point is the same as the maximum signal power, P_R dbm, at the low level point in the AM single sideband suppressed carrier case. This follows from the assumption that peak power in the two systems is the same. The ratio of the rms noise voltage to the rms carrier voltage is then equal to $N + 3 - P_R$ db, and the ratio of the rms noise voltage to the peak carrier power is 3 db less. Therefore, we get

$$\begin{matrix}\text{rms phase modulation} \\ \text{in 3 kc band}\end{matrix} = N - P_R \begin{matrix}\text{db with respect} \\ \text{to one radian.}\end{matrix} \tag{20-23}$$

and, by the familiar relationship between phase and frequency deviation,

$$\begin{aligned}\text{rms frequency modulation in a 3 kc band} &= N - P_R + 20 \log 2\pi f_n \quad \text{db with respect to one radian/sec.} \\ &= N - P_R + 20 \log f_n \quad \text{db with respect to one cps}\end{aligned} \tag{20-24}$$

The ratio of the rms frequency deviation $\Delta F/\sqrt{2}$, produced by the maximum sine wave P_s which the system has to transmit, to the rms frequency deviation in a 3 kc band due to noise is

$$\begin{aligned}\text{Ratio} = (\text{db}) &= 20 \log \Delta F/\sqrt{2} + (P_R - N - 20 \log f_n) \\ &= P_R - N + 20 \log \Delta F/\sqrt{2} f_n \text{ db}\end{aligned} \tag{20-25}$$

This, then, is the signal-to-noise ratio if P_s is taken as the signal. Finally then the noise at zero level is P_s dbm minus this ratio in db or

$$\text{Noise}_{FM} = P_S - P_R + N - 20 \log \Delta F/\sqrt{2} f_n \text{ dbm} \tag{20-26}$$

Comparison of Equations (20-27) and (20-22) shows that the FM advantage is

$$\text{FM advantage} = 20 \log \Delta F/\sqrt{2}\, f_n \tag{20-27}$$

We may therefore note that unless the peak frequency deviation is equal to or greater than the $\sqrt{2}$ 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 noise in the top channel would be 3 db higher than in a single sideband AM system. This is an FM advantage of -3 db. From the considerations of the previous section, pure phase modulation has a signal-to-noise ratio which is 4.8 db better than pure FM. Pure phase modulation has a signal-to-noise advantage of 1.8 db over amplitude modulation when the peak frequency deviation is equal to the frequency of the top transmitted channel.

Random Noise and Interference in Large Index Systems

In the preceding sections we have assumed that practically all of the power in an FM wave is in the carrier and have neglected the effects of sidebands on the noise performance of the system. At this point we consider briefly the effect of noise and interference in

high index systems. The effect of sidebands can be included if Equation (20-12) is rewritten as

$$M(t) = A_c \cos [\omega_c t+\varphi(t)] + \sum_{n=1}^{N} A_n \cos [(\omega_c+\omega_n)t + \theta_n] \qquad (20\text{-}28)$$

Here the only change has been to add the phase modulation $\varphi(t)$ to the carrier term. The derivation may be carried out in the same manner as previously but with a slightly different result. The output may be written as

$$M(t) = A_s(t) \cos [\omega_c t - \varphi(t) + \varphi_s(t)] \qquad (20\text{-}29)$$

where

$$\varphi_s(t) = \frac{1}{A_c} \sum_{n=1}^{N} A_n \sin [\omega_n t - \varphi(t) + \theta_n] \qquad (20\text{-}30)$$

If we now examine the phase distortion term, $\varphi_s(t)$, we see that each of the noise components is actually an FM wave with the same index of modulation as the original signal. Therefore, if the sideband energy of the transmitted FM wave is small, the sideband about each noise component will also be small. The frequency modulation produced by the noise is obtained by taking the derivative of $\varphi_s(t)$.

$$\text{freq. dev.} = \frac{d}{dt} \varphi_s(t)$$

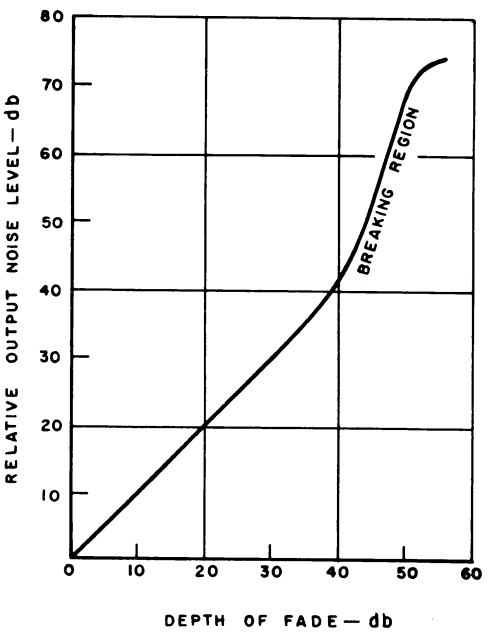
$$= \frac{1}{A_c} \sum_{n=1}^{N} A_n [\omega_n-\varphi'(t)] \cos [\omega_n t-\varphi(t) + \theta_n] \qquad (20\text{-}31)$$

For high index systems where the number of random noise components is small the expression above may lead to results which are quite different from those obtained for a low index system. However, when the interference consists of many components such as is the case with thermal noise, the spectrum of $\varphi_s(t)$ is usually relatively flat with frequency even though it consists of a large number of small FM waves instead of a large number of individual noise components. The frequency deviation obtained by taking the derivative therefore tends to have a triangular spectrum which, for all practical purposes, is the same as the one obtained for a low index signal.

Breaking Region

In the previous sections we have considered the phase and frequency modulation which is produced by random noise only when the total

noise power is much less than the carrier power. As long as 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 about ten times the noise power, this linearity no longer holds and the output signal-to-noise ratio decreases faster than the input signal-to-noise ratio is decreased. In this region, which is referred to as the breaking region, the system **rapidly** becomes **unusable.**



Typical Effect of Breaking Region on Output Noise During Deep Fades

Figure 20-6

In general, then, one would not design a system to operate in this region. However, a system which normally operates in a linear region may have to operate in the breaking region during deep fades in the radio signal. This is illustrated in Figure 6 for a particular system. This figure assumes that the output signal is kept constant during the fade by an automatic volume control. Consequently, the gain increases as the signal fades, and the output noise increases linearly with the depth of fade until the breaking region is reached. The noise then increases at a faster rate.

A more detailed discussion of the breaking region may be found in Reference 1.

References

1 - Frequency Analysis, Modulation and Noise - S. Goldman - McGraw-Hill Book Company, Inc. - 1948.

2 - Modulation Theory - H. S. Black - D. Van Nostrand Company, Inc. - 1953.

3 - Properties of a Sine Wave Plus Random Noise - S. O. Rice - BSTJ vol. 27, pp. 109-157, January, 1948 (or Monograph B1522)

Appendix

In Equation (20-1) the equation for the carrier and interfering sinusoid is written as

$$M(t) = A_c \cos \omega_c t + A_n \cos [(\omega_c+\omega_n)t + \theta_n] \qquad (20\text{-}1A)$$

Use can be made of the trigonometric identity

$$\cos (A+B) = \cos A \cos B - \sin A \sin B \qquad (20\text{-}2A)$$

to expand the interfering sinusoid component. The expression for the signal then becomes

$$M(t) = A_c \cos \omega_c t + A_n \cos \omega_c t \cos (\omega_n t + \theta_n) - A_n \sin \omega_c t \sin (\omega_n t + \theta_n) \qquad (20\text{-}3A)$$

Equation (20-3A) can be factored to give

$$M(t) = A_c \left([1 + \frac{A_n}{A_c} \cos (\omega_n t + \theta_n)] \cos \omega_c t - \frac{A_n}{A_c} \sin (\omega_n t + \theta_n) \sin \omega_c t \right) \qquad (20\text{-}4A)$$

A second trigonometric identity,

$$a \cos A - b \sin A = \sqrt{a^2+b^2} \cos (A + \tan^{-1} \frac{b}{a}) \qquad (20\text{-}5A)$$

can now be used to write

$$M(t) = A_n(t) \cos [\omega_c t + \varphi_n(t)] \qquad (20\text{-}6A)$$

where

$$A_n(t) = A_c \sqrt{[1 + \frac{A_n}{A_c} \cos(\omega_n t + \theta_n)]^2 + [\frac{A_n}{A_c} \sin(\omega_n t + \theta_n)]^2}$$

$$= A_c \sqrt{1 + [\frac{A_n}{A_c}]^2 + 2\frac{A_n}{A_c} \cos(\omega_n t + \theta_n)} \qquad (20\text{-}7A)$$

and

$$\varphi_n(t) = \tan^{-1} \frac{\frac{A_n}{A_c} \sin(\omega_n t + \theta_n)}{1 + \frac{A_n}{A_c} \cos(\omega_n t + \theta_n)} \qquad (20\text{-}8A)$$

As soon as the original equation is written in this form the amplitude and phase modulation become more obvious. A further simplification in the expression for M(t) results if $A_n \ll A_c$. For this condition, a binomial expression of $A_n(t)$ gives

$$A_n(t) \sim A_c [1 + \frac{A_n}{A_c} \cos(\omega_n t + \theta_n)] \qquad (20\text{-}9A)$$

Similarly,

$$\varphi_n(t) \sim \tan^{-1} \frac{A_n}{A_c} \sin(\omega_n t + \theta_n)$$

$$\sim \frac{A_n}{A_c} \sin(\omega_n t + \theta_n) \qquad (20\text{-}10A)$$

since for small angles the tangent of an angle is approximately equal to the angle expressed in radians.
Thus,

$$M(t) \sim A_c[1 + \frac{A_n}{A_c} \cos(\omega_c t + \theta_n)] \cos[\omega_c t + \frac{A_n}{A_c} \sin(\omega_n t + \theta_n)] \qquad (20\text{-}11A)$$

which is the same as Equation (20-2) of the text.

Chapter 21

USE OF THE FOURIER TRANSFORM FOR TRANSMISSION SYSTEM ANALYSIS AND DESIGN

A non-rigorous derivation of the Fourier Transform Pair is made from the Fourier Series to show the extension of the principle of frequency analysis of a periodic time function to the non-periodic case. The analysis is then applied to find the spectra of a rectangular pulse and an impulse. The impulse response of an ideal low-pass filter is derived, and the method for finding the impulse response of any general low-pass or band-pass transmission characteristic is given.

Introduction

The Fourier Transform is one of our most useful tools for analyzing some of the effects which arise in transmission systems. In later chapters this transform will be used to study the effect of transmission deviations in frequency modulation and pulse systems. The present chapter has been included to review some of the important properties of the Fourier Transform and to illustrate its use.

A signal is usually thought of as a function whose value is specified at every instant of time. Such a description specifies its behavior completely. In the case of transmission problems, however, this form of data is not the most convenient one with which to work, because we usually have information about lines and networks in terms of frequency response, rather than in the time domain. We therefore need a method of passing from the time-domain description of the signal to a frequency-domain description, and back again. As one might expect, it is possible to pass from one domain to another by means of mathematical transformations. We are thus enabled to answer such questions as "A pulse (time-domain) is applied to a transmission line whose characteristics we know (frequency-domain); how does the pulse look (time-domain) at the output of the line?"

The duality between frequency and time domains in describing signals and linear networks is a familiar concept to all those who have studied engineering, physics, or mathematics. It is so fundamental that a person often transfers his thinking from one domain to the other without conscious effort. For instance, one might picture a sine wave, of frequency f_o, in the time domain as a snake-like curve which crosses

the time axis $2f_o$ times per second, or in the frequency domain as a narrow spike located at a point $f = f_o$ on the frequency axis and characterized by two numbers giving its amplitude and phase. In this simple case, a method of passing from one of these representations to the other is not difficult to formulate. One can find frequency, amplitude, and phase from a time domain picture of a sinusoidal wave by merely counting and measuring appropriate dimensions. Similarly, the time domain waveform can be constructed if these quantities are given. For more complicated waveforms this transformation is not so simple and a more sophisticated method must be employed to pass from one domain to the other. The Fourier Transform Pair is the mathematical formalization of this useful concept and, as such, is an indispensable tool when dealing with signal or network.

Mathematical and Philosophical Background

The alternative description of a signal in the time and frequency domain is based upon the fact that when sine waves of various frequencies are combined with suitable amplitudes and phases, their sum can be made to approximate any one of a large group of time signals. Similarly, any one of these signals can be decomposed into these component sine waves. A good starting point for a non-rigorous derivation of the Fourier Transform Pair* is the more familiar Fourier Series given by:

$$f(t) = \frac{A_o}{2} + \sum_{n=1}^{\infty} (A_n \cos n\omega_o t + B_n \sin n\omega_o t) \qquad (21\text{-}1)$$

in which

$$A_n = \frac{1}{p} \int_{-p}^{p} f(t) \cos n\omega_o t \, dt, \quad n=0,1,2, \ldots \qquad (21\text{-}2)$$

$$B_n = \frac{1}{p} \int_{-p}^{p} f(t) \sin n\omega_o t \, dt, \quad n=1,2,3, \ldots \qquad (21\text{-}3)$$

---

*A number of elegant deviations of the Fourier Integral and Transform exist in the literature. (e.g., "Modern Analysis", Chapter IX, by Whittaker and Watson, or "Advanced Engineering Mathematics" by C. R. Wylie, Jr., Chapter 5). Presented here is a simpler derivation whose principal purpose is to make less mysterious to the mathematically unsophisticated certain features, such as the concept of negative frequencies and the use of the exponential time function, $e^{j\omega t}$. The derivation given here is not a general one. The case of an even time function is chosen for simplicity and to clarify particular points of interest. A similar procedure can be applied to the general case, however.

Equation (21-1) is the familiar Fourier series, in which

$$2p = \frac{1}{f_o}$$

$$f_o = \frac{\omega_o}{2\pi} = \text{fundamental frequency of periodic time function}$$

$$\frac{1}{f_o} = \text{duration of fundamental period}$$

The validity of this concept is often crudely demonstrated in the laboratory by taking a fundamental frequency and adding to it proper proportions of harmonics to form a square wave or, conversely, a square wave can be passed through a wave analyzer to show that it is made up of a combination of harmonically related sinusoids. There is a rigorous mathematical proof of the validity of the relationship between a periodic time function and its representation in the frequency domain which will not be covered here. For our purpose, it is assumed that the student is familiar with this idea and the Fourier Series will be used to develop the Integral and Transform.

One point should be emphasized. When we say that a periodic time function can be represented as a set of discrete, harmonically related frequency components, we are saying that a statement of the amplitude and distribution of the frequency components is as complete and as accurate a description of a signal as is the time function itself, and that we ought to be able to think of a signal in terms of either, choosing the frequency domain or time domain as convenience dictates. The Fourier Integral and the Fourier Transform will simply extend this concept to cover non-periodic functions. This extension is needed because we transmit not periodic, but non-periodic, functions of time. Telephone signals are typical examples - whether we transmit AM, FM, or PCM, the signals do not repeat periodically, but constantly change their character.

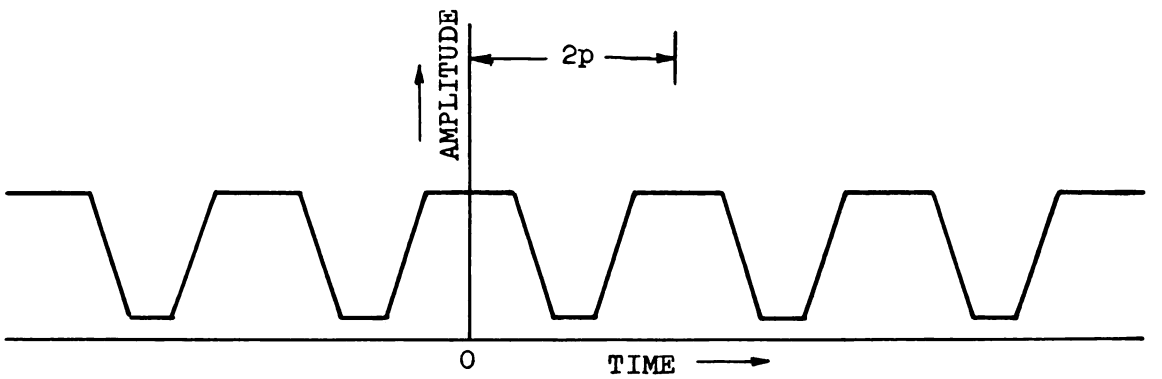
Derivation

Let us assume the special case of an even time function. An even time function is one in which $f(t) = f(-t)$; that is, it is symmetrical about the vertical axis. The integral of an even time function taken between symmetrical limits about the vertical axis equals twice the integral from the zero axis to either of the two limits. This becomes obvious if we remember that $\int_{-T}^{T} f(t)\,dt$ is actually the area under the $f(t)$ curve from -T to T and that if $f(t)$ is symmetrical about the

vertical axis then the area between t = -T and t = 0 equals the area between t = 0 and t = T. (See Figure 1.) Expressed analytically,

$$\int_{-T}^{T} f(t)\, dt = 2 \int_{0}^{T} f(t)\, dt \qquad (21\text{-}4)$$

when f(t) is an even function.



An Even Periodic Function of Time

Figure 21-1

Equation 21-1 shows f(t) to be a function of sine and cosine terms. A cosine curve is an even function [cos ωt = cos (-ωt)]. A sine curve is odd [sin ωt = -sin (-ωt)]. If f(t) is an even function it must consist only of even components. Therefore, for an even function of time f(t), there are no sine terms and Equations (21-1, 2, 3) become (ignoring the d.c. term):

$$f(t) = \sum_{n=1}^{\infty} A_n \cos n\omega_o t \qquad (21\text{-}5)$$

$$A_n = \frac{2}{p} \int_{t=0}^{p} f(t) \cos n\omega_o t\, dt \qquad (21\text{-}6)$$

$$B_n = 0 \qquad (21\text{-}7)$$

Using (21-6) we can write (21-5) as:

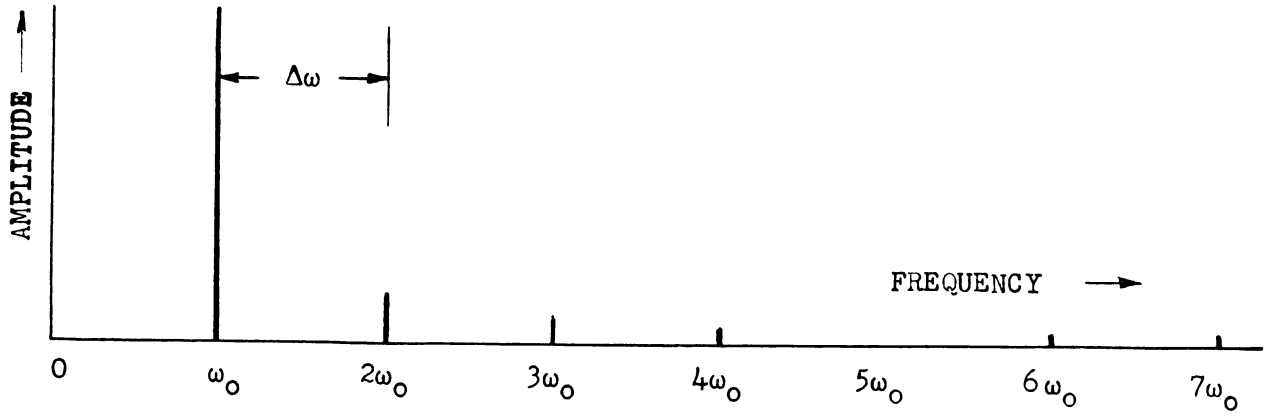
$$f(t) = \sum_{n=1}^{\infty} \left[\frac{2}{p} \int_{t=0}^{p} f(t) \cos n\omega_o t\, dt \right] \cos n\omega_o t \qquad (21\text{-}8)$$

Multiply thru by $\frac{\pi}{\pi}$:

$$f(t) = \sum_{n=1}^{\infty} \left[\frac{2}{\pi} \int_{t=0}^{p} f(t) \cos n\omega_o t\, dt \right] (\cos n\omega_o t) \left(\frac{\pi}{p}\right) \qquad (21\text{-}9)$$

The periodic time function f(t) is thus represented by a Fourier Series which consists, in the frequency domain, of an infinite number of discrete sinusoids. Plotted in terms of amplitude vs frequency, these would be a series of spikes at angular frequencies of $1\omega_o$, $2\omega_o$, $3\omega_o$ .... $n\omega_o$ having amplitudes A_1, A_2, A_3 .... A_n. In addition, of course, there is a dc component of amplitude $A_o/2$.

It is convenient to write ω_n for $n\omega_o$, and to note that $\frac{\pi}{p}$ (the last factor in 21-9) is equal to ω_o, the separation between the discrete cosinusoids, and to call this separation $\Delta\omega$.



Discrete Frequency Components of Periodic Function of Time

Figure 21-2

Substituting $\Delta\omega$ for $\frac{\pi}{p}$ and ω_n for $n\omega_o$ in (21-9), we have

$$f(t) = \sum_{n=1}^{\infty} \left[\frac{2}{\pi} \int_{t=0}^{p} f(t) \cos \omega_n t \, dt\right](\cos \omega_n t)(\Delta\omega) \qquad (21\text{-}10)$$

We have been discussing a periodic function of period 2p. If we let $p \to \infty$, our periodic function of many cycles becomes a periodic time function of one cycle -- in plain, non-mathematical language, a single, non-cyclical pulse. This will introduce some interesting changes in Equation (21-10). We began this demonstration with a periodic time function whose frequency transform was a series of spikes separated by $\omega_o = \Delta\omega = \frac{\pi}{p}$. By letting $p \to \infty$ we make $\Delta\omega \to d\omega$. In other words the frequency spikes have now moved closer together so that they have actually become a continuum. Thus we find that the frequency transform of a non-periodic time function contains energy at <u>all</u> frequencies

and not only at discrete frequencies, as is the case for periodic functions. When we let p approach infinity, we can drop the subscript n, knowing that ω represents any frequency and not only harmonics of some fundamental.

Also, since letting $p \to \infty$ moves the frequency components infinitely close together, it can be shown that the summation from n=1 to $n=\infty$ will approach a definite integral, thus

$$\sum_{n=1}^{\infty} \cos \omega_n t \cdot \Delta\omega \to \int_{\omega=0}^{\infty} \cos \omega t \, d\omega$$

Equation 21-10 can now be written

$$f(t) = \int_{\omega=0}^{\infty} \left[\frac{2}{\pi} \int_{t=0}^{\infty} f(t) \cos \omega t \, dt\right] \cos \omega t \, d\omega \qquad (21\text{-}11)$$

If, for convenience, we define a new quantity*

$$g(\omega) = \frac{1}{\pi} \int_{t=0}^{\infty} f(t) \cos \omega t \, dt \qquad (21\text{-}12)$$

Equation 21-11 becomes

$$f(t) = 2 \int_{\omega=0}^{\infty} g(\omega) \cos \omega t \, d\omega \qquad (21\text{-}13)$$

Equations (12) and (13) closely resemble the Fourier Transform Pair, except that the exponential time function $e^{j\omega t}$ is missing and the concept of negative frequency has not been introduced. These arise from the mathematics involved in converting (12) and (13) to more convenient forms, as follows:

Recall that, **

$$\cos \omega t = \frac{1}{2} (e^{j\omega t} + e^{-j\omega t}) \qquad (21\text{-}14)$$

and substitute this in (21-12)

$$g(\omega) = \frac{1}{2\pi} \int_{0}^{\infty} f(t) \left[e^{j\omega t} + e^{-j\omega t}\right] dt$$

$$= \frac{1}{2\pi} \int_{0}^{\infty} f(t) \, e^{j\omega t} \, dt + \frac{1}{2\pi} \int_{0}^{\infty} f(t) \, e^{-j\omega t} \, dt \qquad (21\text{-}15)$$

---

* The particular choice of $1/\pi$ as the multiplier is an arbitrary one; we could associate the $1/\pi$ factor with the remainder of the expression instead of with $g(\omega)$, or we could split it into two factors, each $1/\sqrt{\pi}$, and associate one with $g(\omega)$ and one with the rest of the expression for f(t). Various authors make different choices here.

**See appended note on negative frequencies.

Consider the first term of (15); we can reverse the limits and change the sign of this first term:

$$g_1 \equiv \frac{1}{2\pi} \int_0^{\infty} f(t)\, e^{j\omega t}\, dt = -\frac{1}{2\pi} \int_{\infty}^{0} f(t)\, e^{j\omega t}\, dt \qquad (21\text{-}16)$$

Since integrating f(t) from ∞ to 0 is the same as integrating f(-t) from $-\infty$ to 0, we can change the sign of t wherever it occurs provided we also change the signs of the limits:

$$g_1 = -\frac{1}{2\pi} \int_{-\infty}^{0} f(-t)\, e^{-j\omega t}\, d(-t) \qquad (21\text{-}17)$$

$$= \frac{1}{2\pi} \int_{-\infty}^{0} f(-t)\, e^{-j\omega t}\, dt$$

We have said f(t) is an even function such that f(t) = f(-t), therefore

$$g_1 = \frac{1}{2\pi} \int_{-\infty}^{0} f(t)\, e^{-j\omega t}\, dt \qquad (21\text{-}18)$$

Going back to equation (15), and writing the first term as in (18)

$$g(\omega) = \frac{1}{2\pi} \int_{-\infty}^{0} f(t)\, e^{-j\omega t}\, dt + \frac{1}{2\pi} \int_{0}^{\infty} f(t)\, e^{-j\omega t}\, dt$$

$$= \frac{1}{2\pi} \int_{-\infty}^{\infty} f(t)\, e^{-j\omega t}\, dt \qquad (21\text{-}19)$$

This is the conventional form of the Fourier Transform, which together with the Fourier Integral (which we shall consider next) makes up the Fourier Transform pair. Before we continue, it will be instructive to pause for a moment to examine $g(\omega)$ as given by (21-19); it has a symmetry property which will prove highly useful.

Consider the function $g(-\omega)$, which we can derive from (21-19) merely by changing the sign of ω wherever it occurs;

$$g(-\omega) = \frac{1}{2\pi} \int_{-\infty}^{\infty} f(t)\, e^{j\omega t}\, dt \qquad (21\text{-}20)$$

while the conjugate of $g(\omega)$, denoted by $g^*(\omega)$, can be written by following the general rule that if we have an expression for a function, the conjugate is obtained by merely changing the sign of j wherever it occurs. Thus, from (21-19):

$$g^*(\omega) = \frac{1}{2\pi} \int_{-\infty}^{\infty} f(t)\, e^{j\omega t}\, dt \qquad (21\text{-}21)$$

Thus

$$g(-\omega) = g^*(\omega) \qquad (21\text{-}22)$$

implying that

$$g_R(\omega) = g_R(-\omega) \qquad (21\text{-}23)$$

and

$$-g_I(\omega) = g_I(-\omega) \qquad (21\text{-}24)$$

all under the stipulation that f(t) be a real function.

Now, consider Equation (13)

$$f(t) = 2\int_0^\infty g(\omega)\cos \omega t \, d\omega \qquad (21\text{-}13)$$

We know that f(t) is a real function. The right-hand side of equation (13) must, therefore, also be real. Since cos ωt is real, it follows that in this particular case $g(\omega)$ is real. Being entirely real, with no imaginary component, $g(\omega)$ for the case we are considering must be an even function of frequency - that is, $g(\omega) = g(-\omega)$, from (21-23). This is a consequence of having chosen f(t) as an even function of time, and should not be taken as a general truth.

Knowing $g(\omega)$ to be an even function in this case, we can manipulate (21-13) as we did (21-12), first expanding it and then manipulating the second term as we did the first in the case of 21-12, and concluding with

$$f(t) = \int_{-\infty}^\infty g(\omega)\, e^{j\omega t}\, d\omega \qquad (21\text{-}25)$$

This is the Fourier Integral, also called the Inverse Fourier Transform. While the above derivation is not a generalized one*, having been chosen to easily demonstrate the ideas involved, the results given by (21-19) and (21-25) are much more general. They apply to any function of time if it is a single-valued function, has only a finite number of discontinuities and a finite number of maxima and minima in any finite interval, and if the integral $\int_{-\infty}^\infty |f(t)|\, dt$ converges. All signals of practical interest in communications satisfy the above conditions. The integral condition might seem to be a serious limitation. However, since all practical sources of voltage and current have finite energy, it follows that the integral must converge.

Summary

To recapitulate: the representation of a signal as a sum of sine waves was first pointed out by Fourier and is embodied as the well known Fourier Series representation of periodic functions.

---

*For example, some readers may note that we could have manipulated the expanded versions of (12) and (13) to change the signs of the $j\omega t$ exponent in (19) and (25). This is a result of having chosen the special case of f(t) an even function.

More generally, the process of decomposing a non-periodic signal is called Fourier Analysis, while the inverse is known as Fourier Synthesis. These processes can be carried out mathematically by evaluating the following definite integrals:

Fourier Transform (Fourier Analysis)

$$g(\omega) = \frac{1}{2\pi}\int_{-\infty}^{\infty} f(t)e^{-j\omega t}\,dt = \mathfrak{F}[f(t)] \qquad (21\text{-}26)$$

Fourier Integral (Fourier Synthesis)

$$f(t) = \int_{-\infty}^{\infty} g(\omega)e^{j\omega t}\,d\omega = \mathfrak{F}^{-1}[g(\omega)] \qquad (21\text{-}27)$$

where the $\mathfrak{F}$ notation is an operational form, ω is the frequency and t the time variable, $f(t)$ is the time function and $g(\omega)$ is its representation in the frequency domain. Together these integrals are known as the Fourier Transform Pair.* They can be used for any function of time which satisfies the conditions stated in the preceding section.

Spectrum of a Rectangular Pulse

The time function shown in Figure 3 will serve to illustrate the use of the transform. The function can be expressed as:

$$\begin{cases} f(t) = 0 & -\infty < t < -\Delta/2 \\ f(t) = E & -\Delta/2 < t < \Delta/2 \\ f(t) = 0 & \Delta/2 < t < \infty \end{cases} \qquad (21\text{-}28)$$

Thus the transform is

$$g(\omega) = \frac{1}{2\pi}\int_{-\Delta/2}^{\Delta/2} E\,e^{-j\omega t}\,dt = \frac{E}{2\pi}\int_{-\frac{\Delta}{2}}^{\frac{\Delta}{2}} e^{-j\omega t}\,dt \qquad (21\text{-}29)$$

It can be shown that

$$\int_{-\frac{\Delta}{2}}^{\frac{\Delta}{2}} e^{j\omega t}\,dt = \frac{2\sin\omega\frac{\Delta}{2}}{\omega} \qquad (21\text{-}30)$$

Multiplying numerator and denominator by $\frac{\Delta}{2}$ to obtain a $\frac{\sin x}{x}$ form we obtain:

$$g(\omega) = \frac{E\Delta}{2\pi}\,\frac{\sin\omega\,\Delta/2}{\omega\,\Delta/2} \qquad (21\text{-}31)$$

---

*In many texts, the $1/2\pi$ factor in 21-26 is replaced by $1/\sqrt{2\pi}$, and $f(t)$ is set equal to $1/\sqrt{2\pi}$ times the right hand side of 21-27. This is merely a change in the arbitrary definition of $g(\omega)$ which was given in Equation 21-12.

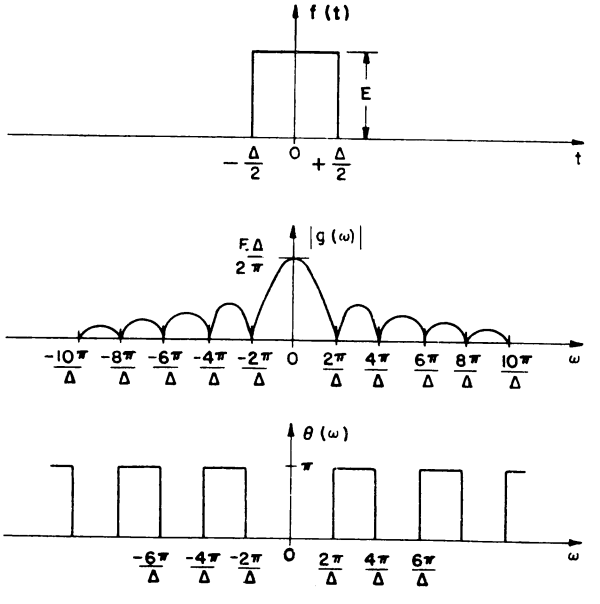
The corresponding frequency and phase spectra are also plotted in Figure 3. Note that here again we have chosen to analyze an even function of time, and in consequence $g(\omega)$ is real. This is a special case; in general $g(\omega)$ is a complex function:

$$g(\omega) = |g(\omega)|\, e^{j\theta(\omega)} = g_R(\omega) + j\, g_I(\omega) \qquad (21\text{-}32)$$

where $g_R(\omega)$ and $g_I(\omega)$ are the real and imaginary parts respectively of $g(\omega)$. In the example above $g(\omega)$ is real, but contains phase reversals in some portions of the spectrum. The frequency spectrum referred to above is the absolute value of $g(\omega)$, while the phase spectrum is the value of the phase angle $\theta(\omega)$ associated with $g(\omega)$ [Note that a phase reversal is the same as a change in sign of $g(\omega)$]. Both amplitude and phase information are necessary to specify a time function since both are contained in $g(\omega)$, which is required for the Fourier synthesis procedure. In fact, a given frequency spectrum can result from any one of an infinite number of time functions, each of which results in a different phase spectrum.

Transmission Characteristics vs Impulse Response

It was pointed out earlier that time functions are sometimes not convenient to work with analytically, and it was implied that in some cases the complex spectrum can be utilized to simplify rather



Non-Periodic Time Function and its Frequency and Phase Spectra

Figure 21-3

complicated problems. The advantages to be had by operating in the frequency domain arise from the simple relation between the input and output of linear networks when specified in that dimension. In a typical problem, the input function has a spectrum $g_i(\omega)$ and the output $g_o(\omega)$. The transmission path can be likewise described by a frequency function which is its transfer impedance, transfer voltage or current ratio, or what is commonly called its "frequency response". This function is denoted by $Y(\omega)$ and can be established by computation from the known circuit constants of the system or experimentally by means of a sine wave test signal input and a suitable output meter to measure amplitude and relative phase.

The relationship between the input and output spectra of a time signal applied to a network is particularly simple. In complex notation:

$$g_o(\omega) = Y(\omega)\, g_i(\omega) \qquad (21\text{-}33)$$

In polar form:

$$\text{Frequency spectrum: } |g_o(\omega)| = |Y(\omega)|\ |g_i(\omega)| \qquad (21\text{-}34)$$

$$\text{Phase spectrum: } \theta_o(\omega) = \theta_Y(\omega) + \theta_i(\omega) \qquad (21\text{-}35)$$

The validity of these relations rests upon the superposition principle since g_o is computed by assuming that it is a linear combination of the responses of the network to the each frequency component in the input wave individually. This observation implies that <u>if the response of a linear system to the gamut of sine wave excitations is known, then its response to any other waveform can be found uniquely</u> by merely decomposing that wave into its Fourier components and computing the response to each individual component. The output waveform, $f_o(t)$, can be found by evaluating the Fourier Integral of $g_o(\omega)$. The convenient generality underlined above is the basis for all sine wave testing techniques used in practice. It should be noted, however, that it is useful only for <u>linear systems</u> since it is only in such systems that superposition is generally valid. In the case of a non-linear device, such as a rectifier, the response to each input waveform must be computed separately and the network's complex frequency response does not allow generalization to include other functions.*

- -

*It has been suggested that a complete specification of a non-linear system can be had by noting its response to an infinitely long sample of random noise, since such a sample will contain all possible waveforms. It is conceivable that practical test procedures for specific devices can be formulated from this proposal.

The network can also be completely described in terms of its "impulse response". An impulse is the result approximated when we narrow a pulse without limit while keeping its area ($E\Delta$ in Figure 3, usually set equal to unity or denoted by δ) unchanged. In the time domain, then, an impulse is a signal having energy but vanishingly small duration*. The corresponding frequency spectrum, which can be found from Equation (21-31) by setting $E\Delta$ equal to unity and letting $\Delta/2$ approach zero, contains all frequencies from $-\infty$ to $+\infty$, of equal phase and of amplitude $1/2\pi$. The impulse response of a network is the time function $y(t)$ that would be found at the output as a result of applying an impulse to the input terminals. Since the time function applied to the input has a flat frequency spectrum we would expect $y(t)$, the time function at the output, to have a spectrum which differed from flatness by the frequency characteristic of the network. In other words, $Y(\omega)$ gives the frequency and phase spectra of $y(t)$. Expressed analytically, an impulse input $1/2\pi$ to a network $Y(\omega)$ produces an output $y(t)$ given by

$$\mathfrak{F}\,[y(t)] = \frac{1}{2\pi}\,Y(\omega)$$

from which it follows that

$$Y(\omega) = \int_{-\infty}^{\infty} y(t)\; e^{-j\omega t}\; dt \qquad (21\text{-}36)$$

and also

$$y(t) = \frac{1}{2\pi}\int_{-\infty}^{\infty} Y(\omega)\; e^{j\omega t}\; d\omega \qquad (21\text{-}37)$$

The impulse response is, of course, a real function of time. Thus the properties of $g(\omega)$ stated in Equations 22, 23, and 24 also hold for $Y(\omega)$. That is,

$$\left.\begin{aligned} Y(-\omega) &= Y^*(\omega) \\ Y_R(\omega) &= Y_R(-\omega) \\ -Y_I(\omega) &= Y_I(-\omega) \\ |Y(\omega)| &= |Y(\omega)| \end{aligned}\right\} \qquad (21\text{-}38)$$

where Y_R and Y_I are respectively the real and imaginary parts of $Y(\omega)$.

---

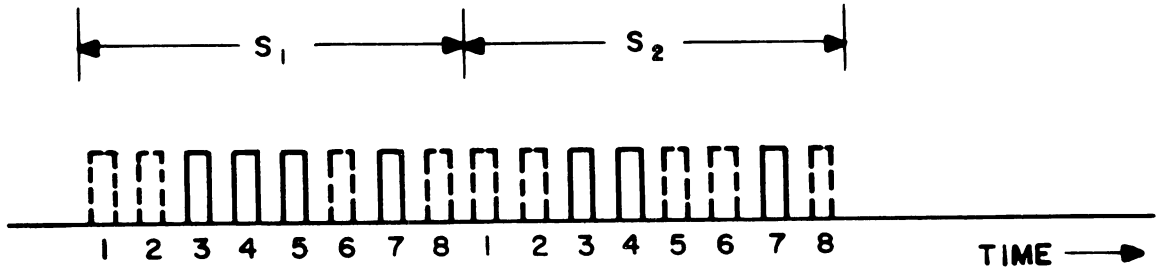
*Similarly we also speak of single frequencies as impulses in the frequency domain - that is, spikes having amplitude but no bandwidth. The mathematician's symbol for an impulse is $\delta(x)$, meaning a function which is zero everywhere except where the argument "x" is zero. Thus $\delta(\omega-\omega_1)$ means a spike at a frequency of ω_1, since when the variable ω is equal to ω_1, the argument is zero.

These are extremely important properties of any physical transmission path. They are expressed in words by saying that the real component of any physical transmission characteristic will display even symmetry about zero frequency, and the imaginary component will display odd symmetry. This fact will be made use of frequently in subsequent discussion.

Impulse Response of Ideal Low-Pass Filter

As an example of the usefulness of the Fourier Transform Pair and the ideas we have been developing, let us now consider a problem in pulse transmission.

Suppose we are transmitting information by Pulse Code Modulation. At the transmitting terminal, we either send or do not send a pulse at time t_1, t_2 etc. At the receiving end, the problem is to tell whether or not a pulse is present at time t_i.



Two Possible Successive PCM Signals

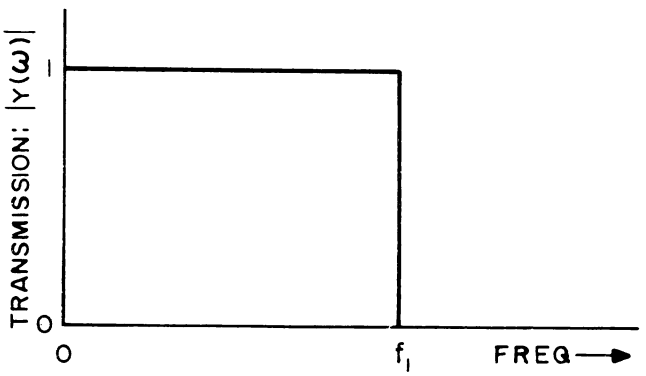
Figure 21-4

For example, the difference between two successive code signals (as illustrated by S_1 and S_2 in Figure 4) might lie in the fact that S_1 has a pulse at position 5, whereas S_2 does not. If anything happens to these signals which will tend to disguise this fact, our receiver will tend to make an error in trying to distinguish between S_1 and S_2.

If our transmitting medium were of unlimited bandwidth, had no delay distortion, and were free of noise, there would be no difficulty. This, of course, is not the case; let us examine first the idealized case of bandwidth limitation alone.

One of the possible sources of error is this: the energy in the fourth pulse position of S_2 may spill over into position 5, where ideally no energy should be. Let us inquire whether it will, and if so, how seriously, and let us first ask this question with reference to the system transmission characteristic shown in Figure 5. This idealized

transmission characteristic has a constant, finite value* of attenuation from dc to f_1, and infinite attenuation above f_1. It has no delay distortion for frequencies from dc to f_1; delay distortion above f_1 is of no consequence, of course. (This is an easy case to analyze first - it happens to be impossible to achieve it, but it can be approached. More achievable characteristics are more complicated to analyze, so we will defer consideration of them till we have looked at this easier case.)



Idealized Low Pass Transmission Characteristic

Figure 21-5

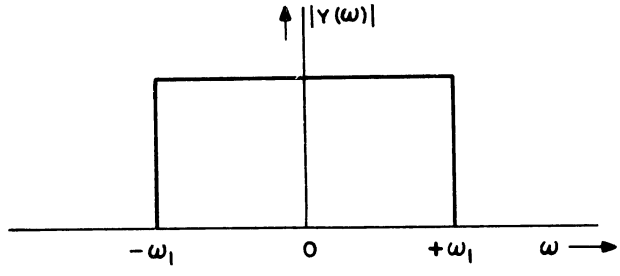
The next question we have to settle is: what assumption shall we make as to the spectrum of the input pulse at position 4? Off-hand one would be inclined to assume a rectangular pulse, though in fact we know that real pulses are never ideally rectangular. We can make our problem even simpler than this, however. Let us compare the spectrum of an impulse (flat vs frequency, with no phase reversals) with the spectrum of a rectangular pulse in the region of $\omega=0$ (almost flat for very low frequencies). We see that if the transmitted bandwidth is small enough compared to the first frequency at which $(\sin x)/x$ becomes zero, the output will be the same whether we take the input to be a narrow rectangular pulse or an impulse. The spectrum of an impulse is so easy to handle analytically that we shall assume the input to be an impulse. If we want to refine our results later we can do so by modifying the input spectrum to have the $(\sin x)/x$ shape, or we can modify our $Y(\omega)$. Which we modify, of the two factors that appear as a product, does not matter.

---

*Any finite value of attenuation would do, since we are not introducing noise yet. Zero db attenuation (unity transmission) is a convenient value to use.

To recapituate, our problem is: what is the signal, as a function of time, at the output of the path having the characteristic shown on Figure 5, if the input f(t) is an impulse -- more briefly, what is the impulse response of such a transmission path?

To answer this, let us first note that we can plot $|Y(\omega)|$ for negative as well as positive frequencies. By the relations of Equations 21-38, the plot would look like Figure 6 (where $\omega_1 = 2\pi f_1$ has been substituted for f_1).



Idealized LP Characteristic (Pos & Neg Freq)

Figure 21-6

From Equation 21-37, the output pulse is

$$y(t) = \frac{1}{2\pi} \int_{-\infty}^{\infty} Y(\omega)\, e^{j\omega t}\, d\omega$$

Obviously, from Figure 6, this (ignoring the constant delay) gives:

$$y(t) = \frac{1}{2\pi} \int_{-\omega_1}^{+\omega_1} e^{j\omega t}\, d\omega \qquad (21\text{-}39)$$

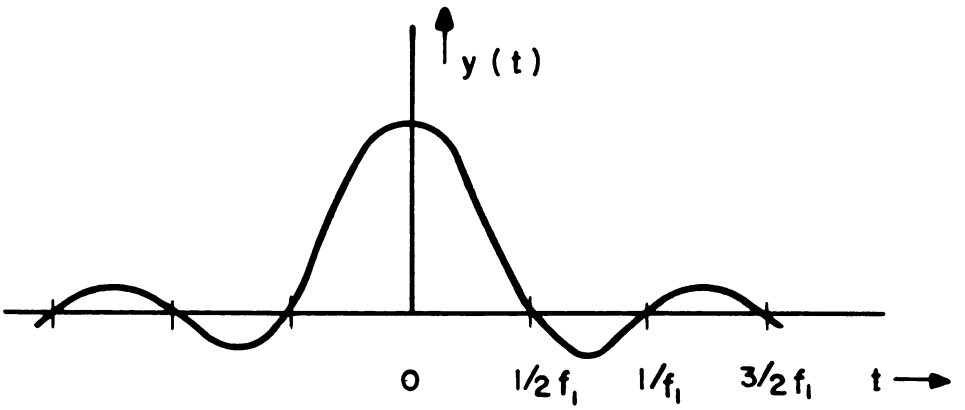
This is very similar to Equation (21-29). The result is:

$$y(t) = \frac{\omega_1}{\pi} \frac{\sin \omega_1 t}{\omega_1 t} \qquad (21\text{-}40)$$

This is plotted on Figure 7*. Clearly the optimum time for the next pulse is at $t = 1/2\, f_1$. As a numerical example, suppose

- -

*t=0 on this plot is arbitrary; for a physical network which approximated the characteristic of Figure 5 and 6, the zero time point would represent the delay of the transmission path.



y(t) At Output of LP Transmission Path

Figure 21-7

the cut-off of the transmission path is at 500 kc - then the interval between impulses should be 1 μs (repetition rate, 1 mc). A shorter interval will tend to make the receiver think a pulse is present when in fact it is not; a longer interval will result in some cancellation when the following pulse is present.

This result - the spacing of pulses to avoid intersymbol crosstalk - is one of the fundamental theorems of pulse transmission.

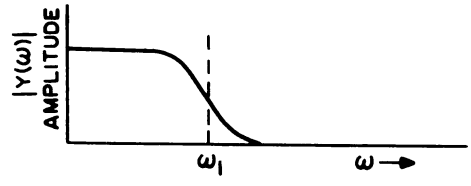
The above example illustrates sufficiently the way in which the Fourier Transform Pair can be used. Starting with an input signal which is a given function of time, we can find the signal (as a function of time) at the output of a network if we know the transmission characteristic of the network. Our data and results may be expressed in very general functional terms in order to display the nature of a problem, or specific formulae may be used to obtain specific numerical results. In any particular case, finding the solution may be easy (as in the illustrative example above) or may involve laborious or clever mathematical manipulation of the specific functions involved in the problem. The basic idea remains the same.

This might be taken as the conclusion of this chapter, since our purpose was to illustrate the basic ideas involved, and there is no end to the illustrative examples that we could go on to consider. There is, however, one type of problem that is of great practical importance in transmission, and that involves a type of manipulation that is often skimmed over by authors who are (properly) concerned

with the results rather than the reader's comprehension of the procedure by which they are obtained. This is the analysis of the bandpass problem. The remainder of this chapter is therefore devoted to a discussion of the band pass case. An interesting application of this occurs when we analyze the transmission of a very short pulse through a low-pass characteristic which has a gradual cut-off.*

Impulse Response of Low Pass Characteristic With Gradual Cutoff

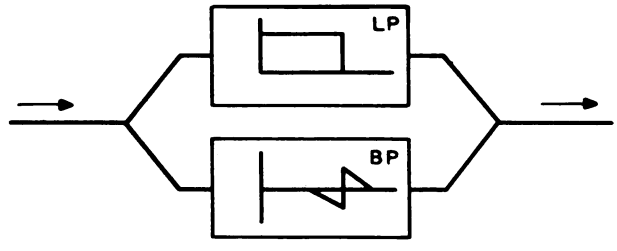
It was stated earlier that the low-pass characteristic of Figure 5 was unrealizable. To make our results on the preceding PCM problem more meaningful, and further illustrate the use of the Fourier Transform Pair, let us consider the impulse response of the transmission characteristic shown in Figure 8.



Idealized LP Characteristic with Gradual Cutoff

Figure 21-8

This can be considered as the sum of two transmission paths as illustrated in Figure 9, where the bandpass path has odd symmetry about ω_1. The total impulse response is the sum of the impulse responses of the two paths; we already know the answer for the low pass path, and need only find the impulse response for the band pass case.



Resolution of Gradual Cutoff Into Component Paths

Figure 21-9

- -

*The following discussion is adapted, with changes, from Monograph 2284, "Theoretical Fundamentals of Pulse Transmission", by E. D. Sunde.

Band-pass Transformations

A band-pass transmission characteristic, specified in terms of amplitude $A(\omega)$ and phase $B(\omega)$, can be written, like any other transmission characteristic

$$Y(\omega) = A(\omega)\ e^{jB(\omega)} \tag{21-41}$$

In dealing with band-pass characteristics, it is convenient to consider the symmetry (or lack of symmetry) of the characteristic about a reference frequency ω_r within the transmitted band. For this purpose we define a new variable u, thus

$$u = \omega - \omega_r \tag{21-42}$$

In terms of the new variable u, the band-pass characteristic can be expressed as

$$G(u) = a(u)\ e^{jb(u)} \tag{21-43}$$

where

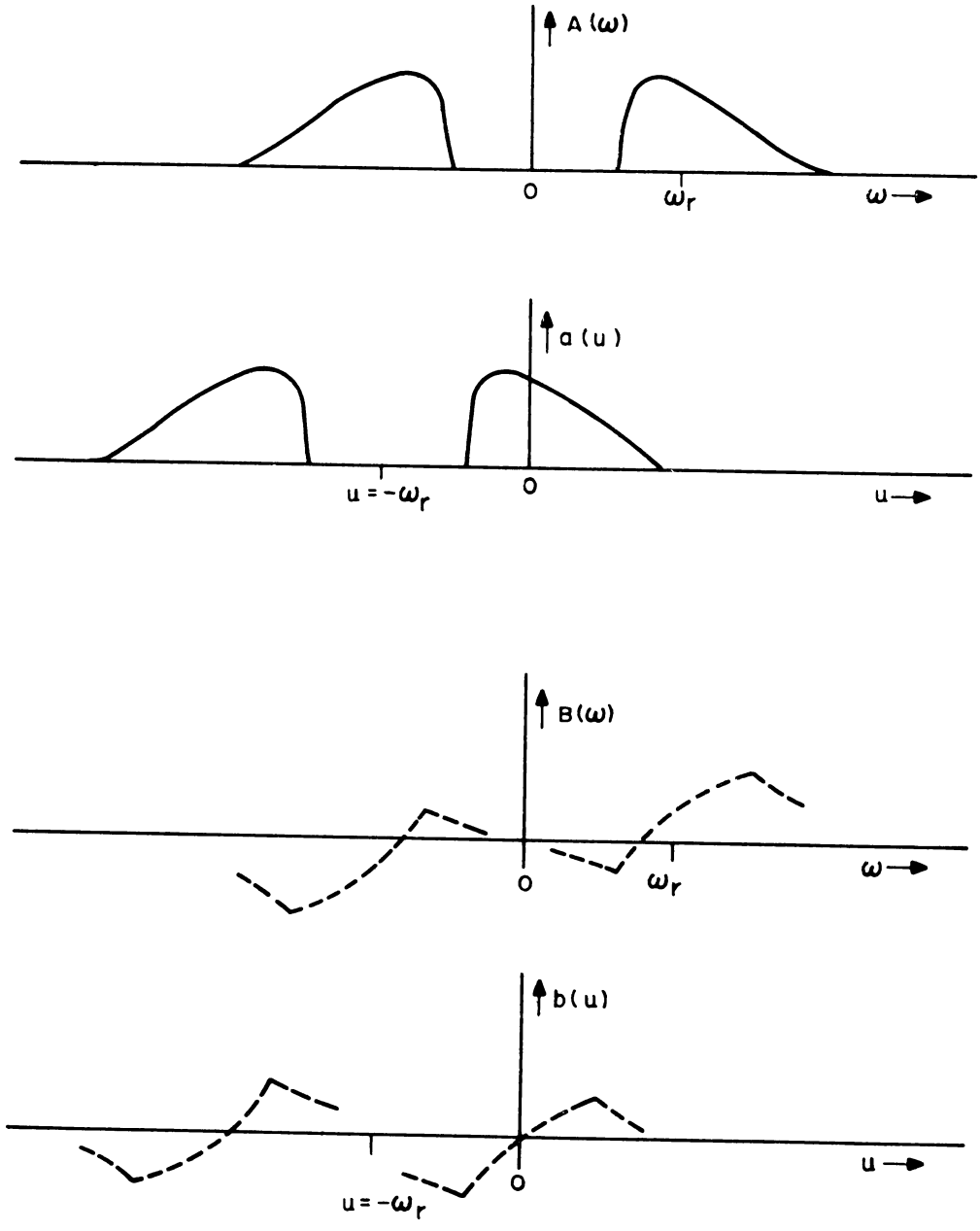
$$a(u) = A(u+\omega_r) \tag{21-44}$$

$$\left.\begin{aligned} b(u) &= B(u+\omega_r) - B(\omega_r),\ u > -\omega_r (\text{or } \omega > 0) \\ &= B(u+\omega_r) + B(\omega_r),\ u < -\omega_r (\text{or } \omega > 0) \end{aligned}\right\} \tag{21-45}$$

$B(\omega_r)$ can, in any particular case, be evaluated as some angle, which for convenience we will define as β. Then

$$\left.\begin{aligned} b(u) &= B(u+\omega_r) - \beta,\ u > -\omega_r \\ &= B(u+\omega_r) + \beta,\ u < -\omega_r \end{aligned}\right\} \tag{21-46}$$

This transformation shifts the transmission characteristic, as shown in Figure 10, so that on the u-axis the point of zero relative frequency (i.e., the u = 0 point) corresponds to the $\omega = \omega_r$ point on the ω-axis. Note that in addition to the lateral shift, we have made a vertical shift in the phase characteristic to make the phase shift zero at u = 0. This is merely for mathematical convenience; it does not change the delay (slope of phase curve) or delay distortion (variation in slope of phase curve) of the path. It should be recalled that the phase characteristic of any physical transmission path must display odd symmetry about zero frequency. This fundamental property applies whether the characteristic is plotted on the ω-axis or, in this case, the u-axis. Therefore, it is necessary to write two relations, as in (21-46), so that when analytically expressing the vertical shift of the phase characteristic the odd symmetry property is maintained.



Translation of $Y(\omega)$ to $G(u)$

Figure 21-10

Note: This sort of frequency shift is often performed without explicitly introducing a new variable such as "u". One could, in the second of the above diagrams, label the horizontal axis as ω rather than u; the function being plotted is then $A(\omega + \omega_r)$. The point $u = -\omega_r$ would then be labeled $\omega = -\omega_r$ Similarly the bottom graph would become a plot of $B(\omega + \omega_r)$ against ω. Taken together they form a plot of $Y(\omega + \omega_r)$; obviously this will, for $\omega = 1$ kc, have the same ordinate value as $Y(\omega)$ had for $\omega = \omega_r + 1$ kc. Occasionally it is then convenient to define a new function such as $H(\omega) = Y(\omega + \omega_r)$.

The impulse response, it will be recalled, is

$$y(t) = \frac{1}{2\pi} \int_{-\infty}^{+\infty} Y(\omega)\, e^{j\omega t}\, d\omega \tag{21-47}$$

$$= \frac{1}{2\pi} \int_{-\infty}^{+\infty} A(\omega) e^{j[\omega t + B(\omega)]}\, d\omega \tag{21-48}$$

$$= \frac{1}{2\pi} \int_{-\infty}^{+\infty} A(\omega) \cos [\omega t + B(\omega)]\, d\omega + \frac{1}{2\pi} \int_{-\infty}^{+\infty} A(\omega)\, j \sin [\omega t + B(\omega)]\, d\omega \tag{21-49}$$

The second of these integrals is equal to zero, since $A(\omega)$ has even symmetry and $\sin[\omega t + B(\omega)]$ has odd symmetry, giving odd symmetry to the whole expression to be integrated. On the other hand, the function to be integrated in the first of these integrals has even symmetry, so the area from $-\infty$ to $+\infty$ is merely twice the area from zero to $+\infty$. Therefore,

$$y(t) = \frac{1}{\pi} \int_{0}^{\infty} A(\omega) \cos [\omega t + B(\omega)]\, d\omega \tag{21-50}$$

[Note that so far this is valid for the general case of the impulse response of any physical network.]

Now let us transform (21-50) into a forw which is more convenient for bandpass problems.

For $A(\omega)$ substitute $a(u)$
For $B(\omega)$ substitute $b(u) + \beta$
For $d\omega$ substitute du
For ωt substitute $ut + \omega_r t$
For $\omega = 0$ substitute $u = -\omega_r$

[At this point it might be noted that it is often advantageous to change the upper limit of integration from infinity to some specific frequency above which the transmission is zero. In some cases, $A(\omega)$ is zero when $\omega > 2\omega_r$, which can lead to a convenient symmetry of the limits of integration later.]

Equation 21-50 then reads:

$$y(t) = \frac{1}{\pi} \int_{-\omega_r}^{\infty} a(u) \cos [ut + \omega_r t + b(u) + \beta]\, du \tag{21-51}$$

But

$$\cos[ut + \omega_r t + b(u) + \beta)] = \cos [\omega_r t + \beta) + (ut + b(u))] \tag{21-52}$$

which is the cos $(\alpha+\beta)$ form, and can be written

$$\cos(\omega_r t+\beta)\cos[ut + b(u)] - \sin(\omega_r t + \beta)\sin[ut + b(u)] \qquad (21\text{-}53)$$

Since $(\omega_r t + \beta)$ is a constant vs ω, we can write the sin and cos functions of this quantity outside the integral sign, and obtain:

$$y(t) = \frac{\cos(\omega_r t + \beta)}{\pi}\int_{-\omega_r}^{\infty} a(u)\cos[ut + b(u)]\,du$$

$$- \frac{\sin(\omega_r t + \beta)}{\pi}\int_{-\omega_r}^{\infty} a(u)\sin[ut + b(u)]\,du \qquad (21\text{-}54)$$

We will be able to see the significance of this expression a little better if we split each integral (from $-\omega_r$ to ∞) into two integrals - one from $-\omega_r$ to zero, the other from zero to infinity. This gives:

$$y(t) = [R_-(t) + R_+(t)]\cos(\omega_r t + \beta)$$
$$+ [Q_-(t) - Q_+(t)]\sin(\omega_r t + \beta) \qquad (21\text{-}55)$$

where we define

$$R_-(t) = \frac{1}{\pi}\int_{-\omega_r}^{0} a(u)\cos[ut + b(u)]\,du$$

$$= \frac{1}{\pi}\int_{0}^{\omega_r} a(-u)\cos[ut - b(-u)]\,du^* \qquad (21\text{-}56)$$

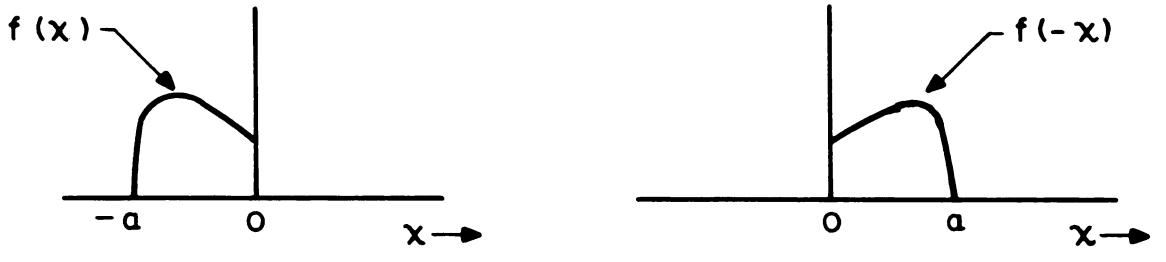
---

*Obtaining (21-56) from the line above involves first recognizing that

$$\int_{-a}^{0} f(x)\,dx = \int_{0}^{a} f(-x)\,dx \qquad \text{(see sketch)}$$

and then, because cos (y) = cos (-y), substituting

cos [ut - b(-u)] for cos [-ut + b(-u)]



Equation (21-58) is obtained in a similar manner.

$$R_+(t) = \frac{1}{\pi}\int_0^{\infty} a(u)\cos[ut + b(u)]\,du \qquad (21\text{-}57)$$

$$Q_-(t) = \frac{1}{\pi}\int_0^{\omega_r} a(-u)\sin[ut - b(-u)]\,du \qquad (21\text{-}58)$$

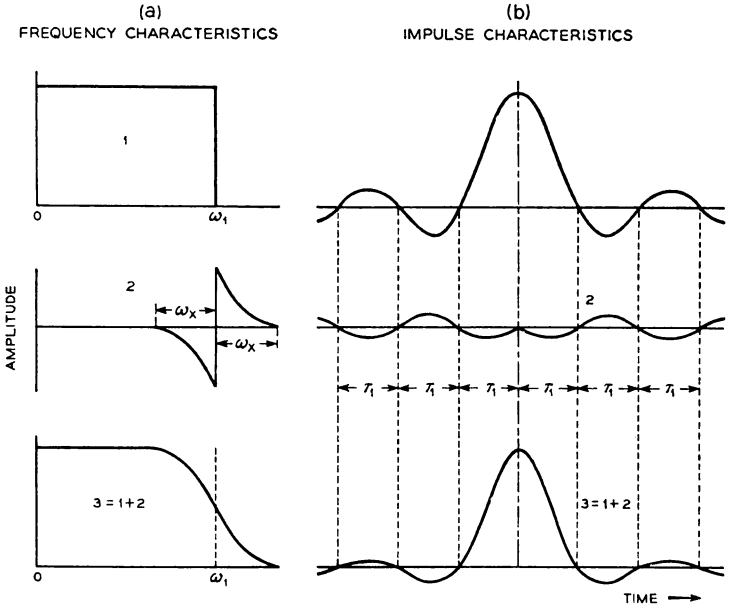
$$Q_+(t) = \frac{1}{\pi}\int_0^{\infty} a(u)\sin[ut + b(u)]\,du \qquad (21\text{-}59)$$

The envelope $\overline{F}(t)$ of the impulse transmission characteristic is given by

$$\overline{F}(t) = [(R_- + R_+)^2 + (Q_- - Q_+)^2]^{1/2} \qquad (21\text{-}60)$$

Comparison of 21-57 with 21-50 shows that R_+ can be identified with the impulse characteristic of a low pass system having the same frequency characteristic above zero as the bandpass system has above ω_r.

Similarly 21-56 can be identified with the impulse characteristic of a low pass system which from 0 to ω_r has the characteristic that the band-pass has from 0 to $-\omega_r$.



a) Idealized transmission characteristic with gradual cut-off, 3, obtained by superposition of characteristic with sharp cut-off, 1, and characteristic, 2, with odd symmetry about ω_1. Linear phase shift assumed.

b) Associated impulse responses.

Figure 21-11

The impulse characteristics Q_- and Q_+ arise from asymmetry in the transmission characteristic with respect to ω_r. Their algebraic sum is not present in low-pass systems, since by definition the amplitude characteristic has even symmetry and the phase characteristic odd symmetry with respect to zero frequency.

The first and second components of (21-55) are referred to as the in-phase and quadrature components of the impulse characteristic of band-pass systems.

Application of Band-Pass Transformation to a Gradual Cutoff Characteristic

The final evaluation of the impulse response of the transmission characteristic shown in Figure 8 is now merely a matter of choosing an appropriate bandpass characteristic for the second network of Figure 9, and going through the mathematics. The analysis is simplified if we let $\omega_r = \omega_1$ and choose a low-pass shape such that the band-pass component has odd symmetry about ω_1 and is defined by a simple equation.

As an illustration of the method, the impulse response of the band-pass component of the transmission characteristic shown in Figure 11 will be found. Let the band-pass characteristic be given by

$$\left.\begin{aligned} A(\omega) &= \tfrac{1}{2}\left[1 - \sin\frac{\pi(\omega-\omega_1)}{2\omega_x}\right] \qquad \omega_1 \le \omega \le \omega_1+\omega_x \\ &= -\tfrac{1}{2}\left[1 + \sin\frac{\pi(\omega-\omega_x)}{2\omega_x}\right] \qquad \omega_1-\omega_x \le \omega \le \omega_1 \\ &= 0 \text{ for all other values of } \omega \end{aligned}\right\} \tag{21-61}$$

In general, to obtain the time function plotted in Figure 11 the delay (slope of the phase curve) for the low-pass and band-pass paths should be equal, and the phase curves linear. For convenience in this problem, however, let $B(\omega)$ equal zero. If follows, then, that

$$\left.\begin{aligned} a(u) &= \tfrac{1}{2}\left[1 - \sin\frac{\pi u}{2\omega_x}\right] \; 0 \le u \le \omega_x \\ &= \tfrac{1}{2}\left[1 + \sin\frac{\pi u}{2\omega_x}\right] \; -\omega_x \le u \le 0 \\ &= 0 \text{ for all other values of } u \\ b(u) &= 0 \text{ for all values of } u \end{aligned}\right\} \tag{21-62}$$

At this point in the problem it is important to recognize some characteristics of the band-pass function being considered. In the first place, the function possesses odd symmetry about the $u = 0$ (or $\omega = \omega_1$) axis. As a result, the summation of the R components in the first term of Equation (21-55) must be zero since this term exists only when the function has an even symmetry component. That this sum will turn out to be zero for this characteristic can be seen by examining Equations (21-56) and (21-57). Both integrals must have the same magnitude because each evaluates the area under the same shape curve (i.e., a curve of the form $a(u) \cos ut$), but since one integral involves $a(-u)$ and the other $a(u)$, and since $a(u)$ is an odd function, integration between the same limits causes the sign of the areas to be opposite. Thus,

$$R_- (t) = -R_+ (t)$$

Similar reasoning leads to the conclusion that

$$Q_- (t) = - Q_+ (t),$$

a condition which must be so for a function such as this which possesses wholely odd symmetry about $u = 0$. Applying these conditions, Equation (21-55) becomes (letting $\omega_1 = \omega_r$),

$$y_b(t) = -2Q_+(t) \sin \omega_1 t$$

$$= -2 \sin \omega_1 t \left[\frac{1}{\pi} \int_0^\infty a(u) \sin ut \, du \right]$$

$$= - \frac{2}{\pi} \sin \omega_1 t \int_0^{\omega_x} \frac{1}{2}\left[1 - \sin \frac{\pi u}{2\omega_x} \right] \sin ut \, du \qquad (21\text{-}63)$$

Here the notation $y_b (t)$ is used to denote that Equation (21-63) gives the impulse response of only the band-pass component of

the transmission characteristic shown in Figure 11. The problem now becomes one of performing the integration and combining terms. Thus,

$$y_b(t) = -\frac{1}{\pi}\sin\omega_1 t\left[\int_0^{\omega_x}\sin ut\,du - \int_0^{\omega_x}\sin\frac{\pi u}{2\omega_x}\sin ut\,du\right]$$

$$= -\frac{1}{\pi}\sin\omega_1 t\left[-\frac{\cos\omega_x t}{t} + \frac{1}{t} - \frac{\sin(\frac{\pi}{2\omega_x} - t)\omega_x}{2(\frac{\pi}{2\omega_x} - t)} + \frac{\sin(\frac{\pi}{2\omega_x} + t)\omega_x}{2(\frac{\pi}{2\omega_x} + t)}\right]$$

$$= -\frac{1}{\pi}\sin\omega_1 t\left\{\frac{1}{t} - \cos\omega_x t\left[\frac{1}{t} + \frac{1}{2(\frac{\pi}{2\omega_x} - t)} - \frac{1}{2(\frac{\pi}{2\omega_x} + t)}\right]\right\}$$

$$= -\frac{1}{\pi}\sin\omega_1 t\left\{\frac{1}{t} - \omega_x\cos\omega_x t\left[\frac{1}{\omega_x t} + \frac{1}{\pi - 2\omega_x t} - \frac{1}{\pi + 2\omega_x t}\right]\right\}$$

$$= -\frac{1}{\pi}\sin\omega_1 t\left\{\frac{1}{t} - \frac{1}{t}\cos\omega_x t\left[\frac{1}{1 - (\frac{2\omega_x t}{\pi})^2}\right]\right\}$$

$$= -\frac{\omega_1}{\pi}\,\frac{\sin\omega_1 t}{\omega_1 t}\left[1 - \frac{\cos\omega_x t}{1 - (\frac{2\omega_x t}{\pi})^2}\right] \qquad (21\text{-}64)$$

Equation (21-64) gives the impulse response of the band-pass characteristic, which is plotted in Figure 11 for the particular case of $\omega_x = \omega_1/2$. The impulse response of the complete filter can be found by adding the impulse response of an ideal low-pass filter to

$y_b(t)$. Equation (21-40) showed the impulse response of an ideal low-pass filter of cut-off frequency ω_1 to be

$$y_1(t) = \frac{\omega_1}{\pi} \frac{\sin \omega_1 t}{\omega_1 t} \tag{21-65}$$

Adding Equation (21-65) to (21-64) gives the impulse response of the gradual cut-off characteristic as

$$y(t) = \frac{\omega_1}{\pi} \frac{\sin \omega_1 t}{\omega_1 t} \frac{\cos \omega_x t}{1 - (2\omega_x t/\pi)^2} \tag{21-66}$$

Note that for the case plotted in Figure 11 the zeros still occur at the same time as for the original sharp low-pass characteristic, but the amplitude of the preceding and succeeding oscillations has been considerably reduced. This effect of a gradual cut-off characteristic is important in reducing interference between adjacent pulses, as will be discussed in greater detail in Chapter 27.

Limits of Integration and Time Varying Spectra in Fourier Analysis

An examination of the Fourier Integral (Equation 21-27) indicates that in order to determine the function of time corresponding to a particular frequency spectrum, it is necessary to know that spectrum from d.c. to infinite frequencies. In the application of Fourier Synthesis in any real situation, the signal under study will always have been generated by a source capable of producing only a finite range of frequencies. Similarly, it will have been carried by a medium capable of transmitting only a finite bandwidth. Hence, it will be necessary to examine the spectrum only in this region, and it can be assumed to be zero outside this region. It might be suspected that such a finite bandwidth would restrict the number of time functions which can be synthesized. In Chapter 25 we will find this is indeed the case. Only those functions can be synthesized whose fastest time-rate of change is of the same order as the rate of change of the highest frequency component that may be present.

If we now look at the Fourier Transform (Equation 21-26), we find that here also the integral extends from minus infinity time to

plus infinity time. It shows that if a particular function is known from the beginning to the end of time, a corresponding time-invarient* frequency spectrum can be found. To say that in a practical situation we do not have sufficient time to examine a signal over all eternity is an understatement. Just as it was possible to limit the bandwidth used in the Inverse transform, we must find a physical basis for limiting the time duration over which a function must be known in Fourier Analysis, if this technique is to be of use in practical situations.

Let us examine this problem in the light of a particular case. It is desired to find the effect of passing a square pulse through a "black box" with a particular frequency characteristic. To do this, the Fourier Transform of the input is obtained, multiplied by the frequency characteristic of the "black box" and the inverse transform of the result gives the resulting output time function. Strictly speaking, the answer obtained in this way is only applicable if the square pulse is the only input to which the device has ever been subjected, and if no future inputs ever occur. Practically, we know that if the pulse under consideration comes sufficiently long after any preceding input and sufficiently precedes any succeeding input, the result obtained will be valid and independent of any of these other signals. This result can be explained on the basis that any real device has a finite "memory" and the effect of any inputs which have long preceded the instant at which the output is under consideration, will have a negligible effect. Physically, this memory can be associated with such facts as that charges on condensers eventially leak out completely, fields around inductors drop back to zero, etc. The above discussion leads to the conclusion that in order to obtain the output at a particular time resulting from an input and a transmission characteristic, it is only necessary to examine the input time function around the time of interest over a period which is equal or longer than the "memory" of the system (which may include human observers

- -

*The phrase "time-invariant" is ambiguous. It could mean that at any time we connect a frequency analyzer to the circuit we would observe the spectrum regardless of the time of connection. This is not what is meant here. Rather, we mean that this spectrum is entirely out of the time domain. It is the spectrum information which would be collected by an analyzer of infinite memory which was connected to the circuit from the beginning of time to the end of time whose record we are forbidden to examine until time equals plus infinity.

as components) involved. The input may be considered as zero at all other times since the result is essentially independent of what goes on at these times.

A mathematical treatment to discover the nature of the "memory" will clarify that concept. In mathematics texts* dealing with Fourier transforms, the convolution theorem is proved. This states that if

$$\mathfrak{F}^{-1}\,[Y(\omega)] = y(t)$$

$$\mathfrak{F}^{-1}\,[G_i(\omega)] = g_i(t)$$

$$\text{then} \quad g_o(t) = \mathfrak{F}^{-1}\,[Y(\omega)\cdot G_i(\omega)] = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} g_i(\tau)y(t-\tau)d\tau \qquad (21\text{-}67)$$

If, in the above, $G_i(\omega)$ is interpreted as the spectrum of an input signal and $Y(\omega)$ as a frequency characteristic of a circuit, then the right hand side of Eq. 21-67 gives the output as a function of time since $Y(\omega)\cdot G(\omega)$ is the familiar expression of the output in the frequency domain. Note that τ is the variable of integration and will not appear in our final result. From what was said on Page 21-12, $y(t)$ is the impulse response of the circuit. In order to give physical meaning to the terms in the integral let us consider $g_i(t)$, the input, as composed of a large number of impulses spaced Δt apart. (By making Δt sufficiently small, the actual function can be approximated with any desired degree of accuracy).

$$g_o(t) = \sum_{n=-\infty}^{\infty} g_i(n\Delta t)\ \delta\ (t-n\ \Delta t)$$

Substituting, in 21-67, this gives for the output

$$g_o(t) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} y(t-\tau) \sum_{n=-\infty}^{\infty} g_i(n\Delta t)\delta(\tau-n\Delta t)d\tau$$

---

*See for example, Page 528 of Guillemin - The Mathematics of Circuit Analysis, Technology Press, Wiley, 1949.

$$= \sum_{n=-\infty}^{\infty} g_i(n\Delta t)\, \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} y(t-\tau)\delta(\tau-n\Delta t)\,d\tau$$

Since $\delta(x)$ = zero for all values of x except x = 0, this equals

$$g_1(t) = \sum_{n=-\infty}^{\infty} g_i(n\Delta t)\, \frac{y(t-n\Delta t)}{\sqrt{2\pi}} \qquad (21\text{-}68)$$

The output at time t is, therefore, made up of the weighted sum of all previous inputs [$y(x) = 0$ for $x < 0$ since it is the impulse response of a network and no real network will have an output preceding an input].

$$\frac{y(t-n\Delta t)}{\sqrt{2\pi}}$$

is called the weighting function and usually it will get smaller as its argument increases, i.e., as the impulse being added comes earlier and earlier relative to the moment in which the output is desired. In the case of RC networks, for instance, it will be of the form $Ae^{-a(t-\tau)}$ where $1/a$ is the usual time constant. From this it can be seen that any input which precedes the moment at which the output is desired by more than a few time constants can be neglected with negligible error.

The concept of "memory" which has been developed in the preceding paragraphs leads to an approach that can be very useful in certain problems. An example will again be used to introduce the important ideas. Assume that the input to a particular device is a square wave from $t = -\infty$ to $t = 0$ and a sawtooth from $t = 0$ to $t = \infty$. It is possible to obtain the Fourier transform for this function by applying Eq. 21-26. The result would be a time invariant frequency spectrum which could be multiplied by the frequency characteristic of the device and subjected to the Inverse transform to get an output function of time which would be valid from $t = -\infty$ to $t = +\infty$. Another approach would be to take two transforms, one for a square wave extending from $t = -\infty$ to $t = +\infty$, the other for the sawtooth - over the same limits. Each of these spectra are now multiplied by the frequency characteristic of the device and the Inverse transform obtained. We can take the result for

the square wave end and apply it from $t = -\infty$ to $t = 0$ and result for the sawtooth from $t = 0$ to $t = \infty$. The combined result would agree with the result obtained by the first method using the single time invariant spectrum except in the region around $t = 0$. In this region, the second method will give an erroneous result since the device will still "remember" the square wave while being fed the sawtooth. A valid result around $t = 0$ can be obtained by finding a third spectrum in the region around the origin extending for a time of the same duration as the "memory". To sum up - we can represent the input either as a single time invariant spectrum or by three sequentially applicable spectra. The three time variant spectra will all be different from each other and from the time invariant spectrum. This example suggests that it is possible to approximate the output corresponding to a particular input by dividing the input duration into periods equal to, or longer than, the "memory" of the system through which they are fed. Each of these inputs can be treated as if it were a regular Fourier spectrum and the corresponding output associated with the proper time period. This approach has physical meaning since this is precisely the way the device "sees" the incoming signals. It is especially useful if the time variant spectra in the time region of interest are more easily obtained then the regular Fourier spectrum.*

To summarize the results of this section we can say that, in practice, we modify the mathematician's Fourier transform and integral as given on Page 21-9. This modification is usually implicit, but is non-the-less a fact. Rather than using the infinite limits given in these equations, we use limits which depend upon the characteristic of the physical components with which we are dealing. In the Fourier integral, instead of Equation 21-27, the finite bandwidth of any real system is taken into account and we can write

$$f(t) = \int_{-\omega_2}^{-\omega_1} g(\omega)e^{j\omega t}\, d\omega + \int_{\omega_1}^{\omega_2} g(\omega)e^{j\omega t}\, d\omega = \mathfrak{F}^{-1}[g(\omega)]$$

---

*The difference between time varying spectra and the regular Fourier spectrum is exemplified by the difference between the instantaneous frequency and the spectrum of an FM wave, i.e., the difference between $\omega_c + \varphi'(t)$ and the expressions involving Bessel Functions - developed in the last chapter. In fact, in the case of low index systems where instantaneous frequency does not change radically in a memory period, the instantaneous frequency can be used in approximate calculations to estimate the output resulting from a given smooth transmission characteristic.

In Fourier analysis, the interest is always in the behavior around some time t. We take into account the finite memory of all real devices and write

$$g(\omega) = \frac{1}{2\pi} \int_{t-\Delta t}^{t+\Delta t} f(t)e^{j\omega t}\, dt = \mathfrak{F}[f(t)]$$

where Δt is chosen large compared to the interval of interest and the time constants of the components involved. In effect, we treat the situation as if no other input has ever or will ever occur.

The idea of the "memory" of components in turn suggests the idea that it is possible to carry out analysis on the basis, not of one time invariant spectrum, but by using a set of different spectra, each of which applies only during a particular time period. This approach is only meaningful when the periods over which each of the time invariant spectra applied are chosen at least as long as the "memory" of the system over which the signal is to be transmitted.

APPENDIX

A Note: Negative Frequencies

It is with (21-14) that we introduce the concept of negative frequencies. This mathematical fiction comes about because we want the eventual expression we are working toward to have the easily manipulated exponential form - yet we are dealing with real functions of time. (The concept of imaginary or complex functions of time has no physical meaning - such functions do not exist in nature or in the laboratory. It might be argued that negative frequencies don't exist either - but we can handle them mathematically, whereas we cannot write equations which are real functions of time on one side and imaginary on the other.)

How does (21-14) introduce negative frequencies? It is because the second term, $e^{-j\omega t}$, can be written $e^{j(-\omega)t}$. This may sound unconvincing, thus stated, but future developments will make it more convincing. As we go on, and simplify (21-15) and then examine the second equation which makes up the Fourier Transform Pair, we find ourselves integrating from $\omega = -\infty$ to $\omega = +\infty$, and talking about $g(-\omega)$. By this time we are certainly talking about negative frequencies - and if we look back to see where they crept in, Equation 21-14 turns out to be the place.

It is obvious that the inclusion of the second term of 21-14 is necessary if the right-hand side is to be real, as cos ωt is. Recall merely that

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

$$e^{-j\omega t} = \cos \omega t - j \sin \omega t$$

The sum of these is obviously a real function of time, 2 cos ωt. The imaginary components disappear. If we tried to work with $e^{j\omega t}$ only, we would have the imaginary function of time j sin ωt implicit in our basic equations. This would make it difficult to work with real functions of time, which are the only kind we have.

One further point: another way of writing cos ωt is

$$R\,[e^{j\omega t}]$$

meaning "real part of the bracketed quantity". Often we get lazy, and omit the "R" and the brackets - which is all right if we remember in time to throw out the imaginary functions of time that arise, and just keep the real ones. This form of representation must be used with care - one can get wrong answers by falling into the trap of multiplying two imaginary (and non-existent) functions of time, thus producing a spurious real function in the final result.

Chapter 22

EFFECT OF TRANSMISSION DEVIATIONS IN PM AND FM SYSTEMS

Transmission deviations in PM and FM systems introduce intermodulation products which appear at baseband and which can be controlled only by equalization ahead of the demodulator. Two methods for analyzing the effects of transmission deviations are presented. Method 1 can be used for either low or high index modulation systems so long as the transmission-frequency characteristic is relatively smooth. Method 2 is limited to low index systems but applies to both smooth and irregular transmission-frequency characteristics.

Introduction

In Chapter 20 it was shown that the addition of thermal noise to an FM wave produced both amplitude and phase modulation of the wave. Transmission deviations in the transmission path also produce both amplitude and phase modulation. Here the term transmission deviation refers to any irregularity in transmission characteristics such that the gain and delay are not the same for all frequency components in the FM wave.

A Qualitative Approach to Transmission Deviations

A qualitative idea of the effect of transmission deviations can be obtained from the following considerations. Assume that an FM wave with many sideband components is applied to a network which has perfect transmission for the entire signal except at the frequency of one of the sideband components such that the amplitude of this particular sideband component is slightly altered. This is equivalent to adding to the applied signal a small extraneous signal at the frequency of this particular component. Hence, the output signal from the imperfect network may be thought of as consisting of the input wave plus a small extraneous signal. Although the source of the extraneous signal is much different from the source of the random noise components, the effect is much the same. Both amplitude and phase modulation of the original wave will occur.

Consider briefly two particular cases which serve to illustrate the above principles. In the first case the modulating wave is a single sinusoid. Under this condition the frequency spectrum of the FM wave consists of a carrier and a number of sideband components separated from the carrier by multiples of the modulating frequency. Transmission deviations at any of these frequencies will distort the signal being transmitted. For example, if a first order sideband component is altered, this is equivalent to adding an extraneous sinusoidal

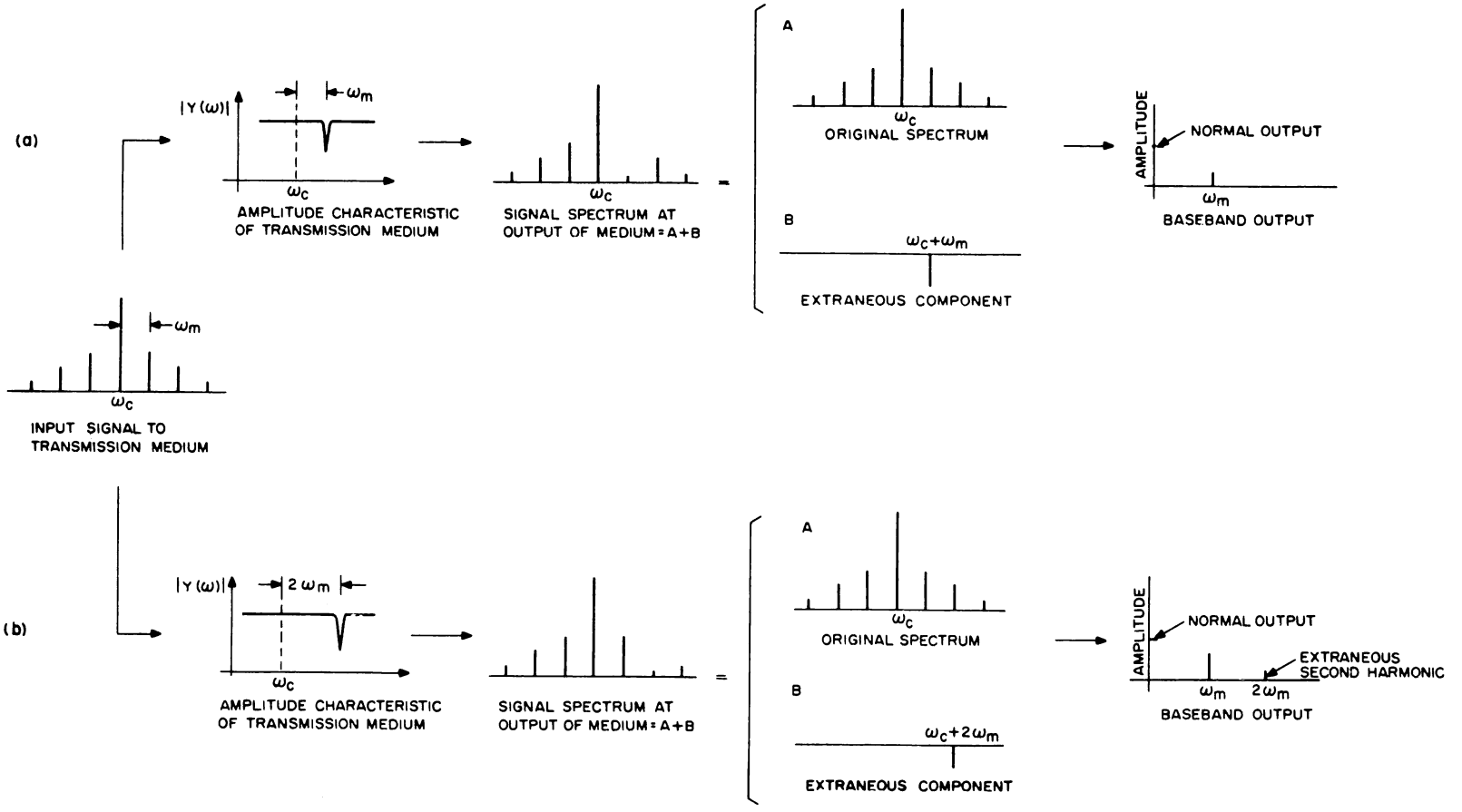
signal at the frequency of this component, as shown in Figure 1a. If the transmission deviation is small the principle effect will be a small undesired phase modulation at a frequency equal to the frequency difference between the carrier and the sideband component which has been altered. In this case the difference frequency would be equal to the modulating frequency, and the sinusoidal baseband output would merely be altered in level and phase. On the other hand, if the transmission deviation were such as to alter a second order sideband component the baseband output will contain an unwanted second harmonic of the modulating frequency. This is shown in Figure 1b.

As a second illustrative case, let the modulating signal consist of two sinusoids. The frequency spectrum of an FM wave when the baseband signal is two sinusoidal tones can be obtained from Equation (19-28). This equation shows that when the two modulating tones have radian frequencies ω_v and ω_w there are sideband components at each of the frequencies $\omega_c + n\omega_v + m\omega_w$. For illustrative purposes consider the case when $n = 1$ and $m = -1$. We then are concerned with a frequency component at $\omega_c + \omega_v - \omega_w$. If the amplitude or phase of this component of the FM wave is altered during transmission this is equivalent to some unwanted phase modulation at a frequency $\omega_v - \omega_w$. At the output of the system, therefore, the two original baseband components would be present as well as an unwanted component at the difference frequency between the two components.

These examples illustrate that transmission deviations in an FM system can introduce frequency components at the output of a system which did not exist at the input. In this sense, then, transmission deviations in an FM system have an effect similar to electron tube nonlinearity in an AM system. It can be shown that these modulation products are a function of the modulation index much as the modulation products in a vacuum tube are a function of signal level. With a given transmission deviation the ratio of the signal to the modulation noise decreases as the index of modulation is increased. After demodulation, equalizers cannot be used to eliminate these modulation products, and it is therefore necessary to equalize the system ahead of the demodulator if the modulation products are to be reduced.

A More Quantitative Approach

Transmission deviations in an FM system produce amplitude and phase modulation of the FM wave which are definite functions of the transmitted signal and the transmission deviations. There are several



Simplified Illustration of Baseband Distortion Caused By A Transmission Deviation

Figure 22-1

methods by which this distortion can be expressed. Two of these are considered in this chapter. The nature of the signal and the transmission deviations determine which of these methods is the more useful in any particular situation.

In the first approach, designated Method 1, it is assumed that the transmission-frequency characteristic is sufficiently smooth that the gain and phase characteristics can be adequately represented by a summation of linear, parabolic, and cubic shape components. Such components are illustrated in Figure 2. If this requirement holds, and, incidently, it does in many of the problems encountered, a series expansion involved in the analysis converges sufficiently rapidly to restrict the number of terms which have to be considered to just those related to the shapes defined. Additional simplifications can then be made if the transmission deviations are small, as is ordinarily the case, but this is not a necessary restriction. For the case of small deviations, however, a simple breakdown of amplitude, phase, and frequency modulation terms arising from each of the component gain and phase shapes can be made, as tabulated in Table 1. The resulting baseband distortion terms are tabulated in Table 2. It is important to emphasize that this method applies so long as the transmission shape is relatively smooth; no restrictions are placed on the index of modulation of the signals.

Method 2, on the other hand, is restricted to problems involving low index systems because of a series expansion made on the modulating function. The method can be used, however, to handle both smooth and irregular (discontinuous) transmission-frequency characteristics. In Method 2 the transmission-frequency characteristic is expressed in terms of symmetric and skew-symmetric components, as illustrated by Figure 3 and discussed in the associated section of the text.

It follows that one type of problem still remains to be handled. This is the case of an irregular transmission-frequency characteristic and a large index of modulation. Various means of handling specific problems have been devised. Method 3 is an example of a technique which is applicable to both high and low index systems. It gives the modulation noise produced by echos in the top baseband channel of an FM system. Since echos produced by impedance mismatches are among the most common sources of transmission deviations, this method is an extremely useful tool. Furthermore, approximate results for deviations due to other causes can also be obtained.

Method 1 - The Amplitude and Phase Modulation Produced by Particular Transmission Deviations

In this section only particular types of transmission deviations will be considered. These will be linear, parabolic, and cubic gain shapes; and parabolic and cubic phase shapes. Linear phase shape is omitted since it is equivalent to constant delay and does not introduce distortion. The input signal will be taken as

$$e_1(t) = A_c \cos\left[\omega_c t + \varphi(t)\right] \tag{22-1}$$

and the transmission characteristic as

$$Y(\omega) = \left[1+g_1(\omega-\omega_c)+g_2(\omega-\omega_c)^2+g_3(\omega-\omega_c)^3\right]\varepsilon^{j[b_2(\omega-\omega_c)^2+b_3(\omega-\omega_c)^3]} \tag{22-2}$$

where

$e_1(t)$ = applied FM or PM wave
A_c = constant amplitude of FM or PM wave
ω_c = carrier frequency in radians per second
$\varphi(t)$ = angle or phase modulation in radians
$Y(\omega)$ = transmission characteristic which is function of radian frequency ω.

Here g_1, g_2 and g_3 are constants which determine, respectively, the amount of linear, parabolic and cubic gain shape. Note that this is not a functional notation as $g(\omega)$ was in the discussion of the Fourier Transform. Similarly b_2 and b_3 are constants which, respectively, determine the amount of parabolic and cubic phase shape.

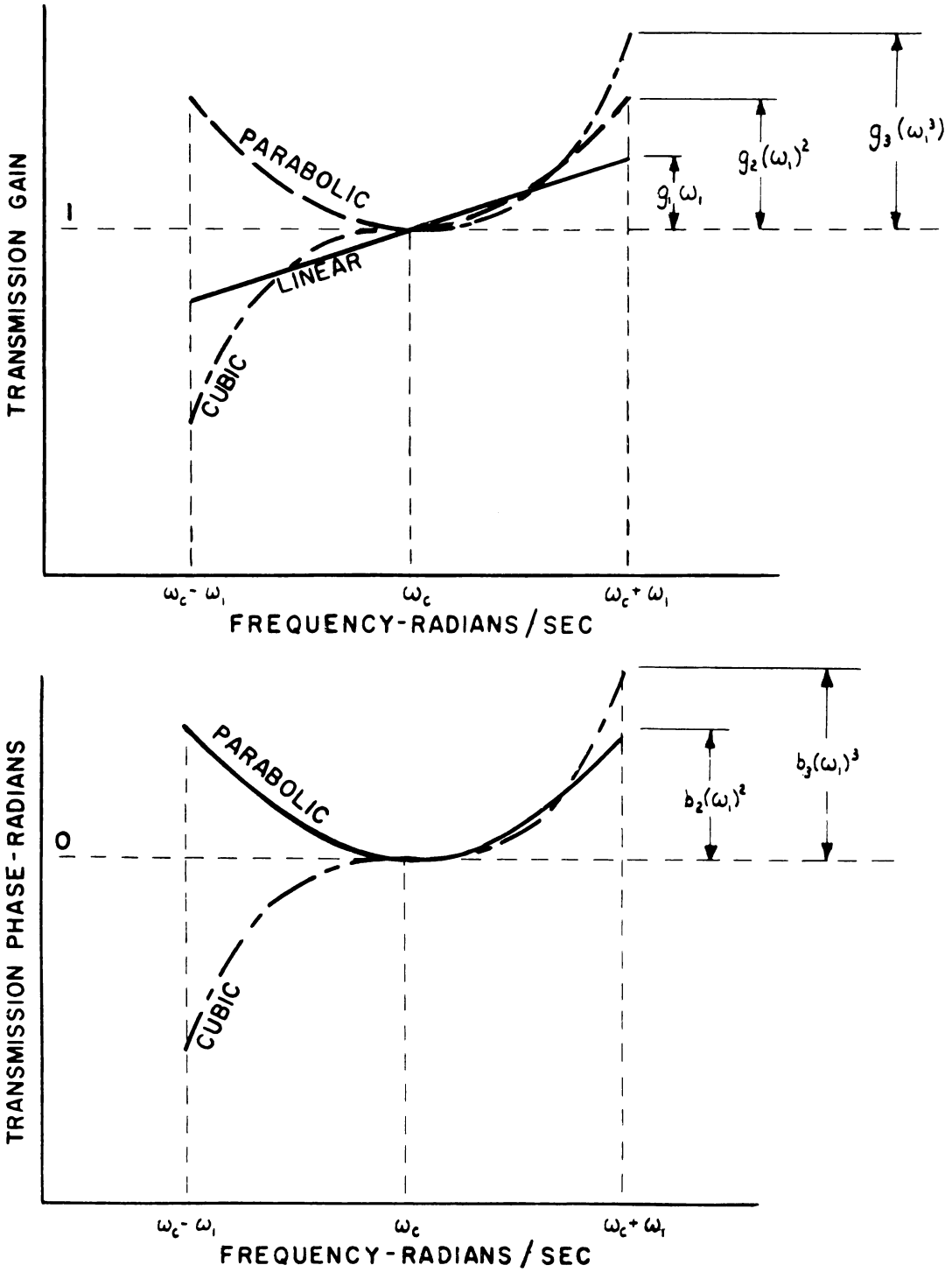
The transmission characteristic is normalized with respect to the carrier frequency such that the transmission at the carrier frequency is unity. Typical transmission characteristics are shown in Figure 2.

Equation (22-1) can also be written in exponential notation as

$$e_1(t) = \left[\begin{matrix}\text{Real}\\ \text{Part}\\ \text{of}\end{matrix}\right] A_c \varepsilon^{j[\omega_c t+\varphi(t)]} \tag{22-3}$$

At this point the bracket indicating that only the real part of the expression should be retained will be dropped in order to simplify the expressions which must be written. It will be re-inserted later in the derivation. This gives

$$e_1(t) = A_c \varepsilon^{j\omega_c t}\, \varepsilon^{j\varphi(t)} \tag{22-4}$$



Particular Transmission Gain and Phase Shapes

Figure 22-2

The spectrum, $G_1(\omega)$, of the input FM wave is given by the direct Fourier transform of $e_1(t)$ as

$$G_1(\omega) = \frac{1}{2\pi}\int_{-\infty}^{\infty} e_1(t)\ \varepsilon^{-j\omega t}\ dt$$

$$= \frac{Ac}{2\pi}\int_{-\infty}^{\infty} \varepsilon^{j\omega_c t}\ \varepsilon^{j\varphi(t)}\ \varepsilon^{-j\omega t}\ dt \qquad (22\text{-}5)$$

This, of course, is one of the equations of the familiar Fourier transform pair. It will be recalled that if the spectrum of the signal were known and we wished to find the time-function, we could use the inverse transform - thus, for the signal we are discussing

$$e_1(t) = A_c\varepsilon^{\ j\omega_c t}\ \varepsilon^{j\varphi(t)} = \int_{-\infty}^{\infty} G_1(\omega)\varepsilon^{j\omega t}\ d\omega \qquad (22\text{-}6)$$

It will be convenient to express equations of this type in the more concise notation:

$$G_1(\omega) = \mathfrak{F}\ [e_1(t)] \qquad (22\text{-}7)$$

$$e_1(t) = \mathfrak{F}^{-1}[G_1(\omega)] \qquad (22\text{-}8)$$

where $\mathfrak{F}[\ ]$ and $\mathfrak{F}^{-1}[\ ]$ denote respectively the direct and inverse Fourier transforms of the bracketed quantities.

If the transmission characteristic of the transmission path to which the input signal is applied is $Y(\omega)$, the spectrum, $G_2(\omega)$, of the output signal is equal to

$$G_2(\omega) = Y(\omega)\ G_1(\omega) \qquad (22\text{-}9)$$

Thus every frequency component of the input signal is multiplied by the transmission at that frequency to obtain the output component at that frequency. Substitution of Equation (22-7) in (22-9) gives

$$G_2(\omega) = Y(\omega)\ \mathfrak{F}[e_1(t)] \qquad (22\text{-}10)$$

The output signal, $e_2(t)$, is given by the inverse Fourier transform of the output spectrum $G_2(\omega)$ as

$$e_2(t) = \mathfrak{F}^{-1}\ [G_2(\omega)] \qquad (22\text{-}11)$$

Substitution of Equation (22-10) into (22-11) gives

$$e_2(t) = \mathfrak{F}^{-1}\ \left[Y(\omega)\ \mathfrak{F}[e_1(t)]\right] \qquad (22\text{-}12)$$

$$e_2(t) = \mathfrak{F}^{-1}\left[Y(\omega)\ \mathfrak{F}[A_c\varepsilon^{j\omega_c t}\varepsilon^{j\varphi(t)}]\right] \qquad (22\text{-}13)$$

An alternative expression for the output signal can be obtained by substitution of Equation (22-9) into (22-11)

$$e_2(t) = \mathfrak{F}^{-1}\left[Y(\omega)G_1(\omega)\right] = \int_{-\infty}^{\infty} Y(\omega)G_1(\omega)\ \varepsilon^{j\omega t}\,d\omega \qquad (22\text{-}14)$$

In this equation ω is simply 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 (22-14) may then be written as

$$e_2(t) = \int_{-\infty}^{\infty} Y(\omega+\omega_c)\ G_1(\omega+\omega_c)\ \varepsilon^{j(\omega+\omega_c)t}\,d\omega \qquad (22\text{-}15)$$

$$e_2(t) = \varepsilon^{j\omega_c t}\int_{-\infty}^{\infty} Y(\omega+\omega_c)\ G_1(\omega+\omega_c)\varepsilon^{j\omega t}\,d\omega \qquad (22\text{-}16)$$

In shorter form this becomes

$$e_2(t) = \varepsilon^{j\omega_c t}\ \mathfrak{F}^{-1}\left[Y(\omega+\omega_c)\ G_1(\omega+\omega_c)\right] \qquad (22\text{-}17)$$

From Equation (22-5) the expression for $G_1(\omega+\omega_c)$ can be obtained by replacing ω by $\omega+\omega_c$.

$$G_1(\omega+\omega_c) = \frac{Ac}{2\pi}\int_{-\infty}^{\infty} \varepsilon^{j\omega_c t}\ \varepsilon^{j\varphi(t)}\ \varepsilon^{-j(\omega+\omega_c)t}\,dt$$

$$= \frac{Ac}{2\pi}\int_{-\infty}^{\infty} \varepsilon^{j\varphi(t)}\ e^{-j\omega t}\,dt$$

$$= \mathfrak{F}[A_c\varepsilon^{\ j\varphi(t)}] \qquad (22\text{-}18)$$

Substitution of Equation (22-18) into (22-17) gives

$$e_2(t) = A_c\varepsilon^{j\omega_c t}\ \mathfrak{F}^{-1}\left[Y(\omega+\omega_c)\ \mathfrak{F}[\varepsilon^{j\varphi(t)}]\right] \qquad (22\text{-}19)$$

This is the desired result. It shows that the effect of a transmission characteristic $Y(\omega)$ on an FM wave can be expressed in terms of the effect of a transmission characteristic $Y(\omega+\omega_c)$ on the modulation term, $\varepsilon^{j\varphi(t)}$, of the FM wave. This is a logical result. This modulation

term is really the same as the actual FM wave 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 $Y(\omega+\omega_c)$.

The normalized transmission characteristic given in Equation (22-2) will now be considered.

$$Y(\omega) = \left[1 + g_1(\omega-\omega_c)+g_2(\omega-\omega_c)^2 +g_3(\omega-\omega_c)^3\right] \varepsilon^{j[b_2(\omega-\omega_c)^2+b_3(\omega-\omega_c)^3]} \tag{22-2}$$

Here g_1, g_2, g_3, b_2, and b_3, are real constants. From Equation (22-2), $Y(\omega+\omega_c)$ becomes

$$Y(\omega+\omega_c) = [1 + g_1\omega + g_2\omega^2 + g_3\omega^3] \varepsilon^{j[b_2\omega^2+b_3\omega^3]} \tag{22-20}$$

If the exponential is expanded using $e^x=1+x+\frac{x^2}{2!}+\frac{x^3}{3!}\ldots$, and only the terms of less than the fourth power of ω are retained, this gives

$$Y(\omega+\omega_c) = [1 + g_1\omega + g_2\omega^2 + g_3\omega^3] [1 + jb_2\omega^2 + jb_3\omega^3]$$

$$= 1 + g_1\omega + (g_2 + jb_2)\omega^2 + (g_3 + jb_3 + jg_1b_2)\omega^3,.. \tag{22-21}$$

At this point this result can be substituted into Equation (22-19). However, we will first write equation (22-21) in terms of the operator, $p = j\omega$.

$$Y(\omega+\omega_c) = 1-jg_1p - (g_2+jb_2)\,p^2 + j(g_3+jb_3+jg_1b_2)p^3 \tag{22-22}$$

From Fourier transform theory we know that multiplication by p in the frequency domain is equivalent to differentiation in the time domain. For example,

$$\mathfrak{F}^{-1}\left[p\,\mathfrak{F}[f(t)]\right] = \frac{d}{dt}[f(t)] = f'(t), \tag{22-23}$$

where f(t) is any function of time. Similarly,

$$\mathfrak{F}^{-1}[p^2\,\mathfrak{F}[f(t)] = \frac{d^2}{dt^2}[f(t)] = f''(t) \tag{22-24}$$

When Equation (22-22) is substituted into Equation (22-19) the result may be written (using φ to represent $\varphi(t)$ for the moment)

$$e_2(t) = A_c\varepsilon^{j\omega_c t}\left[1-jg_1\frac{d}{dt}-(g_2+jb_2)\frac{d^2}{dt^2}+j(g_3+jb_3+jg_1b_2)\frac{d^3}{dt^3}\right]\varepsilon^{j\varphi}$$

$$= A_c\varepsilon^{j\omega_c t}\left\{e^{j\varphi}-jg_1[j\varphi']\varepsilon^{j\varphi}-(g_2+jb_2)[(j\varphi')^2+j\varphi'']\varepsilon^{j\varphi}\right.$$

$$\left.+j(g_3+jb_3+jg_1b_2)\left[-3\varphi'\varphi''+j\varphi'''-j\varphi'^3\right]\varepsilon^{j\varphi}\right\} \qquad (22\text{-}25)$$

The terms may be collected to give

$$e_2(t) = A_c\varepsilon^{j[\omega_c t+\varphi(t)]}[1+P(t)+jQ(t)] \qquad (22\text{-}26)$$

where

$$P(t) = g_1\varphi'(t)+g_2\varphi'^2(t)+b_2\varphi''(t)+3(b_3+g_1b_2)\varphi'(t)\varphi''(t)$$

$$+g_3[\varphi'^3(t)-\varphi'''(t)] \qquad (22\text{-}27)$$

and

$$Q(t)= -g_2\varphi''(t)+b_2\varphi'^2(t)+(-b_3-g_1b_2)[\varphi'''(t)-\varphi'^3(t)]$$

$$-3g_3\varphi'(t)\varphi''(t) \qquad (22\text{-}28)$$

The final result is obtained by taking the real part of the expression in Equation (22-26).

$$e_2(t) = A_c\left\{[1+P(t)]\cos[\omega_c t+\varphi(t)]-Q(t)\sin[\omega_c t+\varphi(t)]\right\} \qquad (22\text{-}29)$$

Equation (22-29) can be written as the amplitude and phase modulation of the original input wave.

$$e_2(t) = A_c\sqrt{[1+P(t)]^2+[Q(t)]^2}\cos[\omega_c t+\varphi(t)+\theta(t)] \qquad (22\text{-}30)$$

where

$$\theta(t) = \arctan\frac{Q(t)}{1+P(t)} \qquad (22\text{-}31)$$

When $P(t) \ll 1$ and $Q(t) \ll 1$, Equation (22-30) can be written approximately as

$$e_2(t) = A_c[1+P(t)]\cos[\omega_c t+\varphi(t)+Q(t)] \qquad (22\text{-}32)$$

Therefore, the amplitude modulation is given approximately by P(t) and the phase modulation is given approximately by Q(t) when the transmission deviations are small. In practical systems this is ordinarily the case,

since large deviations would tend to make the system unstable. In Table 22-1 the various amplitude and phase modulation terms for the case where P(t) and Q(t) are small compared to unity are separated from Equation (22-27) and (22-28) according to the type of transmission deviation which caused the term. The resulting frequency modulation term is, of course, the derivative of the phase modulation term.

Baseband Distortion

If an ideal demodulator is used in either an FM or a PM system the amplitude modulation of the signal may be neglected. In many practical systems a limiter is inserted before the demodulator to suppress the amplitude modulation. The unwanted phase and frequency modulation due to transmission deviations, however, cannot be separated (except by equalization before demodulation) from the desired modulation and are demodulated along with the desired modulation. These distortion terms may be written in terms of the modulating signal V(t) by means of Equations (19-8) and (19-9) which are $\varphi(t) = kV(t)$ for PM systems and $\varphi'(t)=k_1V(t)$ for FM systems. By means of these substitutions the output distortion may be expressed in terms of V(t). This is done in Table 22-2 such that all of the distortion terms are with respect to V(t). To illustrate the procedure used to obtain Table 22-2, consider the phase modulation term which results from a cubic gain transmission deviation. Table 22-1 shows this deviation term to be

$$- 3g_3 \, \phi'(t) \, \phi''(t)$$

Substituting $\phi(t) = kV(t)$, this term becomes

$$- 3g_3 \, kV'(t)[kV''(t)]$$

Since the desired modulation term is $\phi(t)$, the ratio of undesired to desired modulation terms is

$$\frac{-3g_3 \, kV'(t)[kV''(t)]}{kV(t)} = \frac{-3g_3 \, kV'(t) \, V''(t)}{V(t)}$$

Therefore, with respect to the modulating signal V(t), the output distortion arising from a cubic gain transmission deviation is $-3g_3 \, kV'(t) \, V''(t)$, which is the first term entered in the second column of Table 22-2. All of the other terms in this table can be obtained from Table 22-1 in a similar manner.

TABLE 22-1

AMPLITUDE, PHASE, & FREQUENCY MODULATION CAUSED BY PARTICULAR TRANSMISSION DEVIATIONS

TYPE OF TRANSMISSION DEVIATION	RESULTING AMPLITUDE MODULATION = P(t)	RESULTING PHASE MODULATION = Q(t)	RESULTING FREQUENCY MODULATION = Q'(t)
LINEAR GAIN	$g_1 \phi'(t)$		
PARABOLIC GAIN	$g_2 \phi'^2(t)$	$-g_2 \phi''(t)$	$-g_2 \phi'''(t)$
CUBIC GAIN	$g_3 [\phi'^3(t) - \phi'''(t)]$	$-3g_3 \phi'(t)\phi''(t)$	$-3g_3 [\phi'(t)\phi'''(t) + \phi''^2(t)]$
PARABOLIC PHASE	$b_2 \phi''(t)$	$b_2 \phi'^2(t)$	$2b_2 \phi'(t)\phi''(t)$
CUBIC PHASE	$3b_3 \phi'(t)\phi''(t)$	$-b_3 [\phi'''(t) - \phi'^3(t)]$	$-b_3[\phi''''(t)-3\phi'^2(t)\phi''(t)]$
LINEAR GAIN PLUS PARABOLIC PHASE (See Note 1)	$3g_1b_2 \phi'(t)\phi''(t)$	$-g_1b_2 [\phi'''(t) - \phi'^3(t)]$	$-g_1b_2[\phi''''(t)-3\phi'^2(t)\phi''(t)]$

NOTE 1 - This is an interaction term which is in addition to separate distortions produced by linear gain and parabolic phase.

INPUT SIGNAL = $A_c \cos[\omega_c t + \phi(t)]$ OUTPUT SIGNAL = $A_c [1+P(t)] \cos[\omega_c t+\phi(t)+Q(t)]$

NORMALIZED TRANSMISSION = $[1+g_1(\omega-\omega_c) + g_2(\omega-\omega_c)^2 + g_3(\omega-\omega_c)^3]\varepsilon^{j[b_2(\omega-\omega_c)^2 + b_3(\omega-\omega_c)^3]}$

Table 22-1 Amplitude, Phase, & Frequency Modulation Caused by Particular Transmission Deviations

TABLE 22-2

DISTORTION WITH RESPECT TO THE MODULATING SIGNAL V(t)

TYPE OF TRANSMISSION DEVIATION	PM SYSTEM		FM SYSTEM	
	EQUALIZABLE*	NOT EQUALIZABLE*	EQUALIZABLE*	NOT EQUALIZABLE*
LINEAR GAIN				
PARABOLIC GAIN	$-g_2\, V''(t)$		$-g_2\, V''(t)$	
CUBIC GAIN		$-3g_3\, kV'(t)\, V''(t)$		$-3g_3k_1[V(t)V''(t) + V'^2(t)]$
PARABOLIC PHASE		$b_2\, k\, V'^2(t)$		$2b_2\, k_1\, V(t)\, V'(t)$
CUBIC PHASE	$-b_3\, V'''(t)$	$b_3\, k^2\, V'^3(t)$	$-b_3\, V'''(t)$	$3b_3k_1^2\, V^2(t)\, V'(t)$
LINEAR GAIN PLUS PARABOLIC PHASE (See Note in Table 22-1)	$-g_1b_2\, V'''(t)$	$g_1b_2\, k^2\, V'^3(t)$	$-g_1b_2\, V'''(t)$	$3g_1b_2\, k_1^2\, V^2(t)\, V'(t)$
	PM SIGNAL = $A_c \cos [\omega_c t + kV(t)]$		FM SIGNAL = $A_c \cos [\omega_c t + k_1 \int^t V(t)dt]$	

*EQUALIZABLE MEANS THESE DISTORTION TERMS ARE EQUALIZABLE AFTER DEMODULATION

Table 22-2
Distortion, With Respect to Signal Caused by Particular Transmission Deviations

The terms in Table 22-2 are separated into those that can be corrected by equalization after demodulation and those that cannot. The terms which cannot be equalized after demodulation are modulation products. This illustrates an important difference between amplitude and angle modulated systems which was previously mentioned. We shall repeat it here. In angle modulated systems the transmission deviations can cause modulation products and thus introduce distortion at frequencies at which there is no signal. In this sense then, transmission deviations in an angle modulated system produce distortion similar to that produced by vacuum tube nonlinearities. Since this sort of distortion cannot be equalized after demodulation, any equalization which is done must be ahead of demodulation if it is to reduce the modulation products in the output.

In Table 22-2 another important observation may be made. Those distortion terms which may be equalized are the same for both PM and FM systems and do not depend on the constants k and k_1. Since the index of modulation is proportional to k and k_1 it is therefore obvious that those terms which may be equalized are a function only of the transmission deviations and the signal to be transmitted. Only the terms which are not equalizable depend on the index of modulation. For example, second order modulation terms, with respect to V(t), are proportional to the index of modulation. The third order modulation terms, with respect to V(t), are proportional to the square of the index of modulation.

We might summarize Table 22-2 as follows if we omit interaction terms.

Transmission Shape	Order of Distortion
Linear gain (alone)	None
Parabolic gain	1st order
Cubic gain	2nd order
Parabolic phase (linear delay)	2nd order
Cubic phase (parabolic delay)	1st and 3rd order

The first order distortion terms are those that are equalizable such as $-g_2V''(t)$ and $-b_3V'''(t)$. Second order distortion terms would include $-3g_3kV'(t)V''(t)$ and $b_2kV'^2(t)$.

We can also note that linear and cubic transmission shapes have odd order symmetry about the carrier frequency and make the following general statements.

Even order gain and delay shapes cause odd order distortion.
Odd order gain and delay shapes cause even order distortion.
An obvious exception is linear gain which produces no distortion.

An Illustrative Example - Sinusoidal Modulation

We shall assume that the baseband signal consists of 1 mc and 4 mc sinusoidal waves.

$$V(t) = A_v \cos \omega_v t + A_w \cos \omega_w t \ , \qquad (22\text{-}33)$$

where

$\omega_v = 2\pi \times 10^6$ rad per sec

$\omega_w = 8\pi \times 10^6$ rad per sec

If this signal is applied to a phase modulator the output PM wave becomes

$$\text{PM signal} = A_c \cos [\omega_c t + k A_v \cos \omega_v t + k A_w \cos \omega_w t] \qquad (22\text{-}34)$$

We shall take $k = 1/2$, and $A_v = A_w =$ one volt so that the peak phase deviation is equal to 1 radian. We can now consider what happens when this wave is subjected to a particular transmission deviation. For illustrative purposes we shall consider only parabolic gain and parabolic phase. We shall take the gain to be down 1 db at a frequency 10 mc from the carrier and the phase error at this same frequency to be 0.1 rad. From this information we can solve for g_2 and b_2.

$$20 \log [1 + g_2 (2\pi \times 10^7)^2] = -1 \text{ db} = 20 \log .891$$

$$1 + g_2 (2\pi \times 10^7)^2 = .891$$

$$g_2 = - \frac{.109}{(2\pi \times 10^7)^2} \qquad (22\text{-}35)$$

$$b_2 (2\pi \times 10^7)^2 = .1 \text{ rad}$$

$$b_2 = \frac{.1}{(2\pi \times 10^7)^2} \qquad (22\text{-}36)$$

The output of the system can be obtained directly from Table 2.

$$\text{Output} = V(t) - g_2 V''(t) + b_2 k V'^2(t) \qquad (22\text{-}37)$$

From Equation (22-33) $V'(t)$ and $V''(t)$ become

$$V'(t) = - A_v\omega_v \sin \omega_v t - A_w\omega_w \sin \omega_s t$$

$$V''(t) = - A_v\omega_v^2 \cos \omega_v t - A_w\omega_w^2 \cos \omega_s t \qquad (22\text{-}38)$$

Substitution of Equation (22-33) and (22-38) into (22-37) gives

$$\begin{aligned}\text{Output} = {} & A_v \cos \omega_v t + A_w \cos \omega_w t \\ & + g_2 A_v\omega_v^2 \cos \omega_v t + g_2 A_w\omega_w^2 \cos \omega_w t \\ & + b_2k A_v^2\omega_v^2 \sin^2 \omega_v t + b_2k A_vA_w \omega_v \omega_w \sin \omega_v t \sin \omega_w t \\ & + b_2k A_w^2\omega_s^2 \sin^2 \omega_w t\end{aligned}$$

The final result becomes

$$\begin{aligned}\text{Output} = {} & .999 \cos \omega_v t + .983 \cos \omega_w t \\ & - .00025 \cos 2\omega_v t - .004 \cos 2\omega_w t \\ & + .002 \cos (\omega_v+\omega_w)t - .002 \cos (\omega_w-\omega_v)t \\ & + .00425\end{aligned}$$

Here we can see that the transmission deviation has changed the levels of the two sine waves which were transmitted. In addition second harmonic and sum and difference frequency terms have been introduced. An equalizer could be used to correct the levels of the transmitted sine waves but the rest of the terms would still remain.

Differential Gain and Phase

In discussing the transmission of NTSC color TV signals (Chapter 16), it was pointed that the transmission (gain or phase) of the color subcarrier might be affected by the presence of other components of the signal - e.g., by the instantaneous magnitude of the luminance components. If this occurs, we have the type of distortion called "differential gain" and/or "differential phase". The depth and hue of the color is then incorrectly reproduced at the receiver. In video and AM systems, this type of distortion is produced by intermodulation. We have seen that transmission deviations in angle modulated systems produce distortions of the signal similar to those produced by intermodulation in AM. Not surprisingly, then, we find that transmission deviations in angle-modulated systems produce differential gain and phase.

Consider now the effect of parabolic phase (linear delay) distortion in an FM system when the baseband signal consists of two sinusoids. Let ω_v represent a component of the luminance signal, and ω_w the color subcarrier. Thus ω_v will now be taken as the angular frequency corresponding to 15.75 kc/s, for example, and ω_w as the angular frequency corresponding to the 3.58 mc/s color subcarrier.

From Table 2, the output of an FM system which has parabolic phase distortion will be the applied signal plus a distortion term given by

$$D = 2b_2 k_1 V(t) V'(t)$$

In this case

$$V(t) = A_v \cos \omega_v t + A_w \cos \omega_w t \tag{22-39}$$

whence $$V'(t) = -A_v\omega_v \sin \omega_v t - A_w\omega_w \sin \omega_w t$$

Taking only the terms which give rise to components around ω_w, we have

$$D \doteq 2b_2 k_1 [(- A_v A_w \omega_v \sin \omega_v t) \cos \omega_w t - (A_v A_w \omega_w \cos \omega_v t) \sin \omega_w t]$$

The total output of interest is $A_w \cos \omega_w t + D$, or

$$A_w \{[1+P] \cos \omega_w t + Q \sin \omega_w t\}$$

where in this case

$$P = - 2b_2 k_1 A_v \omega_v \sin \omega_v t$$

$$Q = - 2b_2 k_1 A_v \omega_w \cos \omega_v t$$

Consider now the differential gain and phase. Since the distortion has been assumed small, the amplitude of the color sub-carrier output will be very closely given by the amplitude of the cos $\omega_w t$ term. In other words, it will be proportional to $A_w[1+P]$, and the ratio of this to the magnitude of $A_w \cos \omega_w t$ when the high level signal ω_v is absent will be simply $[1+P]$. The differential gain is therefore given by $20 \log_{10} [1+P]$.

In the case of differential phase, it will be clear that if P is small compared with unity the phase of the ω_w output is given by $\frac{Q}{1+P}$. Since the distortion is small, P may be neglected in comparison with unity so that the differential phase is given by $\Delta\varphi = Q$.

The same sort of analysis can be carried out for other types of deviations. We obtain (ignoring the interaction term arising from linear gain plus parabolic phase):

$$P = -k_1 A_v [2b_2\omega_v \sin \omega_v t - 3g_3(\omega_v^2 + \omega_w^2) \cos \omega_v t]$$
$$-k_1^2 A_v^2 [3b_3\omega_v \sin 2\omega_v t]$$

$$Q = -k_1 A_v [2b_2\omega_w \cos \omega_v t + 6g_3\omega_v\omega_w \sin \omega_v t]$$
$$-k_1^2 A_v^2 [\tfrac{3}{2} b_3\omega_w \cos 2\omega_v t] \qquad (22\text{-}40)$$

From these equations it follows that the differential gain ΔG is

$$\Delta G = 20 \log_{10}\{1 - k_1 A_v [2b_2\omega_v \sin \omega_v t - 3g_3(\omega_v^2+\omega_w^2) \cos \omega_v t]$$
$$-k_1^2 A_v^2 [3b_3\omega_v \sin 2\omega_v t]\} \text{ db} \qquad (22\text{-}41a)$$

and the differential phase $\Delta\varphi$ is

$$\Delta\varphi = \{-k_1 A_v [2b_2\omega_w \cos \omega_v t + 6g_3\, \omega_v\omega_w \sin \omega_v t]$$
$$-k_1^2 A_v^2 [\tfrac{3}{2} b_3\omega_w \cos 2\omega_v t]\} \text{ Radians} \qquad (22\text{-}41b)$$

Example

As an example, the effect of linear delay shape on differential phase will be computed in detail using magnitudes appropriate for the TJ system.

The first step is to determine the magnitude of $k_1 A_v$. Without pre-emphasis $k_1 A_v$ would correspond to 4 mc/s, this being the nominal peak frequency deviation for the TJ system. With pre-emphasis, the magnitude of the signal at ω_v is reduced by 14 db. (This assumes the same amount of pre-emphasis as is used on TD2.) When the pre-emphasis network is inserted, however, it is usual to increase the gain at the input to the system by about 4 db in order to re-establish the same peak frequency deviation. (The procedure involved here is similar to the 4.8 db re-adjustment of signal level that we went through on page 20-15.) The net result of this is that $k_1 A_v$ corresponds to 4 mc/s reduced by 10 db, i.e., $k_1 A_v = \frac{2\pi \times 4 \times 10^6}{3.16}$ radians/sec.

We now determine what value of b_2 results in 1 mμsec of delay at the edge of the FM spectrum. By definition the phase deviation due to $b_2\omega^2$ and the derivative of this gives the delay, i.e., delay = $2b_2\omega$. Using a value of ω corresponding to 10 mc/s and equating the delay to 1 mμsec gives

$$b_2 = \frac{10^{-9}}{2x2\pi x10^7} = .0796 \times 10^{-16}$$

The differential phase is given by:

$$\Delta\varphi = -2b_2 k_1 A_v \omega_w \cos \omega_w t$$

and the peak-to-peak value of this is

$$4b_2k_1A_v\omega_w = 4 \times .0796 \times 10^{-16} \times \frac{8\pi x10^6}{3.16} \times 2\pi \times 3.58 \times 10^6$$

$$= 56.8 \times 10^{-4} \text{ radians}$$

$$= 0.32 \text{ degrees}$$

Thus parabolic phase or linear delay results in 0.32 degrees of differential phase per mμsec of delay deviation at the band edge.

Similar computations for the other coefficients yield the results tabulated below.

(a) Differential Gain

Type of Transmission Deviation	Diff. Gain/db of Deviation at ± 10 mc/s	Diff. Gain/mμsec. of Deviation at ± 10 mc/s
Parabolic phase	-	2.0×10^{-4} db
Cubic Phase	-	0.26×10^{-4} db
Cubic Gain	0.10 db	-
Quartic gain	0.012 db	-

(b) Differential Phase

Type of Transmission Deviation	Diff. Phase/db of Deviation at ± 10 mc/s	Diff. Phase/mμsec. of Deviation at ± 10 mc/s
Parabolic phase	-	0.32°
Cubic phase	-	0.02°
Cubic gain	0.005°	-
Quartic gain	0.0015°	-

As an indication of what the results mean in terms of delay requirements, suppose that the requirement for differential phase on four TJ end-links with a total of 40 repeaters (10 per link) is 2°. Suppose also that half of this is allocated to baseband equipment and half to FM components. Then the differential phase due to delay distortion must be less than 1°. Suppose also that the equalization is such that the linear delay deviations are random so that the distortion contributions from individual repeaters can be added on an RMS basis. The differential phase requirement per repeater will therefore be $1/\sqrt{40} = .158°$.

From the result just established this would permit a maximum linear delay deviation per repeater of approximately 0.5 mμsec.

Noise with Telephone Loading - An Illustrative Example

The distortion produced by transmission deviations becomes increasingly difficult to deal with as the modulating signal becomes more complex. The exact type of solution which was illustrated in the previous section can be obtained easily only when the modulating signal consists of a few sinusoidal components. The same type of solution becomes unmanageable when we consider the modulating signal to be a large number of frequency-multiplexed telephone channels. We shall therefore illustrate another procedure which can be used in such cases.

We shall assume that the baseband signal consists of 1000 telephone channels in a band which extends from zero to 4 mc. This signal is transmitted over a pure FM system. The peak frequency deviation will be taken as 4 mc. In the FM transmission path we shall assume that there is a parabolic phase deviation which is equal to 0.1 radian at a frequency 10 mc away from the carrier. The modulation noise in a telephone channel caused by this transmission deviation then can be found by the following procedure.

The total output signal can be obtained from Table 22-2.

$$\text{Output} = V(t) + 2b_2 k_1 V(t) V'(t)$$

$$= V(t) + b_2 k_1 \frac{d}{dt} [V^2(t)] \qquad (22\text{-}42)$$

In this form we shall be able to make use of methods developed for analyzing modulation noise in amplitude modulation systems. The expression above very closely resembles the expression for the second order modulation produced by electron tubes. The principal difference is that here the derivative of the squared term is required. By the methods referred to it is possible to find the modulation noise in any particular telephone channel due to a distortion term

$$b_2 k_1 V^2(t).$$

If the effect of the derivative in Equation (22-42) is then included the final answer is obtained. We know, however, that taking the derivative of a function of time is equivalent to multiplying its voltage spectrum by $j\omega$. Therefore, the noise voltage in a single telephone channel due to the distortion term $b_2 k_1 V^2(t)$ can be multiplied by the center frequency (in the baseband signal) of the channel in radians per second to obtain the effect of the derivative in Equation (22-42). We shall now illustrate this procedure for a single telephone channel.

From the previous illustrative example we may obtain the value of the constant b_2.

$$b_2 = \frac{.1}{(2\pi \times 10^7)^2} \qquad (22\text{-}43)$$

We may recall that the instantaneous frequency deviation $\varphi'(t)$ may be written (see Equation 19-9) as

$$\varphi'(t) = k_1 V(t) \qquad (22\text{-}44)$$

From this it follows that the peak frequency deviation is equal to k_1 times the peak voltage $V(t)$. In an actual system k_1 is a physical property of the FM modulator, and the peak voltage must be adjusted to give the desired peak frequency deviation. Here, however, we are free to select arbitrary values for k_1 and the peak voltage provided their product is equal to the peak frequency deviation. We will, therefore, take the peak voltage to be 1 volt, from which

$$k_1 = 2\pi \times 4 \times 10^6 \text{ rad/sec per volt } * \qquad (22\text{-}45)$$

With b_2 and k_1 known the rest of the problem consists of finding the modulation noise in the telephone channel of interest due to $V^2(t)$ and then multiplying this result by $b_2 k_1 \omega_i$ where ω_i is the center frequency of the telephone channel. We shall consider the top channel since the multiplier ω_i will be largest for this channel and the modulation noise will tend to be worst there.

The modulation noise in dba in a telephone channel due to modulation products of a particular type "x" (where x may be A+B, A-B, 2A-B, etc.), can be found by using an equation due to W. R. Bennett:

$$W_x = T_a - S_x + H_x + 10 \log N_{xp} - \rho_x + \eta_x V_o + .115 \lambda_x \sigma^2 \qquad (22\text{-}46)$$

Here W_x is the noise in dba at zero db transmission level, and H_x is a measure of the non-linearity of the system, discussed further below. The rest of the terms, for our purposes, can be lumped together as a constant for any particular frequency allocation and type of product. Equation (22-46) then becomes $W_x = H_x + K_x$. For this problem only A+B products are important, and the value of K_x is found to be 68.7 db.

H_x is defined as the power in dbm of the "x" type product, measured at the zero transmission level point at the output of a system,

---

*If we choose or are given some other voltage level point, the value of k_1 is correspondingly changed; the zero db transmission level point results are of course unchanged.

and formed by the intermodulation of fundamentals each of which is zero dbm at the zero level point.

The magnitude of an $\alpha + \beta$ product is a function of the magnitudes of the sine waves which caused it. If the applied signal is taken as

$$\text{applied signal} = A \cos \alpha + B \cos \beta \tag{22-47}$$

and the output as

$$\text{output} = V(t) + a_2 V^2(t), \tag{22-48}$$

then the amplitude of the $\alpha + \beta$ product is given by the product $a_2 AB$. We shall now determine the actual values of A, B, and a_2 to find H_x.

From the preceding discussion, we note that a_2 is equal to $b_2 k_1 \omega_i$. For the moment, let us for brevity merely write a_2; we shall evaluate it later.

From the discussion of the second illustrative example of noise in FM systems it will be recalled that the required load carrying capacity for a system of 1000 channels is +24.5 dbm. In that discussion it was pointed out that a 24.5 dbm sine wave at zero transmission level will therefore have a peak value equal to the maximum signal that the system must be designed to transmit. A 24.5 dbm sine wave at zero transmission level is equivalent then to a sinusoidal voltage V(t) which has a 1 volt peak,* and an $\alpha + \beta$ product formed by 24.5 dbm fundamentals will have $(24.5 + 20 \log a_2)$ dbm of power. An $\alpha + \beta$ product formed by zero dbm fundamentals will have a power 2 x 24.5 db lower than a product formed by 24.5 dbm fundamentals, or $(-24.5 + 20 \log a_2)$ dbm. Therefore, we find that H_x is $(-24.5 + 20 \log a_2)$ dbm.

Substitution in Equation (22-46) then gives

$$W_x = (-24.5 + 20 \log m_2 + 68.7) \text{ dba} = (44.2 + 20 \log m_2) \text{dba} \tag{22-51}$$

Finally, evaluating a_2 we obtain:

$$a_2 = b_2 k_1 \omega_i = \frac{.1}{(2\pi \times 10^7)^2} \times (2\pi \times 4 \times 10^6) \times (2\pi \times 4 \times 10^6) = .016 \tag{22-52}$$

$$20 \log b_2 k_1 \omega_i = -35.9 \text{ db} \tag{22-53}$$

---

*Recalling that we evaluated k_1 in terms of a signal of one volt peak value. A different choice would call for corresponding change at this point in the argument.

Whence: $W_x = 44.2 - 35.9 \text{ dba} = 8.3 \text{ dba}$ (22-54)

This is the noise in the top channel, at the zero db transmission level point, caused by the parabolic phase deviation.

Extension of the Method to Other Types of Deviations

The method which was illustrated in the previous section can be extended to other types of transmission deviations in FM systems by noting that the distortion terms which appear in Table 22-2 can be written in another form. For example, in the previous illustration we had to note that

$$2V(t)\,V'(t) = \frac{d}{dt}\,[V^2(t)]. \tag{22-55}$$

Similarly, we have for other terms in Table 22-2,

$$3[V(t)V''(t)+V'^2(t)] = 3\,\frac{d}{dt}\,[V(t)V'(t)]$$

$$= \frac{3}{2}\,\frac{d^2}{dt^2}\,[V^2(t)], \tag{22-56}$$

and

$$3V^2(t)V'(t) = \frac{d}{dt}\,[V^3(t)] \tag{22-57}$$

However, in PM systems some of the distortion terms such as that for parabolic phase,

$$b_2\,k\,V'^2(t),$$

cannot be handled by this method. In this case we note that, when V(t) is a baseband signal consisting of frequency-multiplex telephone channels, V'(t) cannot be assumed to consist of a large number of tones all of equal level. Therefore, not all of the products of a particular type falling in a particular channel will have the same power, and the method of counting products would have to be modified to take into account this power difference. This will not be considered further here.

Method 2 - The Amplitude and Phase Modulation Produced by Transmission Deviations in Low Index Systems

In the preceding sections the response to an FM wave applied to a transmission shape $Y(\omega)$ was given in Equation (22-19) as the real part of

$$e_2 = A_c\varepsilon^{j\omega_c t}\,\mathfrak{F}^{-1}\left[Y(\omega+\omega_c)\;\mathfrak{F}\,[\varepsilon^{j\varphi}]\right] \tag{22-19}$$

We then proceeded to expand $Y(\omega+\omega_c)$ in a power series and derive the amplitude and phase modulation produced by each term in the expansion. This method is particularly applicable when the transmission deviations are rather smooth shapes and the power series expansion converges rapidly. Although Equation (22-32) and resulting Tables 22-1 and 22-2 apply to small transmission deviations, this is not a necessary restriction. Method 1 can be used whenever the power series representation of the transmission characteristic converges rapidly.

When the transmission characteristic is discontinuous (for example, an ideal band pass filter) the power series approach is not applicable. Such characteristics and their effects can, however, be analyzed by the following alternate method, provided the index of modulation is small - i.e., when φ is unity or less.* The method, originally developed by Mr. S. Doba, depends upon resolving $Y(\omega+\omega_c)$ into its symmetric and skew symmetric components; let us first review the standard methods for doing this.

Resolution of Transmission Characteristics Into Components

Generally, it is possible to separate any complex function $q(x)$ into a sum of symmetric and skew symmetric components about any point $x = x_o$. (By the symmetric component we mean a component which has even real and odd imaginary symmetry about the reference axis; by skew-symmetric we mean having odd real and even imaginary symmetry about the reference axis.) To see how this is accomplished, consider the three vector model shown in Figure 3. The center vector represents $q(x_o)$ while the upper and lower vectors are $q(x_o-\Delta x)$.

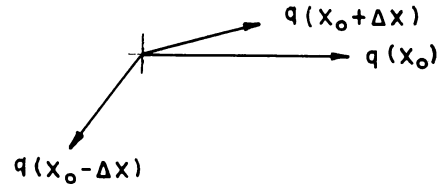
If for example, the function under discussion is a transmission characteristic, the magnitude and phase of the transmission at carrier frequency can be represented by a vector; this would be our $q(x_o)$ vector. At some higher frequency $x_o+\Delta x$ we might have a different gain and phase shift; this can be represented in magnitude and phase by the $q(x_o+\Delta x)$ vector. Similarly we can plot a vector representing transmission at the frequency $x_o-\Delta x$.

The symmetric (q_s) and skew symmetric (q_a) components of these vectors are also shown in Figure 3. For the former, their real parts are equal and their imaginary parts are conjugates, while for the latter, their real parts are negatives and the imaginary parts are equal. Observe that vectorial addition of $q_s(x_o+\Delta x)$ and $q_a(x_o+\Delta x)$ gives us our original $q(x_o+\Delta x)$ vector, similarly $q_s(x_o-\Delta x)$ and $q_a(x_o-\Delta x)$ add to give $q(x_o-\Delta x)$.

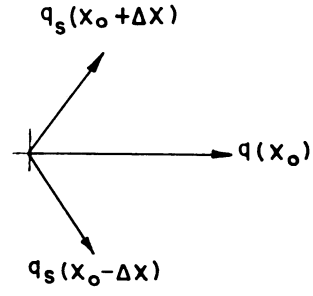
- -

*The analysis of high index systems which have discontinuities in the transmission characteristics is very difficult, and we shall not attempt to consider it here.

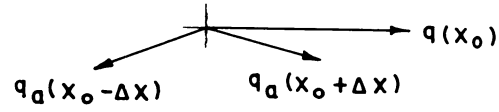
ORIGINAL VECTORS



SYMMETRIC COMPONENTS



SKEW-SYMMETRIC COMPONENTS



Resolution of Vectors into Symmetric and Skew-Symmetric Components

Figure 22-3

Observe also the even and odd symmetry of the real and imaginary components; q_s and q_a have, respectively, the symmetries relative to x_o that real and imaginary parts of physical transmission characteristics have relative to zero frequency.

The q_s and q_a vectors may be found from the q vectors by the following operations

$$q_s(x_o+\Delta x) = \frac{1}{2}\left[q(x_o+\Delta x) + q^*(x_o-\Delta x)\right] \tag{22-58}$$

and

$$q_a\ (x_o+\Delta x) = \frac{1}{2}\left[q(x_o+\Delta x) - q^*(x_o-\Delta x)\right]$$

where q_s and q_a are respectively the symmetric and skew symmetric components, and $q^*(x_o-\Delta x)$ is the conjugate of $q(x_o-\Delta x)$. Note that if $q(x_o)$ is not a positive vector, this relation does not give the correct components. However, it can be assumed without loss of generality in Eq. (22-58) that $q(x_o)$ is a positive real number.

For future use, let us note at this point that we could, if we chose, postulate some other function of x, say $p(\Delta x)$, such that*

$$p(\Delta x) = q(x_o+\Delta x) \tag{22-59}$$

whence

$$\begin{aligned} p_s(\Delta x) &= q_s(x_o+\Delta x) \\ p_a(\Delta x) &= q_a(x_o+\Delta x) \end{aligned} \tag{22-60}$$

Let us now consider the transmission characteristic $Y(\omega+\omega_c)$. Recall that this represents a plot, centered about zero frequency, of a characteristic $Y(\omega)$ originally centered about ω_c. As before, we will eliminate any linear phase shift component of $Y(\omega+\omega_c)$, since constant delay introduces no distortion. Doing this will, in practice, simplify the mathematics without changing the results. An ideal transmission characteristic will then be one for which the transmission at all frequencies of interest is equal to the transmission at ω_c - that is, one which satisfies the relation

$$\text{Ideal:} \quad \frac{Y(\omega+\omega_c)}{Y(\omega_c)} = 1 \tag{22-61}$$

---

* Compare with $H(\omega)$, Figure 21-10.

We can now define normalized symmetric and skew-symmetric components of the transmission characteristic. At any pair of frequencies equally spaced ω above and below the carrier, the function $Y(\omega+\omega_c)$ will, in the general case, assume values differing in magnitude and angle from the ideal relationship shown in Equation 61. We can plot vectors representing $Y(\omega+\omega_c)/Y(\omega_c)$ for frequencies spaced from the carrier by $-\omega$, 0, and $+\omega$, similar to those shown at the top of Figure 3 for analogous values of Δx. For $\omega=0$, the center vector, we find, of course, that $Y(\omega+\omega_c)/Y(\omega_c) = 1 + j0$. The symmetric and skew-symmetric components of the vectors for frequencies ω above and below the carrier can be written in a manner which explicitly displays the deviations from ideal, and which will be found in the original work by Doba and others in this field:

$$\frac{Y(\omega+\omega_c)}{Y(\omega_c)} = 1 + \bar{y}(\omega) - j\hat{y}(\omega) \qquad (22\text{-}62)$$

Here $1 + \bar{y}(\omega)$ is analogous to $p_s(\Delta x)$ above, differing only in that we did not normalize or subtract linear phase in that case. Note that $1 + \bar{y}(\omega)$ represents a resultant vector made up of two components, the first expressing ideal transmission and the second, $\bar{y}(\omega)$, giving the symmetric component of the deviation from the ideal case. Similarly, $-j\,\hat{y}(\omega)$ is analogous to $p_a(\Delta x)$; note the -j, however, which rotates our original skew-symmetric component definition by 90°. By analogy with Equation 58 we can write (letting $Y(\omega_c) = Y_c$)

$$1 + \bar{y}(\omega) = \frac{1}{2}\left[\frac{Y(\omega+\omega_c)}{Y_c} + \frac{Y^*(-\omega+\omega_c)}{Y_c}\right]$$

$$-j\,\hat{y}(\omega) = \frac{1}{2}\left[\frac{Y(\omega+\omega_c)}{Y_c} - \frac{Y^*(-\omega+\omega_c)}{Y_c}\right]$$

and since $Y^*(-\omega) = Y(\omega)$

$$\bar{y}(\omega) = \frac{1}{2Y_c}\left[Y(\omega+\omega_c) + Y(\omega-\omega_c)\right] - 1$$

$$-j\,\hat{y}(\omega) = \frac{1}{2Y_c}\left[Y(\omega+\omega_c) - Y(\omega-\omega_c)\right] \qquad (22\text{-}63)$$

With these definitions, $\bar{y}(\omega)$ and $\hat{y}(\omega)$ would both be identically zero for an ideal transmission characteristic. We may also note that

$$\bar{y}^*(-\omega) = \bar{y}(\omega)$$

$$\hat{y}^*(-\omega) = \hat{y}(\omega)$$

Thus, $\bar{y}(\omega)$ and $\hat{y}(\omega)$ each have same symmetry properties relative to zero frequency as a physical transmission path. (This was not true of q_a. It is true of $\hat{y}$ because of the 90° rotation to which attention was called above.) It can be shown that when a real function of time f(t) is applied to a network with this symmetry the response will also be a real function of time. This property will allow us to separate the distortion terms which are to be obtained into their real and imaginary parts.

Evaluation of Noise

The expression for $Y(\omega+\omega_c)$ given in Equation (22-62) will now be substituted into Equation (22-19) to obtain

$$e_2 = A_c Y_c \, \varepsilon^{j\omega_c t} \, \mathfrak{F}^{-1} \left[[1+\bar{y}(\omega)-j\hat{y}(\omega)] \, \mathfrak{F} \, [\varepsilon^{j\varphi}] \right]$$

$$= A_c Y_c \, \varepsilon^{j(\omega_c t+\varphi)} \left\{ 1 + \varepsilon^{-j\varphi} \mathfrak{F}^{-1} \left[[\bar{y}(\omega)-j\hat{y}(\omega)] \, \mathfrak{F} \, [\varepsilon^{j\varphi}] \right] \right\}$$

$$= A_c Y_c \, \varepsilon^{j(\omega_c t+\varphi)} \, [1 + P(t) + jQ(t)] \qquad (22\text{-}64)$$

where

$$P(t) = \mathrm{Re} \left\{ \varepsilon^{-j\varphi} \mathfrak{F}^{-1} \left[(\bar{y} - j\hat{y}) \, \mathfrak{F} \, [\varepsilon^{j\varphi}] \right] \right\} \qquad (22\text{-}65)$$

and

$$Q(t) = \mathrm{Im}(\varepsilon^{-j\varphi} \, \mathfrak{F}^{-1} \, | \, (\bar{y}-j\hat{y}) \, \mathfrak{F} \, [\varepsilon^{j\varphi}] \, | \,) \qquad (22\text{-}66)$$

These expression should be compared with those given in Equations (22-27) and (22-28).

When the index of modulation is low the exponentials in Equations (22-65) and (22-66) can be readily expanded into a rapidly convergent series by using the identity

$$\varepsilon^x = \sum_{n=0}^{\infty} \frac{x^n}{n!}$$

We get

$$P+jQ = [1-j\varphi - \frac{\varphi^2}{2}\cdots] \, \mathfrak{F}^{-1} \left\{ (\bar{y}-j\hat{y}) \, \mathfrak{F} \, [1+j\varphi - \frac{\varphi^2}{2} - j\frac{\varphi^3}{6}\cdots] \right\} \qquad (22\text{-}67)$$

By their definition $\bar{y}(0)$ and $\hat{y}(0)$ are zero. It therefore follows that $\bar{y}\ \mathfrak{F}[1]$ and $\hat{y}\ \mathfrak{F}[1]$ are zero. If we retain no terms higher than the third order in φ, the remaining distortion terms may be written as

$$\begin{aligned} P + jQ = \mathfrak{F}^{-1} \Big\{ & \bar{y}\ \mathfrak{F}[j\varphi] - \bar{y}\ \mathfrak{F}\left[\frac{\varphi^2}{2}\right] - \bar{y}\ \mathfrak{F}\left[j\frac{\varphi^3}{6}\right] \\ & -j\hat{y}\ \mathfrak{F}[j\varphi] + j\hat{y}\ \mathfrak{F}\left[\frac{\varphi^2}{2}\right] + j\hat{y}\ \mathfrak{F}\left[\frac{\varphi^3}{6}\right] \Big\} \\ & - j\,\varphi\,\mathfrak{F}^{-1} \Big\{ \bar{y}\ \mathfrak{F}[j\varphi] - \bar{y}\ \mathfrak{F}\left[\frac{\varphi^2}{2}\right] \\ & \qquad -j\hat{y}\ \mathfrak{F}[j\varphi] + j\hat{y}\ \mathfrak{F}\left[\frac{\varphi^2}{2}\right] \Big\} \\ & - \frac{\varphi^2}{2}\,\mathfrak{F}^{-1} \Big\{ \bar{y}\ \mathfrak{F}[j\varphi] - j\hat{y}\ \mathfrak{F}[j\varphi] \Big\} \end{aligned} \tag{22-68}$$

These terms may be separated to give

$$\begin{aligned} P = {} & \mathfrak{F}^{-1}\,\hat{y}\ \mathfrak{F}[\varphi] - \mathfrak{F}^{-1}\,\bar{y}\ \mathfrak{F}\left[\frac{\varphi^2}{2}\right] + \varphi\,\mathfrak{F}^{-1}\bar{y}\mathfrak{F}[\varphi] \\ & - \mathfrak{F}^{-1}\,\hat{y}\ \mathfrak{F}\left[\frac{\varphi^3}{6}\right] + \varphi\,\mathfrak{F}^{-1}\,\hat{y}\ \mathfrak{F}\left[\frac{\varphi^2}{2}\right] - \frac{\varphi^2}{2}\,\mathfrak{F}^{-1}\,\hat{y}\ \mathfrak{F}[\varphi] \end{aligned} \tag{22-69}$$

$$\begin{aligned} Q = {} & \mathfrak{F}^{-1}\,\bar{y}\ \mathfrak{F}[\varphi] + \mathfrak{F}^{-1}\,\hat{y}\ \mathfrak{F}\left[\frac{\varphi^2}{2}\right] - \varphi\,\mathfrak{F}^{-1}\,\hat{y}\ \mathfrak{F}[\varphi] \\ & -\mathfrak{F}^{-1}\,\bar{y}\ \mathfrak{F}\left[\frac{\varphi^3}{6}\right] + \varphi\mathfrak{F}^{-1}\,\bar{y}\ \mathfrak{F}\left[\frac{\varphi^2}{2}\right] - \frac{\varphi^2}{2}\,\mathfrak{F}^{-1}\,\bar{y}\ \mathfrak{F}[\varphi] \end{aligned} \tag{22-70}$$

The terms in P and Q have the following interpretation; expressions such as $\mathfrak{F}[\varphi]$ and $\mathfrak{F}[\frac{\varphi^2}{2}]$ are the Fourier spectra of the modulating function $\varphi(t)$ and of the square of $\varphi(t)$ respectively. Expressions such as $\bar{y}\ \mathfrak{F}[\varphi]$ and $\hat{y}\ \mathfrak{F}[\varphi]$ are the spectrum of the modulating function after passing through the symmetric and skew symmetric components of the transmission characteristic $y(\omega)$. Finally, the addition of the inverse operator, such as $\mathfrak{F}^{-1}\,[\bar{y}\ \mathfrak{F}\,[\varphi]\,]$, yields the function of time which would occur if the modulating function $\varphi(t)$ were passed through the symmetric component of the transmission characteristic $y(\omega)$. Thus, by evaluating the operation of the transmission characteristic on the modulating function and its powers, the effects of deviations of the transmission medium from flat amplitude and linear phase can be determined.

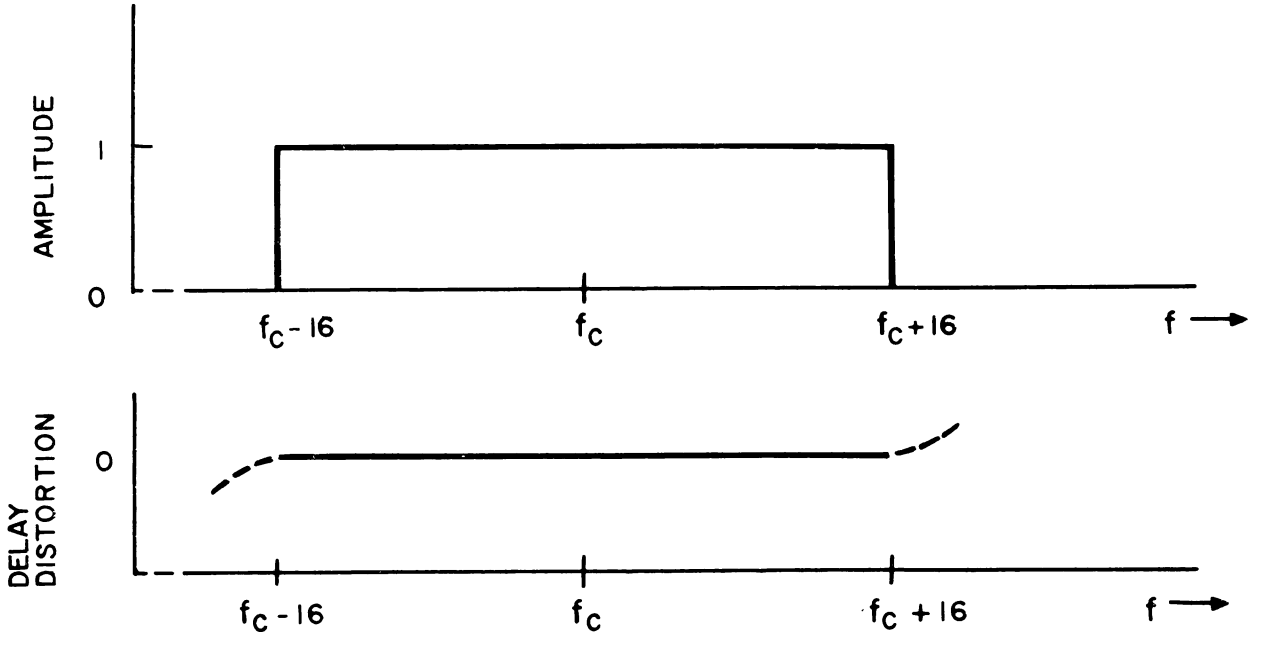
Illustrative Example : Method 2

As an example of the use of the $\bar{y}$ and $\hat{y}$ method, we consider a rather simple case -- and one which is of real practical interest -- which could not be handled by the power series method of representing transmission characteristics.

A signal M(t) is to be transmitted by a low-index phase-modulation system whose transmission characteristic is distortionless over the band $f_c \pm 16$ mc, where f_c is the carrier frequency. Outside this band the system does not transmit at all, being band-limited by ideal filters whose in-band phase distortion has been perfectly equalized. The input signal M(t) has a power spectrum that is flat vs frequency from dc to 10 mc, with no components above 10 mc. The baseband output signal is limited to 10 mc by a low pass filter. Using Method 2 of Chapter 22, plot:

a) $\bar{y}$ and $\hat{y}$ vs freq.
b) a qualitative evaluation of the distortion products vs baseband output frequency, assuming perfect limiting.

The transmission characteristic given is:



Transmission Characteristic of Illustrative Example

Figure 22-4

The component shapes $\bar{y}(\omega)$ and $\hat{y}(\omega)$ may be found by using Equation 22-63.

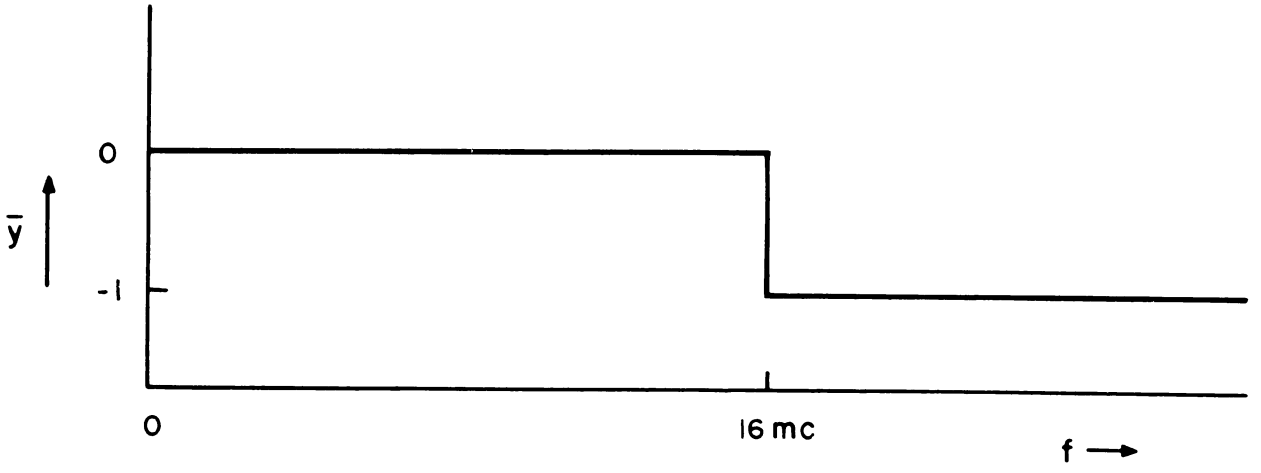
$$\bar{y}(\omega) = \frac{1}{2Y_c}\left[Y(\omega+\omega_c) + Y(\omega-\omega_c)\right] - 1$$

$$= \frac{1}{2(1)}\left[1+1\right] - 1 = 0 \text{ for } 0 < f < 16 \text{ MC}$$

$$= \frac{1}{2}\left[0+0\right] - 1 = -1 \text{ above } 16 \text{ MC}$$

$$-j\,\hat{y}(\omega) = \frac{1}{2Y_c}\left[Y(\omega+\omega_c) - Y(\omega-\omega_c)\right]$$

$$-j\,\hat{y}(\omega) = \frac{1}{2(1)}\left[1-1\right] = 0 \text{ for all frequencies}$$



Plot of $\bar{y}(\omega)$

Figure 22-5

The next step is to consider Equations 22-69 and 22-70 which give the distortion products P(t) and Q(t). P(t) is the amplitude modulation, which will be removed by the limiter, and therefore need not be evaluated. Q(t) is the phase modulation distortion produced by $\bar{y}$ and $\hat{y}$, and is to be determined from Eq. 22-70:

$$Q(t) = \mathfrak{F}^{-1}\,\bar{y}\,\mathfrak{F}\left[\varphi(t)\right] + \mathfrak{F}^{-1}\,\hat{y}\,\mathfrak{F}\left[\frac{\varphi^2(t)}{2}\right]$$

$$-\varphi\,\mathfrak{F}^{-1}\,\hat{y}\,\mathfrak{F}[\varphi(t)] - \mathfrak{F}^{-1}\,\bar{y}\,\mathfrak{F}\left[\frac{\varphi^3(t)}{6}\right]$$

$$+\varphi\,\mathfrak{F}^{-1}\,\bar{y}\,\mathfrak{F}[\frac{\varphi^2(t)}{2}] - \frac{\varphi^2}{2}\,\mathfrak{F}^{-1}\,\bar{y}\,\mathfrak{F}[\varphi(t)]$$

We observe that we will have to know something about $M^2(t)$ and $M^3(t)$,-- i.e., $\varphi^2(t)$ and $\varphi^3(t)$. A more detailed discussion of the spectra of these functions will follow; for the moment it will suffice to say that the spectrum of $M^2(t)$ extends from dc to 20 mc (since M(t) goes from dc to 10 mc) and that $M^3(t)$ extends from almost dc up to 30 mc.

Evaluating each of the 6 terms of the above equation:

1st term $\quad \mathfrak{F}^{-1}\, \bar{y}\, \mathfrak{F}[\varphi(t)] = 0$

since up to 16 MC, $\bar{y}$ is zero, and for all frequencies above 10 MC, $\varphi(t)$ is zero.

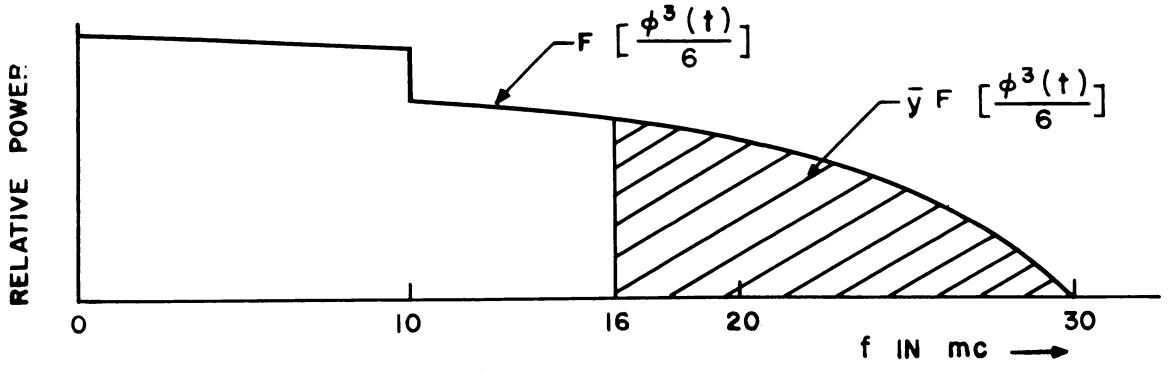
2nd term $\quad \mathfrak{F}^{-1}\, \hat{y}\, \mathfrak{F}[\frac{\varphi^2(t)}{2}] = 0$

since $\hat{y} = 0$ for all freq.

3rd term $\quad -\varphi(t)\, \mathfrak{F}^{-1}\, \hat{y}\, \mathfrak{F}[\varphi t] = 0$

since $\hat{y} = 0$ for all freq.

4th term $\quad -\mathfrak{F}^{-1}\, \bar{y}\, \mathfrak{F}[\frac{\varphi^3(t)}{6}]$



Plot of Fourth Term

Figure 22-6

This product is zero below 16 MC, since $\bar{y} = 0$ below 16 MC. The shaded area shows the product between 16 and 30 MC. Since the baseband output is limited to 10 MC there will be no contribution from this term at receiver output.

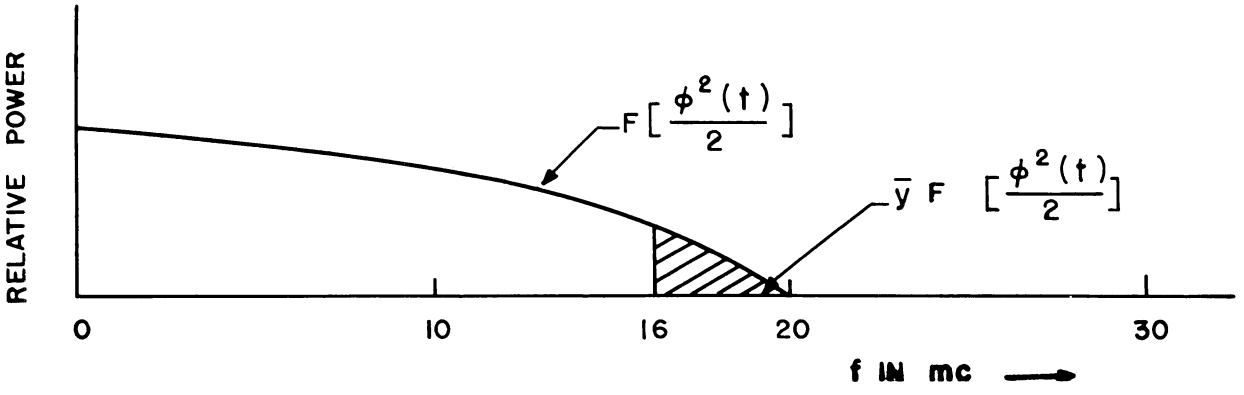
5th term $\quad \varphi(t)\, \mathfrak{F}^{-1}\, \bar{y}\, \mathfrak{F}[\frac{\varphi^2(t)}{2}]$

The shaded area shows the frequency region where $\bar{y} = -1$ and hence $\bar{y}\, \mathfrak{F}[\frac{\varphi^2(t)}{2}]$ is not zero. It is a shaped noise-like spectrum over

the band 16 to 20 MC. The time function corresponding to this frequency function is

$$\mathfrak{F}^{-1}\left[\bar{y}\ \mathfrak{F}\ \frac{\varphi^2(t)}{2}\right]$$

This is multiplied by $\varphi(t)$, which gives second order A-B modulation products which fall above 6 MC. Since the baseband is limited to 10 MC, the modulation products of interest extend from 6 MC to 10 MC.



Plot of Fifth Term

Figure 22-7

6th term $\quad \frac{-\varphi^2(t)}{2}\ \mathfrak{F}^{-1}\ \bar{y}\ \mathfrak{F}\ [\varphi(t)] = 0$

This term is zero since $\bar{y} = 0$ for all frequencies where $\varphi(t)$ is not equal to zero.

To sum up, then, we see that the fifth term is the only one which really contributes in this particular case. The nature of its contribution has been qualitatively discussed above; a quantitative solution would be the next objective in actual practice. This would involve merely a special and somewhat tedious computation of the intermodulation products obtained by multiplying two spectra,-- one flat, the other, shaped. The usefulness of the $\bar{y}\ \hat{y}$ method lies in the fact that it has thus reduced the esoteric to the mundane.

Note on $M^2(t)$ and $M^3(t)$: It was noted above that a knowledge of the spectra of $M^2(t)$ and $M^3(t)$ is needed for a complete solution of the problem. These could be obtained from the methods of Chapter 12. It would be necessary to

a. Replace M(t) by a series of sine wave signals of random phase and uniform amplitude, postulating one such tone per channel.
b. For each frequency of interest, get $\mathfrak{F}[M^2(t)]$ by counting second order products and adding their

powers, with due attention to their relative magnitudes (e.g., the fact that A+B products are 6 db hotter than 2A products). Similarly find $\mathcal{F}[M^3(t)]$ by considering third order products, including the compression term.

At this point it is necessary to warn the student that the example given here is an extremely simple and idealized one. Firstly, we have made the unrealistic assumption that there is no in-band phase distortion. Secondly, the case studied was one where the transmission characteristic had perfect symmetry about the carrier frequency. These two conditions made the determination of $\bar{y}$ and $\hat{y}$ very easy and resulted in very simple expressions for these functions. Unfortunately, neither of these conditions is liable to be satisfied in most practical problems.

Method 3 - Noise Produced in the Top Baseband Channel of an FM System by Echos at Radio Frequencies

An examination of FM radio systems shows that perhaps the majority of unequalized transmission deviations are due to echos caused by impedance mismatches in the transmission path. Figure 8 provides a means for calculating the noise in the top baseband channel of an FM system caused by such echos. This figure has been taken from Reference 2. It applies to high as well as to low index systems and echo delays of any magnitude, but does assume that the top baseband frequency is many times the lowest baseband frequency. The analysis which produces Figure 8 makes use of the fact that a signal made up of very many voice channels approaches thermal noise in character. The signal can, therefore, be thought of as consisting of flat random noise covering the same frequency band and having the same average power as the multichannel telephone signal.

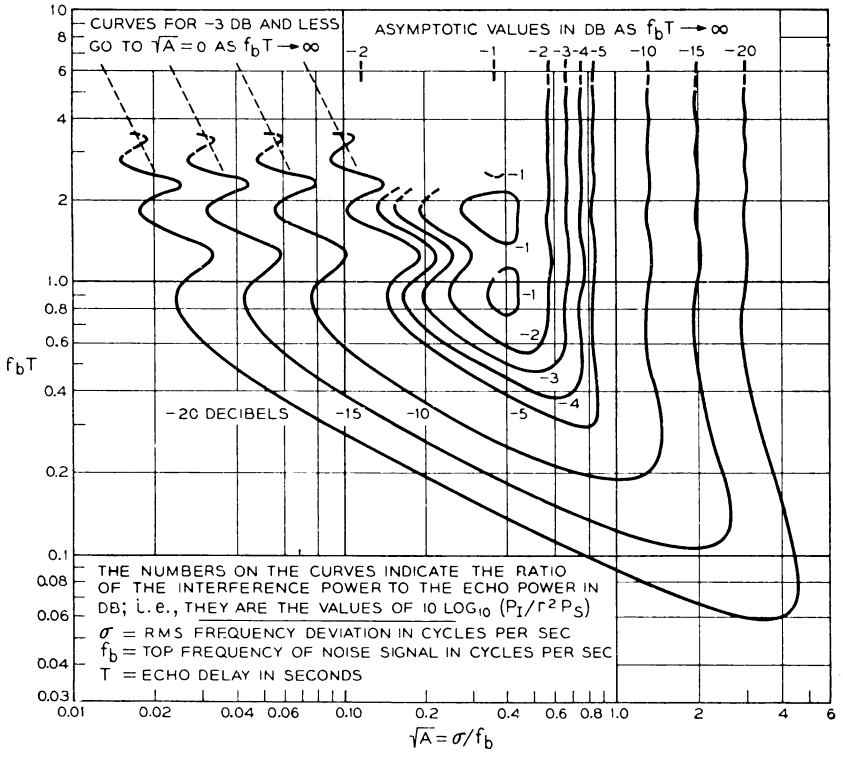
The method of Figure 8 is most easily demonstrated by an example. Assume that a particular system has an echo that is 50 db below the signal power and which lags the signal by 0.3 microseconds. Let the top baseband frequency be 1 mc and the peak frequency deviation be 4 mc. For thermal noise, the commonly used peak factor is 4.* Since the theory on which

- -

*More precisely, the peak values of random noise will exceed values four times the rms values less than 0.01% of the time. It is interesting to note that approximately the same value for peak factor can be obtained from Figure 12-2. As N grows infinitely larger, Δ_c approaches 10 db. This means the system must be capable of carrying a sine wave with an rms value 10 db above the average total talker power. Such a sine wave has a peak $\sqrt{2}$ times the rms value. The system is thus designed to withstand peaks which are $10 + 20 \log \sqrt{2} = 13$ db above the rms value of the load. This is roughly a peak factor of 4.

Figure 8 is based assumes that the multichannel load has the same characteristic as thermal noise, we convert the peak frequency deviation to rms deviation by dividing by this factor. Therefore, σ = 1 mc. This gives σ/f_b = 1 and f_bT = 0.3 cycles. Figure 8 gives a value of about -7 db for this point. Therefore, the noise in the top baseband channel will be 7 more db below the baseband signal than the echo was below the radio signal or -50 - 7 = -57 db below the signal. This can be converted to a more usefull number. Using a value of -11 dbm as the power of the average power talker (see Page 12-3) and an activity factor of 25%, the average signal power in a one way telephone channel will be -11 - 10 log 4 = -17 dbm. Then the noise at zero level due to the echo will be -17-57 = -74 dbm or 8 dba.

Strictly speaking, Figure 8 applies only to a pure FM signal. The presence of pre-emphasis may reduce the actual noise produced by an echo 3 or 4 db below the value calculated.



Contours of Constant Interference in the Top Channel of a Multichannel FM System

Figure 22-8

On page 16-18 in the television chapter it was shown that a sinusoidal gain or phase ripple is indistinguishable from a pair of leading and lagging echos. This suggests that the method described here can be used to estimate the effect of any shape of transmission deviation. The shape of interest can be expressed as a Fourier series. To each term in the series, a corresponding echo can then be associated and its effect found by use of Figure 8. When the echos so obtained are sufficiently small, the noise resulting from each echo can be added to that due to the others on a power basis. The result will be the noise in the top channel only, but in an FM system, this will be the one most severely degraded by transmission deviations. It is therefore, the one that sets the requirement.

Bibliography

1 - S. O. Rice, "Distortion in a Noise-Modulated FM Signal by Nonlinear Attenuation", BSTJ, vol. 36, pp. 879-890, July, 1957 (also Monograph 2859).

2 - Bennett, Curtis, Rice, "Interchannel Interference in FM and PM Systems under Noise Loading Conditions", BSTJ, vol. 34, pp. 601-636, May, 1955.

Chapter 23

FREQUENCY ALLOCATION

Available bandwidth, desired baseband width, carrier frequency deviation, required signal quality, cost per telephone channel, and protection against fading are important factors in determining the frequency allocation plan of a multichannel microwave system. In addition, interferences between channels which arise within the system often can be minimized by a suitable frequency allocation. These subjects are discussed, and the application of the principles to the TD-2, TH, and TJ systems are illustrated.

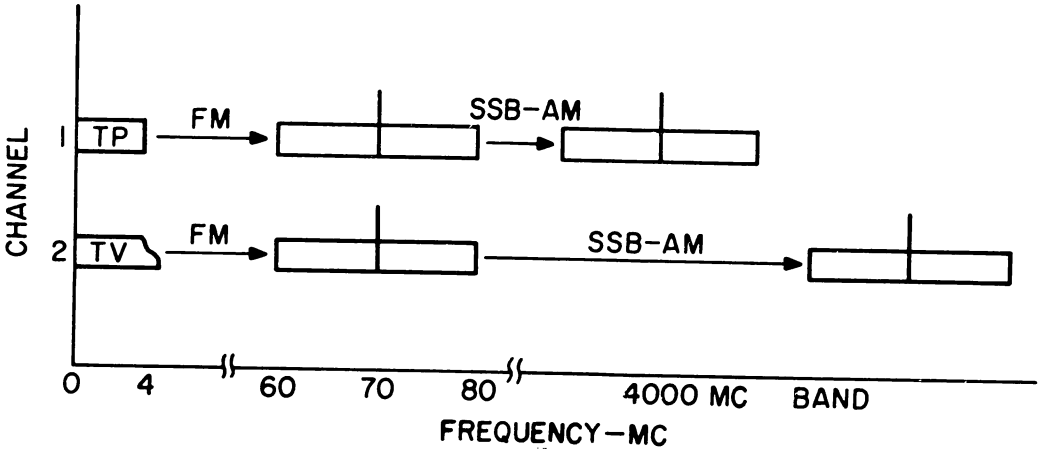
Introduction

One of the important problems which influences both the design and maintenance of multichannel microwave systems is the choice of appropriate microwave, intermediate, and beat oscillator frequencies. As will be seen in this chapter a carefully chosen frequency allocation can do much to reduce the design problems in other parts of the system. A poor choice, on the other hand, may degrade the system performance, increase the difficulty of design, and ultimately increase the manufacturing, installation and maintenance costs of the system.

In order to outline the general problem we shall refer to Figure 1 at the end of Chapter 17 which shows a typical microwave system. The transmitting equipment for two channels is shown in the upper left corner. A frequency-multiplexed telephone signal is applied to Channel 1 and a television signal to Channel 2. Although the existence of other channels is apparent from the figure, we shall concern ourselves with the two already mentioned. The baseband signals are fed to FM transmitting terminals from which are obtained two separate FM signals at an intermediate frequency. From here each FM signal goes to a radio transmitter where it is "beat" with the output of a beat frequency oscillator in a converter. This process is essentially single-sideband amplitude modulation and simply translates the entire FM signal up to an appropriate microwave frequency. If the frequency translation is sufficiently different for the two channels, the two signals may be combined by microwave filters and delivered to an antenna with the assurance that the two signals may be separated at a distant receiving terminal without excessive interchannel interference. The frequency translation which we have just described is shown in Figure 1.

At the receiving terminal a similar process takes place in the reverse order. The two FM signals are separated by microwave filters and sent to separate converters where again appropriate beat frequencies are used to provide translation of the FM signal back to the intermediate frequency.

A number of FM signals can be combined at microwave frequencies by the method just described and can thereby be transmitted by a single antenna. This results in overall economy but at the same time permits possible interferences among the individual channels. As more channels are added the sources of interference tend to increase. Thus we find that the frequency allocation chosen must be a suitable compromise between economy and performance, and at the same time meet the Bell System needs. Some of the factors which influence this compromise will be considered qualitatively in the following sections.



Steps in Formation of Microwave Radio Signal

Figure 23-1

First, we shall consider some of the factors which determine the channel bandwidth and number of microwave channels which a given system shall have. We shall find that we are limited to some extent by FCC regulations, that certain desirable baseband widths influence our design, and that other factors such as the frequency deviation we use, the signal quality required, the cost per telephone channel, and protection against fading, must also be considered. For particular systems some of these factors may be more important than others.

Next we shall consider some typical types of interference which may occur in multi-channel microwave systems. We will show how certain frequency allocation techniques are helpful in minimizing these interferences.

Finally, the frequency allocations used in the TD-2, TJ, and TH Systems will be briefly reviewed.

Factors Which Determine the Bandwidth and Number of Microwave Channels

Available Microwave Bandwidth

One of the more obvious limitations to the number of microwave channels we may have is the bandwidth made available to us by the Federal Communications Commission. At the present time there are three such bands. These are the 500 mc bands at 4,000 and 6,000 mc and a 1,000 mc band at 11,000 mc, as listed below.

Band	Band Edges in Megacycles
4 KMC	3,700 - 4,200
6 KMC	5,925 - 6,425
11 KMC	10,700 -11,700

In each case the bandwidth is approximately 10% of its center frequency. These bandwidths set an upper limit to the capacity of our microwave systems.

Desirable Baseband Width

Since we are limited in the maximum bandwidth we may use our problem is simplified to deciding whether to use a few very broad microwave channels or a greater number of narrow ones. An important consideration here is the desirable baseband width, since this will establish a minimum limit to the width of an acceptable microwave channel. As an approximation we can make use of the rule-of-thumb given in Chapter 19 which states that the bandwidth should be at least twice the sum of the baseband width and the peak frequency deviation. We thus conclude that the minimum bandwidth for a microwave channel will be greater than twice the top baseband frequency.

The choice of a baseband width sometimes depends on a number of factors; in other cases it may be relatively clear-cut. For example, in the TD-2 and TJ Systems one of the baseband signals is a 4 mc television signal. These require a 4 mc baseband width. The TH System has been provided with a 10 mc baseband width so that the Bell System will be able to accommodate 10 mc television signals which are sometimes proposed for closed theatre television loops. In each of these systems the television signals have determined the baseband width. Frequency-multiplexed telephone signals, or in the case of TH, a combined telephone and 4 mc television signal, are then selected so that they can utilize the same baseband widths.

One might ask why we don't make the baseband width even wider than is required for these television signals. Several factors are involved here. Among them are convenience in growth, the difficulty in providing wider bandwidth circuits, the problem of providing protection against fading, and certainly the fact that in order to drop telephone channels in an FM system the entire channel has to be brought to baseband. From this latter fact alone we may conclude that baseband signals comprised of thousands of telephone signals would not be practical except on the most heavily loaded routes. This follows from the fact the FM terminals become more expensive as they are required to handle larger signals and, in addition, those signals which are not dropped are degraded by the additional demodulation and modulation processes. More microwave channels with fewer telephone channels on each make it possible to drop a small portion of the total load without degradation to the rest.

Another consideration may be that of growth. A microwave system which utilizes a bandwidth of 500 mc has an extremely large capacity. For example, the TH System will probably have six active channels each capable of handling over 2,000 telephone channels. On some routes only a portion of this capacity may be required when the system is first installed. If a system has several microwave channels the equipment for the individual channels may be obtained as needed. This provides for convenience in growth.

Frequency Deviation

Frequency deviation has already been introduced by means of the rule-of-thumb as a factor in the determination of microwave bandwidth. Frequency deviations which are small compared to the top baseband frequency have only small effect on the bandwidth required but have a poor signal-to-noise performance when compared to AM systems with the same power. (For more discussion of this point the reader is referred to the section on FM Advantage in Chapter 20.) Large frequency deviations, on the other hand, will require much greater bandwidths and are generally undesirable for several reasons. One is, of course, the uneconomic use of the microwave bandwidth available. Others include the difficulty of obtaining linear FM modulators capable of large frequency deviations and the difficulty of providing IF Circuits with adequate bandwidth. For the TJ, TD-2, and TH Systems, peak frequency deviations of about 4 mc have been adopted. It is obvious that these make a significant contribution to the required channel width.

Signal Quality

In the preceeding sections we made use of a rule-of-thumb to determine the required bandwidth of a single channel. Although this rule is an excellent guide in system design the reader is aware that FM signals have sidebands which theoretically extend to infinite frequency. Any rule-of-thumb is therefore a compromise between quality and economy and some distortion is always introduced when the higher order sidebands are removed by filters or by other band limiting circuits. On the other hand, if we fail to band-limit two FM signals which are to be placed in adjacent microwave channels, the higher order sidebands will extend into the adjacent channel and cause interference. By studies of the particular signals to be transmitted in adjacent channels with specified spacing it is possible to choose a filter shape which will minimize the total signal degradation due to these two causes. Even with this optimization, however, as the two microwave channels are spaced closer together the signal quality will decrease and more complex filters will be required.

Cost

In order to establish a microwave route a series of tower locations need to be obtained. New roads often have to be constructed. Towers and antennas have to be built and installed. These costs tend to be fixed and independent of the number of channels. As the number of telephone channels is increased the cost per channel is reduced. Hence, it is desirable to provide capacity for as many telephone channels as possible. This can be accomplished by moving the adjacent channels closer together. Beyond a certain point, however, more complex filters are required, as mentioned above, and the tolerances on the various IF and microwave frequencies used in the system must be made smaller, thereby causing the system to become more complex and expensive. In general then we find 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 will tend to increase after a certain point. Theoretically, there is an optimum point of operation where the total cost per channel is a minimum. In practice we would like to operate somewhere near this point.

What has just been said applies in particular to our long haul systems designed for heavily loaded routes. For lightly loaded routes where the additional channels could not be used even if available the situation is different. We then find it desirable to move the adjacent channels apart so as to swap unneeded bandwidth for a reduction in filter and equipment complexity.

Protection Against Deep Fades

As was pointed out in Chapter 18, microwave channels are subject to fades of various magnitudes. The loss which occurs during small fades can be compensated by suitable automatic gain controls. Deep fades, however, may seriously degrade the noise performance. Fortunately, these fades tend to be selective, and if we have several microwave channels in our system, one or two of them may be reserved as a protection channel. When a regular channel fails, due to a fade, or perhaps some equipment failure, its signal may be transferred to a space channel by rapid switching techniques so that there is very little signal degradation. Such protection greatly increases the dependability of our system, and is very important on heavily loaded routes.

We must realize, however, that such protection reduces the capacity of our system. It is economically desirable then, to let a single channel act as a spare for several active ones.

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. Let us now take up the second problem of this chapter: sources of interference in multi-channel microwave systems as related to the choice of frequency allocation.

Interference in Microwave Channels

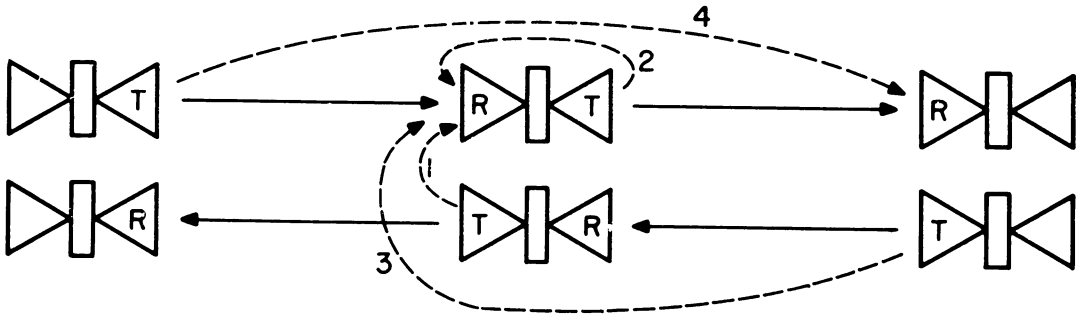
In this section we shall be concerned with various types of interference which may occur in a multichannel microwave system. These may be separated into the four main types listed below.

1. Same channel interference.
2. Image channel interference.
3. Adjacent channel interference.
4. Tone interference.

Each of these will be considered separately. External interferences are assumed to be kept within limits by FCC regulations and will not be discussed here.

Same Channel (Co-Channel) Interference

The problem of same channel interference is illustrated in Figure 2. Here we have shown, in block form, three consecutive repeaters and have indicated four typical interference paths. Separate receiving and transmitting antennas are shown although in some systems only a single antenna may be used. The two most serious interference paths are those labeled "1" and "2". Here we have high level signals



Same Channel Interference

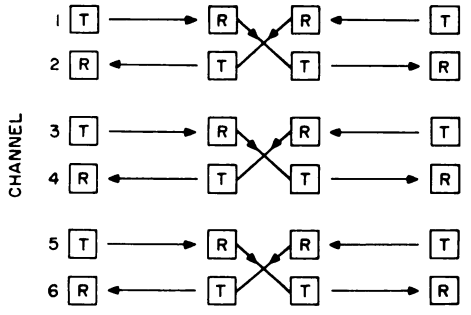
Figure 23-2

from transmitting antennas interfering with low level signals at the receiving antenna. 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. Additional loss can be introduced if we polarize one of the signals vertically and the other horizontally. In a practical situation, however, this is still likely to be insufficient and we avoid the problem by making the transmitting and receiving frequencies at a given repeater different. Such an arrangement is shown in Figure 3a and is known as a two <u>frequency allocation</u>.

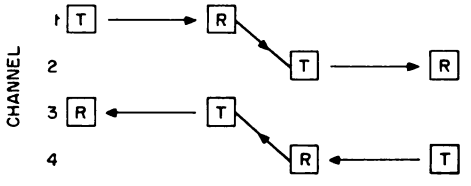
The interference from the path labeled "3" is not so large. Here, if using a two frequency allocation, we have two signals being received on the same frequency and we may expect that normally they will be at about the same levels at the receiving antenna. In this case the interference will be reduced by the front-to-back ratio of a single antenna which may be about 65* db for a delay lens or horn type antenna and about 50 db for a paraboloid antenna. Further advantage of up to about 10 db can again be obtained by using different polarizations. This arrangement can be made to work and is used in some systems. However, one must be careful to make sure the interference will not be excessive during a fade on the desired channel. A <u>four-frequency allocation</u> such as shown in Figure 3b avoids this problem and is sometimes used.

- -

*Actually about 80 db but limited by foreground reflection to this value.



a. TWO - FREQUENCY ALLOCATION



b. FOUR - FREQUENCY ALLOCATION

Figure 23-3

Another source of same channel interference is shown by the path labeled "4" in Figure 2. This is avoided by putting successive repeaters slightly out of line.

Image Channel Interference

Image Channel Interference is illustrated in Figure 4. Here we have two signals with carrier frequencies of 11,000 mc and 11,140 mc. We separate these signals with filters and apply them to converters to be translated down to an IF frequency of 70 mc. Suppose we use a beat oscillator frequency of 11,070 mc for the 11,000 mc signal. As long as the filters are ideal there does not seem to be any problem. Suppose, however, that filter 1 inadequately rejects the 11,140 signal. Then this signal would also beat with the 11,070 mc tone to give an unwanted 70 mc IF interference. This is known as image channel interference, where the image channel is the channel which differs in frequency from the beat frequency by the same amount as the desired channel but is on the other side of the beat frequency.

One way to avoid image channel interference is to leave the image channel empty. Otherwise, one must make sure that the filtering is adequate to prevent excessive interference even if a deep fade occurs on the desired channel. Polarization may again be helpful here.

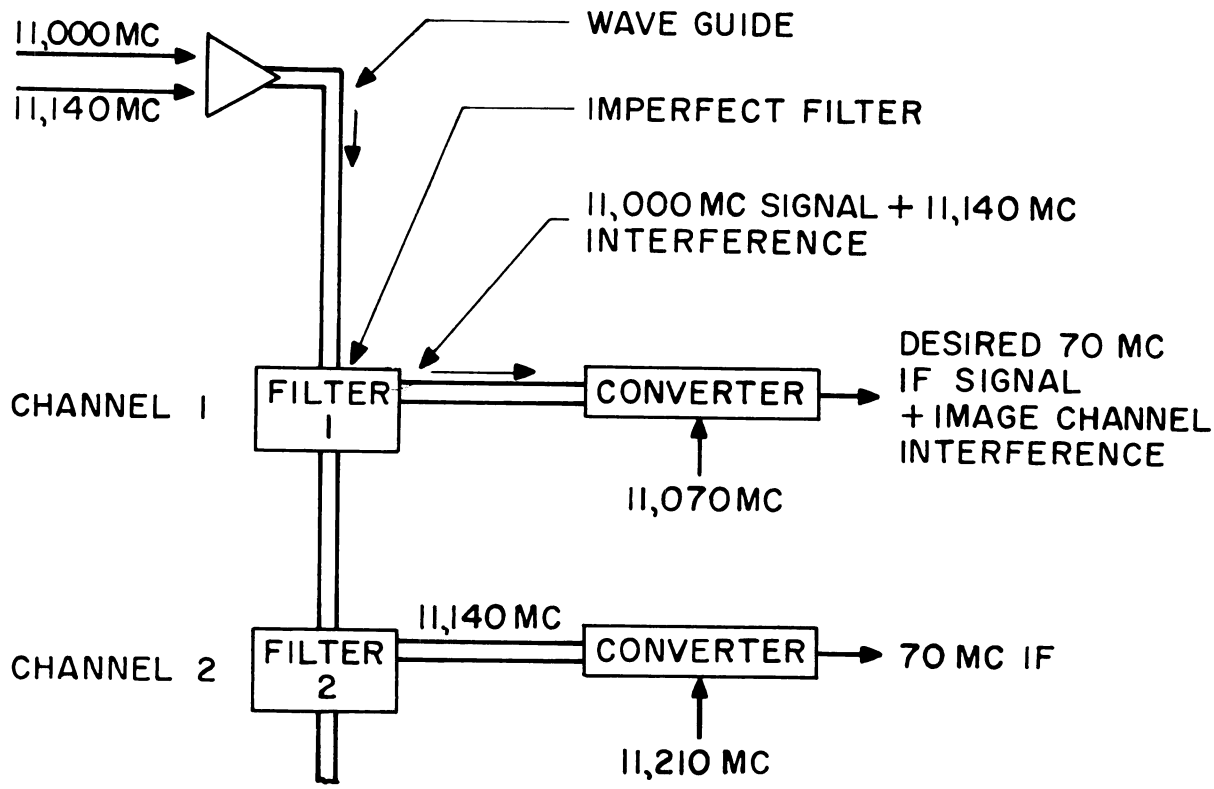
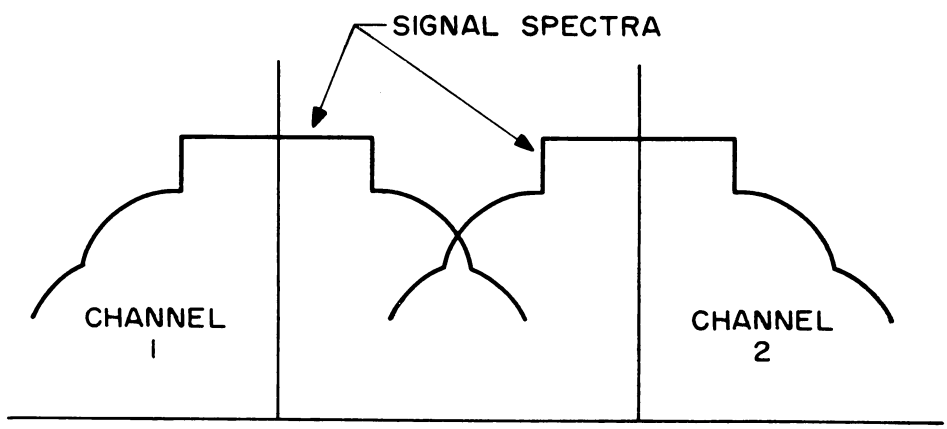


Image Channel Interference

Figure 23-4

Adjacent Channel Interference

Adjacent channel interference was mentioned briefly in a previous section and occurs when two FM channels are placed close together in frequency so that the sidebands from one extend into the other. In Figure 5 we show how this can happen by making use of the spectra of



Adjacent Channel Interference

Figure 23-5

a carrier which has been phase modulated by a baseband signal consisting of random noise. We can prevent the interference by removing the higher order sidebands with filters before the two signals are combined. However, this cannot be done without causing some distortion of the signal. Here, considerable advantage can be obtained if different polarizations can be used for the two signals. In an actual system design one would have to make a study of the signals involved in order to determine an acceptable spacing for adjacent channels.

Tone Interference

Finally, we shall consider those interferences which, although they may be caused in several different ways, have this characteristic in common: they are all essentially single-frequency interferences. The more important sources of such tones are "same channel interference", "image channel interference", and interference from beating oscillators.

"Same channel" and "image channel" interferences 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 - as, for example, the protection or spare channel carrier). In such cases, the cross-talk will consist primarily of the high-level carrier component, and may be treated as a single frequency interference.

At first glance it might seem that "same channel" interference would not be very serious, since nominally the interfering tone will have the same frequency as the carrier of the disturbed channel. Both signals, however, are likely to be not exactly at their nominal values, and a frequency difference of as much as several megacycles may actually exist.

A typical problem will illustrate the situation. Suppose the desired signal has a low index of modulation, (for example, one radian) 0 dbm of power, and a carrier frequency of 6,000 mc. The interference consists of a -40 dbm tone at 6002 mc. The problem is basically of the same type considered in Chapter 20 and may be solved as follows. The interference is 40 db below the desired signal and will produce 2 mc amplitude and phase modulation of the desired carrier.

peak phase dev = -40 db with respect to one radian
= .01 radian

peak frequency dev = peak phase deviation X frequency of phase deviation.
= .01 x 2 = .02 mc.

The amplitude modulation can be removed with a limiter and when the remaining signal is applied to a PM or FM demodulator there will be a single frequency interference at 2 mc in the baseband output.

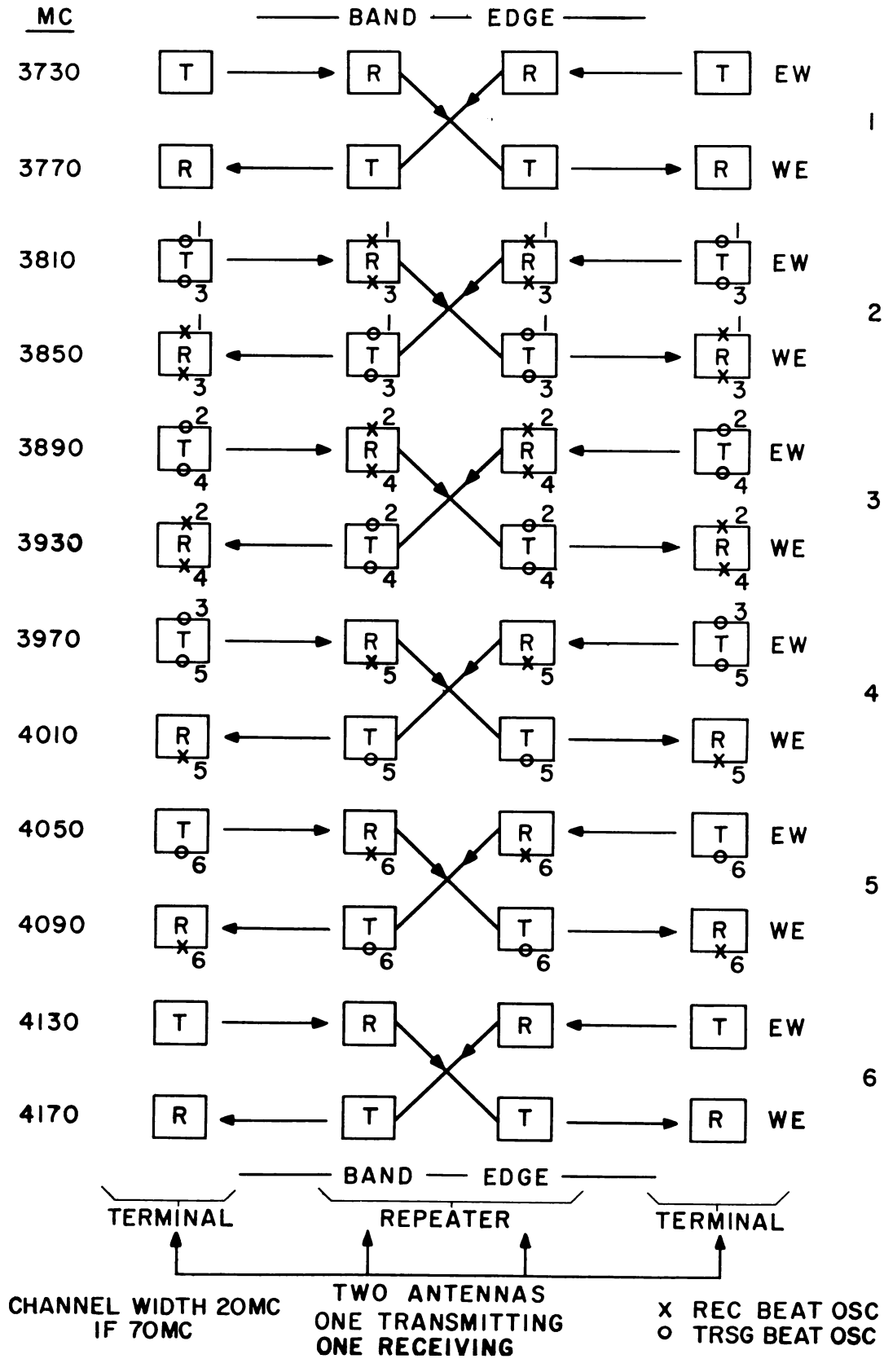
If the frequency allocations have been judiciously selected, "same channel" interference at radio frequencies is rare. It might occur, for example, when we have freak transmission conditions which result in "overreach" from a distant repeater at a time when the disturbed repeater was subject to a moderate fade. "Same channel" crosstalk at IF is, however, a much more probable occurrence, since repeater stations and terminals are rich in 70 mc signals.

The discussion above applies equally well to image channel interference. Other tone interferences may arise from the beating oscillators. Unless the frequency allocation is carefully planned, a beating oscillator frequency for one channel may fall in the band of another microwave channel. An extremely large amount of filtering will then be required to keep the high-level beating oscillator output from "leaking" out of the converter where it is used and getting into the channel at the same frequency. If the IF frequency, the microwave channel frequencies, and the beating oscillator frequencies are selected so that the beating oscillator frequencies fall between the microwave channels considerable advantage may be obtained.

Another source of tone interferences will be briefly mentioned. These are higher order products, possibly 4th or 5th order, which may be produced in a converter if extraneous tones from other microwave channels or other beating oscillators are present. Those products which fall in the frequency band of the desired output then constitute tone interferences. The problem may be difficult to handle analytically, in which case measurements on a given converter may be required to ensure good design. In general, for satisfactory performance we find that all extraneous tones should be at least 20 db below the carrier of the desired signal.

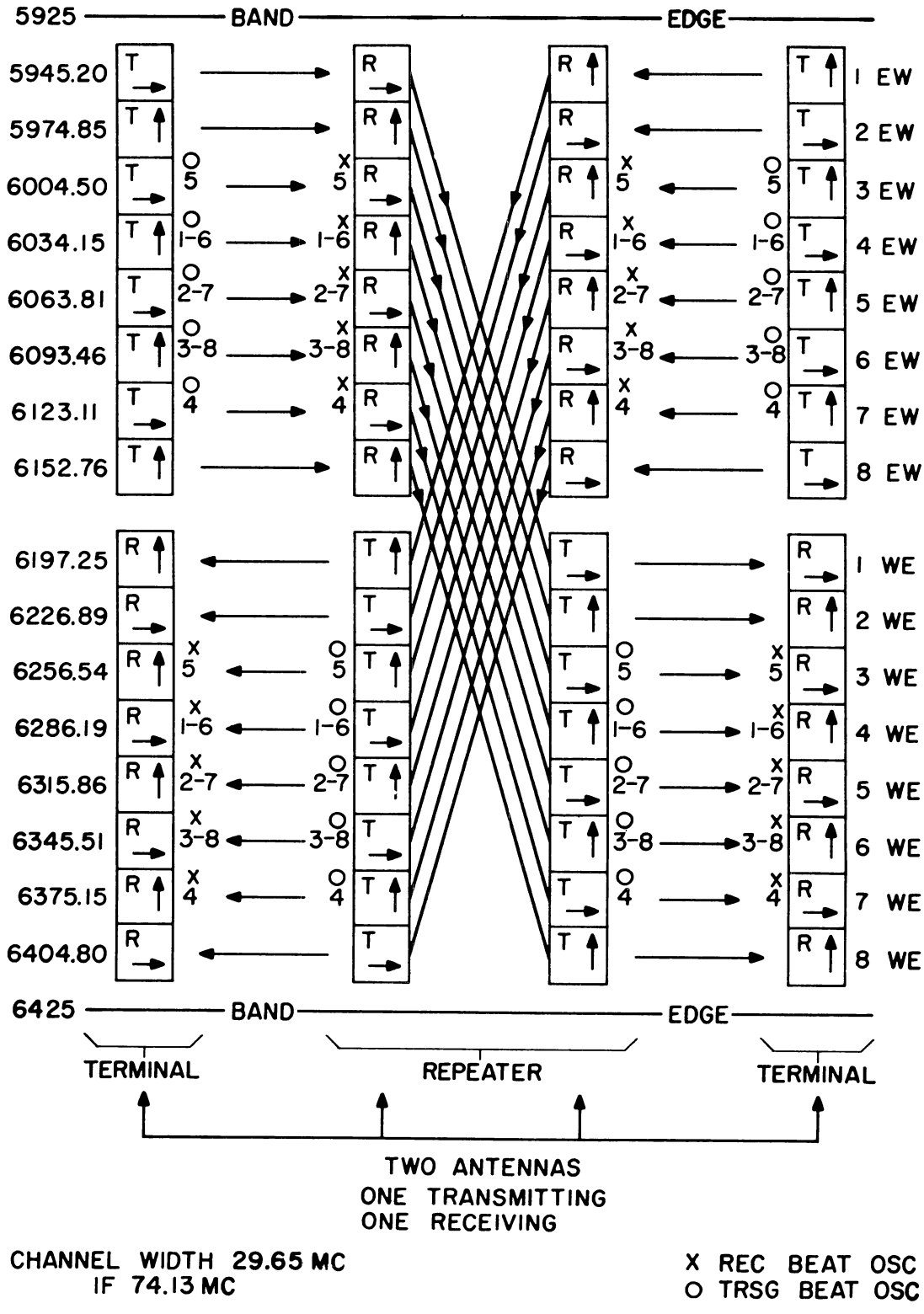
Frequency Allocations for TD-2, TH, and TJ Systems

Since the discussion of frequency allocation has been qualitative throughout it may be well at this point to illustrate how some of the concepts which we have talked about have influenced actual system designs. This will be done with reference to Figures 6, 7, and 8 which show the frequency allocations for the TD-2, TH and TJ Systems. On these diagrams, the vertical scale represents frequency and the height of the small boxes indicates the bandwidth of a single channel. These boxes are arranged in four columns and represent three different physical locations. The center two columns show the channel arrangement and frogging which occurs at a single repeater. The two outer columns represent terminal locations or half of the adjacent repeater. Each box is marked T or R to indicate "transmitting" or "receiving".



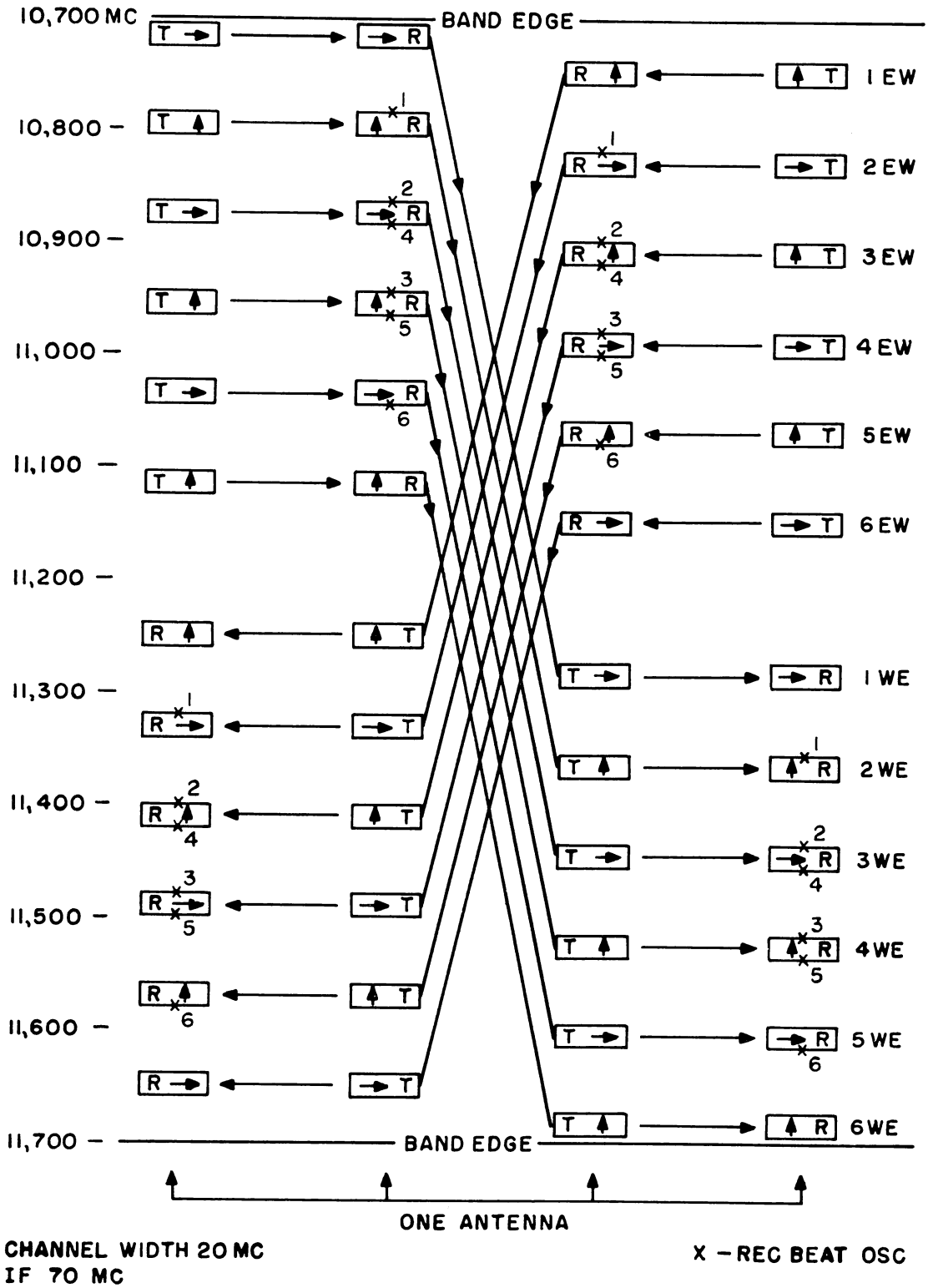
Frequency Allocation Scheme
TD-2 System

Figure 23-6



Frequency Allocation Scheme
TH System

Figure 23-7



Frequency Allocation Scheme
TJ System

Figure 23-8

Small crosses and circles are used to indicate respectively the transmitter and receiver beat oscillator frequencies. For example, in Figure 6 if we look at the box in the upper left hand corner we see that it represents transmission at a carrier frequency of 3730 mc. Since a 70 mc IF frequency is used we need a beat oscillator frequency which is either 3660 or 3800 mc. In the figure we see a small circle at a frequency of 3800 mc marked with a small "1" to indicate that this, rather than 3660, has been chosen as the frequency of the transmitting beat oscillator for Channel 1.

At the bottom of each of these figures, the number of antennas used at a repeater or terminal location (at the points indicated by the arrows) is noted. Thus, we see that the TD-2 system uses four antennas at a repeater location. Separate antennas are used for transmission and reception in each direction.

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 we shall therefore include only brief descriptions. The individual diagrams should be studied and compared.

TD-2 System

Six two-way microwave channels are provided, one of which is intended as a protection channel. Since two frequencies are used for each channel this is a two frequency allocation. Frequency frogging is shown by the diagonal lines in the center of Figure 6. This frogging takes place in two steps. The signal is first brought to 70 mc (IF), amplified, and then put back at the new microwave frequency.

Separate antennas are used for the two directions of transmission. The frequencies of the beat oscillators are indicated by small crosses and circles. Notice that this frequency allocation scheme has kept these frequencies at the edge of the microwave channels and at the same time kept the image channels empty. For example, consider the channel at 3730 mc with its beat oscillator at 3800 mc. The image channel would be at 3870 which is an unoccupied space. On spur routes an allocation is used which places the channels in the slots between channels on the main routes. Placing the beat oscillators at the edge of the channels instead of in the middle of the slots minimizes the chance of interference from these beat oscillators into the spur route channels.

TH System

The TH System has more channels of greater width than the TD-2 System, thereby attaining a more efficient use of the available bandwidth. Adjacent channels are closer together and opposite polarizations for adjacent channels are used to reduce adjacent channel interference. The beat oscillator frequencies are selected so that they fall between the channels. An IF of 74.13 mc was chosen so as to make this possible. Again separate antennas are used for transmission and reception.

TJ System

The TJ System uses a four-frequency allocation and provides space between adjacent channels. This is possible since the TJ System is designed to be for lightly loaded routes and operates at 11,000 mc where a 1,000 mc bandwidth is available. Large signal capacity is therefore not required and the bandwidth can advantageously be used by providing wide separation between channels so as to reduce filter complexity and requirements on frequency stability. A less expensive system is thereby achieved. A single antenna in each direction is used for both transmission and reception.

Chapter 24

ILLUSTRATIVE RADIO SYSTEMS DESIGN PROBLEM

The material studied in the previous chapters is applied to the design of an 11 kmc 250 mile system capable of carrying 100 4kc telephone channels per broadband microwave channel. The allocation of noise and modulation requirements is illustrated, and methods for handling non-linearity in the baseband amplifiers and the FM modulators and demodulators are discussed. The problem serves to review the previous work and illustrate some of the aspects of microwave radio system design.

Introduction

In this chapter we hope to illustrate how the material presented in the previous chapters can be applied to a radio system design problem. Obviously, in order to do so, the problem we pick must be simple. Many considerations of importance in a practical system will have to be omitted or treated in a qualitative manner. Nevertheless we believe the problem will be helpful in tying together some of the separate topics presented in the previous chapters.

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 objectives and at the same time he must try to minimize the overall cost to the ultimate customer. Among other things, the system should be easy to maintain in service and must be 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 one might expect, 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. The value of these preliminary calculations is discussed in some detail at the beginning of Chapter 6 where the emphasis is primarily on AM systems. It will be well worthwhile to read that section again, keeping in mind that although the components and important system parameters may be different for AM and FM systems, many of the design procedures are similar.

With this brief introduction we now consider a simplified problem. In doing so we must realize that any single problem cannot be typical in all respects to all possible designs. Some systems are designed for long routes, perhaps 4000 miles, others are short haul systems. Not all radio systems operate at the same microwave frequencies or require the same capacity. With this in mind the reader should treat this chapter as an illustration of certain procedures which are helpful, rather than as an outline for all system designs.

Statement of the Problem

Assume we are asked to design a low cost radio system to operate in the 11 KMC band given the following specifications.

Maximum length	250 miles.
Capacity	100 single sideband suppressed carrier telephone channels.
Baseband	100 - 500 kc.
Growth	4-6 microwave channels on a route should be planned for.
Dropping Points	This system is planned for operation on routes where the telephone channels will frequently have to be dropped.
Maximum Repeater Spacing	25 miles (See Note 1 below).
Noise Performance	23 dba at -9TL for fades of less than 35 db in a single link; 30 dba at -9TL for a 40 db fade in a single link (see Note 2 below).

Note 1

This specification is based on a study of fading which for the purpose of this problem has already been completed. The results of this study indicate that because fading is more severe in the 11 KMC band than at lower frequencies a maximum repeater spacing of 25 miles should be used.

Note 2

The requirement on noise of 29 dba at the -9TL for a 4000 mile system is often interpreted as follows. If the telephone channels at the end of a 4000 mile system are sampled for noise at the -9TL during the busy hour, the rms value of the samples should not exceed 29 dba. This assumes a distribution and permits some samples to exceed

29 dba. A second point on this distribution is then needed to specify the allowable dispersion of the samples. Typically this is that the noise shall not exceed 40 dba at -9TL for more than .01% of the samples.

The 250 mile end-link system we are considering here must have better noise performance than a 4000 mile system. A decision has been made in this case that end-link systems in tandem should have performance equal to 32 dba. This explains the 23 dba requirement. However, from the previous discussion we know that occasionally more noise can be permitted. In our system this will occur during heavy fades. The second noise requirement has therefore been established based on a study of the frequency of occurrence for fades of various depths.

A Preliminary Calculation

At this point it will be helpful if we proceed with a preliminary calculation of the system noise performance. In order to do so it will be necessary to make certain assumptions regarding the transmitter output power, the antenna gain, the receiver noise figure, and the peak frequency deviation. This can be done by making a study of existing devices and systems. We may assume that some improvement will be possible in some of the devices if we decide to use them in our system. Suppose our study establishes the following facts.

Output power - A klystron is available which has 0.5 watts of output power. A new design might yield a 2 watt output, but some additional development would be required.

Antenna gain - In order to have a low cost system, simple paraboloid antennas appear desirable. At 11 KMC a 4' paraboloid has a gain of 40 db and a beam angle of 1.5°. A larger antenna would have more gain but a smaller beam angle and would probably require a more stable tower, which would make the system more costly.

Receiver noise figure - A noise figure of 15 db seems feasible. By additional development effort this might be reduced slightly.

Frequency deviation - Peak deviations of about 4 mc have been used in other systems, so we might take this as a first guess. This is not a sacred number, however, and some change appears possible if it proves desirable.

Channel Bandwidth - We can make an assumption based on the rule of thumb given in Chapter 19. For a 4 mc frequency deviation and .5 mc top baseband frequency a band of 2 (4 + .5)=9 mc

will be required. We will consider a 10 mc bandwidth for this preliminary calculation.

We now can proceed with our calculation of noise performance using the values arrived at on the preceeding page.

Receiver Input Power

Receiver input power = output power - free space path loss + antenna gains - wave guide losses

Assumed output power = 1/2 watt or 27 dbm

Free space path loss (From Figure 18-5) = 145 db

Antenna gains (Assume two 4' paraboloids) = 80 db

Assumed waveguide losses = 5 db

From which

Receiver input power = 27 - 145 + 80 - 5 = -43 dbm

Calculation of Noise

At this point we shall use the procedure developed in Illustrative Example 2 of Chapter 20 to determine the noise in a telephone channel. We first check to see that we are not in the breaking region. The noise power in a 1 cps band is -174 dbm. For a 10 mc bandwidth and a 15 db noise figure the total noise power becomes

$$\text{Noise power} = -174 \text{ dbm} + 10 \log 10^7 + 15$$
$$= -89 \text{ dbm}$$

The ratio of the signal power to the noise power in the absence of a fade is

$$S/N = -43 - (-89) = 46 \text{ db}$$

If we now assume that the breaking region starts when the S/N ratio becomes less than 10 db we will not be in the breaking region for fades of less than 36 db. With this established we now proceed with the calculation of noise in the worst channel.

Frequency Deviation Produced by Noise

At this point an examination of the index of modulation would show that it is greater than unity and that the carrier component of the signal is small compared to the total signal. Nevertheless we shall complete the noise calculation just as we would for a low index system.

This will simplify the work and will introduce very little error in the final result. Hence, we use the result of the receiver input power calculation.

Carrier power = -43 dbm

The noise power in a 1 cps band is -174 dbm plus 15 db due to the noise figure.

Noise power in 1 cps band = -159 dbm

The rms phase deviation produced by two one cycle bands of noise, one ω_n radians per second above and the other ω_n radians per second below the carrier is $a_n\sqrt{2}/A_c$ radians where a_n = rms noise voltage in a one cycle band and A_c is the peak carrier voltage.

$$20 \log \frac{a_n\sqrt{2}}{A_c} = 20 \log a_n - 20 \log \frac{A_c}{\sqrt{2}}$$

$$= -159 - (-43)$$

$$= -116 \text{ db with respect to 1 radian}$$

At the output of the discriminator the noise due to the two 1 cycle bands will both fall in a 1 cycle band at ω_n and will therefore be produced by the phase deviation computed above. The total noise in a 3 kc band will be the sum of the powers in 3000 such 1 cycle bands.

rms phase deviation in 3 kc bandwidth $= -116 + 10 \log 3000$

$= -81$ db with respect to 1 radian

$= .9 \times 10^{-4}$ radians

The rms frequency deviation in a 3 kc band is given, to a close approximation, by the product of the phase deviation and the center frequency of the 3 kc band. Since we will be interested in the noisest channel we will consider the frequency deviation in a 3 kc band at the top bansband frequency.

rms frequency dev. in 3 kc band at .5 mc $= .9 \times 10^{-4} \times .5 \times 10^{6} = 45$ cps

Determination of P_s

We now find the value of P_s which is the power in dbm at OTL of the maximum sine wave which the system must handle without overloading. For $V_o = -12.5$ vu, $\sigma = 5$ vu, and $N_a = .25 \times 100$, the average power of the telephone signal during the busy hour at zero level is given by Equation (12-2) as

$$\text{avg power} = V_o + .115\,\sigma^2 + 10 \log N_a - 1.4 \text{ dbm}$$

$$= -12.5 + 2.9 + 14 - 1.4 \text{ dbm}$$

$$= 3 \text{ dbm}$$

The value of P_s is obtained by adding to the average power a peak factor which is given in Figure (12-2).

$$P_s = 3 + 15 = 18 \text{ dbm}$$

Noise in dba

The maximum frequency deviation for the system considered here is 4 mc. If this were produced by a sinusoidal signal the rms frequency deviation would be $2\sqrt{2}$ mc. Thus, a signal with 18 dbm of power at OTL is equivalent to an rms frequency deviation $2\sqrt{2}$ mc. Conversely, the noise power in a 3 kc band at OTL in the top channel is therefore

$$\text{Noise power in 3 kc band} = 18 \text{ dbm} + 20 \log \frac{45 \text{ cps}}{2\sqrt{2} \times 10^6 \text{ cps}}$$

$$= 18 - 96 =$$
$$= -78 \text{ dbm at OTL}$$
$$= -87 \text{ dbm at -9TL}$$
$$= -5 \text{ dba at -9TL}$$

If we now ignore any possible differences in link lengths and assume ten 25 mile links in a 250 mile system, the total thermal noise becomes

$$\text{Total noise} = -5 \text{ dba} + 10 \log 10$$
$$= 5 \text{ dba at -9TL}$$

This meets our requirement with considerable margin. We now examine the effect of a fade in a single link.

Effect of Fading in a Single Link

A 35 db fade in a single link would bring the noise from that link from -5 dba up to 30 dba. Under this condition the noise contribution from the other links becomes negligible. However, 30 dba is 7 db above the requirement of 23 dba. A check of the S/N ratio shows that for a 35 db fade the S/N ratio is reduced from 46 to 11 db which is very close to the breaking region. A 40 db fade would put us well into the breaking region. This is shown on Figure 1 by the solid line. The two requirements are indicated by the small squares.

We must now consider the possible changes which we might make so as to improve the performance. A 7 db improvement would just meet the 23 dba requirement but would not permit a share of the requirement to be given to modulation noise. We shall therefore attempt to attain about 10 db of improvement for a 35 db fade condition. This will also

give us additional margin against "breaking", which now occurs for a 36 db fade. Possibilities we might consider are pre-emphasis, an increase in the frequency deviation, or an increase in the S/N ratio at the receiver input. The latter might be accomplished by increasing power output, a reduced noise figure, or by an increase in antenna gain. Suppose we now consider these possibilities individually.

(1) Pre-emphasis could be used to obtain phase modulation. An improvement of about 4 db might be attained in the output noise. The breaking point would still occur for a 36 db fade.

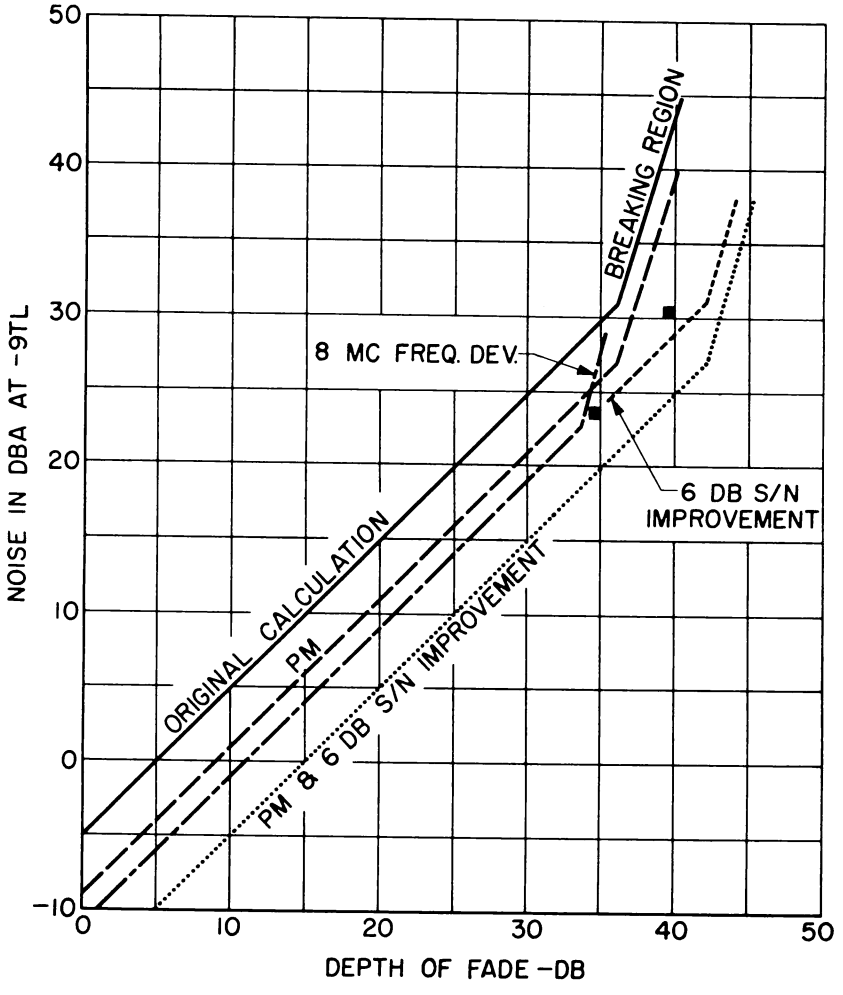
(2) An increase in the frequency deviation to 8 mc would increase the necessary channel bandwidth from 10 mc to 18 mc. The breaking point would be lowered by 10 log (18/10) = 2.5 db and would occur for a 33.5 db fade. In this respect, therefore, we would have degraded the system. For fades less than this, however, a noise improvement of 6 db would be attained.

(3) A 6 db increase in the S/N ratio at the receiver input would improve the noise performance by 6 db when we are below the breaking point, and would also change the breaking point from a 36 db fade to a 42 db fade.

The effect of these changes is shown by the dash lines in Figure 1. The use of an increased frequency deviation does not seem desirable. It requires a wider bandwidth and moves the breaking point in the wrong direction. A combination of pre-emphasis to obtain phase modulation and a 6 db improvement in the S/N ratio at the receiver input would allow us to meet our requirements. To get a 6 db S/N improvement we could change the output power of the klystron from .5 watts to 2 watts, increase the gain of each antenna by 3 db, improve the noise figure by 6 db, or use some combination of these parameters. We probably would not be able to attain 6 db improvement in the noise figure, and higher gain antennas would probably require a more rigid tower. It therefore seems desirable to attempt the development of a 2 watt klystron.

Combining the advantages of using phase modulation and the successful development of a 2 watt klystron, our thermal noise performance can be summarized as follows:

Fade	Thermal Noise at -9 TLP
0 db	-5 dba (10 links contribute)
35 db	20 dba (1 link dominates)
40 db	25 dba (1 link dominates)



Noise vs Fade for Various Design Choices

Figure 24-1

Frequency Allocation

We are now considering a low cost system for use in the 11 KMC band. A 1000 mc bandwidth is available and only moderate growth has to be provided for. Our situation is similar to that of the TJ System for which a frequency allocation was discussed in the previous chapter. Our channel width can be less because we are not asked to transmit a television signal. We need to provide for approximately the same amount of growth. Without going into more detail we shall assume that we can use a frequency allocation very similar to that for the TJ System shown in Figure 23-8.

Baseband vs IF Repeaters

For a short haul system such as this the relative advantages and disadvantages of baseband vs IF repeaters should be considered. Since frequent drops are required many of the repeaters will have to be terminal points. This is a strong point in favor of baseband repeaters since every baseband repeater consists of a pair of terminals. In addition, the transmitting terminal of a baseband repeater type system is simpler in that it does not require the carrier supplies and microwave frequency amplifiers used in IF repeater systems. These factors make the baseband repeater desirable.

On the other hand the baseband repeater system has certain disadvantages. The FM wave is demodulated at each repeater whether or not the baseband signal is needed at that point. The signal degradation is therefore greater than in a system with IF repeaters. However, this is not necessarily serious for short systems. Other disadvantages include the fact that frequency control is likely to be poorer in baseband systems and that the modulation index may be subject to greater variability since it may be changed at each repeater point.

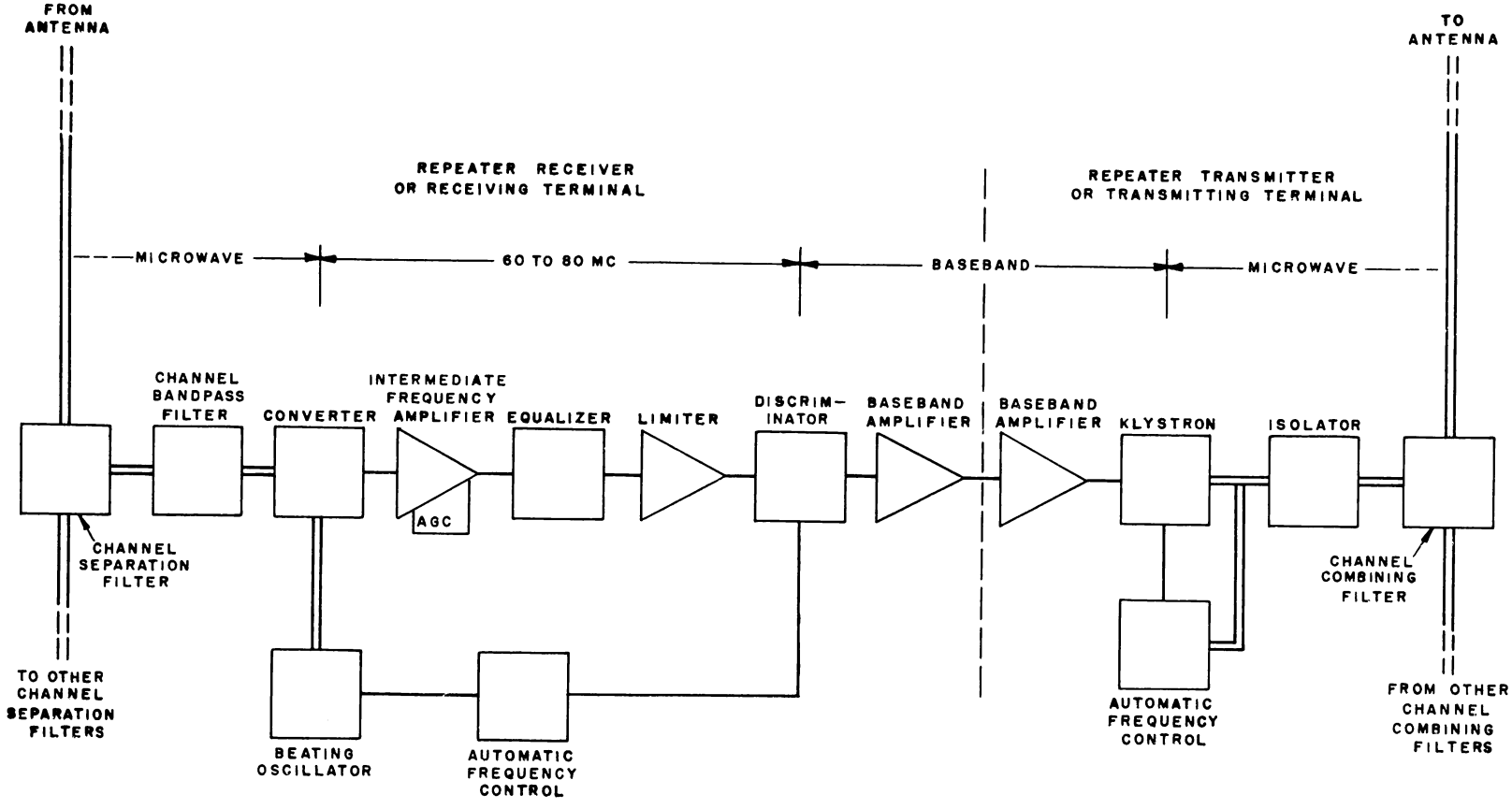
From these considerations we conclude that baseband repeaters can be used in short haul systems. Some decrease in cost and some increase in signal degradation may be expected. For the purpose of our problem we shall assume that the reduction in cost is of primary importance and that we shall use baseband repeaters. The block diagram of our repeater would probably closely resemble the typical baseband repeater shown in Figure 2.

Modulation Noise

The next topic we shall consider is that of modulation noise. Typically this may be caused by transmission deviations in the FM portions of the system, non-linear FM modulators and demodulators, and (just as in AM systems) by non-linear baseband amplifiers. Other sources are possible and should be examined in an actual system design*. For illustrative purposes we shall restrict ourselves to those listed above.

- -

*For example, we have ignored the effects of imperfect limiting, whereas in fact limiters never completely eliminate the amplitude modulation caused by imperfect modulators, and by transmission deviations. It might be argued that this effect is included in the list above, by implication. True, but it is an aspect which might have been overlooked. A fundamental feature of transmission system design is involved here. Of course we pay major attention to obviously major effects - but we can never be confident that we have taken note of all the subtle ways in which imperfections and interferences can arise - and often such an overlooked effect turns out to be quite serious.

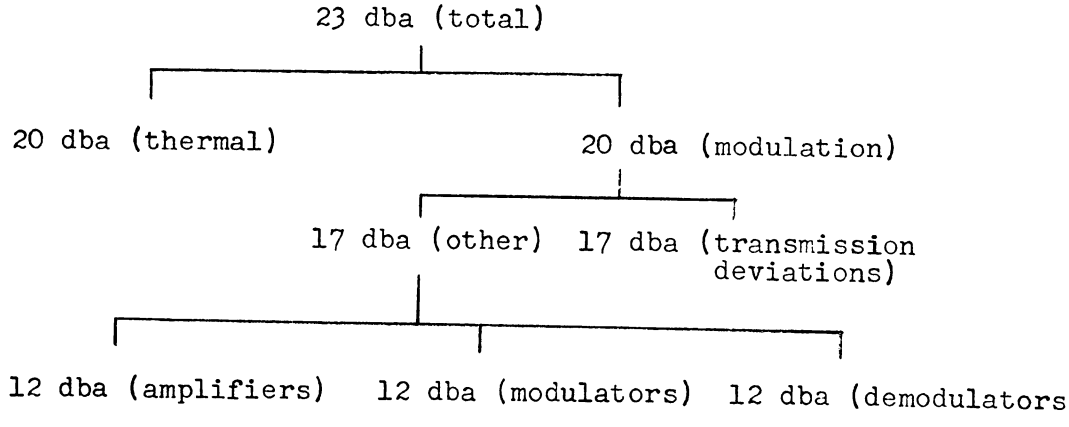


Block Diagram of Typical Baseband Repeater

Figure 24-2

For a 35 db fade, the total noise requirement is 23 dba at -9TL. Since we expect only 20 dba from thermal noise, another 20 dba can be allotted to modulation noise. This in turn must be divided among the several contributors listed above. The actual division should be made in accordance with the difficulty with which the individual requirements can be met. In a preliminary calculation such as this a tentative division often has to be made, subject to later revision as the separate problems are investigated. Suppose in our case we expect transmission deviations to be the most severe problem and that the others are about equally difficult. We might then make a tentative division as follows where we assume that the contributions from the separate sources will add randomly.

Tentative Division of Noise Requirement



Each of these requirements is for a complete 250 mile system which typically might contain ten 25 mile links. From them we want to get requirements on the individual amplifiers, modulators, demodulators, and transmission shapes. We shall consider them separately.

Amplifiers

The transmitting and receiving terminals each contain a baseband amplifier, making a total of 20 for a complete system. These 20 amplifiers may be considered as an AM system with a total requirement of 12 dba. The methods developed in the first half of this course may then be applied to determine the required second and third order modulation performance of each of the amplifiers.

Transmission Deviations

In Chapter 22 a problem was considered in which the noise in dba caused by a particular transmission deviation was calculated. Here the problem is just the opposite. We are given a requirement on noise in dba and must find out how good the transmission quality has to be. Obviously, the same methods can be used. We might start by allocating the total requirement for transmission deviations among the various portions of the system through which the FM signal is transmitted. A further division can be made to the particular transmission shapes such as parabolic phase and cubic phase terms. Each of these can then be treated as separate problems. If some of the required transmission shapes appear more difficult to realize than others a new division of the requirement can be made which takes this difficulty into account by giving that shape a larger portion of the total requirement.

FM Modulators and Demodulators

So far in this course we have been relatively unconcerned about practical devices used in radio systems. We have talked of ideal FM modulators and demodulators just as we often talk of ideal amplifiers. However, in previous chapters we have indicated that actual amplifiers are not linear and have developed methods for analyzing their performance in AM systems. Similar methods can be developed to handle the non-linear modulators and demodulators (e.g., klystrons and discriminators) actually encountered in FM system design. In many cases these methods are almost identical to those used for amplifiers or for transmission deviations. For example, the non-linear characteristic of an FM modulator can often be expressed by a power series as

$$\varphi'(t) = a_1 V(t) + a_2 V^2(t) + a_3 V^3(t) \ldots$$

where

$\varphi'(t)$ = instantaneous frequency deviation
$V(t)$ = baseband signal applied to modulator
a_n = coefficient of the nth order modulation term

Such an equation can be obtained from either a theoretical knowledge of the modulator performance or by measurements. In the latter case the equation is the result of fitting a curve to steady state measurements of the output frequency for various values of dc voltage applied to the modulator input. A word of caution is in order here, however. One would expect such a "steady state" equation to be

valid when V(t) is allowed to vary at a slow rate. It does not necessarily follow, however, that the instantaneous frequency deviation can be represented by the same equation when V(t) contains high frequency baseband components. For instance, we might expect "sluggishness" or undesirable transient effects. This is a problem which must be investigated when any particular FM modulator is under study. However, in many cases it can be shown that all of the important modulation terms are contained in the "steady state" equation for non-linearity. When this is the case, the non-linearity of an FM modulator can be treated in exactly the same manner as the non-linearity of a baseband amplifier.

The FM demodulator can, in many cases, be handled in a similar manner. In order to illustrate the procedure we shall consider a very simple type of demodulator. We shall assume that the FM wave,

$$f(t) = \cos [\omega_c t + \varphi(t)],$$

is applied to a pentode such that the plate current is

$$i(t) = i_o \cos [\omega_c t + \varphi(t)]$$

We now assume that the load impedance for the tube can be expanded in a power series.

$$Z(\omega) = R_o[1+g_1(\omega-\omega_c)+g_2(\omega-\omega_c)^2+g_3(\omega-\omega_c)^3]e^{j[b_2(\omega-\omega_c)^2+b_2(\omega-\omega_c)^3.]}$$

The voltage across the load impedance is given as follows

$$e(t) = F^{-1} (Z(\omega) F[i(t)])$$

$$= i_o R_o[1+P(t)] \cos [\omega_c t + \varphi(t) + Q(t)]$$

where P(t) and Q(t) are given in Table (22-1). Whereas in the FM portions of the system we usually neglect P(t) and concern ourselves with Q(t), here we do the opposite. This becomes obvious from the following considerations. The amplitude modulation P(t) may be written as

$$P(t) = g_1\varphi'(t) + g_2\varphi'(t) + g_3 [\varphi'^3 (t) - \varphi'''(t)]$$

$$+b_2\varphi''(t) + 3b_3\varphi'(t)\, \varphi''(t) + 3g_1 b_2\, \varphi'(t)\, \varphi''(t)$$

where in an FM system $k_1\varphi'(t) = V(t)$. Hence, the first term in P(t) above is proportional to the baseband signal. This suggests that an FM demodulator can be constructed by making the constant, g_1, large so

that we get a large amount of amplitude modulation proportional to the baseband signal. An envelope detector can then be used to recover the baseband signal. The other terms in P(t) represent distortion terms which can be evaluated by the methods described in Chapter 22.

The Final Stages of Design

We have now arrived at a point where we can turn our attention to the individual blocks in the block diagram. Input and output levels and linearity requirements can be specified. This permits development work on these parts to proceed toward specific goals. As this work progresses the overall system design will from time to time have to be reviewed to make sure that the various assumptions which have been made are still valid. Some changes in the division of requirements or in the input and output levels may be desirable as the final design is gradually achieved.

Chapter 25

THE PHILOSOPHY OF PULSE CODE MODULATION SYSTEMS

The use of digital instead of analog methods of transmission offers great advantages. In digital systems, the message is coded into a signal which can be regenerated rather than merely amplified at each repeater. Although a channel of much greater bandwidth is required, such systems can operate with a low signal-to-noise ratio in the channel, since the channel noise does not reach the subscriber. Transmission by pulse code modulation involves sampling, quantization, coding, time division multiplex transmission, recognition, regeneration, and, ultimately, decoding. This chapter introduces these ideas as a preliminary to the more detailed discussion found in subsequent chapters.

Introduction

In the past chapters attention has been directed toward various amplitude, frequency, and phase modulation systems. All of these use a sinusoidal carrier whose amplitude, phase, or frequency is continuously varied in accordance with the modulating function.

In this and the following chapters, we shall discuss the use of a series of pulses instead of a sinusoidal carrier to carry the information contained in the modulating function. There are many possible ways of modifying the characteristics of a train of pulses to convey information. One can vary pulse amplitudes so that the envelope follows the modulating function (Pulse Amplitude Modulation, or PAM). Alternatively, one could vary pulse position (Pulse Position Modulation, or PPM) or the pulse duration (Pulse Duration Modulation, or PDM). Another interesting method which may be important in the future is delta modulation. Delta modulation, as well as PPM and PDM, are briefly described in the Appendix to this chapter. In this and the following few chapters, however, we shall examine the advantages to be obtained from, and the problems associated with, the particular type of pulse transmission known as Pulse Code Modulation. But before we go into any detail on PCM, it will be profitable to define some of the general ideas and vocabulary involved.

The development of PCM has been closely associated with the advent of information theory, which for the first time has given us a clear way of thinking about the fundamentals of our job. One of the important ideas involved is that of sampling.

Sampling

Our basic job is to transmit a "message", which we can think of as a voltage which varies continuously with time. This is our modulating function. In AM or FM systems, as we have seen, we continuously

vary the carrier in accordance with the modulating function. The first point to recognize when we consider the feasibility of pulse modulation is that the continuous transmission of information about the modulating function is unnecessary.

In any physically realizable transmission system the message or modulating function is limited to a finite frequency band. Such a function can assume only a finite number of independent values in a finite time. In fact, if the highest frequency component is f_c cycles per second, the time function cannot assume more than $2f_c$ independent values per second. For this reason the amplitudes at any set of points in time spaced τ_c seconds apart, where $\tau_c=1/2f_c$, specify the message completely.* Hence, to transmit a band-limited message** of duration T, we do not need to send the entire continuous function of time. It suffices to send the finite set of $2f_cT$ independent values obtained by sampling the instantaneous amplitude of the signal at a regular rate of $2f_c$ samples per second. (If it surprises the reader to find that $2f_cT$ pieces of data will describe a continuous function completely over the interval T, it should be remembered that the $2f_cT$ coefficients of the sine and cosine terms of a Fourier series do just this, if, as we have assumed, the function contains no frequencies higher than f_c.)

The basic theorem which we have been discussing here is called the sampling principle which, in a restricted form, states:

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

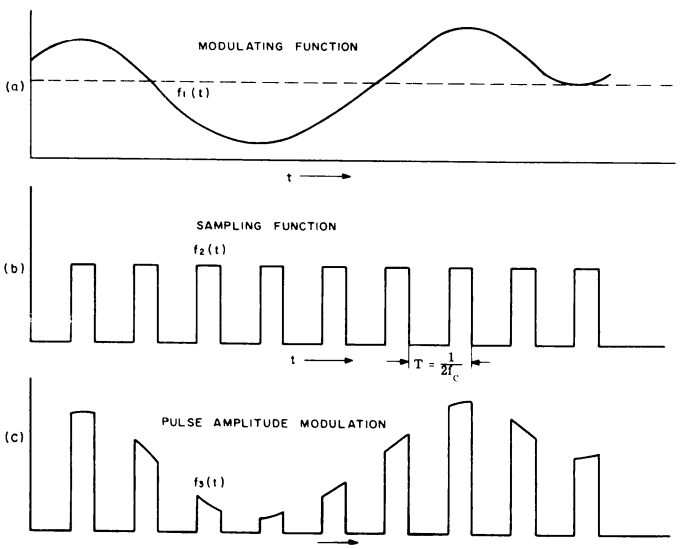
The application of the sampling principle reduces the problem of transmitting a continuously varying message to one of transmitting a finite number of amplitude values.

---

*A simple proof of this is given in the Appendix to Monograph B1611. This paper, by Oliver, Pierce and Shannon, which originally appeared in the Proceedings of the Institute of Radio Engineers, is an excellent discussion of the philosophy of PCM, and much of this text has been abstracted from it.

**A useful distinction is made in this chapter, as well as the ones to follow, between the terms "message" and "signal". We define a signal as a coded (AM, FM, or PCM, for example) representation of the message we really have to deliver.

The process of sampling is illustrated in Figure 1. The function $f_1(t)$ illustrated in Figure 1(a) is assumed to contain no frequencies above f_c. Figure 1(b) shows a sampling function $f_2(t)$. The sampling frequency is $2f_c$. The result of sampling is shown in Figure 1(c). This function, $f_3(t)$, is defined analytically as the product $f_1(t)\ f_2(t)$ and is a form of pulse amplitude modulation. Note that this is not instantaneous sampling, since the $f_2(t)$ pulses have duration. Instantaneous sampling can, of course, never be realized in a physical system. The effects of sampling for non-zero time (so-called "natural sampling", as against instantaneous sampling) are discussed in the next chapter.



Sampling

Figure 25-1

Reconstruction

Let us now proceed to the receiving end of the system. The PAM signal, $f_3(t)$, may be transmitted to the receiver in any form which is convenient or desirable from the transmission standpoint. At the receiver the incoming signal is then operated on to recreate the original PAM sample values so that they appear in their original time sequence at a rate of $2f_c$ pulses per second. To reconstruct the message it is merely necessary to generate from each sample a proportional impulse, and to pass this regularly spaced series of impulses through an ideal low-pass filter of cutoff frequency f_c. Except for an over-all time delay and possibly a constant of proportionality, the output of this filter will then be identical to the original message.

Ideally, then, we could achieve perfect reproduction of a message if we could transmit information giving us exactly the instantaneous amplitude of the message at intervals spaced $1/2f_c$ apart in time.

Quantization

It is, of course, impossible to transmit the exact amplitude of a modulating function. In conventional AM or FM, or in a system using pulse height or position to carry information, some error will always occur. Noise, distortion, and crosstalk will affect the modulated wave so that the recovered message will not exactly duplicate the information in the original message. In the systems noted - AM, FM, PAM, or PPM - the error increases as we go through successive repeater sections, since additional noise is added to the signal as it passes through each repeater section.

This situation is analagous to the accumulation of small errors in a long series of slide-rule operations. It suggests the weakness of an analog method of transmission in which the transmitted signal can assume a continuum of values. But suppose we consider instead a digital system. Instead of attempting the impossible task of transmitting the exact value of a sample, let us limit ourselves to certain discrete amplitudes of sample size. Then, when the message is sampled, the amplitude nearest the true amplitude is sent. When this is received and amplified, it will have an amplitude a little 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, we can tell accurately which discrete amplitude the signal was supposed to have. Then the signal can be reformed, or a new signal created, which again has the amplitude originally sent.

Representing the message by allowing only certain discrete 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. By quantizing we limit our "alphabet". If the received signal lies between a and b, and is closer (say) to b, we guess that b was sent. If the noise is small enough, we shall always be right.

- -

*The magnitude of this noise, and the use of instantaneous compandors to reduce it, will be considered in more detail in the next chapter.

Coding

A quantized sample could be sent as a single pulse which would have certain possible discrete amplitudes, or certain discrete positions with respect to a reference position. However, if many allowed sample amplitudes are required, one hundred, for example, it would be difficult to make circuits to distinguish these, one from another. On the other hand, it is very easy to make a circuit which will tell whether or not a pulse is present. Suppose, then, that several pulses are used as a code group to describe the amplitude of a single sample. Each pulse can be on (1) or off (0). If we have three pulses, for instance, we can devise a code to represent the amplitudes shown in Table I.

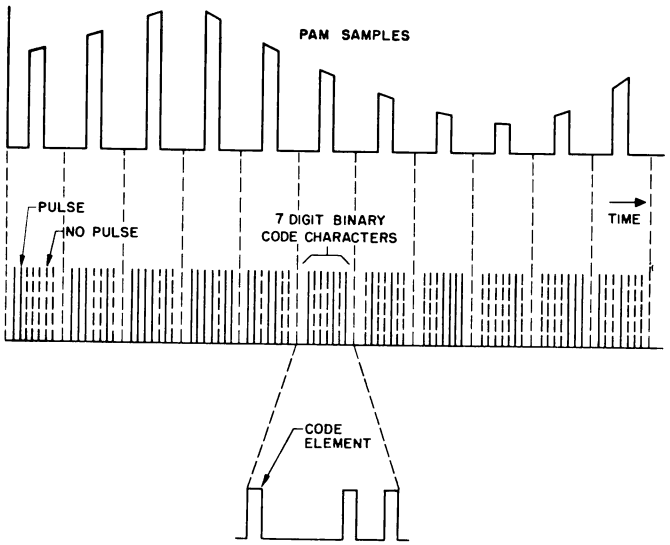
TABLE I

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

These codes are, in fact, just the numbers (amplitudes) at the left written in binary notation. In this notation, the place-values are 1, 2, 4, 8, -; i.e., a unit in the righthand column represents 1, a unit in the middle (second) column represents 2, a unit in the left (third) column represents 4, etc. 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. Figure 2 illustrates the coding of a PAM signal into seven digit code.

It is possible, of course, to code the amplitude in terms of a number of pulses which have allowed amplitudes of 0, 1, 2 (base 3 or ternary code), or 0, 1, 2, 3 (base 4 or quaternary code), etc., instead of the pulses with allowed amplitudes 0, 1 (base 2 or binary code). If ten levels were allowed for each pulse, then each pulse in a code group would be 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. As we shall see, however, binary code (0,1) seems to offer the most advantages, and present development is proceeding along binary lines.



Binary Pulse Coding

Figure 25-2

Decoding

To decode a code group of the type just described, one must generate a pulse which is the linear sum of all the pulses in the group, each multiplied by its place value (1, b, b^2, b^3,. . .) in the code. This can be done in a number of ways. For example, we might mention what is perhaps the simplest way which has been used. This involves sending the code group with "units" pulse first, and the pulse with the highest place value last. The pulses are then stored as charge on a capacitor-resistor combination with a time constant such that the charge decreases by the factor 1/b between pulses. After the last pulse, the charge (voltage) is sampled. Such a method, while feasible, has the disadvantage that the most significant digit is in its fastest decay period when used, leading to large errors. More advantageous types of decoders are discussed in the next chapter.

Time Division Multiplex

In an earlier chapter, as part of our discussion of applications of the Fourier Transform Pair, it was pointed out that there is an optimum rate for the transmission of short pulses through a band-limited medium. We found then that for a low-pass characteristic

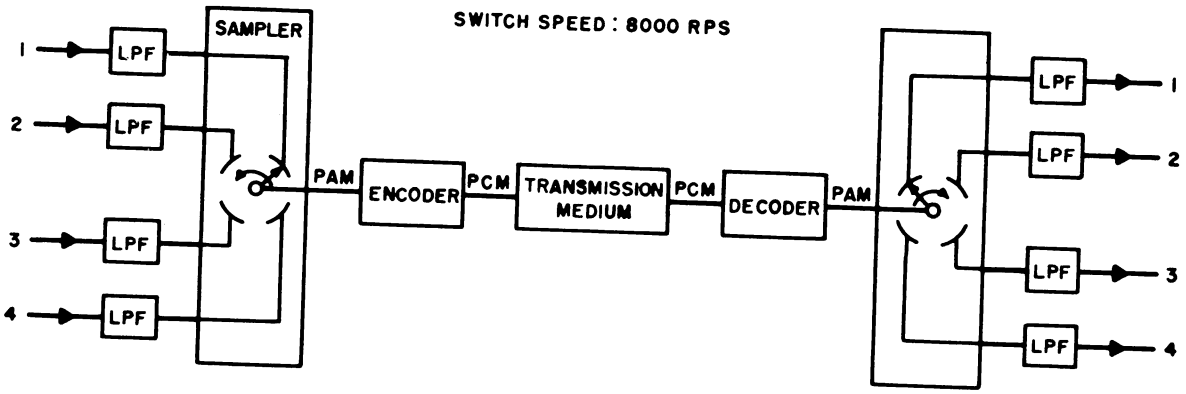
which transmits up to some frequency f_1 cps, we can send $2f_1$ pulses per second. Thus a 750 kc channel could carry 1.5 million pulses per second. Consider the transmission of 4 kc telephone messages by 8 digit* binary PCM over a channel which has a bandwidth of 750 kc/s.

We have seen that our sampling rate should be 8000 times per second or one sample per 125 microseconds. Each sample will result in one code character consisting of eight code elements (1's or 0's). If we can send 1.5 million pulses per second, eight pulses can be sent in 5.33 microseconds. If we send only the information pertaining to one message, our pulse pattern vs time would consist of an 8-pulse character, taking 5.33 microseconds, then idle time for about 120 microseconds, followed by another 5.33 microseconds of use, and so on. Obviously, we are not using our channel very efficiently. On the other hand, if we send code characters from other channels during our idle time we ought to be able to transmit, not one, but about 24 telephone messages over our 750 kc channel. Interleaving signals on a time basis in this way is called Time Division Multiplex.

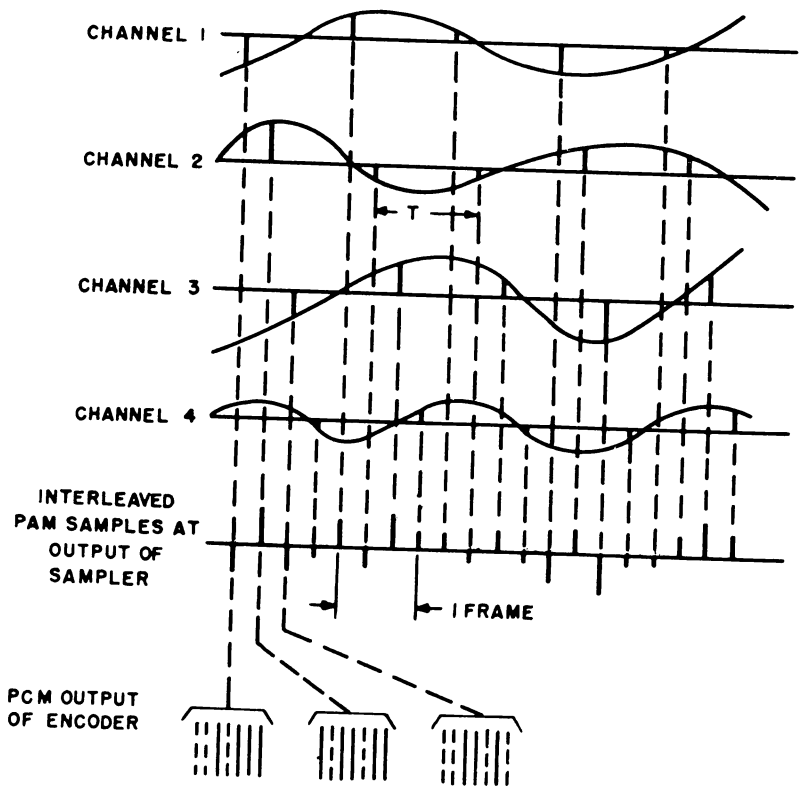
An illustration of a time division multiplex PCM system is shown in Figure 3. Although a four channel system was chosen for convenience, the concepts associated with this system pertain equally well to a system involving any number of channels. Figure 3(a) shows four messages in the form of time-varying voltages which are to be transmitted over a common channel. Each message is band-limited to 4 kc by a low-pass filter in the message path. Our problem is to sample each message, interleave the resulting PAM signals, and, finally, encode the PAM signals into binary PCM. For purposes of graphically illustrating the principle, the sampling mechanism is shown as a switch rotating at the required 8000 cycles per second sampling rate. Such a switch combines the functions of sampling and interleaving. Figure 3(b) illustrates the interleaved PAM samples so obtained at the sampling circuit output. These samples are fed into an appropriate encoder circuit which produces a seven-digit PCM encoding of the incoming PAM signals. The encoder output is shown in Figure 3(b). At the receiving end the inverse functions of decoding and sample separation are performed, as illustrated by Figure 3(a).

---

*Actually we are thinking here of seven digits to represent the message sample, and an eighth pulse for supervisory and signalling purposes.



(a) ELEMENTS OF PCM SYSTEM



(b) FORMATION OF PAM AND PCM SIGNALS

Time Division Multiplex

Figure 25-3

Some useful terms can be defined from Figure 3. The sampling interval, T, shown for message channel 2, is the time between successive samples of the message voltage in a channel. A frame represents the output obtained from one complete cycle of the sampling circuit. The frame interval is equal in duration to the sampling interval and is the time required to obtain one complete frame. The frame rate is the reciprocal of the frame interval and is equal to the rate at which the frames are generated, which, of course, equals the sampling rate. Thus, for the system illustrated, the frame rate is 8000 frames per second, and the frame interval (or sampling interval) equals 125 microseconds. Notice, too, that the time interval for each code character (representing a sample from each message channel) must be equal to or less than one-fourth of the frame interval, if four message channels are to be accomodated.

Time division implies switching at precisely fixed times in order to separate the messages at the receiver. Therefore, additional time, within the frame interval, is usually allocated to some sort of timing or gating signal.

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 frame 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 in a frame is increased. This means that frequency limitations of the transmission medium inevitably restrict the number of message channels which can be included in a frame.

In general terms: if f_c cycles per second is the highest frequency in our message, and n the number of code elements per code character, then we require approximately nf_c cycles per second of bandwidth per message, plus an allowance for gating time. This is n times the bandwidth required for direct transmission or for single sideband AM. In terms of our preceding discussion, we used n=8, and found 24 channels consistent with a 750 kc bandwidth. This is about eight times the 96 kc bandwidth required for the single sideband AM transmission of 24 channels.

Summary

Thus far we have merely described some interesting possibilities. We have pointed out that with no penalty we can sample instead of transmitting continuously. Also, we have recognized that since we will always have errors anyway, we may gain some advantage by

taking one error in the beginning -- the error we get by quantizing or rounding off -- instead of accumulating errors in each repeater section. Finally we saw that once we have quantized, we can code -- and, incidentally, interleave code symbols from a number of channels on a time division multiplex basis.

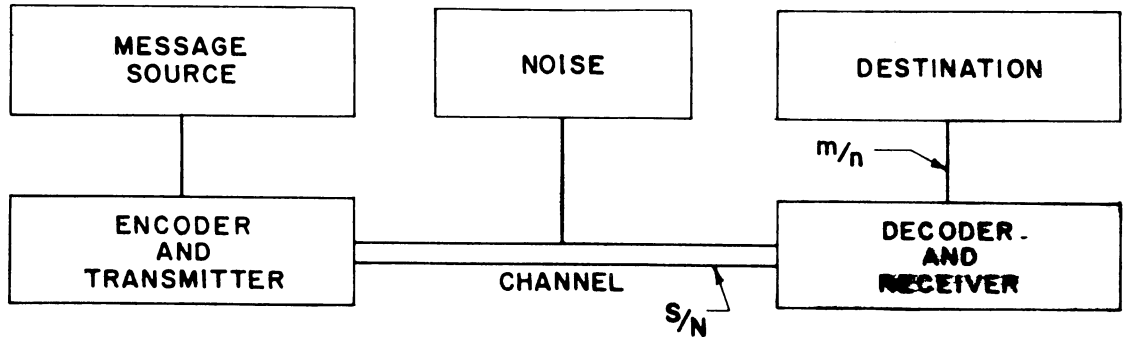
All these ideas, however, seem to have merely resulted in a communication system which uses much more bandwidth than we used before. This brings us to stressing one point that has been mentioned glancingly, but which deserves more emphasis: -- the tremendous difference between code signals and continuous signals, or between analogue (slide-rule) and digital (computing machine) techniques. This lies in the fact that it is possible to regenerate a fresh, noise-free symbol in a repeater. Analogue signals can be amplified -- along with any accrued noise. Digital signals can be recreated, and the noise accrued in the previous line section need not be transmitted.

We have not yet shown that these are anything more than interesting ideas, or that an improved or more economical method of transmitting telephone messages is inherent in them. The final answer depends on our success in developing practical circuitry, but first we should ask whether the ideas would theoretically lead to an improved system. To answer this question we need to examine transmission systems in general terms.

Parameters of Transmission Systems

A block diagram of a general transmission system is shown In Figure 4. The message source in this system produces information in the form of an electrical wave. The message may be a continuous function of time, as for example, the output of a telephone transmitter, or it may itself be discrete, as in data transmission. The function of the system is to transmit this message to the destination with the required fidelity. In order to accomplish this function, the message is coded into a signal which is then sent over the channel. The coder may consist merely of a simple amplitude, frequency, or pulse modulator or it may be a more complex scheme in which a continuous message is converted to a discrete form. In general, the bandwidth and complex spectrum of the signal will be different from those of the message. The receiver consists of a decoder which, in a typical case, may be a demodulator and/or a circuit for converting the incoming discrete signals into a form for use by the destination. At the output

of the decoder, therefore, the message reappears -- generally speaking, it may be in either discrete or continuous form but typically it has the same form as the original.



Generalized Transmission System

Figure 25-4

The point to be emphasized is that during its passage through the channel, the signal does not, in general, have the form of the message. This leads us to draw an important distinction between "signal-to-noise" ratio and "message-to-noise" ratio. The rms noise and signal in the channel determine the channel signal-to-noise ratio, an important parameter of the system. In the case of discrete signals, such as a series of binary pulses, a more convenient parameter is the error probability, P_e. This expresses the probability that an incoming pulse will be misinterpreted by the receiver, so that (for binary pulses) no pulse would appear at the decoder output where a pulse was present in the original message. Factors such as channel noise, timing errors, and transmission deviations are important in determing P_e. The parameter at the decoder output which corresponds to the channel signal-to-noise ratio is the message-to-noise ratio. It is important to realize that the signal/noise ratio, S/N, before the decoder need not be the same as the message-to-noise ratio, m/n. Indeed in the cases of greatest interest to the system engineer it is greatly different.

A case in point is a wideband FM system. Here, the signal-to-noise ratio before the limiter and discriminator is typically 20-40 decibels less than the m/n in the audio section. (Note also that the audio bandwidth is considerably less than the r-f bandwidth). Similarly, the selectivity of a tuned r-f section in an AM receiver eliminates a great deal of noise lying outside of the band occupied by the signal. Later in this chapter, this effect will be explored more

fully. For the present, it may be satisfying philosophically to the student to note that such improvement is not unreasonable from the following viewpoint. If the form of the signal (an FM wave, or an AM wave limited to a certain band, in the cases cited above) is known a priori, then it is possible to build a receiver which gives a large response to that particular signal and a much smaller response to any other. That is to say that the receiver would be selectively sensitive to a particular class of signals, and all transmitted signals are required by proper coding to fall in that class. In the voice-frequency field, the compandor illustrates this principle. Knowing that all messages lie within a certain vu range, the expandor is selectively sensitive to them. Noise and crosstalk outside of the message volume range are recognized as being not part of the message, and are not passed on to the destination.

The systems which most effectively utilize this principle do so by trading bandwidth for S/N. That is, increased bandwidth is utilized to transform the message into signals easily recognized by the selective receiver. Corresponding reduction of S/N requirements is thereby achieved.

More often than not the message is intended for a human destination.* The overall system performance then depends upon the subjective reaction of that destination. It is possible, therefore, to define a "subjective signal-to-noise ratio" or a subjective fidelity which can serve as a suitable overall measure of the communication process. It should be realized, however, that the connection between this measure and m/n can be quite complex, particularly if the human ear or eye is the ultimate destination. Many years of empirical and theoretical study have gone into efforts to explain and specify what kind of errors, omissions, or aberrations in the message stimuli can be tolerated by the human perceptual apparatus without undue effect on the impressions gained, for such tolerances bear directly upon transmission requirements.

To summarize, then, the parameters which serve to measure the performance of a general transmission system are the channel signal-to-noise ratio (or error probability) and bandwidth, the post-decoder

- -

*In some cases, of course, the destination will be an electronic or mechanical device which operates in a well-defined way on the message. Examples are, for instance, electronic computers or switching networks. For such destinations, the subsequent discussion is not appropriate.

message-to-noise ratio, and, finally, a measure of subjective fidelity at the output.

S/N vs m/n In PCM Systems

With this philosophical background and definitions of terms, let us now proceed to our question: What are the advantages of PCM? First, how about performance in terms of S/N vs probability of error?

To detect the presence or absence of a pulse reliably requires a certain signal-to-noise ratio. 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. Let us assume that we have an ideal detector, and that the noise is "white" noise (i.e., noise with a uniform power spectrum and Gaussian amplitude distribution as, for example, thermal noise). If the expected pulse magnitude is V_o, and if the signal when sampled exceeds $V_o/2$, we say a pulse is present. This result will be in error, however, if the noise at that instant exceeds $V_o/2$ in the positive direction when no pulse is present, or in the negative direction when a pulse is present. As the signal power P_s is increased the probability of error due to white noise decreases very rapidly, so that if P_s/N is large enough to make the signal intelligible at all, only a small increase will make the transmission nearly perfect. An idea of how rapidly this improvement occurs may be had from Table II. The last column in the table assumes a pulse rate of 10^6 per second.

TABLE II

Peak Signal Power to Average Noise Power $\frac{P_s}{N}$	Probability of Error	Average Time Between Errors (Rounded Off) if Pulse Rate is 10^6 per Second
13.3 db	10^{-2}	10^{-4} sec
17.4 db	10^{-4}	10^{-2} sec
19.6 db	10^{-6}	1 sec
21.0 db	10^{-8}	3 min
22.0 db	10^{-10}	3 hours
23.0 db	10^{-12}	12 days

Clearly, there is a fairly definite threshold below which the interference is serious, and above which the interference is negligible. This threshold occurs when the peak signal power is about 20 db greater than the average noise power, at which time the average signal power (about half the pulses present, half absent) exceeds the noise power by 17 db.* This gives an error rate of about 10^{-6}, which is believed to be satisfactory for telephone service. Comparing this figure of 20 db with the 60- to 70- odd db per repeater section required for high-quality straight AM transmission of speech, it will be seen that PCM requires much less signal power, even though the noise power is increased by the n-fold increase in bandwidth.

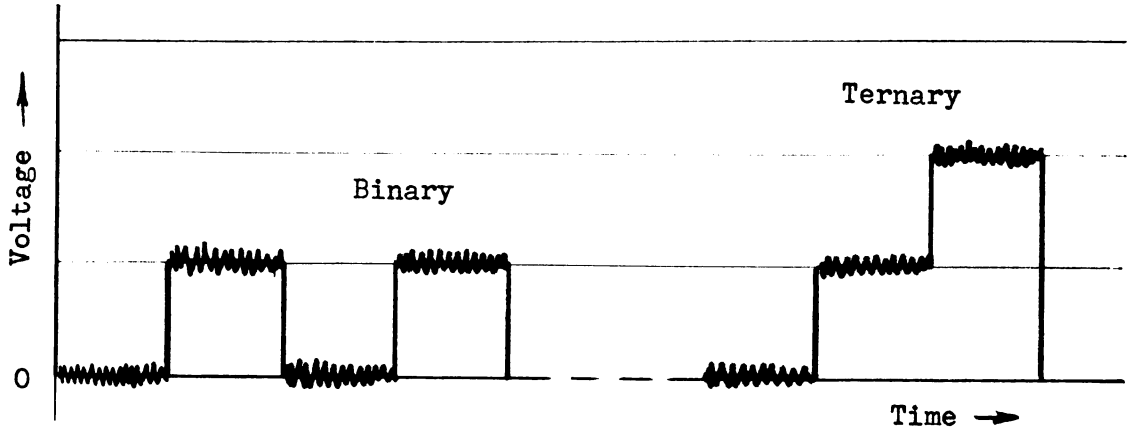
The above discussion has assumed an on-off (base 2) system. If pulses are used which have b different amplitude levels (i.e., a base b system), then a certain amplitude separation must exist between the adjacent levels to provide adequate noise margin. A greater total signal power is then needed to obtain the same (small) error probability, assuming, of course, the same repeater spacing for binary and b-base systems. This is one reason for using binary PCM (0,1) rather than some other base. (Another very important reason for preferring binary PCM is the simplicity of the binary recognition circuitry, of course.) The greater power of ternary, as against binary, PCM is illustrated by Figure 5. Obviously, the separation between steps must be about the same for the same error probability if the noise is the same. Thus, if all code elements are equally likely, the power is increased as the number of steps (that is, b) is increased.

On the other hand, the ternary signal carries more information per pulse than the binary signal does. In other words, increasing "b" decreases the bandwidth required to transmit a given amount of sample information per second (fewer pulses required per second, hence less bandwidth) but calls for an increase in total signal power relative to the noise. This is one example of the process of trading bandwidth for S/N.

So far we have seen that PCM requires more bandwidth and less power than is required with direct transmission of the signal itself, or with straight AM. We have, in a sense, exchanged bandwidth for

---

*The comment that, with half the pulses present and half absent, average power is half the peak power, implies that we are considering the signal only during the time slots in which a pulse might appear. The actual duty cycle of the signal is of no consequence in defining the probability of error, since errors can occur only during the periods when we are making decisions.



Power in Binary and Ternary PCM

Figure 25-5

power. Has the exchange been an efficient one? Are good **message-to-noise** ratios in the recovered signal feasible in PCM? And how sensitive to interference is PCM? We shall now try to answer these questions.

Channel Capacity

A good measure of the bandwidth efficiency is the information capacity of the system as compared with the theoretical limit for a channel of the same bandwidth and power. The information capacity of a system may be thought of as the number of independent symbols or characters which can be transmitted without error in unit time. The simplest, most elementary character is a binary digit, and it is convenient to express the information capacity as the equivalent number of binary digits per second,(Bits/sec), C, which the channel can handle. Shannon and others have shown that an ideal system has the capacity,

$$C = W \log_2 (1 + \frac{S}{N}) \qquad (25\text{-}1)$$

where

W = bandwidth in cps
S = average signal power *
N = white noise power

This equation states that for a channel of bandwidth W and with a signal to noise ratio S/N, there exists a coding scheme which will

- -

*In terms of the symbols used in Table II, $S = \frac{P_s}{2}$.

permit information to be sent over the channel at a rate of C bits/second without error. (i.e., infinite message-to-noise ratio or zero error probability). Two channels having the same C have the same capacity for transmitting information, even though the quantities W, S, and N may be different.

The equation does not specify how such an ideal coding of the message may be realized, and, indeed, in any practical system yet proposed there will be a finite probability of error for a finite transmission rate. In the present context, the thing of greatest interest that Shannon's result shows is that bandwidth can be traded logarithmically with signal-to-noise for a fixed rate of information transmission, hence for fixed P_e. Suppose, for instance, that the signal bandwidth in an ideal system were doubled, then the signal-to-noise requirement to transmit the message information, in total, would be approximately halved in decibels. Information theory shows that this trade is the most favorable from a quantitative point-of-view that can be made.

Shannon's result applies strictly only in the case of ideal coding (zero error), but a similar equation can be derived for general (binary, ternary, etc.) PCM. It turns out to have the same form as Equation (1). Thus, we find that in PCM systems, bandwidth can be traded logarithmically for S/N for a fixed P_e, just as in an ideal system.

Let us compare the performance of a particular binary PCM system with Shannon's ideal system. Suppose we take an error probability of about 10^{-6} as a satisfactory objective for telephone service. Table II has told us that to achieve this we must have a peak signal-to-average noise ratio in the channel of

$$\frac{P_s}{N} = 20 \text{ db}$$

or an average signal-to-average noise ratio

$$\frac{S}{N} = 17 \text{ db}$$

Furthermore, we see from our previous discussion that we can transmit information in binary PCM at a rate of two bits per cycle of bandwidth (for example, 1.5 million pulses per second for a 750 kc channel).

In Shannon's ideal system, on the other hand, we find from Equation 1 that an error-free transmission of 2 bits/cycle of bandwidth (C/W = 2) can theoretically be obtained if

$$\log_2 \left(1 + \frac{S}{N}\right) = 2$$

from which

$$\frac{S}{N} = 3$$

or

$$10 \log \frac{S}{N} = 5 \text{ db}$$

From the analysis we see that for telephone transmission, where an error probability of 10^{-6} is close enough to perfection, binary PCM requires a $\frac{S}{N}$ ratio about 12 db better than Shannon's ideal code.

Other broadband systems in which bandwidth can be traded for S/N do so on a less favorable basis. FM and certain other pulse systems are examples; in these the information capacity is proportional to log W instead of directly proportional to W as in PCM. For a sufficiently wide band, PCM is certain to be the better system.

Message-to-Noise Ratio*

There are several types of noise introduced by a PCM system. One of these is the quantizing noise mentioned in the section on quantization. This is a noise introduced at the transmitting end of the system and nowhere else. Another is so-called "lower side-band noise" which arises from the fact that non-ideal filters are used at the terminals; this will be explained in detail in the next chapter. Filter requirements are specified so that this noise is small. Finally, there will be "false pulse noise", caused by the incorrect interpretation of the intended amplitude of a pulse by the receiver or by any repeater. False pulse noise may arise anywhere along the system, and is cumulative. However, as we have seen earlier, this noise decreases so rapidly as the

---

*Companding is ignored in this discussion, since the comparison is made between peak message voltage and rms noise. The effects of companding are discussed in the next chapter.

signal power is increased above threshold that in any practical system it would be made negligible by design. As a result, the message-to-noise ratio in PCM systems is set by the dominant quantizing noise.

If the message voltage is large compared with a single quantizing step, the errors introduced in successive samples by quantizing will be substantially uncorrelated. The maximum error which can be introduced is one-half of one quantizing step in either direction. All values of error up to this maximum value are equally likely. It can be shown that the rms error introduced is, therefore, $\frac{1}{2\sqrt{3}}$ times the height of a single quantizing step.* When the message is reconstructed from the decoded samples (containing this quantizing error), what is obtained is the original message plus a noise having a uniform frequency spectrum out to f_c and an rms amplitude of $\frac{1}{2\sqrt{3}}$ times a quantizing step height. The ratio of peak-to-peak message voltage to rms noise voltage is, therefore,

$$R = \frac{b^n}{\frac{1}{2\sqrt{3}}} = 2\sqrt{3}\, b^n$$

since b^n is the number of levels. Expressing this ratio in db, we have

$$\begin{aligned} 20 \log_{10} R &= 20 \log_{10} 2\sqrt{3} + n(20 \log_{10} b) \\ &= 10.8 + n(20 \log_{10} b) \end{aligned} \qquad (25\text{-}2)$$

In a binary system, b=2, and

$$20 \log_{10} R \cong (10.8 + 6n) \text{ db}$$

In examining the above expression, let us remember that n, the number of digits, is a factor relating the total bandwidth used in transmission to the bandwidth of the message to be transmitted, i.e., $W=nf_c$. It is something like the index of modulation in FM. Now, for

- -

*See Appendix II of Monograph B1611.

every increment equal to f_c added to the bandwidth used for transmission, n may be increased by one, and this increases the message-to-noise ratio by 6 db. In other words, in PCM, the message-to-noise ratio in db varies linearly with the number of digits per code group, and hence with the bandwidth. Of course, as the bandwidth is increased the noise power increases. Hence, a proportional increase in signal power is required to stay adequately above threshold and obtain substantially error-free transmission so that quantizing noise remains dominant. As long as the error rate is thus kept low, we see that a 6 db message-to-noise ratio improvement is obtained for each digit we add to our binary code.

Ruggedness

One important characteristic of a transmission system is its susceptibility to interference. We have seen that noise in a PCM channel produces no effect unless the peak amplitude is greater than half the separation between pulse levels. In a binary (on-off) system, this is half the pulse height. Similarly, interference such as stray impulses, or pulse crosstalk from a near-by channel, will produce no effect unless the peak amplitude of this interference plus the peak noise is half the pulse height. The presence of interference thus increases the threshold required for satisfactory operation. But, if an adequate margin over threshold is provided, comparatively large amounts of interference can be present without affecting the performance of the circuit at all. A PCM system, particularly an on-off (binary) system, is therefore quite "rugged".

When a number of wire or radio communication routes must converge on a single terminal, or follow similar routes between cities, the ruggedness of the channels is a particularly important consideration. If the susceptibility of the channels to mutual interference is high, many separate frequency bands will 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 frequency occupancy of PCM is exceptionally good and its other transmission advantages are then obtained with little, if any, increase in total bandwidth.

Conclusions

This concludes our examination of PCM in the abstract. It appears that: PCM offers a greater improvement in ruggedness

and message-to-noise ratio than other systems, such as FM, which also depend upon the use of wide bands.

By using binary (on-off) PCM, a high quality signal can be obtained under conditions of noise and interference so bad that it is just possible to recognize the presence of each pulse. Further, by using regenerative repeaters which detect the presence or absence of pulses and then emit reshaped, respaced pulses, the initial signal-to-noise ratio can be maintained through a long chain of repeaters.

PCM lends itself to time-division multiplex.

PCM transmitters and receivers can be expected to be somewhat more complex than are those used for some other forms of modulation. However, the nature of the circuitry is such that although complex, it may be cheaper than, for example, AM terminal equipment giving comparable quality of transmission. Furthermore, a greater proportion of the terminal equipment can be common to all channels, as we shall see.

In all, PCM seems ideally suited for multiplex message circuits, where a standard quality and high reliability are required.

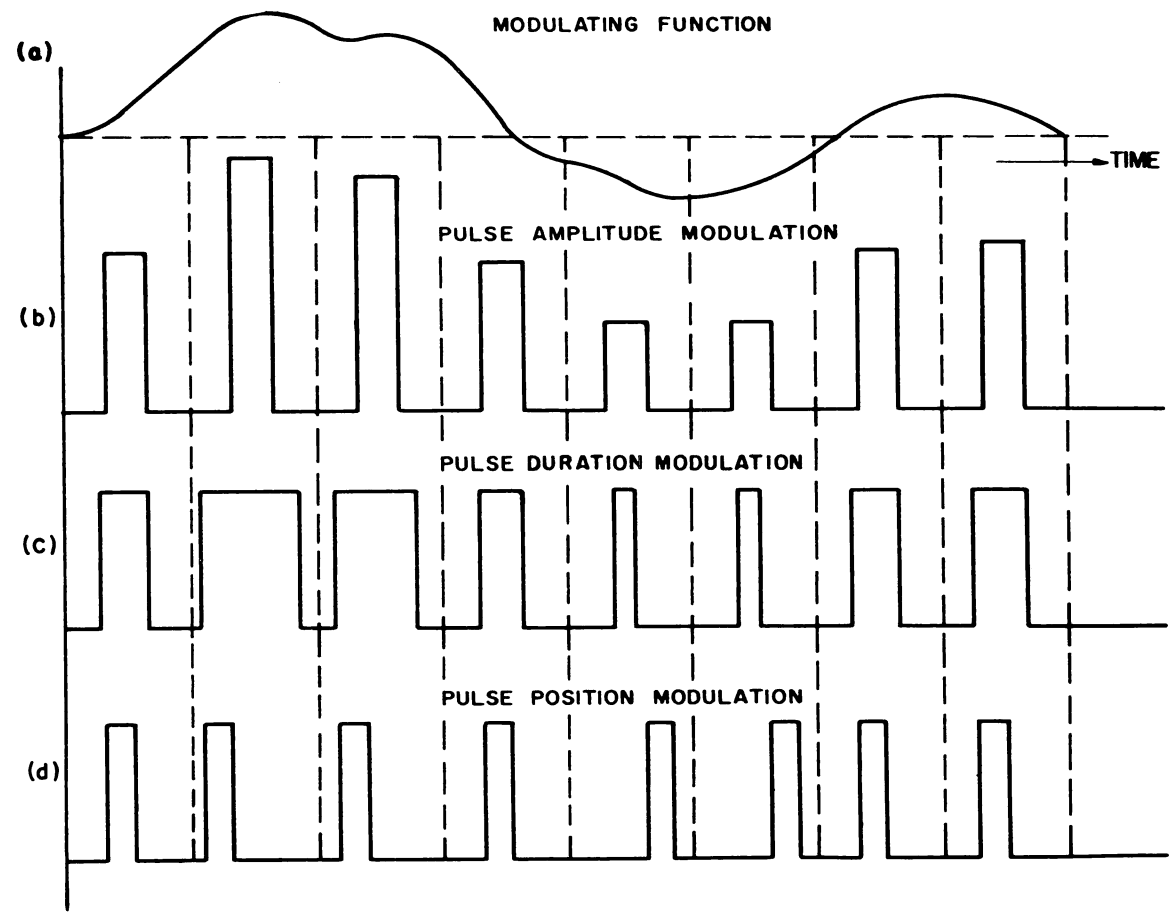
In the following chapters, we shall examine specific practical problems in more detail.

APPENDIX

In the preceding text, mention has been made of pulse modulation systems other than PCM. Some of these are illustrated below.

Unquantized Pulse Modulation Systems

Unquantized or continuous pulse modulation systems employ pulses, the modulated parameter of which varies as a continuous function of the amplitude of the modulating signal. Some of the more common pulse carriers are shown in Figure 6. The modulating wave form is shown in Figure 6a. It is assumed that this waveform is sampled in accordance with the sampling principle to produce the pulse amplitude modulated (PAM) signal shown in Figure 6b. By a suitable encoding process the amplitude modulated pulses may be changed to duration-modulated pulses as shown in Figure 6c or to position modulated pulses as shown in Figure 6d. In pulse duration modulation (PDM) systems the modulating function may vary



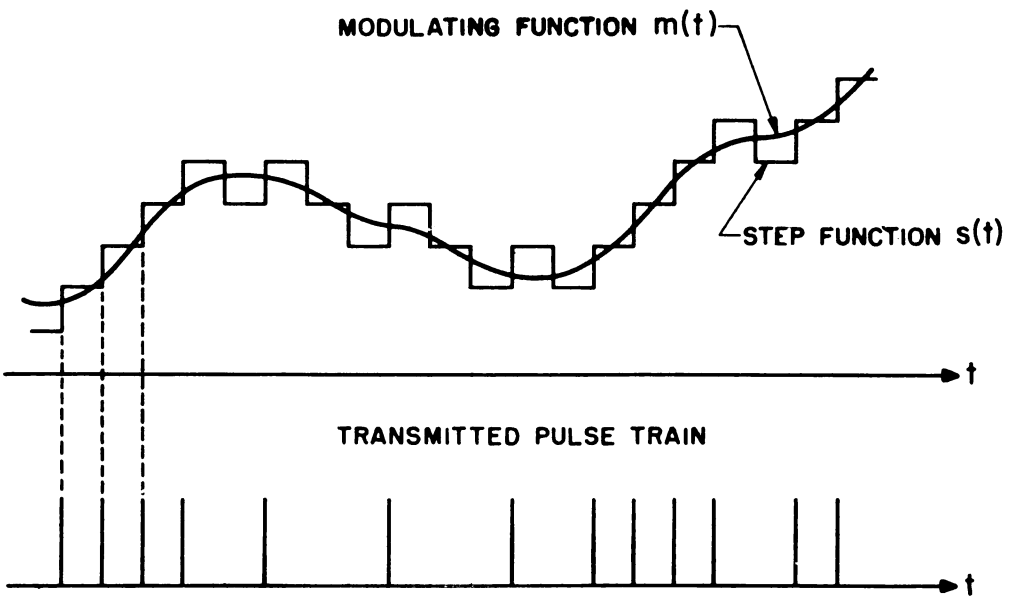
Unquantized Pulse Systems

Figure 25-6

the time of occurrence of the leading edge, the trailing edge, or both edges as shown in Figure 6c. In pulse position modulation (PPM) systems the modulating function varies the time position of a pulse from the position it would occupy in an unmodulated pulse train.

Delta Modulation

Another form of quantized pulse modulation which might be mentioned in passing is called delta-modulation (DM). The waveforms of a typical DM system are shown in Figure 7. The modulation function m(t) is approximated by a function S(t) which is made up of a sum of discrete step functions equal in amplitude and equidistant in time. The difference m(t) - S(t) determines the polarity of the steps in S(t). The delta code consists of binary pulses, one for each positive step and no pulse for a negative step. In the receiver a simple integrating circuit reconstructs the step functions from the binary pulse train. The reconstructed function is then applied to a low pass filter to recover the original signal.



Delta Modulation

Figure 25-7

Chapter 26

PREPARATION AND PROCESSING OF SIGNALS IN PCM

Coding a message for transmission involves a number of steps. First, the message must be sampled. In order to better understand the sampling process and the establishment of filter requirements for the system, the spectra obtained for both instantaneous and natural sampling of the message are studied. Circuits used to sample the message are then briefly described. The next step, quantizing, introduces the dominant source of noise in the PCM systems. Types of encoders which perform the quantizing operation and convert the message samples into binary PCM are described. The concluding problem concerns that of timing and framing the multiplex signal so that it can be recovered at the receiving terminal. A fortunate fact, from the economic standpoint, is that most of the complex circuitry which is involved is common to all channels.

Introduction

In the previous chapter it was shown that PCM affords a favorable means for trading bandwidth for reduced S/N requirements. Furthermore, PCM promises economic savings in the transmission of both speech and video information over existing facilities. These savings accrue because a PCM terminal will probably cost much less than a corresponding frequency multiplex terminal. Terminal costs dominate for short haul systems. PCM is, therefore, an attractive way of increasing the channel capacity of the exchange area trunk plant. Much of the terminal equipment in a time division system is common to all channels, so that consequently the terminal cost of a PCM system varies slowly with the number of channels. A frequency division terminal, on the other hand, increases in size or complexity, and therefore cost, almost linearly with the number of channels. It follows that the savings offered by pulse systems increase with the number of channels as long as terminal costs dominate.

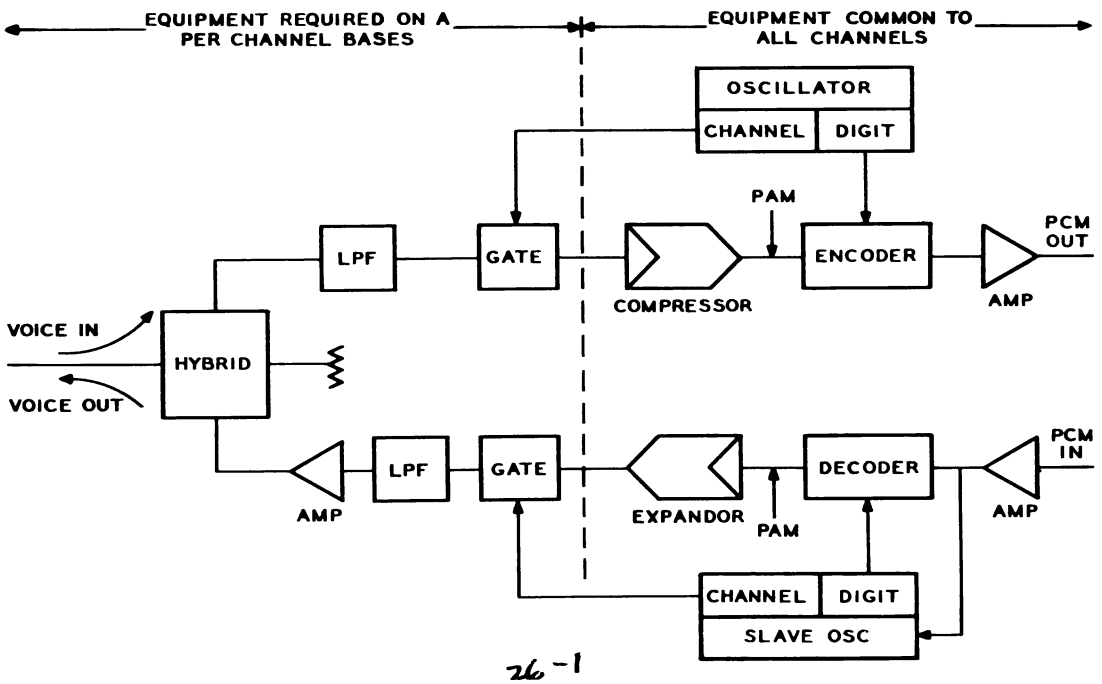
In view of both the economic and technical advantages, a more detailed discussion of PCM is in order. This chapter will be concerned with the functions that must be performed in such a system to prepare the signal for binary transmission and the complementary processes involved

in recovering the message from a PCM signal. Succeeding chapters will deal with the details of pulse transmission.

Terminal Operations

The operations required in any PCM terminal are depicted in block diagram form in Figure 1. They include:

1. Sampling (and reconstruction)
2. Time Division Multiplexing (and de-multiplexing)
3. Companding (compression and expansion)
4. Coding (encoding and quantizing; decoding)
5. Timing (to control the above processes).



PCM Terminal Components

Figure 26-1

Sampling converts the message into a series of equally spaced pulses whose amplitudes are related to the amplitudes of the original message at fixed intervals of time determined by the sampling frequency*.

- -

*PAM signals are considered throughout, although PDM (pulse duration modulation) and PPM (pulse position modulation) may also be quantized to yield PCM.

The PAM samples from each channel are multiplexed in time and then compressed prior to encoding. The compressor is utilized to reduce the effects of round-off error (commonly called quantization error, or quantization noise) when the continuous range of amplitudes is approximated by a discrete set of levels for representation by a binary signal. The sampled, multiplexed and compressed signal is then encoded into a series of binary (on or off: 1 or 0) pulses by the encoder. The binary pulse train may then be transmitted directly by a cable or may be used to pulse-modulate a high frequency carrier.

In the receiving portion of the terminal the processes are reversed. The PCM signal is decoded into a PAM signal and expanded to make the overall compandor characteristic linear. At this point, except for the inherent quantization errors introduced when the continuous signal was represented by a finite number of steps (finite number of binary digits), the signal should be a sampled version of the original message. It remains, therefore, to reconstruct the original message by passing the PAM signal through a low pass filter.

Obviously all of the preceeding operations on the signal must be programmed to occur at the proper instants in time. This is the function of the timing or control block. In addition, the receiving terminal must be kept in synchronism with the terminal that transmitted the information. This is the function of the framing circuitry. In the following pages each of the terminal functions will be examined in turn and the questions that have arisen in this "broad brush" treatment will be answered in detail. For example:

1. What does the spectrum of a sampled function look like?
2. What do the compressor and expander do to minimize quantizing noise?
3. How many binary digits are required for high quality speech transmission?

Emphasis will be placed on theory and concept, rather than on detailed instrumentation. As the exposition proceeds, the similarity between a PCM terminal and a special purpose digital computer will become apparent. In fact the pulse circuitry developed for digital computers (and electronic switching systems) carries over into PCM. This fact, plus the advent of solid state devices, has been an important contributing factor to the present attractiveness of PCM.

Sampling Principle

The first step in converting a message which is a real, continuous function of time to a PCM signal involves sampling it at appropriate instants or intervals. As pointed out in the preceding chapter, the average frequency with which this sampling must occur is dependent on how fast the time function changes in amplitude. This is determined by the highest signal frequency present in the spectrum of the time function. To repeat, Nyquist's sampling theorem, stated in a restricted form widely used for practical purposes, says:

> If a message that is a real, continuous function of time, with a band limited spectrum, is sampled at regular intervals and at a rate slightly higher than twice the highest frequency present, the samples contain all of the information in the original message.

It has also been pointed out in preceding material that theoretically the original message can be reconstructed by passing the train of samples through an ideal low pass filter. Notice the words "band-limited" and "ideal filter" in these statements. Suppose we do not want to pay for filters with infinite attenuation and zero delay distortion? What problems then arise? In order to answer this question we must first examine the spectrum of the sampled message.

Instantaneous Sampling and Reconstruction

Sampling can be viewed as the process of multiplying the message function of time, $f(t)$, by a train of impulses, $C(t)$. It will be recalled that the spectrum of a periodic function of time consists of discrete frequencies which are harmonics of the repetition frequency. The envelope of the amplitudes of these components is the same as the envelope of the spectrum of a single pulse. Recall also that the envelope of a single impulse is flat vs. frequency and that all of the spectrum components have the same phase. It follows, therefore, that the spectrum of a train of impulses must be a series of discrete frequencies, all of equal amplitude and phase, which are harmonics of the repetition frequency. We are going to consider the train of impulses as a "carrier" which is to be modulated by the message function. Let T represent the impulse period (so that $1/T$, of course, equals the impulse repetition rate). If we consider the impulses to have zero duration and an area equal in magnitude to T, the Fourier series

representation for this "carrier" is given by the real part of*

$$C(t) = \sum_{n=-\infty}^{\infty} e^{j\,n\,\omega_o t} \qquad (26\text{-}1)$$

where

$$\omega_o = \frac{2\pi}{T}$$

The output of the impulse modulator, which performs the operation of multiplying the message function by the impulse carrier, can be written as

$$f_o(t) = f(t) \cdot C(t)$$

$$= f(t) \sum_{n=-\infty}^{\infty} e^{j\,n\,\omega_o t} \qquad (26\text{-}2)$$

Since f(t) is independent of n, it can be included in the summation, so that Equation 2 can be expanded into the form

$$f_o(t) = f(t) + f(t)\, e^{j\omega_o t} + f(t)\, e^{j2\omega_o t} + \ldots. \qquad (26\text{-}3)$$

(Note that the full series includes negative as well as positive values of n.) Let us now obtain the spectrum of the output function, $f_o(t)$, by taking the Fourier transform of Equation 3, term by term. Thus,

- -

*This is merely the general Fourier series representation in complex form: i.e.,

$$C(t) = \sum_{n=-\infty}^{\infty} a_n e^{j\,n\,\omega_o t}$$

where

$$a_n = \frac{1}{T} \int_{-\frac{T}{2}}^{\frac{T}{2}} C(t)\, e^{-j\,n\,\omega_o t}\, dt$$

In this case a_n=1 for all values of n since we are dealing with δ functions of area = T.

$$F_o(\omega) = \frac{1}{2\pi}\int_{-\infty}^{\infty} f_o(t)\, e^{-j\omega t}\, dt$$

$$= \frac{1}{2\pi}\int_{-\infty}^{\infty} f(t)\, e^{-j\omega t}\, dt$$

$$+ \frac{1}{2\pi}\int_{-\infty}^{\infty} f(t)\, e^{-j(\omega-\omega_o)t}\, dt$$

$$+ \frac{1}{2\pi}\int_{-\infty}^{\infty} f(t)\, e^{-j(\omega-2\omega_o)t}\, dt + \ldots \qquad (26\text{-}4)$$

Now, the Fourier transform of the message f(t) is just

$$F(\omega) = \frac{1}{2\pi}\int_{-\infty}^{\infty} f(t)\, e^{-j\omega t}\, dt \qquad (26\text{-}5)$$

Look at, for example, the third term of (4). Observe that it differs from (5) only by the change in variable from ω to $\omega-2\omega_o$, and that for the general term the change in variable is from ω to $\omega-n\omega_o$. Since we can write,

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

we can also write

$$\frac{1}{2\pi}\int_{-\infty}^{\infty} f(t)\, e^{j(\omega+n\omega_o)t}\, dt = F(\omega+n\omega_o)*$$

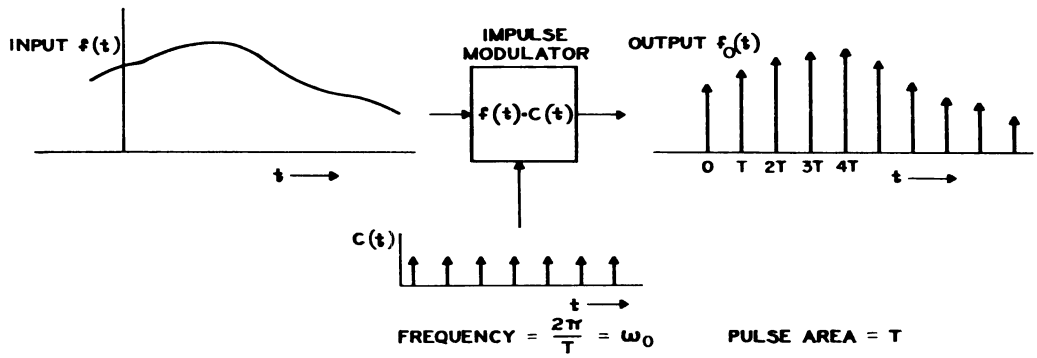
where F is the same function in both expressions. In other words, then, the spectrum of the modulator output, $f_o(t)$, is

$$F_o(\omega) = \sum_{n=-\infty}^{\infty} F(\omega+n\omega_o) \qquad (26\text{-}6)$$

where $F(\omega)$ is the transform of the message.

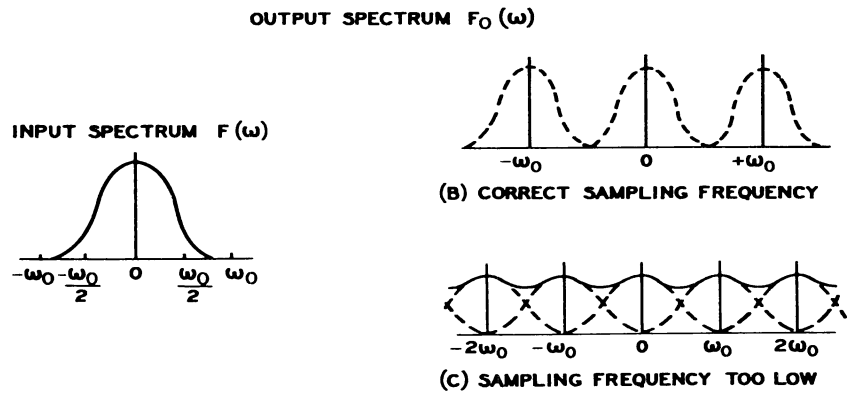
---

*Since we are going to let n assume all values from $-\infty$ to $+\infty$ in accordance with (2), it is immaterial whether we write $(\omega-n\omega_o)$ or, more conveniently, $(\omega+n\omega_o)$.



(A) SAMPLING BY IMPULSE MODULATION

TIME PLOTS



(B) CORRECT SAMPLING FREQUENCY

(C) SAMPLING FREQUENCY TOO LOW

FREQUENCY PLOTS

Instantaneous Sampling

Figure 26-2

Figure 2 illustrates the analysis we have just made. It can be seen from either Equation 5 or Figure 2 b that the output spectrum is periodic on the frequency scale with period ω_o, the sampling frequency. A pair of sidebands around dc, ω_o, $2\omega_o$, and so on through each harmonic of the sampling frequency has been produced.

The question now arises as to whether the sampling process has properly preserved all of the information content of the message, or has it introduced ambiguities which make it impossible to recover the message. As will be seen, this depends on the degree to which the message is truly band-limited.

Consider the case illustrated in Figure 2 b in which the output consists of a number of isolated "mountains" each of which is identical with $F(\omega)$, the message spectrum, except for a translation in frequency. It is clear that the original message can be recovered without distortion by merely passing the output spectrum through a low-pass filter with a cut-off at one-half ω_o. Accordingly, we can then expect that all of the information in the original message has been preserved, as stated in the Sampling Theorem.

When, however, the input spectrum is appreciable at frequencies beyond half the sampling frequency, we get a picture like that shown in Figure 2 c. We can regard this as a result of a sloppy transmitting filter, or of too low a sampling frequency. In either case, the various sidebands overlap and the output from the modulator is ambiguous. A low-pass filter energized by $f_o(t)$ will, in the case of overlapping sidebands, be unable to reconstruct the original message. There will be no way of determining, for example, whether an output component at $\omega=0.6\ \omega_o$ was produced by an input component at $\omega=0.6\ \omega_o$ or one at $0.4\ \omega_o$, since the latter component beating with the fundamental of the sampling frequency will also produce a component at $0.6\ \omega_o$.

In practice, as we shall see, non-ideal band-limiting filters result in our always having some of the ambiguity referred to above and, therefore, some distortion of the message. We can also anticipate that even when the input spectrum is adequately band-limited by the transmitting filter, we may have a distortion because of receiving filter imperfections. Consider the case where the sampling frequency is 8 kc. If the receiving filter does not cut-off sharply at one-half the sampling frequency, instead of getting just the spectrum "mountain" around zero frequency, we get some of the lower sideband of the spectrum around 8 kc. The subscriber will hear this inversion of the message components in the 4 to 8 kc band. This is referred to as a "lower sideband" distortion product and will be discussed more fully under filter requirements.

Natural Sampling and Reconstruction

A type of sampling in which the resultant samples have finite widths and non-flat tops is known as "natural sampling". In the preceeding discussion it has been assumed that impulses have been used to obtain instantaneous sampling. Impulses can never be realized physically, and it is of interest to see how samples of non-zero duration can be

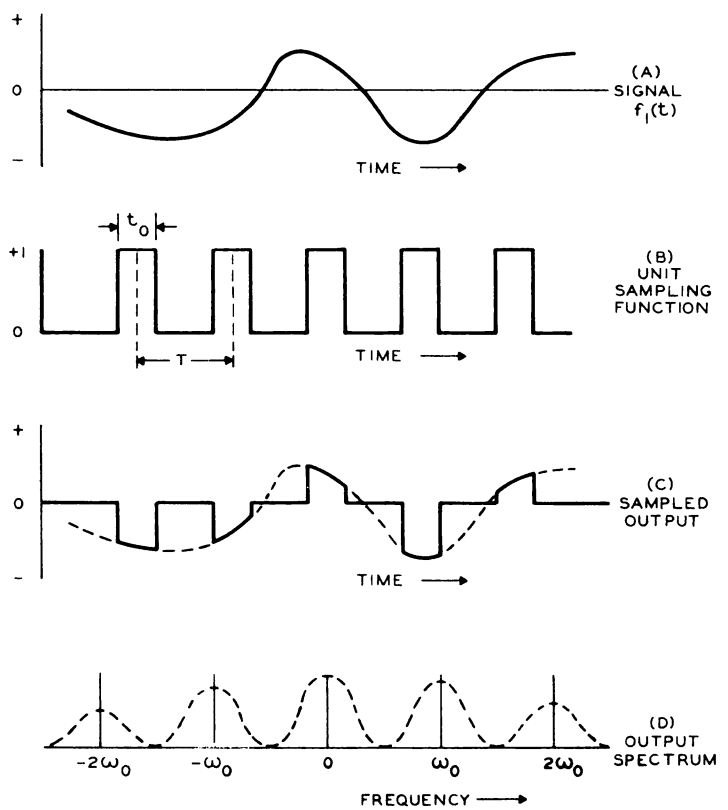
utilized to preserve all of the information in the original message. When the modulating wave or sampling function consists of a periodic train of rectangular pulses of unit height and finite width t_o, as shown in Figure 3b, it can be represented by the real part of the following expression:

$$C(t) = \frac{t_o}{T} \sum_{n=-\infty}^{\infty} C_n e^{jn\omega_o t} \tag{26-7}$$

where

$$C_n = \frac{\sin n \frac{t_o}{T} \pi}{n \frac{t_o}{T} \pi} = \frac{\sin \frac{n \omega_o t_o}{2}}{\frac{n \omega_o t_o}{2}} \tag{26-8}$$

$$\omega_o = \frac{2\pi}{T}$$



Natural Sampling

Figure 26-3

Let us now take the product of C(t) and the message f(t), which, as in the previous section, is equivalent to modulating the chain of pulses by the message. This yields the PAM output wave of Figure 3c. Again, if we take the Fourier transform of the product we obtain the spectrum of the output signal as

$$F_o(\omega) = \frac{t_o}{T} \sum_{n=-\infty}^{\infty} C_n F(\omega+n\omega_o) \qquad (26\text{-}9)$$

As shown in Figure 3d, the spectrum resulting from natural sampling is similar to that of instantaneous sampling except for the envelope introduced due to the finite width of the pulses. Each "mountain" is identical with the spectrum around $\omega=0$ except for a multiplying factor C_n. If the conditions of the sampling theorem are satisfied we can reconstruct the original message by passing the signal $f_o(t)$ through an ideal low-pass filter.

It should be remembered that when a PAM signal consisting of natural samples is encoded for PCM, each sample must be characterized by a single value. That value will generally be some sort of average of the non-flat topped PAM pulse. Thus, when such a signal is decoded one would not expect the recovered PAM signal to have the spectrum of the original train of natural, slope-topped samples of Equation (26-9). In practice, the duration of the natural sample is so short that the assumption of instantaneous sampling gives sufficient accuracy. The main point of this discussion is to show that natural sampling, as well as instantaneous sampling, preserves the information of the original message.

Reconstruction of Sample Holding

In the preceding sections, methods have been considered for sampling the message to produce a PAM signal containing all of the message information. We have examined the pulse-train spectrum and seen that the message can be reconstructed by passing the sample pulses through a low-pass filter. Another approach to reconstruction is the "sample holding" scheme. In this approach it is assumed that at the receiver the PAM samples of the message are sufficiently narrow to be considered as a train of impulses. The instantaneous samples are then passed through a "hold" circuit which extends the width of the sample for a time $t_o < T$, as illustrated by Figure 4a. The output of the hold circuit is, therefore, a train of amplitude modulated flat top pulses.

The holding circuit has a transfer (i.e., output/input) characteristic which, in the time domain, is illustrated by Figure 4b.

We need to know the frequency characteristic of the holding circuit to determine the effect of such a circuit on the spectrum of the input impulse train. If we apply the Fourier transform to the time domain characteristic of 4 b, we find the frequency characteristic of the transfer function is*

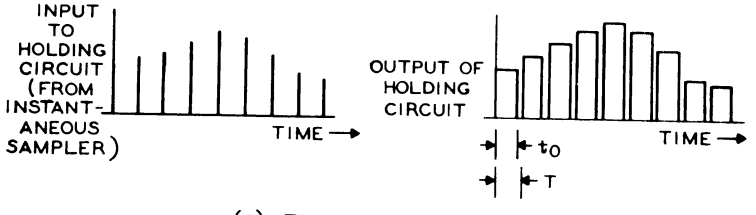
$$Y(\omega) = \frac{1-e^{-j\omega t_o}}{j\omega} = A(\omega)\ e^{-jB(\omega)} \tag{26-10}$$

A little algebraic manipulation will give

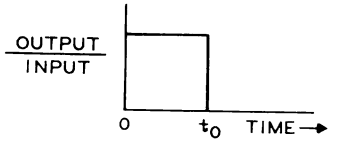
$$A(\omega) = \left| t_o \frac{\sin \omega t_o/2}{\omega t_o/2} \right| \tag{26-11}$$

and

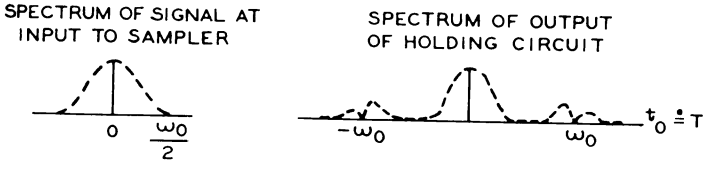
$$B(\omega) = -90° + \tan \frac{\sin \omega t_o}{1 - \cos \omega t_o} \tag{26-12}$$



(A) TIME PLOTS



(B) TRANSFER CHARACTERISTIC OF HOLDING CIRCUIT, IN TIME DOMAIN



(C) FREQUENCY PLOTS

Impulse Sample and Hold

Figure 26-4

*We have dropped the constant factor $\frac{1}{2\pi}$ for convenience, together with any gain or attenuation constant associated with the holding circuit.

The envelope delay can be determined by differentiating (12) with respect to ω. This yields

$$\frac{d\theta}{d\omega} = -\frac{t_o}{2} \qquad (26\text{-}13)$$

Since the spectrum of the output of the hold circuit is the product of the input spectrum and $Y(\omega)$, it can be concluded from (11) and (13) that the effect of the hold circuit is to introduce amplitude distortion of the spectrum of the message and to delay each sample by an amount $\frac{t_o}{2}$. The amplitude distortion can be overcome, of course, by the introduction of an equalizer whose frequency response is the inverse of (11).*

Let us examine the amplitude spectrum of the output of the hold circuit. It is obtained by taking the product of (11) and (6) and is

$$A(\omega)\, F_o(\omega) = t_o \left| \frac{\sin \omega t_o/2}{\omega t_o/2} \right| \sum_{n=-\infty}^{\infty} F(\omega + n\omega_o) \qquad (26\text{-}14)$$

If we make t_o almost equal to T (i.e., holding takes place for a time almost as long as the spacing between samples) and plot the spectrum given by (14) around d.c., $+\omega_o$, and $-\omega_o$, (that is, for n=0, n=+1, n= -1) we obtain Figure 4c. This can be compared with Figure 3d to show the difference between the spectrum for natural sampling and flat-topped samples obtained by holding. Holding has the advantage of increasing the energy in the samples. It also increases the attenuation to unwanted sidebands by virtue of the $\frac{\sin x}{x}$ modification in the spectrum. The amount of attenuation to unwanted sidebands is, of course, dependent upon the length of time the sample is held. This benefit is limited by the necessity for providing some guard space between samples to allow for transient decays and the necessity for equalizing the higher frequency message components.

Some of the advantages of the holding process discussed above are obtained in practice at the receiving terminal in the exchange area PCM system now being developed. Here sampling is at an 8 kc rate, and 24 channels are time division multiplexed. The interval between samples is therefore 125 μs, but the code group representing one sample can be

- -

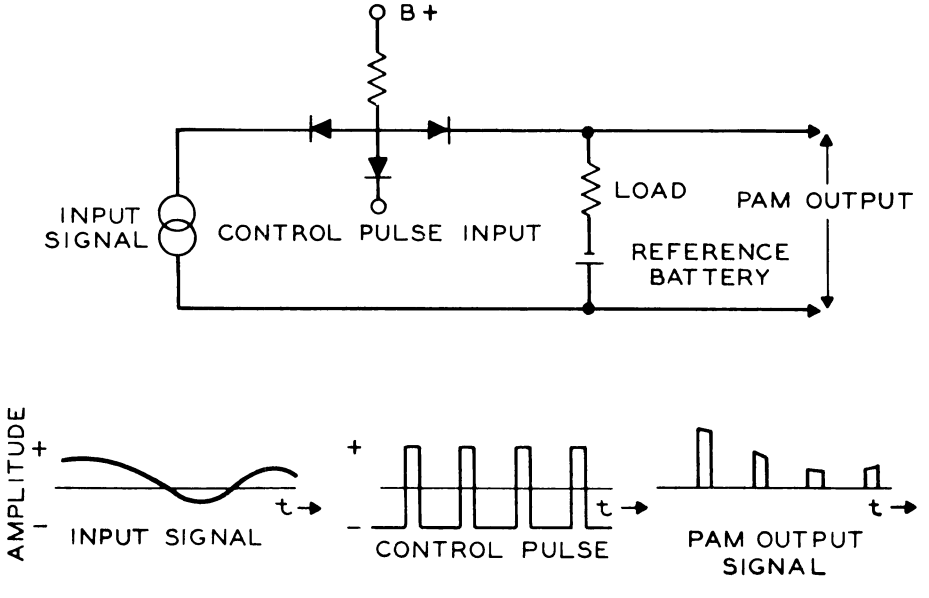
*The effect introduced by the hold circuit is similar to that caused by the finite size of a scanning aperture in television and telephotography and it has therefore become customary to refer to the amplitude distortion in sampling theory as the "aperture effect".

only about 5 μs if 24 messages are to be multiplexed. After the message sample information has been recovered from the multiplexed code signal, one would anticipate a PAM signal of 5 μs on, 120 μs off. By applying this to a capacitor to obtain storage action, some of the benefit of holding is obtained. The time constant of the storage circuit is adjusted so that almost complete decay of one sample has occurred by the time the next sample for that channel arrives 120 μs later.

Sampling Gates and Clampers

So far we have looked upon sampling simply as a means of producing a series of equally spaced pulses whose amplitudes vary in accordance with the amplitude of the modulating function. Since the message voltage varies both plus and minus about zero, it is natural to think of the PAM output of a sampling gate as consisting of both positive- and negative-going pulses (as, for example, in Figure 3 of Chapter 25). However, a relatively simple (therefore, inexpensive) sampling gate to use in the exchange area PCM system is one which produces unipolar pulses, i e., pulses having the same polarity, at its output. Before considering the disadvantages of a unipolar pulse output, let us briefly look at the type of circuit used to sample the message and produce this output.

An example of such a sampling gate (a modified Lewis gate) is shown in Figure 5. When the gate is in the disabled state, it must present a large value of attenuation to any signal appearing at its input. Varistor gates can be built to give 90-120 db of attenuation to dc



Modified Lewis Gate

Figure 26-5

voltages if good components and a low impedance output load are used. Somewhat less attenuation to ac signals is obtained because of the capacitive reactance of the diode impedance. The sampling gates, along with the filters to be discussed in the next section and the hybrid networks, are used on a per-channel basis, but the equipment which follows can be common to all channels. Therefore, in determining the output impedance it should be kept in mind that the sampling gates from all of the channels will be connected together, in general, to form the time division multiplex signal which is operated upon by the common channel equipment. The output load impedance will then depend upon the condition of all of the sampling gates in the terminal.

In its transmitting condition, the gain (or loss) of the gate should be stable with time, temperature, and supply voltage. Another consideration of importance arises in high speed switching of a number of channels. Care must be exercised to use varistors which recover rapidly from the enabled state so that the varistor discharge current from one gate will not add to the new sample current from a second gate and cause crosstalk between the message channels.

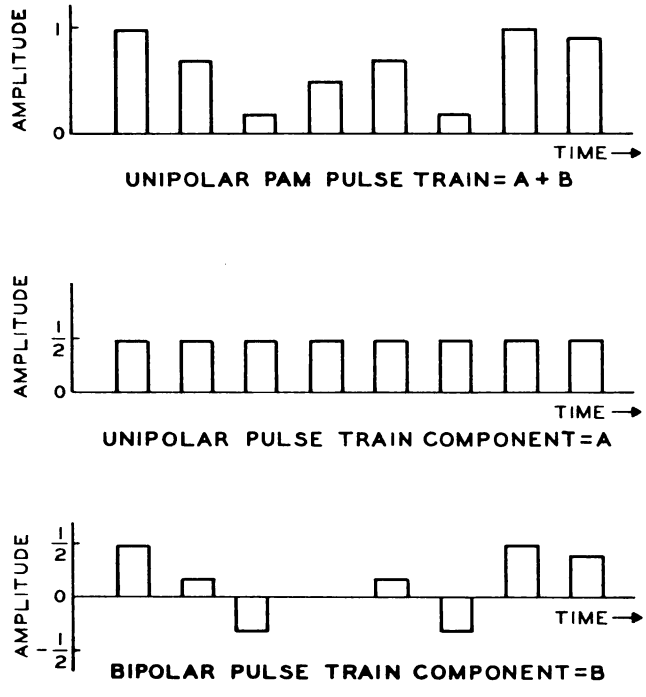
A time division system which has unipolar pulses at the output of the sampling gates has single frequency tones associated with it at all integral multiples of the sampling frequency.* This can be seen from Figure 6, which shows that a unipolar PAM signal can be broken into the sum of a bipolar PAM signal and a series of equal amplitude unipolar pulses appearing at the sampling rate. The bipolar PAM signal represents the message voltage variation and is the output which would be obtained from, for example, a simple switch-type sampler. It is the second component, the string of unipolar pulses, which gives rise to the tone at the sampling frequency and all its odd harmonics, even when no message is present. These tones appear at the receiver after passage through the

---

*A time division system which employs bipolar pulses formed by an ideal product-type modulator (i.e., one such that $f_o(t) = f(t) \cdot C(t)$, as used in the previous sections) does not have any tones at the sampling frequency or its harmonics. This can be seen from Equation (6), for example. At zero frequency, $F(\omega)$ is zero unless the incoming message to the sampling gate is dc biased, which is not generally true. Therefore, $F_o(\omega)$ must be zero at zero frequency, the sampling frequency, and all harmonics of the sampling frequency.

reconstituting filter,* amplifier, and loop. In addition, tones are sent back to the transmitter through the transmitting filter and loop, but since no amplifier is present in this path these tones are considerably weaker than those which appear at the receiver. Thus, we see that one of the disadvantages of using a simple sampling gate which produces unipolar pulses is the presence of unwanted tones which must be held to tolerable magnitudes by the system filters.

There are other disadvantages in the use of a sampling gate which gives unipolar pulses at its output. For example, dc rather than ac amplifiers are required in the terminal equipment which follows. Furthermore, the power handling capacity of the amplifiers for the unipolar pulses would be considerably greater. If, for example, the unipolar pulses vary between 0 and V volts, the peak pulse power will be proportional to V^2. On the other hand, if we let the pulses in a bipolar pulse system vary V volts peak-to-peak, the peak power is proportional to $V^2/4$, or one fourth that of the unipolar system. For these



Composition of Unipolar PAM Signal

Figure 26-6

*We are assuming non-ideal filters, as discussed more fully in the next section.

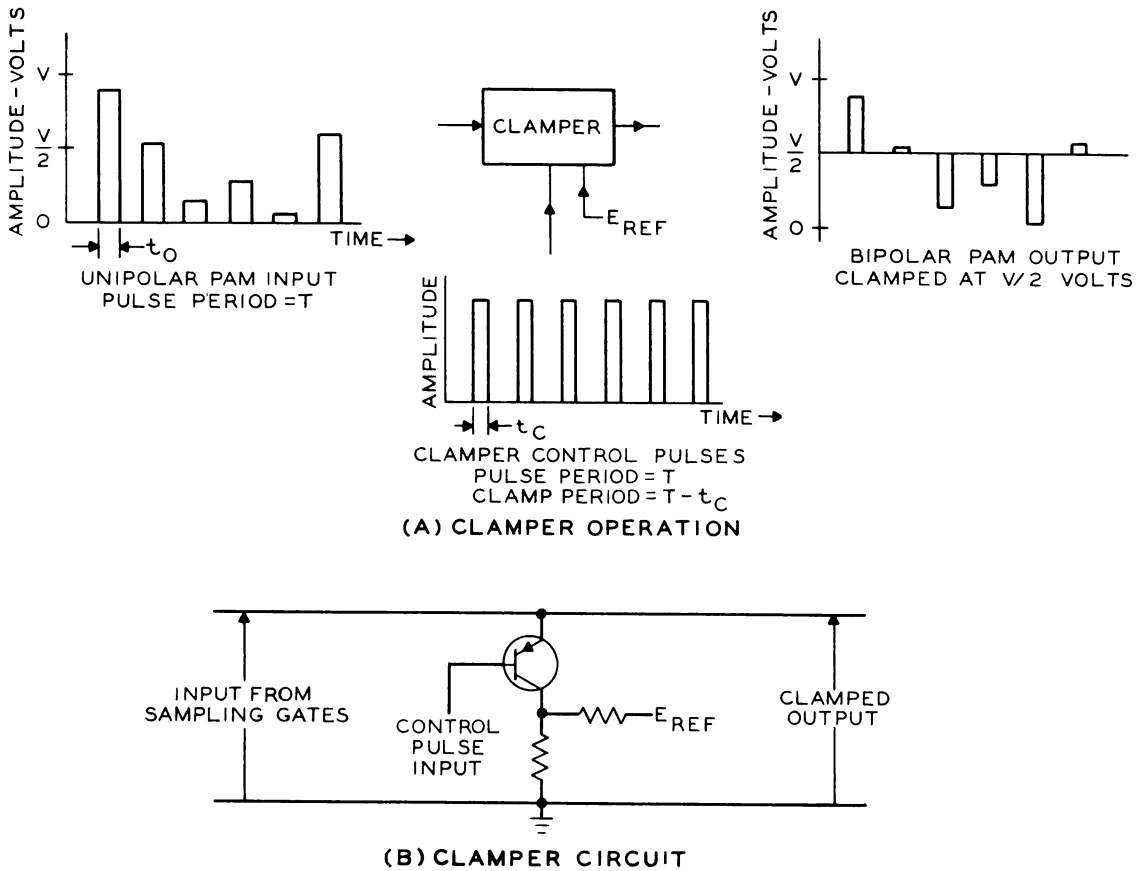
and other reasons the use of a unipolar output pulse sampling gate can be justified only because of its economy, which is an important consideration since there must be one sampling gate per channel. However, these same considerations lead us to introduce a means for converting the unipolar pulse train to a series of bipolar pulses.

One method of forming bipolar from unipolar pulses is shown in Figure 7a. Here we see that if we hold (or clamp) the voltage at V/2 volts between pulses we have, in effect, converted the unipolar pulse train into a bipolar pulse train varying V volts peak-to-peak about a dc bias of V/2 volts. The circuit which performs this holding or clamping function is a clamper, a circuit for which is shown in Figure 7b. The clamper could be used on a per channel basis, of course, but considerable economy results if the outputs of all the sampling gates are tied together first and the resulting time division multiplex PAM signal is fed into a common clamper. A well designed and controlled clamper can be made to introduce about 40 db attenuation to the sampling frequency tone and its harmonics by the conversion of the unipolar into bipolar PAM. In addition, the clamper can be used to "clean up" a pulse train by using sharp rectangular control pulses and extending the clamp period into the input pulse period (i.e., by making $t_c < t_o$). The clamped output pulses will then have sharp leading and lagging edges, regardless of the shape of the incoming pulses.

Effect of Non-ideal Filters

In our earlier discussions, we assumed ideal low pass filters to restrict the signal spectrum to less than half the sampling frequency at the transmitting end. The same idealization was used in describing the reconstruction of the signal from the PAM samples. As we have now pointed out, this abstraction cannot be realized in practice. Furthermore the elaborate filters required to approach the ideal would seriously impair the economic advantage of PCM terminals. Therefore, physical realizability and economics dictate filters which result in some overlap of the sidebands about harmonics of the sampling frequency. This represents a transmission impairment for a PCM system. In addition, this impairment will accumulate when several systems are placed in tandem, or when some channels in a PCM carrier system are dropped and the remaining channels resampled for further transmission. Each successive stage of sampling and reconstruction introduces additional distortion.

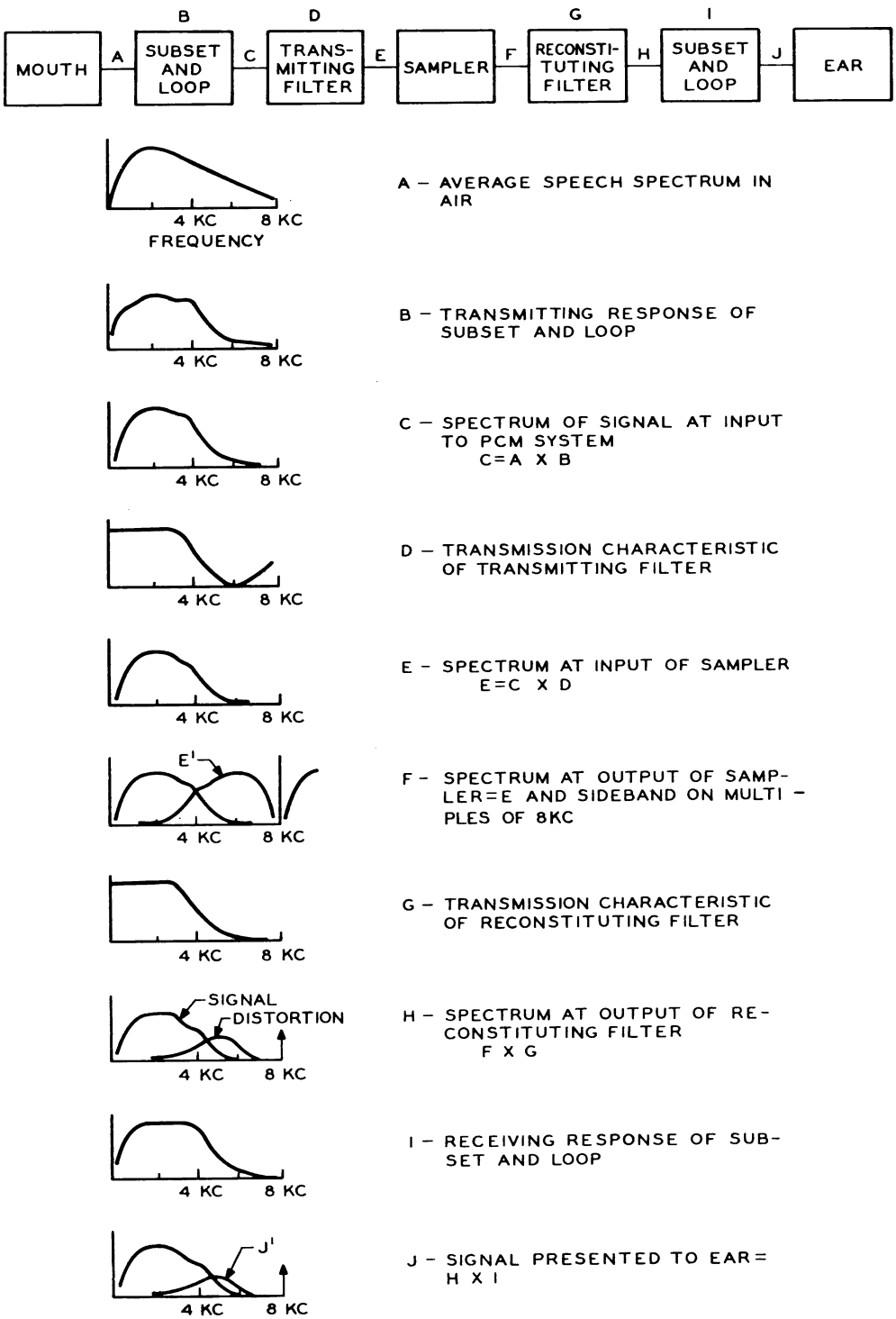
Let us trace the spectrum of a speech message from its emission by a subscriber, through the sampling and reconstruction process,



Clamper

Figure 26-7

to the ear of the receiving subscriber. The successive steps are shown on Figure 8. Spectrum (A) is the average speech spectrum in air. It is operated on by the transmitting response of the subset and loop (B) to produce spectrum (C) which has smaller high frequency components than (A). Spectrum (C) is that of the signal at the input to the PCM system. This spectrum is operated on by the characteristic of the transmitting filter (D) to give spectrum (E) at the input to the sampler. Spectrum (E) has again smaller high frequency components than (C). The spectrum of the samples (F), at the output of the sampler, consists of spectrum (E) as well as tones at integral multiples of the sampling frequency and images of (E) as upper and lower sidebands on these tones. Now we encode, transmit, and decode. Assume this is done perfectly. At the receiving end of the system, the re-constituted samples are passed



Spectra in a PCM System

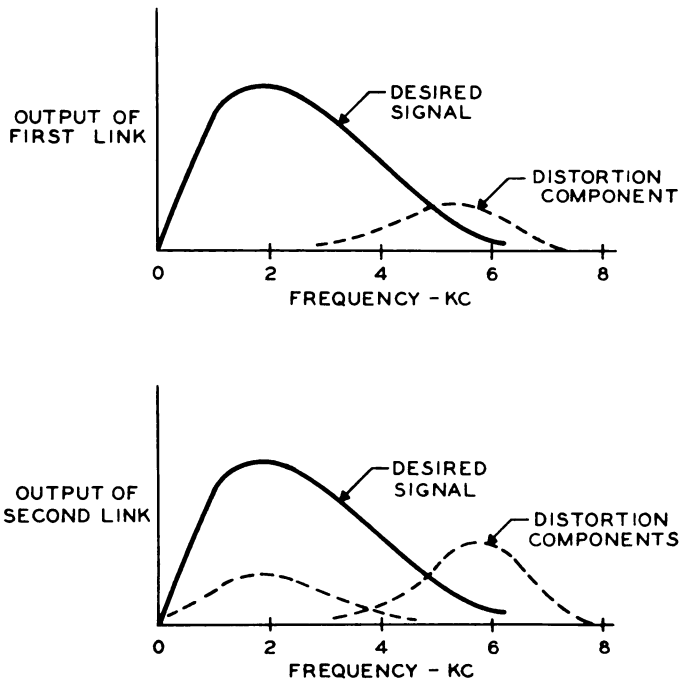
Figure 26-8

through the low-pass reconstituting filter (G) plus an amplifier which restores the power lost in the sampling process. Note that instead of being an ideal filter with a steep cut-off at 4 kc, this filter (G) cuts off so slowly that some 8 kc (sampling frequency) gets through. Spectrum (H) is spectrum (F) operated on by the reconstituting filter (G); presumably, the only appreciable distortion components are the first lower sideband on the sampling frequency and the sampling frequency tone itself. Spectrum (H) is operated on by the receiving subset and loop (I) to give spectrum (J) which is presented to the ear.

Spectrum (J) shows that even with non-ideal filters the spectrum of the original message can be recovered. However, this is accompanied by a residual lower sideband (J') and the sampling frequency tone, both of which constitute distortion or noise. These distortion components impair the quality of the original message and must be held to tolerable magnitudes by the system filters. Note that it is the entire lower sideband which produces distortion and not just the overlap in the 0 to 4 kc region, since the subscriber can hear, to some extent, up to 8 kc or so. Furthermore, since the sampling process has caused the interfering sideband to be inverted, components which originally fell in the 4 to 8 kc region (including, for example, single frequency tones associated with various systems in the telephone plant) will now fall in the 4 to 0 kc region where the ear is much more sensitive to them. Although such components might have been tolerable originally in a regular voice frequency connection, for example, they may become intolerable when translated lower in frequency.

When time division systems are used as blocks in a multi-link system, additional distortions related to that from the lower sideband are produced. Consider, for example, the message which goes into the input of the second link. This message will have the spectrum (H) of Figure 8. Since (H) has more energy above 4 kc than the original message spectrum (C), the spectrum at the output of the second link must contain more lower sideband type noise than the output of the first link. Figure 9 illustrates this point. Obviously, additional links cause the lower sideband noise to increase still further.

Let us now consider how the lower sideband noise is measured and briefly examine the filter requirements to meet our transmission objectives.



Spectra in Multi-Link PCM

Figure 26-9

Filter Requirements

It is of some interest to study how the effect of the lower sideband noise is currently being evaluated in order to set the requirements on the filters for the exchange area PCM system. A time division system has a number of sources of noise: quantizing noise (from the quantizer and from dc bias of the samples), compandor tracking noise, digit errors, etc. It is often convenient to treat these noise sources in terms of an equivalent amount of thermal noise, and to add these equivalent noise powers together to determine the overall system noise performance. At the moment, lower sideband noise in the exchange area PCM system is being handled in just this manner.

In Chapter 2 it was shown how noise in telephony is ultimately evaluated in terms of its interfering effect on speech. A convenient point at which to analyze the noise in a PCM system is the input to the receiving loop. However, the noise power at this point is important only with respect to its interfering effect on the subscriber. Therefore, if we are to express the lower sideband noise in terms of an equivalent thermal noise at the input to the receiving loop, we must

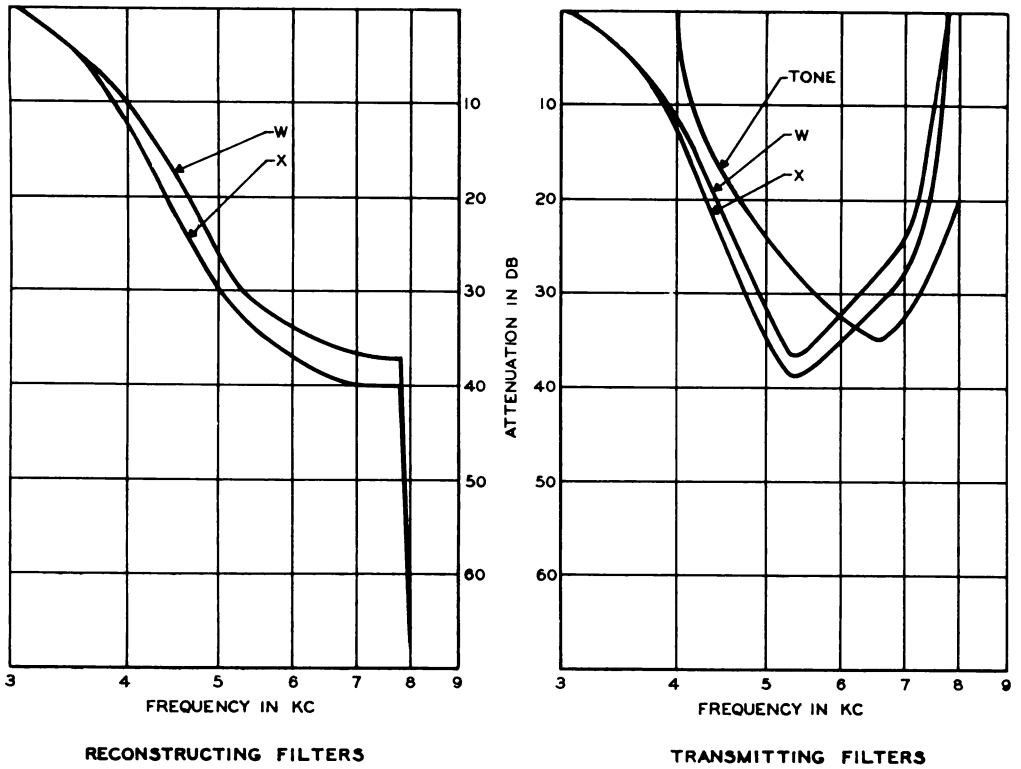
first weight the lower sideband noise by a suitable weighting function. In this case the weighting function must give the effect of the average loop, subset, and ear.* An assumption is now made: weighted noises with equal powers have equal interfering effect. Therefore, the equivalent noise power is that amount of 0 to 4 kc flat thermal noise at the input to the receiving loop which, when properly weighted, produces the same power as the lower sideband produces when similarly weighted. Whether this procedure for handling the lower sideband noise is conservative or not, only experiment can tell. In any event, the speech-correlated character of the lower sideband has been ignored, so that conservative requirements on the system filters should be made.

An objective measure of the effect of the lower sideband is the amount by which the lower sideband noise, at the input to the receiving loop, increases the noise power from all other sources already present at this same point. The sampling gate is a linear device in that the lower sideband produced is always the same fraction of the signal regardless of the volume of the message. Therefore, the loudest talkers will produce the lower sidebands having the highest powers. It is desirable that the degradation caused by the lower sideband be held as small as possible since the other noises present also cause degradations. As it turns out, it is easier and cheaper to control lower sideband noise in comparison with the other sources of noise. Therefore, the permissible degradation (or increase) in system noise caused by the lower sideband has been set at 0.5 db. Since the amount of lower sideband power is dependent upon the talker volume, this 0.5 db degradation must be assigned to a high volume talker. The present objective is to have the lower sideband noise of one talker in a thousand (the 99.9% talker) increase the noise at the input to the receiving loop by 0.5 db. This corresponds to a -1 vu talker at this point. Obviously, talkers having higher volume than this will cause more than 0.5 db degradation. However, the effect of the lower sideband is almost negligible for an average talker (-19 vu for local calls).

Figure 10 shows pairs of requirements on filter characteristics derived to meet the lower sideband noise objective discussed above. The pair labelled W would have characteristics which meet objectives for a single link system. If this objective is specified for a 5 link system,

---

*Currently, T1U (500-type subset) weighting is being used.



Filter Requirements

Figure 26-10

then the characteristics labelled X are required in each link. Obviously, if we hold the noise objective the same but increase the number of links, the filter requirements become more stringent. In all cases the lenient requirements on the transmitting filter loss beyond 5 kc are attributable to the fact that the speech spectrum falls off quite rapidly in this region. In practice the transmitting filter will more than meet the requirements of Figure 10 above 6 kc.

Looking at Figure 1, we see that any 8 kc tones coming out of the amplifier which follows the receiving low-pass filter can get across the hybrid and into the transmitting side of the circuit if the hybrid balance is poor. It is not surprising that we actually find relatively poor balance at so high a frequency. Such unwanted tones should not be allowed to reach the sampler if unwanted sampler distortion products are to be avoided. This effect may dictate additional filtering, at the sampling frequency, in the transmitting filter.

Single frequency tones above the speech band, which are produced by various systems in the telephone plant, may be coupled to the input of a time division system. As previously pointed out, 8 kc

sampling will translate these tones into the speech band since some lower sideband will fall in the 0 to 4 kc range. No requirements can be given on the transmitting filter, however, unless the degree of coupling of these tones into the system is known.

For the sake of comparison with the W and X filter requirements, a tone attenuation requirement for the transmitting filter can be derived by a somewhat different approach. Time division systems should be able to accept all tones considered tolerable for other systems in the plant. Therefore, let us assume that there are tones, in the 4 to 8 kc range, which are associated with the incoming message and which are just tolerable to the subscriber. These tones will no longer be tolerable after translation unless sufficiently attenuated by the transmitting filter. Using data on levels of tolerable tones in the plant, we can derive the amount of attenuation needed to make 4 to 8 kc tones tolerable after translation. The results are shown by the characteristic labelled "tone" in Figure 10. Observe that these requirements govern above 6200 cps, since above this frequency they are more stringent than the W or X curves based on lower sideband noise arising from inverted speech.

It should be emphasized that the procedures specified above for deriving filter requirements for time division systems have not been completely verified by experiment. Ultimately it will be necessary to determine by subjective tests the numerical equivalence between thermal noise and lower sideband. Furthermore, this will be done for different bands of noise, filters, and volume levels.

Quantization

The result of sampling and multiplexing is simply a pulse amplitude modulation (PAM) signal. It is significant that this signal will, in principle, contain all of the information that was in the original message. Just as the original message amplitudes generally occupy a continuous range, so will the sample amplitudes occupy a continuous range and thus take on an infinite number of possible values. We now propose to approximate this infinite number of signal amplitudes by a finite number of discrete values via the process of quantization in the encoding operation. In brief, the signal range is divided into a number of smaller ranges, and a discrete number is assigned to represent each minor range. This, in essence, is what we do when we round off a number to a fixed number of decimal digits. Loss of information is

inherent in this process, but this loss may be made negligibly small through the use of a sufficiently large number of minor ranges, or quantum steps. The loss of information so entailed is a small price to pay for the ultimate advantages that a PCM signal has to offer.

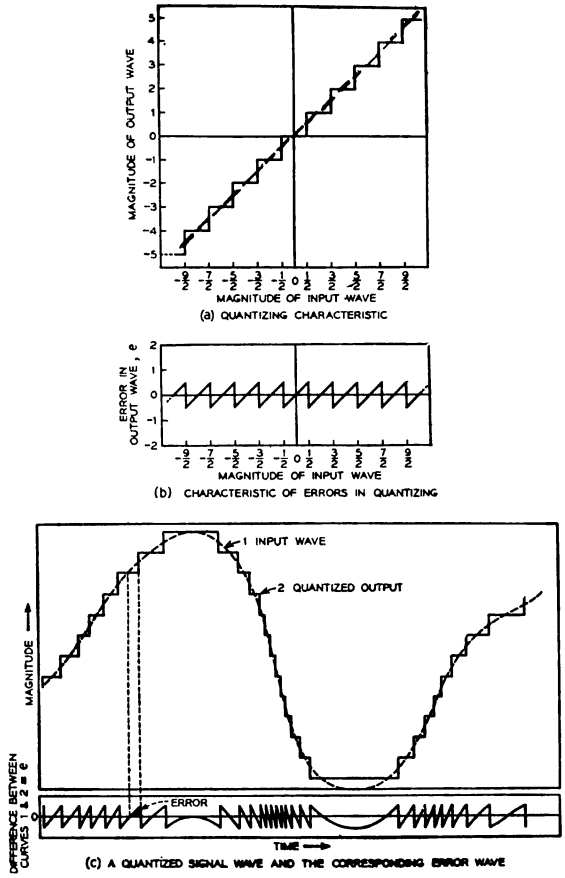
Quantization may be described graphically as shown in Figure 11, wherein a dashed straight line, representing the relation between input and output samples in a linear continuous system, is replaced by a flight of equal steps. The midpoints of the treads fall on the straight line, and the height of the step is the quantum. In Figure 11b, where the error due to quantization is plotted as a function of the input signal, note that the maximum instantaneous error is always minus or plus half a quantum step, regardless of signal strength. This error may, of course, be made arbitrarily small by merely using a large number of steps to cover the total range. However, since the number of steps used will directly affect the complexity of the encoding operation as well as the bandwidth requirement of the PCM transmission medium, it will be wise to use only as many steps as are really necessary.*

In Figure 11c, the error signal (original signal minus quantized signal) which in this case arises from quantizing a continuous input signal, is plotted as a function of time. Observe that the quantized output can be considered the sum of the original signal and the negative of the error signal. It is the presence of this error signal which gives rise to quantization distortion or quantization noise, which may, with qualification, be treated like any other noise that interferes with a signal. As usual, the signal-to-quantum noise ratio will be a useful criterion by which to judge transmission quality.

We note, from the preceding discussion, that the quantum noise is a function of the step size but independent of the signal magnitude. It follows that if equal-size quantum steps are used throughout the signal range, the signal-to-quantum noise ratio will be large for strong signals but very small (unfavorable) for weak signals. Clearly,

- -

*It should be recalled that if b is the base and n the number of digits, b^n represents the number of levels or quantum steps that are available for encoding the PAM signals. Also, the bandwidth required to transmit each message is nf_c, where f_c is the highest frequency component in the original message.

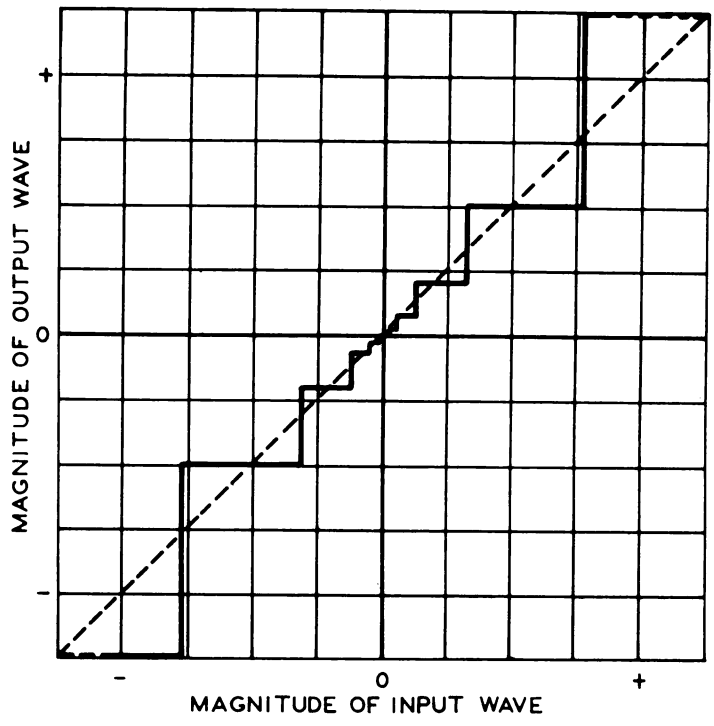


Quantizing Errors - Uniform Quanta

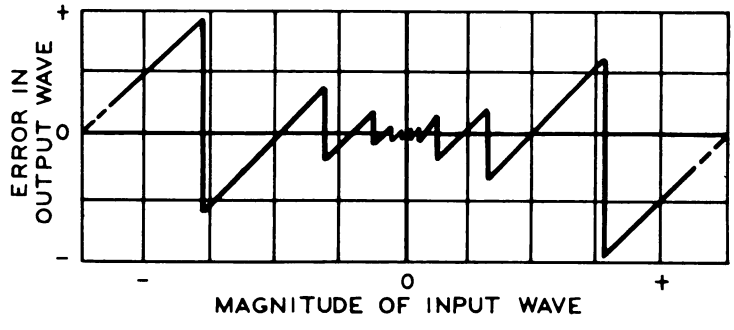
Figure 26-11

a more attractive situation will result if the step size is varied, rather than held uniform, throughout the signal range. That is, the relationship between input and quantized output would better be a tapered staircase function, as shown in Figure 12. Many small steps are provided in the weak signal range but only a few large steps are supplied for strong signal peaks. When compared with a uniform quantizer having the same number of steps, such a tapered quantizer will yield signal-to-quantum noise ratios that are markedly improved for weak signals, through slightly impaired for strong signals.

Although direct instrumentation of tapered or non uniform quantization is quite feasible, it is usually easier to achieve the same result indirectly by judiciously distributing the signal amplitudes over a given number of equal steps (rather than distributing the same number



(A) NONUNIFORM QUANTIZING CHARACTERISTIC

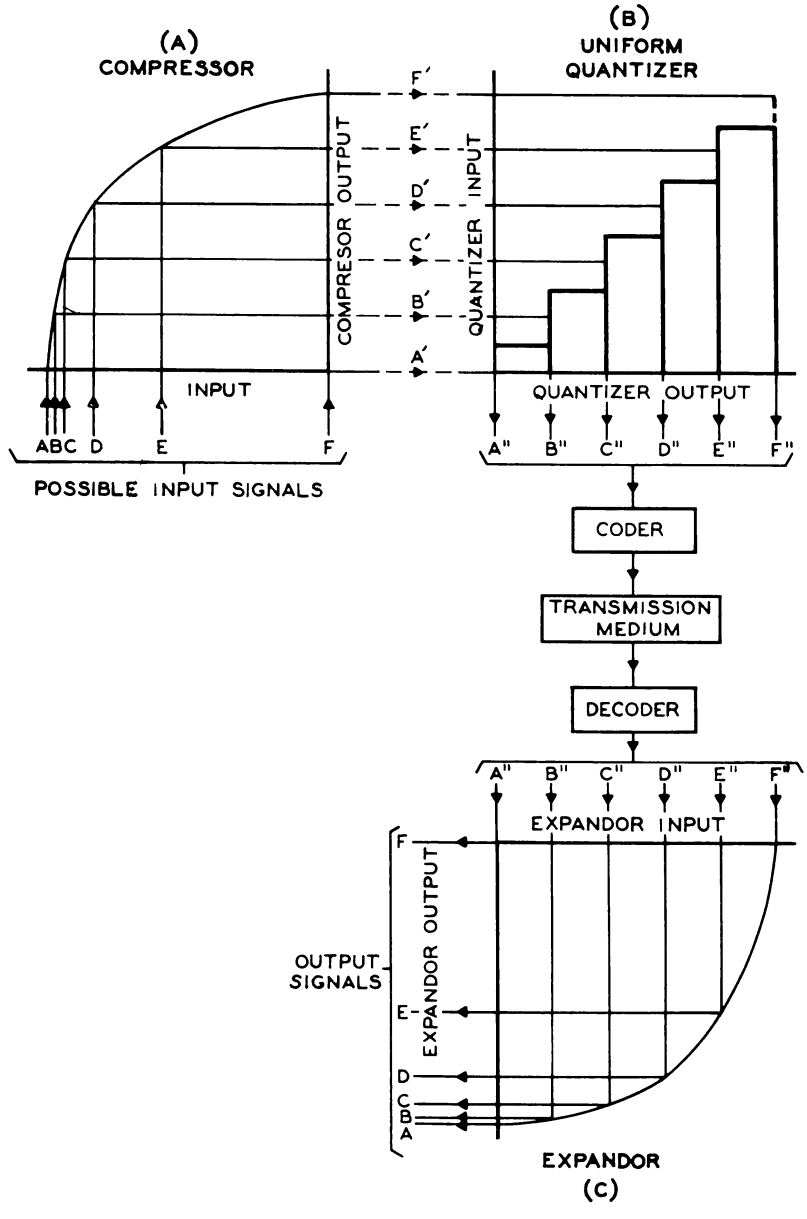


(B) CHARACTERISTIC OF ERRORS IN NONUNIFORM QUANTIZING

Quantizing Errors - Non-Uniform Quanta

Figure 26-12

of unequal steps over a given signal range.) This is usually accomplished via sequential operations of compression, uniform encoding and decoding, and expansion. Such a scheme is illustrated graphically in (a), (b) and (c) of Figure 13. For simplicity, only first quadrant variations are shown, but it should be recognized that the same general action will take place in the third quadrant since the other half of each curve has odd symmetry. Compression is accomplished in (a) by



Compression and Quantization

Figure 26-13

passing the signal through a non-linear transducer that modifies the distribution of amplitudes by preferential amplification of weak signals. When the result of (a) is introduced to the uniform quantizer (b), the weaker signals traverse many more quantum steps than they would have without preferential amplification. Their quantized approximations are

correspondingly improved at the output of (b), since a greater number of discrete levels have been utilized to define them. In fact, signals A, B, and C would all be represented as zero without compression since they are all less than half a quantum step. Strong signals do not fair so well, however, since their peaks are subjected to much less gain in (a) and therefore utilize fewer quantum steps in (b). Fortunately, most of the information in speech messages (or their PAM equivalent) is confined to amplitudes near zero, so that strong signal peaks can stand considerable impairment without having a significant effect on over-all quality. The net effect of introducing compressor (a) before uniform quantizer (b) is to greatly increase the effective signal-to-quantum noise ratio at the output of (b). Although the quantized output of (b) may thereafter be subjected to any of several linear operations, such as encoding and decoding, it must eventually be affected by the non-linear operation of expansion shown at (c), if the original message is to be recovered. The expandor in (c) must have an inverse characteristic to the compressor in (a) if the overall system is to be linear. This combination of compressor and expandor is commonly called a compandor and the net function performed is referred to as companding. All of these circuits must be "instantaneous" in the sense that they must respond to the instantaneous value of short pulses. More precise terms of "instantaneous compressor", "instantaneous expandor", and "instantaneous compandor", are therefore often used to avoid confusing these circuits with slow-acting, syllabic companding circuits used on a per channel basis in some AM frequency multiplex systems.

Instantaneous Compandors

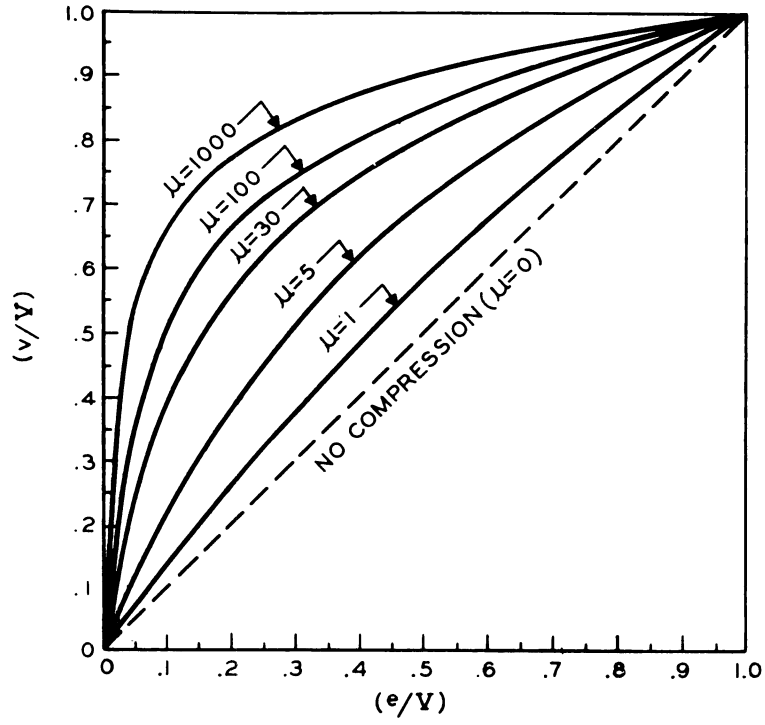
Regardless of whether the signal-to-quantum noise improvement is obtained by applying a linear amplitude range to a non-linear staircase, or by applying a non-linear amplitude range to a linear staircase, there is a preferred law that the non-linear characteristic should follow.

We shall restrict our attention to the properties of the logarithmic type of compression characteristic given by

$$v = \frac{V \log \left[1 + \left(\frac{\mu e}{V}\right)\right]}{\log \left[1 + \mu\right]} \qquad 0 \leq e \leq V \qquad (26\text{-}15)$$

where v is the output voltage in response to an input e, μ is the degree of compression, and V is the maximum value of the input signal. For

negative values of e replace V by -V. Typical compression characteristics are shown in Figure 14.



Typical Compression Characteristics

Figure 26-14

The improvement afforded by companding can be defined as

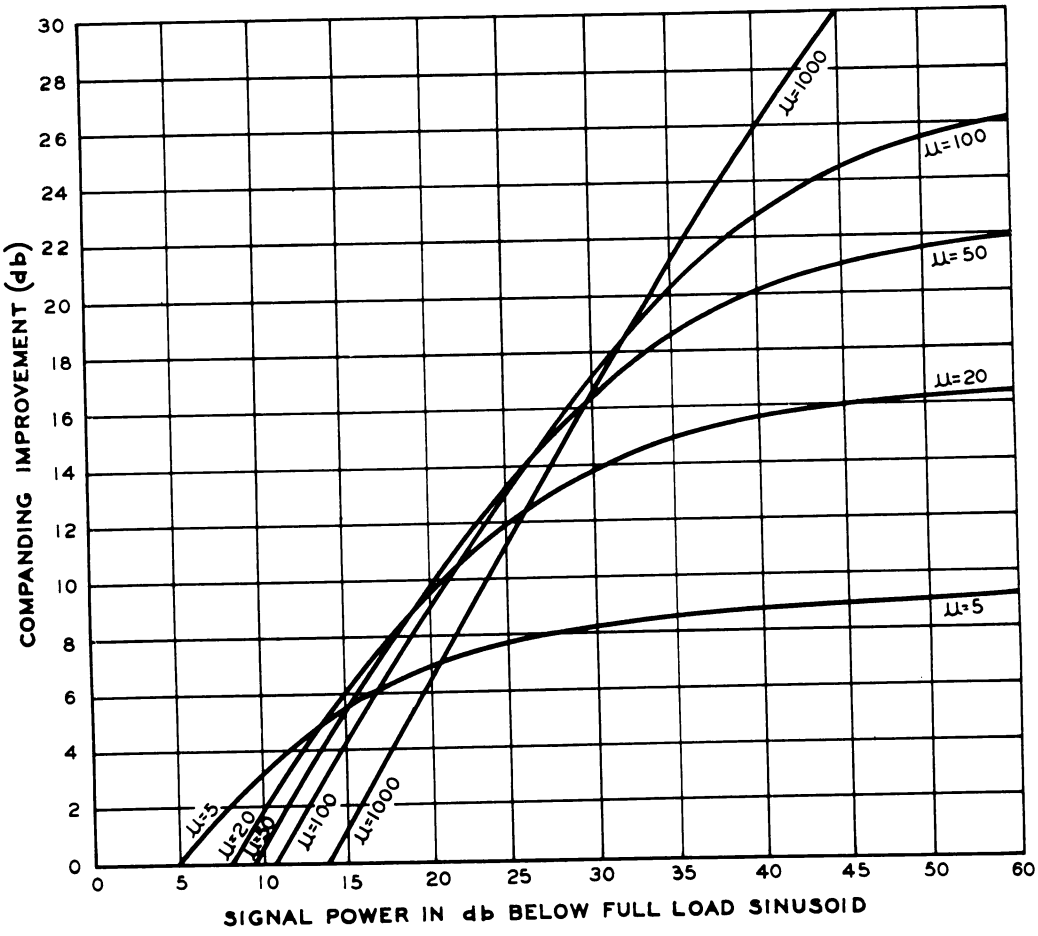
$$I^2 = \frac{\text{Mean square error voltage without companding}}{\text{Mean square error voltage with companding}}$$

The above ratio is plotted against signal power for various values of μ in Figure 15. As noted previously the improvement is greatest for the weak signals, while the strong signals (bull talkers) suffer some impairment.

Another measure of the efficacy of companding can be obtained by comparing the number of binary digits per sample required without companding with the number required with companding for the same quantizing error power for small signals. Since the quantizing error power is inversely proportioned to the square of the number of levels, i.e., $(b^n)^2 = 2^{2n}$, this power will be reduced 6 db for each additional digit.* If this 6 db

- -

*See Chapter 25, page 18.



Companding Improvement of Quantization Noise

Figure 26-15

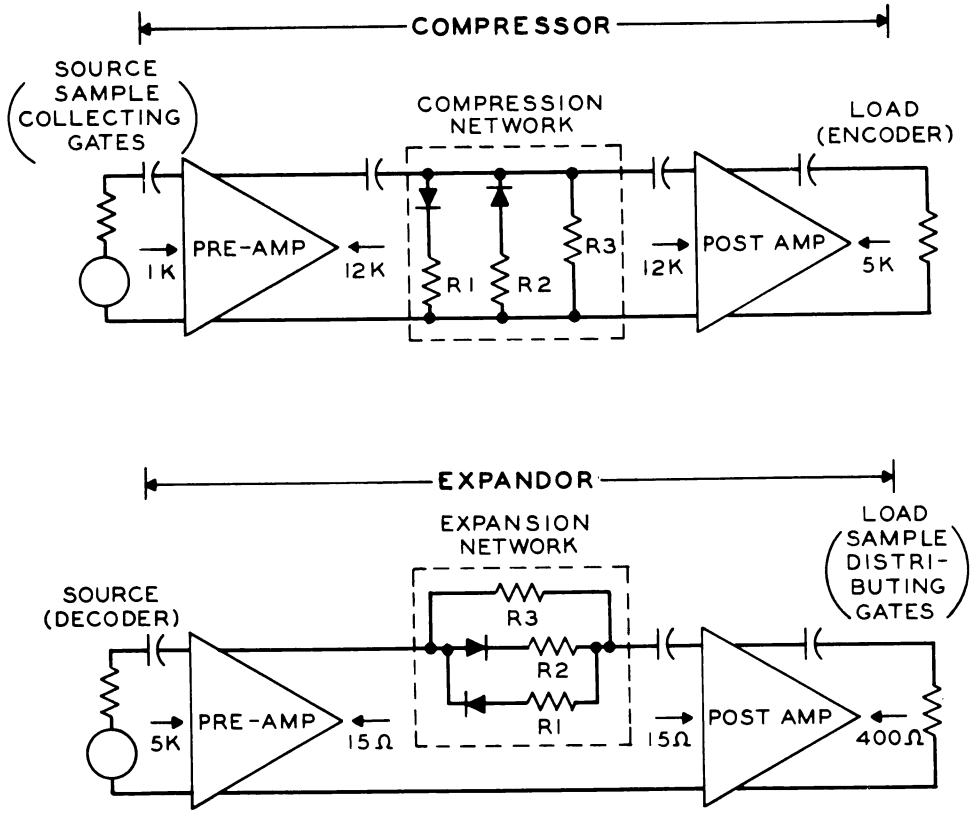
per digit improvement is compared with the roughly 24 to 35 db improvement for weak signals afforded by the compression characteristics of Figure 14 for $100 \leq \mu \leq 1000$, it can be seen that companding is equivalent to the addition of four to six digits. This amounts to an increase in the number of quantizing levels by a factor of between $2^4=16$ and $2^6=64$. For example, a 7 digit code group (128 quantizing levels) with a compression characteristic given by $\mu = 100$ is considered adequate by subjective standards for transmission of speech by PCM. Without the compandor, 11 digits would be required to maintain the same quantizing error power for weak signals. This would increase the pulse repetition

frequency by a factor of 11/7 (about 1.6 times) with an attendant increase in the speed required for encoding and decoding and an increase in the bandwidth required for transmission over the line.

The building blocks that make up an instantaneous compressor are shown schematically in Figure 16. The heart of the compressor in the non-linear network made up of diodes and resistors. The voltage-current characteristic of a semiconductor diode is given by

$$V = \left(\frac{kT}{q}\right) \log \left(1 + \frac{I}{I_s}\right) \qquad (26\text{-}16)$$

which is of the same form as the previously described compandor characteristic. It is necessary to handle bipolar PAM signals from the output of the clamper circuit, thereby dictating the use of two diodes, one for each polarity. Resistors R_1, R_2 and R_3 provide a means for adjusting the circuit to compensate for the fact that all diodes are not identical.



Block Schematic of Compressor and Expandor

Figure 26-16

To avoid masking of the non-linearity, the compression network must be operated between high impedances as shown in Figure 16. Since the expansion network is the inverse of the compressor network, it is operated in series with a generator and load of low impedances. The proper impedances facing the non-linear network are maintained by the pre- and post- amplifiers. These amplifiers must have sufficiently good high-frequency transmission to handle the narrow pulses from all of the channels. In addition, the low frequency cutoff should be sufficiently low to prevent appreciable crosstalk on the tail of one pulse with its neighbors, since if low frequency components are not adequately transmitted, overshoot and overhang will occur.

Encoding and Decoding

Thus far, all of the tandem operations described for processing the speech signals are common to both PAM and PCM systems. The remaining function peculiar to PCM is the process of encoding the PAM samples into a binary code group.

Coding speed is limited by the number of decisions which must be made concerning any signal and by the subsidiary actions taken because of those decisions. We can distinguish three different classes of encoders on the basis of how the encoder phrases the questions leading to its final decision. The three classes are

1. Level-at-a-Time Encoders
2. Digit-at-a-Time Encoders
3. Word-at-a-Time Encoders

Level-at-a-Time Encoders

In a level-at-a-time encoder the value of each code combination is compared in turn with the compressed PAM signal to arrive at a level which most closely approximates the input amplitude. Say, for example, that an amplitude level corresponding to code number 27 (out of a possible 128) is presented. The level-at-a-time encoder might be thought of as asking, in sequence: "Is it level #1? Is it level #2? Is it level #3..." until it gets a "yes". This is a rather slow process since time must be allowed for one comparison for each of the 2^n levels in an n digit system. In terms of the sort of system we have been discussing, we want to code a sample into seven digits in about 5 μs. This is less than one microsecond per decision, so that speed of operation is quite important. Therefore, this method of encoding is confined to relatively low speed systems and will not receive further consideration here.

Digit-at-a-Time Encoders

These encoders require only n decisions, as against the 2^n (maximum) required in level-at-a-time circuits. The penalty is a modest increase in the complexity of the subsidiary actions. Recall the significance of the digits in a seven-digit binary code:

Position:	1	2	3	4	5	6	7
Value of Pulse:	64	32	16	8	4	2	1

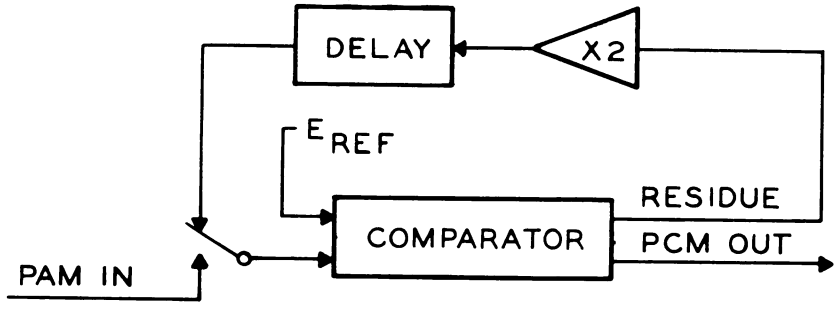
The decision sequence in the digit-at-a-time encoder for code #27 is: "is the amplitude greater than 64? (No) Is it greater than 32? (No) Is it greater than 16? (Yes) Is the remainder 27-16 greater than 8? ..."

All of the known practical types of digit-at-a-time coders operate on the basis of determining the most significant digit first. This type of coding can be viewed as a geometric progression in which it is first determined in which half of the entire group of possible codes the signal sample should be placed. Then it is determined in which quarter of the selected half the signal sample should be placed. Next the eighth of the selected quarter is determined, and so on. The process is carried on until the location is specified to the desired degree of fineness. As we have seen, in order to determine the location to one part of 128, seven separate selections have to be made. One additional selection would permit selection to one part in 256, etc. With this type of coder, the digits can be transmitted as determined, so it is not necessary to wait until the end of the coding operation to send out the code "word" or "character".

There are many types of digit-at-a-time encoders, but only two types will be discussed here: the "recycling encoder" and the "network encoder". Both of these belong to a general class of "sequential comparison encoders". Examples of their operation are given a little later. The basic difference between these two types is the method by which the function of memory or storage is obtained. In the case of the network encoder most of the essential functions are performed by a digital process, whereas in the case of the recycling encoder these functions are analog. It would be expected that these analog functions could be performed with fewer number of component parts than required for digital storage. On the other hand, we would also expect that it would be more feasible to obtain the necessary accuracy of the system by digital means.

The Recycling Encoder

A block schematic of the recycling encoder is shown in Figure 17. In this circuit two outputs are available from the comparator. One is a digital output used for the PCM signal. The second is an analog residue which is multiplied by a factor of 2, delayed one digit of time in the delay network, and then applied to the input to the comparator for determination of the next digit.



Block Schematic of Recycling Encoder

Figure 26-17

Consider, as an example, the encoding of a signal voltage of 25.2 units into a five-digit code (providing 32 levels).

Multiplication Method, Recycling Encoder

Arithmetic	PCM Output	Action of the Control Circuits
25.2 - 16 = 9.2	1	Subtract
18.4 - 16 = 2.4	1	Subtract
4.8 - 16 < 0	0	Don't Subtract
9.6 - 16 < 0	0	Don't Subtract
19.2 - 16 = 3.2	1	Subtract

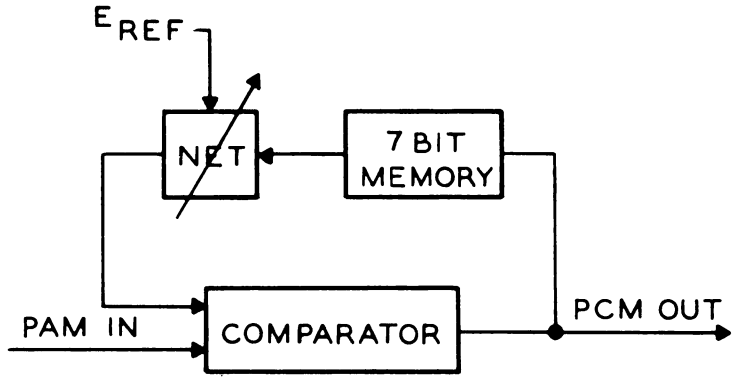
Logic: Positive residue - PCM output; subtract and multiply the residue by 2.
Negative residue - No PCM output; no subtraction but multiply by 2.

The 11001 code has a value of 16+8+0+0+1, or 25. The fact that quantizing has introduced an error of 0.2 is obvious.

The Network Encoder

In the case of the network encoder, only a digital output is taken from the comparator. This is used in the formation of the PCM

code character and is also stored in the 7-digit memory circuit. The memory circuit operates a binary-weighted network, the output of which adds algebraically with the PAM signal input to determine the subsequent comparison to be made. Figure 18 illustrates the network encoder in block schematic form.



Block Schematic of Network Encoder

Figure 26-18

For this type of encoder, the coding of 25.2 into a five-digit code would be accomplished in the following steps:

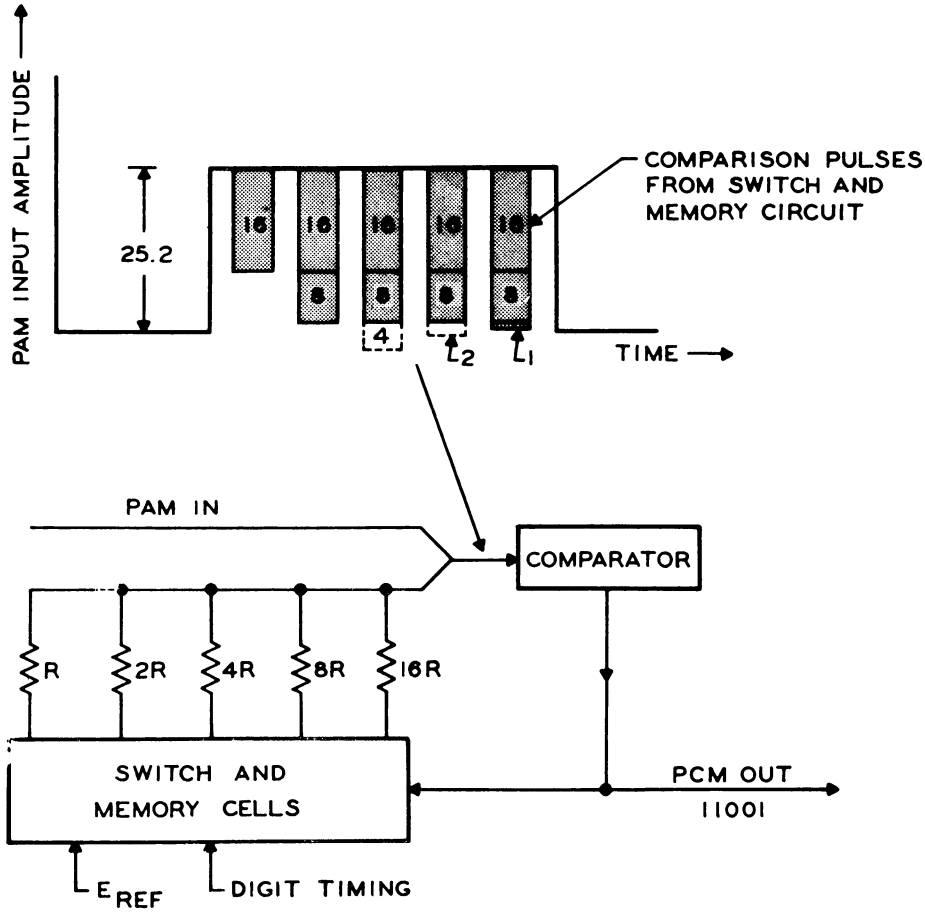
Subtraction Method, Network Encoder

Arithmetic	PCM Output	Action of the Control Circuits
25.2 - 16 = 9.2	1	Subtract
9.2 - 8 = 1.2	1	Subtract
1.2 - 4 < 0	0	Don't Subtract
1.2 - 2 < 0	0	Don't Subtract
1.2 - 1 = 0.2	1	Subtract

Logic: Positive residue - PCM output and subtract.
Negative residue - No PCM output and no subtraction.
The residue is obtained by subtracting 16, 8, 4, 2, 1 at the respective steps of the process.

A recent comparison of recycling encoders and network encoders indicates that, with currently available components and techniques, the network encoder is more feasible. This picture may change as a result of future development, particularly in the component field. However, in view of its present state of development we will concentrate our attention on the network encoder.

The functional operation of the network encoder can be most easily described by the simple diagram in Figure 19. Again, we will use a PAM signal of 25.2 units amplitude to demonstrate the action of the encoder. From the diagram we see that the signal is applied at the input of the comparator and combined with the output of a binary network. This network is operated by switches controlled by memory cells and digit timing. In the case of the first comparison a timing pulse causes the binary network to subtract sixteen units from the PAM signal. Since this residue is positive, the comparator gives a "1" at its output which is transmitted as the first digit of the PCM signal and is also stored in the memory cells. With a stored digit indicating "16", the network will proceed, subtracting the "16" again and again for each suceeding comparison. As it accumulates additional digits they, too, will be subtracted repeatedly in subsequent comparisons. Thus, for the next digit a total of



Network Encoder

Figure 26-19

16+8 = 24 units are subtracted. Again, the residue is positive and the comparator gives a "1" at its output which is transmitted as the second digit of the PCM signal and which is also stored in the memory cells. For the third digit, then, a total of 16+8+4 = 28 units are subtracted. Here, however, the result is negative. Therefore, the resulting output code is "0", and the four is not used again in the subsequent subtractions from the PAM pulse. This process continues until the final comparison is made for the least significant digit. At this time the voltage available from the binary weighted network is equal to the PAM signal to within one quantum step.

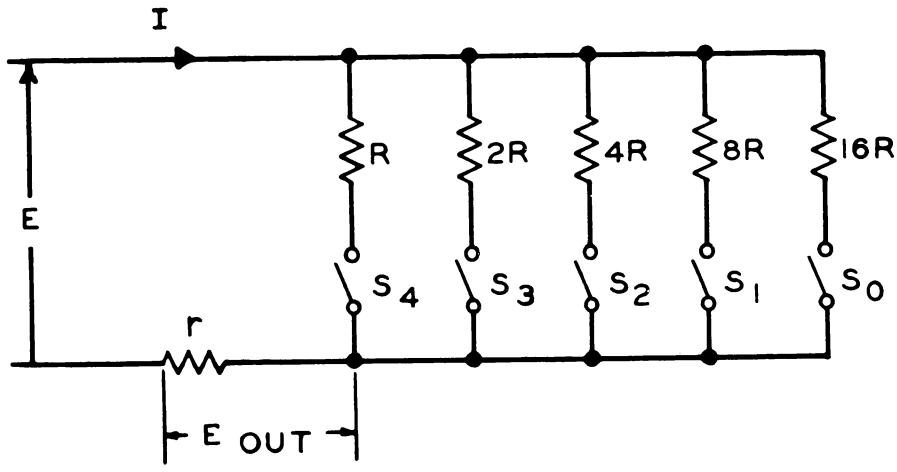
Word-at-a-Time Encoders

The word-at-a-time encoders are faster than the previous types of encoders since only one question has to be answered. The question is essentially "What word (code combination) does this signal level represent?" In order to answer this question, all possible words must be stored internally in the coder and the one that most closely approximates the analog signal selected. Probably the best known of the word-at-a-time coders makes the analog to binary conversion by means of a beam coding vacuum tube. This coder is described in considerable detail in two BSTJ papers. The interested reader is referred to these papers (listed in the references at the end of the chapter) for further details.

Digit-at-a-Time Decoders

Figure 19 can be modified to form a digit-at-a-time decoder by using the combination of digit timing and PCM input signal to connect the reference voltage to the proper resistors in the binary weighted resistance network. This type of decoder simply adds up the weighted value of each digit which is present as a pulse, so that the output will be the sum of the total weighted values used. In this case the comparator is not required. The decoder might then take the simplified form shown in Figure 20. The combination of timing control and the PCM input signal determines the particular set of switches which are closed. In this figure a 5 digit system is assumed. When a particular switch is closed current flows through the pick-off resistor r, and through the binary weighted resistance network. If the binary signal represents the amplitude level V given by

$$V = \sum_{n=0}^{4} a_n 2^n$$



Simplified Decoder - 5 Digits

Figure 26-20

where the a_n's are either zero or one depending upon the presence or absence of the n^{th} digit, we can adopt the convention that switch S_n is closed when $a_n = 1$ and switch S_n is open for $a_n = 0$, where n=0,1,2,3,4. In this case current I of Figure 18 is given by

$$I = \frac{E}{\frac{1}{\frac{a_o}{16R} + \frac{a_1}{8R} + \frac{a_2}{4R} + \frac{a_3}{2R} + \frac{a_4}{R}} + r} \qquad (26\text{-}17)$$

If $r \ll R/2$, then

$$I = \frac{(a_o + 2a_1 + 4a_2 + 8a_3 + 16a_4)}{16R} E \qquad (26\text{-}18)$$

and the voltage drop across r is given by

$$E_{out} = \frac{E}{16R} [\, a_o + 2a_1 + 4a_2 + 8a_3 + 16a_4 \,] \qquad (26\text{-}19)$$

$$= K [\, a_o + 2a_1 + 4a_2 + 8a_3 + 16a_4 \,] \qquad (26\text{-}20)$$

where $K = \frac{E}{16R}$

The example previously used to illustrate the encoding operation can be applied in reverse to the decoder. Assume the binary sequence 11001 is applied to the decoder. In this case switches S_4, S_3 and S_0 will be closed, so that a_4, a_3, and a_0 will be unity while a_2 and a_1 are zero. With these values in Equation (26-20) we get 25k units of voltage for the decoded PAM sample. After each PAM sample is extracted, all switches are opened and the decoder is then ready to operate on the next word. It should be pointed out that this is only one form of resistance network decoder. There are, in fact, many other types of resistance networks that can be used in performing the decoding function.

Terminal Control

All of the preceding discussions have been concerned with the information handling circuit functions required to convert voice and signalling information into discrete quanta or bits for transmission over telephone lines. In this section we will briefly describe the timing control functions required to program these operations. Methods for achieving synchronism and framing of receiving and transmitting terminals will also be outlined.

The control circuitry may be divided into four major blocks:

A. Master oscillator
B. Digit Counter
C. Channel Counter
D. Framing System

Master Oscillator, Digit Counter, and Channel Counter

The basic digit rate of a PCM system is determined by the system requirements. For example, consider the case of 24 voice channels, each encoded into 7 digits, with an eighth digit allocated to high speed signalling. It follows that 24 x 8 = 192 time slots per frame are required to encode the voice and signalling information. If an extra time slot is provided per frame for synchronizing the receiving and transmitting terminals, a total of 193 time slots per frame must be transmitted. Since the frame rate equals the sampling rate, 193 time slots per frame multiplied by an 8000 cps sampling rate gives a basic pulse repetition rate of 1.544 mcps.

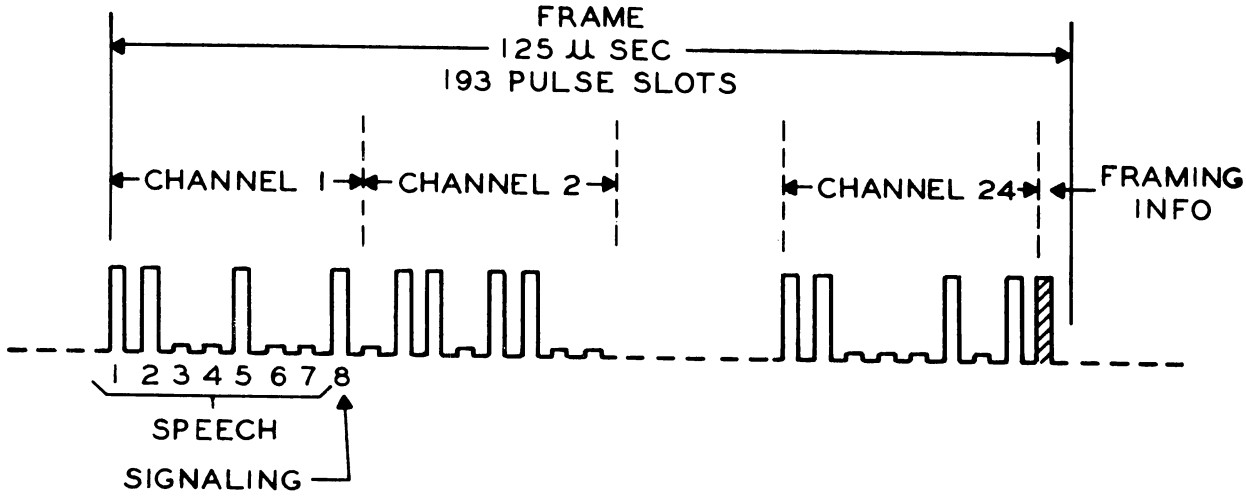
In a transmitting terminal, a 1.544 mcps crystal controlled master oscillator is used as the source of precise timing signals. It is convenient to think of the master oscillator as accurately marking off 1,544,000 time slots per second from which control and timing information

in the form of pulses can be obtained for various points in the system. Part of the output of the master oscillator is fed into the digit counter. The function of the digit counter is to count off time slots at the master oscillator output and supply one pulse every eighth slot. These pulses correspond to the basic system sampling frequency of 192 kc (24 PAM signals every 125 μs = 192 kc) and, as such, provide digit timing information to the encoder, signalling gates, and channel counter. In addition, the digit counter provides the framing pulse (alternately on and off) transmitted at the end of each frame. The channel counter operates from the output of the digit counter to provide one pulse every 193 time slots for operation of the individual channel sampling gates.

Framing System

At this point, before discussing the framing system, it is worthwhile to examine a typical pulse pattern which we might expect to find in a 24 channel system. Drawn on Figure 21 is the pulse pattern corresponding to a single frame -- i.e. speech samples and signalling which result from sampling each channel once. Certain points might be noted in connection with the question of what pulse trains are possible. Some of the things to consider are:

1. The first seven pulses may represent either a speech signal or silence. Zero and 127 would represent extreme voltage excursions of loud talkers in the negative or positive direction. Silence is represented by a value of 63. (In the absence of speech, noise might cause the value to jitter around 63.)
2. For reasons to be discussed in the next chapter, we want to limit the maximum number of pulses that can occur in sequence. We do this by never transmitting the value 127, which would call for seven pulses in a row in one channel slot. This turns out to limit the maximum number of pulses in sequence that can occur to thirteen. (See Figure 21.)
3. The eighth pulse (on- or off-hook signalling pulse) may be present or absent whether or not speech is, since we may use it for dialling, with no speech signal present, and must be able to talk through an on-hook signal when we reach an intercept operator.
4. For revertive pulsing, the seventh digit may be used for signalling (as well as the eighth) while the connection is being set up. When the connection has been established, the seventh digit can return to its role of transmitting message information.



Digit Allocation

Figure 26-21

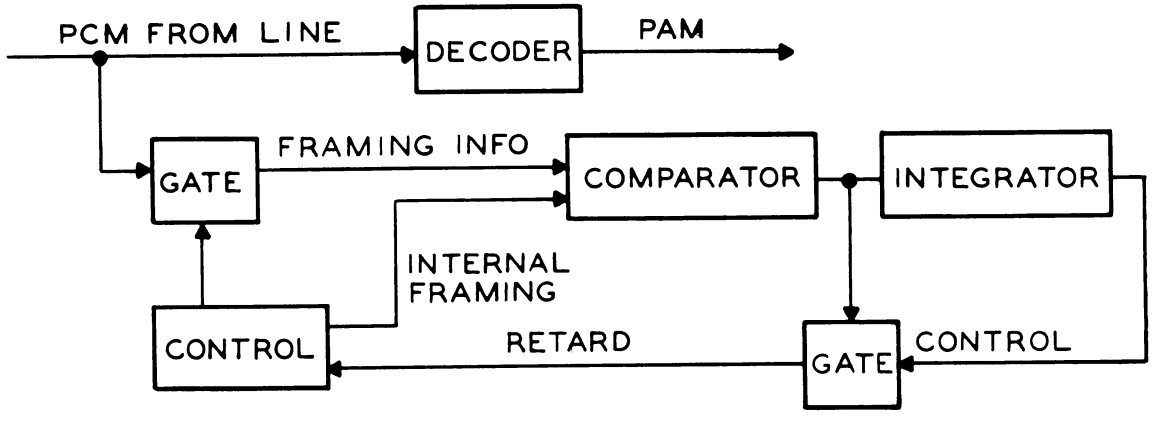
Thus, in Figure 21 note the 8 pulses that are allocated to Channel 1. The first 7 pulses represent the speech sample -- in this case indicating 64 + 32 + 4 = 100 level. The eighth pulse represents the signalling information. After the pulses corresponding to the 24 channels occur, an additional pulse (shown shaded) is added. This is the framing pulse and is required to permit synchronization of the two ends of the system. In ordinary operation, signals obtained from Channel 1 are properly reconstituted on the other end of Channel 1. Provisions must be made to restore proper operation when the system "goes out of frame."

To perform the act of framing, two types of systems are available. In the case of a "forward acting" system, a definite signal is sent prior to sampling the first channel. Thus, the system is brought into synchronism every frame. For a system utilizing only one polarity of pulse (which allows very simple repeaters) a fairly long and complicated pulse pattern would be necessary. Such an approach would then be fairly wasteful in bandwidth and require considerable apparatus.

An alternative approach has the receiving terminal monitor the incoming signal to determine whether the system is in frame. When lack of synchronization is observed, the system then hunts in an orderly manner for the condition of proper synchronization. The process of restoration is such a "backward acting system" is similar to that experienced in a television receiver when, after disturbance due to noise,

the picture slowly rolls back into vertical synchronization. As shown in Figure 21, this requires one additional pulse slot per frame, which is pulsed every second frame. On the average, reframing time may be expected to be about 25 ms, but may be as long as 50 ms in unfavorable cases. During this time the subscriber hears noise; however, loss of frame is expected to be a rather rare occurrence.

Circuitry in the receiver capable of recognizing the absence of synchronization and performing the correcting action is shown in Figure 22. When the control indicates that a frame pulse is expected, it operates a gate and sends the pulse received over the line into a comparator. On the other input to the comparator, the control places a signal corresponding to the pulse or space expected for this frame. If a match is obtained in the comparator, the system is probably in frame and no action need be taken. If the comparator indicates a mis-match, the system is definitely not in frame unless noise on the line has affected the framing pulse. Because of the possibility of noise on the line, one does not wish to initiate action immediately, but rather wait until several mis-matches have been obtained. For this purpose, an integrating device is provided. When one is sure, however, that the system is indeed out of frame, correcting action should be taken on every mis-match. The integrator, therefore, operates a gate which feeds the output of the comparator directly into the control so as to retard the system one pulse slot at a time until synchronism is restored.



Framing Circuit - Receiver

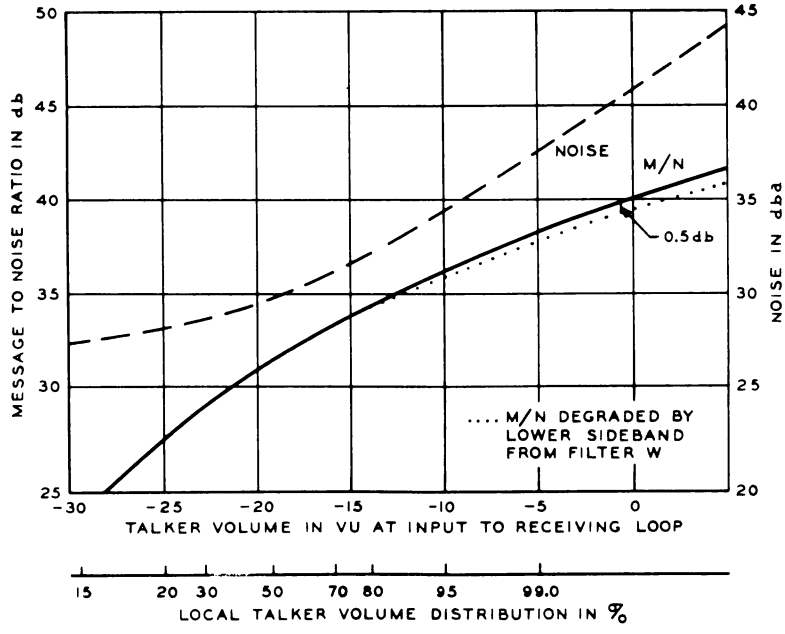
Figure 26-22

Estimated System Noise Performance

We have pointed out that quantizing noise is the dominant source of noise in a PCM system, with lower sideband noise giving some additional degradation of the loud talker. Let us now consider in more detail the sort of noise performance we might expect in the 24 channel exchange area PCM system.

Some subjective comparisons between an exploratory single link PCM system and a standard voice frequency circuit have been made. The same speech signal was fed into both systems with a variable, properly weighted thermal noise source included in the voice circuit. This noise source was gated so as to be present only during signal transmission to make the noise in the standard circuit behave like the quantizing noise in the PCM circuit. Each test subject was told to adjust the noise source until he considered the noise in the standard circuit to be equivalent to the noise in the PCM system.

The results of these tests are shown in Figure 23. The solid line shows the message-to-noise ratio, as a function of talker volume, which was obtained when all sources of noise except lower sideband noise were present. We see that the message-to-noise ratio increases with volume, as we would expect. However, the increase is not linear because of the effect of increased quantization noise as the talker volume increases.



S/N and N vs. Talker Volume

Figure 26-23

In the discussion of filter requirements it was stated, as an objective, that the lower sideband noise would be allowed to degrade the system noise by 0.5 db for the -1 VU talker. Therefore, if the weighted thermal noise equivalent to the lower sideband noise is added on this basis, the dotted curve of Figure 23 results. It is clear that the lower sideband noise causes negligible degradation to the average volume talker.

The message-to-noise data obtained from these subjective tests provides enough information to plot the noise in dba as a function of talker volume. This is shown by the dashed curve of Figure 23. We see that the noise is around 28 dba for weak volume talkers, but exceeds 40 dba for loud talkers. At first glance, 40 dba might seem like an excessive amount of noise. However, we must remember that in PCM the quantizing noise is only present while conversation is being transmitted. Noise during silent intervals is about twice as objectionable as noise accompanying speech. Therefore, in conventional systems, the noise requirement is primarily set by the permissible noise in silent intervals and is in terms of an absolute quantity of noise. In contrast to this, when noise is only present during speech, it is the message to noise ratio which is the important quantity. Generally, if the message-to-noise ratio is of the order of 20 db, the transmission quality is satisfactory. We see that this requirement is met for the weak talkers, and, therefore, we should expect the transmission to be of more than adequate quality for the average and loud talkers.

Conclusions

In preparing and processing signals for transmission by PCM, the tandem operations on the signal include filtering, sampling, compression, and encoding. These functions are under the control of a local clock. Reception of the signal is effected by operating on the PCM pulse train with decoding, expanding, demultiplexing, and filter networks. The receiving terminal timing circuitry is locked to the transmitter. These terminals bear striking similarities to internally programmed digital computers since they contain the essential ingredients of these computers, i.e., input-output equipment (4 wire terminating set), arithmetic unit and memory (encoder and decoder) and a local clock for internal program control. Communication between these computers is in the binary language. In order to have the computers communicate with one another, a transmission path matched to this language must be provided. This problem of transmitting binary pulse trains will be covered in detail in the next two chapters.

References

Sampling and Reconstruction

Modulation Theory - H. S. Black, D. Van Nostrand 1953.

Time Division Multiplex Systems - W. R. Bennett - BSTJ, v 20, 199-221 April 1941. BTL Monograph 1291.

Sampling Gates

Semiconductor Diode Gates - L. W. Hussey - BSTJ, v. 32, 1137-1154, Sept. 1953 - BTL Monograph 2197.

Companding

Spectra of Quantized Signals - W. R. Bennett - BSTJ, v. 27, 446-472, July 1958 - BTL Monograph 1586.

Companding in PCM - P. A. Reiling - Bell Lab. Record - v. 26, 487-490, December 1948.

Quantization Distortion in a PCM System with Non-Uniform Spacing of Levels - P. F. Panter and W. Dite - Proceedings of IRE - January 1951.

Encoding

Decoding in PCM - R. L. Carbrey - Bell Lab. Record - v. 26, 451-456 - November 1948.

Electron Beam Deflecting Tube for Coding in PCM - R. W. Sears - BSTJ, v. 27, 44-57 - January 1948.

Control

Timing Control for PCM - A. E. Johanson - Bell Lab. Record, v. 27, 10-15, January 1949.

Synchronization for the PCM Receiver - J. M. Manley - Bell Lab. Record - v. 27, 62-66 - February 1949.

Coding by Feedback Methods - B. D. Smith - Proceedings of IRE - August 1953.

General

Television by Pulse Code Modulation - W. M. Goodall - BSTJ, v. 36, 33-49, January 1951.

Pulse Code Modulation - H. S. Black and J. O. Edson - Trans. AIEE - v. 66, 895-9, 1947.

The Philosophy of PCM - B. M. Oliver, J. R. Pierce, and C. E. Shannon. BTL Monograph B-1611.

Experimental Multichannel Pulse Code Modulation System of Toll Quality - L. A. Meacham and E. Peterson - BSTJ, v. 36, 1324-1331 - November 1948.

Chapter 27

PULSE TRANSMISSION AND RESHAPING

Regeneration - the pay-off in PCM - is dependent on recognition, which in turn is dependent on the degree to which pulses are distorted in transmission. It is found that a loss characteristic with a gentle roll-off is preferable to a sharp cut-off. The fact that dc and very low frequencies are not transmitted calls for limiting the maximum number of pulses that can occur in sequence, and for the use of dc restoration circuits. Fine structure deviations across the transmission band, which cause echoes, must be kept small to avoid excessive intersymbol interference.

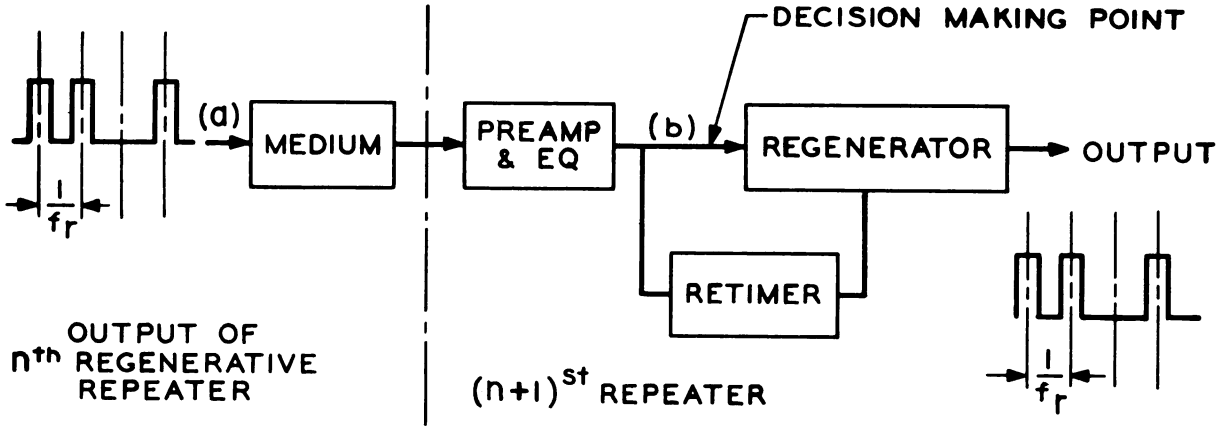
Introduction

In this chapter, transmission-frequency characteristics will be examined to determine the type of frequency response to which we must equalize a medium for satisfactory pulse transmission. Several types of imperfections in the characteristics of the medium will be investigated to determine their effect on the ability to transmit pulses with a low frequency of errors. Most of the theoretical considerations are applicable to data transmission over ordinary telephone lines as well as to PCM. Attention will be focused largely on the impact of these considerations on the transmission of baseband binary PCM.

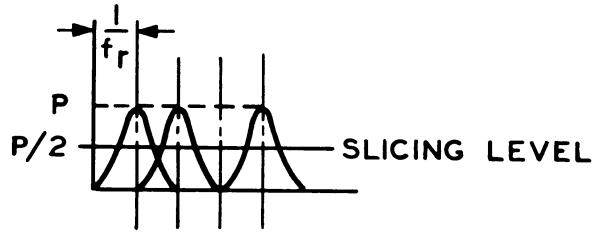
Before the aim of this chapter can be attacked, it will be necessary to bring the problem into better focus, define terms, and review some fundamental properties of transmission systems. To this end we examine the block schematic of a PCM link consisting of one regenerative repeater and the transmission medium between repeaters, as shown in Figure 1. The pulse train entering the medium is assumed to be a binary pulse train made up of rectangular pulses and spaces occurring at a pulse repetition frequency approximately equal to f_r. In the ideal case the pulse repetition frequency would be f_r, and each pulse would be of standard width and height. Imperfections in the previous regenerative repeater, as well as noise, cause the pulses to deviate from the desired spacing, width, and amplitude. The function of the next regenerative repeater is to accept this imperfect pulse train after it is dispersed by the transmission medium and generate a pulse train which is close to being a replica of that originally sent from the transmitting terminal. This process has been referred to as "the payoff in PCM".*

---

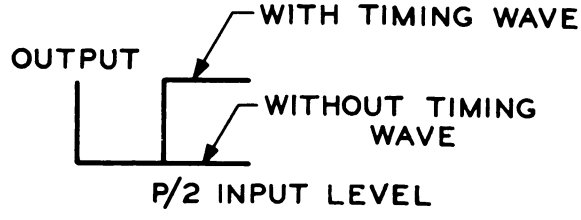
*B. M. Oliver, J. R. Pierce and C. E. Shannon, "The Philosophy of PCM", Proc. IRE, Nov. 1948.



(a) BLOCK SCHEMATIC OF PULSE TRANSMISSION REPEATER SECTION IN PCM



(b) PULSE TRAIN AT (b)



(c) TRANSFER CHARACTERISTICS OF IDEAL REGENERATOR

PCM Transmission

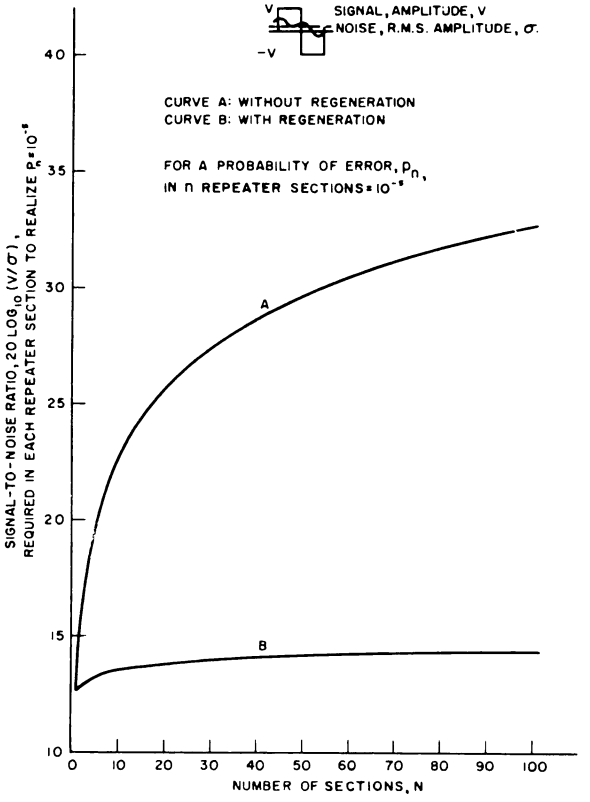
Figure 27-1

In the ideal case, with perfect regenerators, there would be no variation of the regenerated pulses from their desired positions (i.e., "jitter") and each pulse would match the standard in shape. Under these conditions the overall system requirements would be almost identical with those of each repeater section. This is contrasted with AM frequency multiplex systems where the transmission requirements of a repeater to repeater link are considerably more stringent than those of the overall system. Actually the S/N ratio at each repeater in a PCM system must increase slightly as the number of regenerative repeaters (or system length) increases. This is illustrated in the lower curve of Figure 2* for the case where we want to maintain a probability of error due to random noise at 1 in 10^5 pulses. The upper curve of this figure shows the increase in S/N (peak power to mean-square noise) ratio required assuming amplification without regeneration as a function of the number of such repeaters. In deriving these curves it has been assumed that the noise power increases linearly with the numbers of links. This figure gives a quantitative picture of the advantage of regeneration. Of course, we must pay for this advantage with a large increase in bandwidth over that required for AM systems. Chapter 25, however, showed that the trade is efficient.

It should be emphasized that the results shown in Figure 2 were derived under the assumption of ideal linear repeaters as well as ideal regenerative repeaters. Practically realizable regenerative repeaters may require an increase in S/N ratio over the theoretical minimum for several reasons. First, intersymbol interference between transmitted pulses reduces the amount of noise that can be tolerated in a repeater section. This interference can be described as a

---

*It should be pointed out that this figure is consistent with Table II of Chapter 25. Recall that in Table II we assumed a unipolar pulse of amplitude V, with the sampling or slicing level set at V/2. In this case, an error will occur whenever the noise amplitude exceeds V/2 and is of proper polarity. On the other hand, in Figure 2, we show a bipolar pulse code, in which 1 is indicated by a positive pulse of amplitude V and 0 by an equal amplitude negative going pulse. In this system we in effect sample about zero volts; i.e., we ask the question, is the polarity at the time of sampling positive or negative? It follows that an error will not occur until the noise amplitude exceeds V volts and is of proper polarity. Thus, a one link system using a bipolar pulse code, as in Figure 2, requires a signal-to-noise ratio 6 db less than the unipolar pulse system of Table II, for the same error probability.



Noise Advantage of Pulse Regeneration

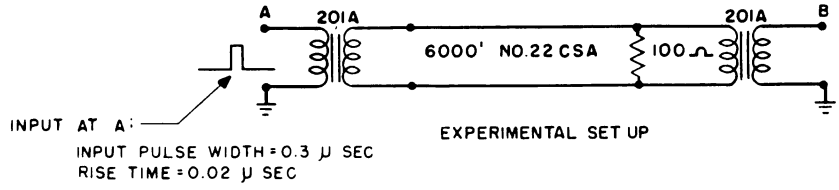
Figure 27-2

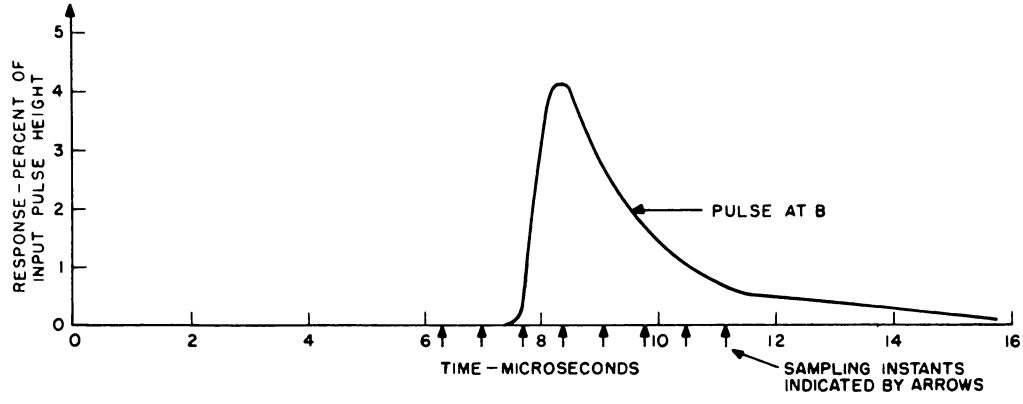
noise impairment. Secondly, it will be shown later in the text that errors in the pulse position from the ideal, hereafter called timing errors, can accumulate over a regenerative repeatered medium. Third, imperfections in repeaters as well as noise can produce variations in pulse width and height from the ideal which can also be translated directly into a noise impairment. Finally, external interference such as crosstalk from the other systems in the same environment and impulse noise (generated at central offices and picked up on cable) require a considerable portion of the margin against error.

Thus far we have indicated broadly what benefits the process of regeneration can reap. Let us return to Figure 1 to examine the operation of a regenerative repeater a little more closely. This will be a preliminary examination. A more detailed discussion will be made in the next chapter. The pictoral display on Figure 1 shows the functions performed in a regenerative repeater along with some waveforms that are

typical of those encountered in transmitting pulses over bandwidth limited systems. For purposes of discussion, the regenerated pulses at (a) are assumed to be rectangular and the 1's and 0's occur at a pulse repetition frequency of f_r. In most pulse systems the transmitted pulse occupies only a fraction of its allotted time slot in order to provide some guard space to reduce intersymbol crosstalk - i.e., to permit transients to decay. For example, in the 24 channel PCM exchange area system, the pulse repetition frequency f_r is about 1.5 mc and each transmitted pulse occupies one-half of a time slot, i.e., $\frac{1}{2f_r}$ =.33 μsec.

Leaving the (n-1) repeater, then, we have (or assume) practically perfect rectangular pulses separated by a guard space. When we transmit these through the medium, the various components of the spectrum suffer amplitude and phase distortion because of the finite bandwidth of the medium. Figure 3 gives a quantitative picture of the effects of bandwidth limitations on pulse response. It shows the response of 6,000' of 22 gauge CSA cable, terminated in 100 ohms, to a rectangular pulse of .33 μsec width. It can be seen that peak pulse





Response of Cable to 0.3 Microsecond Pulse

Figure 27-3

amplitude has been severely attenuated and the transmitted pulse dispersed over many time slots, thereby interfering with other pulses that may precede or follow this pulse. The response of the medium to a binary pulse train would be composed of the superposition of the time functions of Figure 3. Recognition and regeneration of the parent pulse train from this distorted copy without preliminary processing would not be practicable. Therefore, the first function required of a regenerative repeater is that of reshaping the signal in preparation for making a decision as to the presence or absence of a pulse. The combined transmission-frequency characteristic of the medium plus equalizer will be the main topic of this chapter. As a sneak preview to this discussion, the equalized pulse at the decision making point in the repeater might look like that shown in Figure 1b.

Assuming that positive and negative noise amplitudes are equally likely, a decision as to the existence of a pulse would be based on whether or not the received signal is greater or smaller than half the peak pulse amplitude.* This amplitude level is known as the slicing level. The input-output characteristic describing the transfer properties of an ideal regenerator is shown in Figure 1c. When the input amplitude exceeds the slicing level a new pulse of the proper shape is generated. This nonlinear function of a PCM repeater is the important regeneration property. Various types of regenerators and their deviations from the ideal regenerator are classified and discussed in the next chapter.

In addition to reshaping the signal, it is necessary to maintain the correct time relationship between signal pulses to minimize their mutual interaction. This is the function of the retimer shown in Figure 1. Timing information can be extracted from either the incoming or regenerated train by means of a tuned circuit or band pass filter. The output of the timing circuit can then be used to control the regenerator, as shown in Figure 1. It is seen that the functioning of

- -

*It should be pointed out that although half amplitude is the preferred slicing level for baseband pulses this is not the case for carrier pulses. W. R. Bennett has shown that for carrier pulses the probability that some noise of a given power will reduce the signal pulses below half amplitude is less than the probability that some noise will exceed half amplitude. This comes about from the fact that for effective cancellation there must be a 180° phase relationship between noise and pulse carrier. For this reason the slicing level should be set slightly above half amplitude for a carrier pulse system.

the regenerator is dependent on both the input information bearing signal and the clock wave derived by the retimer. Several schemes by which these two functions control the regenerator, as well as the problem of extracting the timing wave, will be examined in Chapter 28.

The important conclusion to be reached from the above discussion and Figure 1 is that the operation of a regenerative repeater can be described by the 3 R's: Reshape, Retime, and Regenerate. Alternatively, the process of regeneration can be viewed as a 3-D problem: that of maintaining the correct pulse amplitude, pulse width and spacing between pulses.

In the discussion which follows, attention will first be concentrated on the process of reshaping. Several transmission-frequency characteristics of the combined medium and equalizer of Figure 1 and their imperfections will be studied. We will indicate the effects of various methods of retiming and regeneration on equalization requirements. It should be emphasized that the development of pulse systems is a relatively new field and many of the associated transmission problems have neither been solved nor completely defined. Analytical techniques for handling the syntheses of these discrete systems are in the development stage.

Amplitude-Phase Relations of Transmission-Frequency Characteristics

We have seen that, in pulse modulation systems, pulses originate at the transmitting end in various combinations, or in varying amplitude, duration, or position, depending on the type of system. Pulses thus modulated to carry information may be transmitted directly or undergo a second modulation process suitable to the transmission medium. The received pulses will differ in shape from the transmitted pulses because of bandwidth limitations, noise, and other system imperfections. The performance of the system in the absence of noise could be predicted if the pulse transmission characteristic (that is, the shape of the received pulse for a given applied pulse) were known. The problem then becomes one of analyzing system performance in the time domain. If we were forced to confine ourselves to the time domain we would find the solution to such problems, and the prediction of system behavior, quite difficult.

A more useful approach is to determine the system performance in terms of the transmission-frequency characteristic (that is, the steady state response expressed as a function of frequency). One reason

for this is that the transmission-frequency characteristics of various existing facilities and their components are known, and for new facilities can be determined more readily (by calculation or measurement) then the pulse-transmission characteristic. However, a more fundamental reason is that the transmission-frequency characteristics of various system components connected in tandem or parallel can readily be combined to obtain the overall transmission characteristic. It is possible, therefore, to analyze complicated systems with the transmission-frequency characteristic as a basic parameter, and to specify requirements that must be imposed on the transmission-frequency characteristic of the system and its components for a given transmission performance.

It will be recalled from the work with the Fourier Transform that the transmission-frequency characteristic of a transmission system can be written as

$$Y(\omega) = G(\omega)\ \varepsilon^{-j\theta(\omega)} \tag{27-1}$$

where $G(\omega)$ is the amplitude-frequency and $\theta(\omega)$ is the phase-frequency characteristic. When a number of networks are connected in series, as is usually the case in transmission systems, the resultant transmission characteristic is

$$\begin{aligned} Y(\omega) &= Y_1(\omega)Y_2(\omega)\ldots.Y_n(\omega) \\ &= (G_1G_2\ldots G_n)\varepsilon^{-j(\theta_1+\theta_2+\ldots+\theta_n)} \end{aligned} \tag{27-2}$$

where Y_1, Y_2 ... Y_n are the transmission characteristics of the individual networks with the same impedance terminations as encountered in the series arrangement, i.e., as measured in place or with equivalent terminations. The requirements for so-called distortionless transmission through a system are that the attenuation or gain, given by the amplitude-frequency characteristic, must be the same at all frequencies, and that the phase-frequency characteristic must be linear with frequency. The latter requirement, expressed in different terms, states that the transmission delay must be the same for all frequencies. Information about relative delay of signal components is important in transmission in general, but in pulse transmission it is crucial.

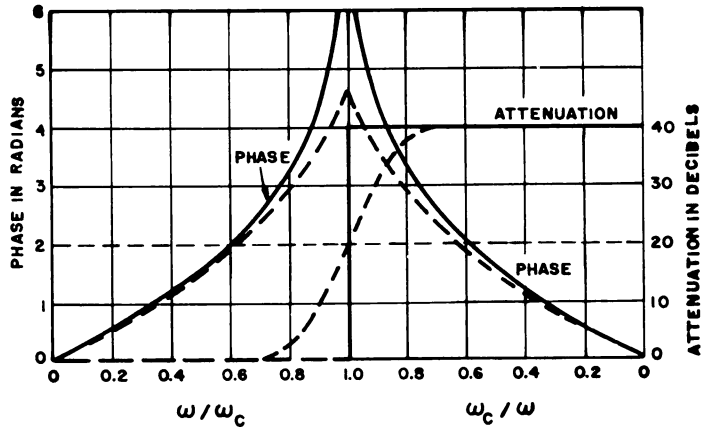
The phase characteristic $\theta(\omega)$ can, in general, be regarded as the sum of three components. The first is a minimum phase shift component which bears a definite relation to the amplitude characteristic of the system. The second is a linear component, ωt_d, which represents a constant delay t_d for all frequencies. The third component is a varying phase characteristic which may be present in a transmission system or which may be intentionally inserted for phase equalization so as to obtain an overall linear phase characteristic.

Let us examine in somewhat more detail the nature of the minimum phase component of $\theta(\omega)$. A complete mathematical treatment of attenuation (gain)-phase relationships will not be given here because the detailed results of the theory are not going to be used in the succeeding sections, and, furthermore, the theory has been treated rather exhaustively in the references given at the end of this chapter. For our purposes here it is sufficient to state that if the amplitude-frequency characteristic $G(\omega)$ is known over the entire range of frequencies, then the phase-frequency component $\theta(\omega)$ is uniquely determined. Conversely, if $\theta(\omega)$ is known over the entire range of frequencies, $G(\omega)$ is uniquely determined. There are mathematical relations which permit us to find one characteristic when the other is known. The actual transmission network we are dealing with may have more phase shift than predicted by these relations (this would involve the second and third components of $\theta(\omega)$ referred to previously), but it can never have less. Therefore, the phase characteristic uniquely determined from a particular amplitude characteristic is known as the minimum phase characteristic.

From the relations between the amplitude and phase characteristics we learn, for example, that the minimum phase associated with a constant attenuation slope of 6n db per octave, where n is any number, is just $n \frac{\pi}{2}$ radians. Furthermore, we find that the phase must change most rapidly in the vicinity of changes of slope in the amplitude characteristic, and that the minimum phase at any given frequency is influenced appreciably only by the changes in attenuation near that frequency. These principles are extremely useful in making rough estimates of the phase associated with a particular amplitude characteristic.

Let us now look at some minimum-phase characteristics which are associated with particular transmission line attenuation characteristics. As an example, the solid curve in Figure 4 shows the attenuation

shape for low-pass filter having 40 db attenuation above the cutoff frequency ω_c. It is seen that the minimum-phase characteristic associated with such filter approximates a linear variation only at very low and very high frequencies and is infinite at the cutoff frequency. On the other hand, a gradual cutoff in the attenuation characteristic modifies the phase characteristic, as shown by the dashed curves in Figure 4. Since the attenuation and minimum-phase characteristics are uniquely related we should expect to be able to produce a linear phase characteristic in the transmission band by suitably shaping the attenuation characteristic. This case will now be considered further because transmission systems with a linear phase characteristic in the pass band are so important in pulse transmission (just as in television transmission).

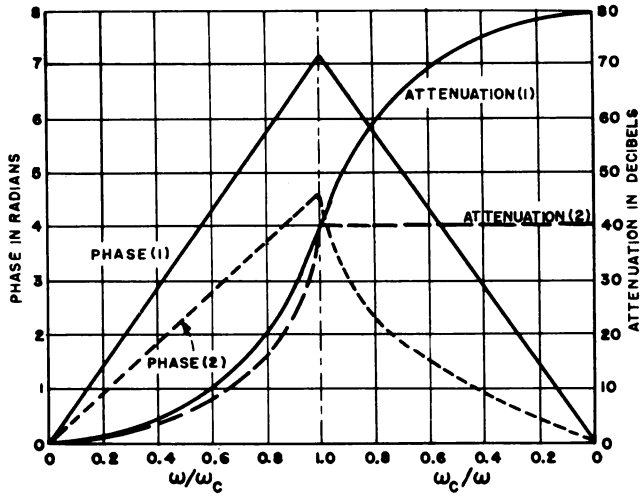


Minimum Phase Low-Pass Filters

Figure 27-4

In Figure 5 the linear minimum-phase characteristic given by the solid line leads to the attenuation characteristic shown. Other attenuation characteristics with linear phase between $\omega/\omega_c = 0$ and 1 are possible, of course, if other variations in the attenuation or phase characteristic for $\omega/\omega_c > 1$ are assumed. For example, the combination of linear phase for $\omega/\omega_c < 1$ and constant attenuation for $\omega/\omega_c > 1$ results in the overall phase-attenuation characteristic shown by the dashed curves of Figure 5. It will be noticed that there is comparatively minor difference between the attenuation characteristics in the pass-band ($\omega/\omega_c < 1$) for the two cases illustrated, so that the reshaping of the attenuation characteristic above ω_c in going from curve

(1) to curve (2) has had a relatively minor pass-band effect. The transmission-loss characteristic shown by the solid curve of Figure 5 represents a close approximation to the type of characteristic employed in some pulse transmission systems and will be discussed more fully in a later section.



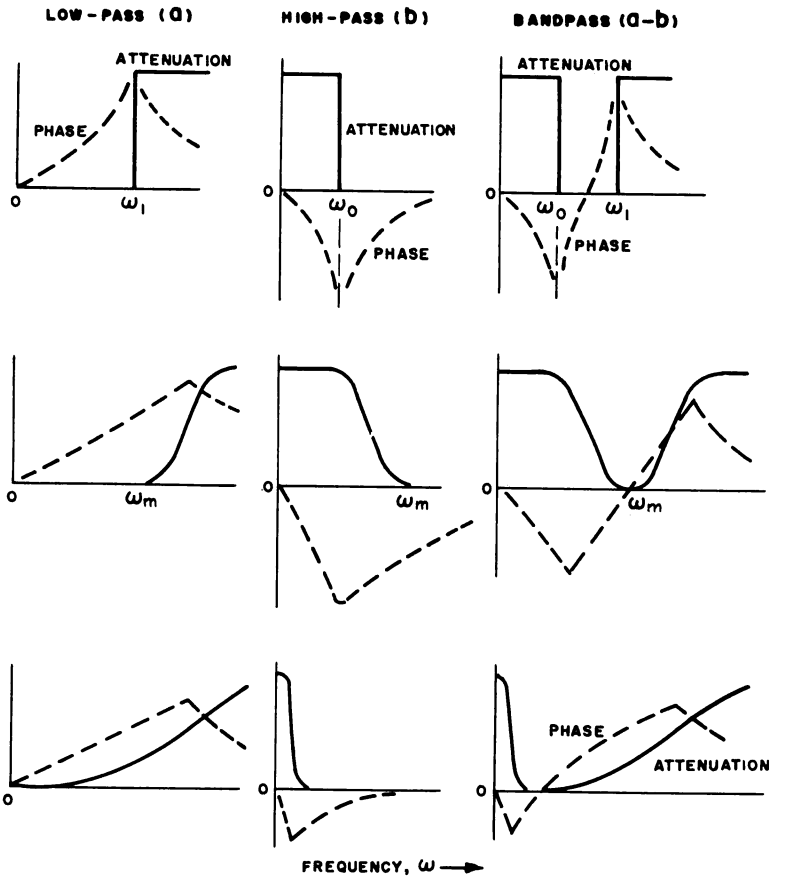
Low-Pass Filter Characteristics for Linear In-Band Phase Shape

Figure 27-5

Band-pass characteristics can be thought of as having been obtained by connecting low-pass and high-pass networks in tandem. The resultant attenuation and phase characteristics are obtained by adding the low and high pass attenuation and phase characteristics, as illustrated in Figure 6. In the second case shown, the bandpass characteristic is assumed to have a linear phase characteristic in the transmission band, in which case the attenuation characteristic will not be symmetrical about the midband frequency unless the latter is high in relation to the bandwidth. The third case illustrates the type of bandpass characteristic encountered in wire systems with a low low-frequency cutoff. There will then be phase distortion at the low end of the band, since it is not feasible with a fairly sharp low frequency cutoff to obtain a linear phase characteristic in the transmission band.

In concluding this general discussion, it should be pointed out that if the amplitude characteristic of a transmission system is modified, there will also be a modification of the phase characteristic. It can be shown, for example, that any cosine-type ripple

introduced in the amplitude characteristic produces a corresponding sine-type ripple in the phase characteristic. Since, in general, any modification in the amplitude characteristic may be represented by a Fourier cosine series, the modification in the phase characteristic will be the corresponding Fourier sine series. These facts will be useful in considering the effect of deviations in the gain and phase from the ideal.



Low- and High-Pass Components of Various Band-Pass Transmission Frequency Characteristics

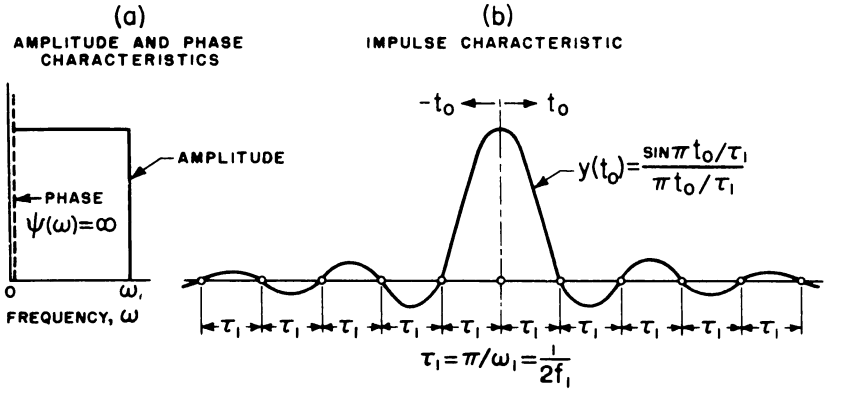
Figure 27-6

With these fundamental principles in mind, let us now consider the impulse response of several transmission-frequency characteristics in order to determine the most desirable shape for the combined medium and equalizer of Figure 1.

Impulse Response of an Ideal Sharp Cut-off Characteristic

In many papers on pulse transmission theory, particularly those dealing with transmission capacity of idealized transmission

systems, an ideal low-pass transmission-frequency characteristic is assumed, with constant amplitude and delay in the transmission band and an abrupt transition to infinite attenuation at the top frequency. Such a characteristic is shown in Figure 7. The minimum phase associated with a sharp cut-off characteristic having finite attenuation outside the transmission band has already been discussed in connection with Figure 4. If we now go to the ideal low-pass characteristic having zero transmission above the pass-band, as in Figure 7, the minimum phase component goes to infinity and the characteristic can no longer be physically realized since it would require infinite transmission delay. It can, however, be approximated by providing sufficiently elaborate phase equalization and allowing a large but finite (rather than infinite) out-of-band attenuation.



Impulse Response of Ideal Low-Pass Filter

Figure 27-7

Since our object will be to show that sharp cut-off characteristics are undesirable in pulse transmission systems, let us assume, for analysis purposes, the characteristic sketched in Figure 7. Let $G(\omega) = 1$ between $\omega = 0$ and ω_1, and let $G(\omega)$ equal zero everywhere else. Furthermore, let us assume that $\theta(\omega) = \omega t_d$, although we now appreciate that this is not compatible with our first assumption. It was shown in the chapter on the Fourier Transform that the response of this characteristic to an impulse is given by

$$y(t) = \frac{\delta\omega_1}{\pi} \frac{\sin \omega_1 t_o}{\omega_1 t_o} \qquad (27\text{-}3)$$

where δ is the area of the impulse and $t_o = t - t_d$.

The resultant pulse transmission characteristic is shown in Figure 7, with the factor $\frac{\delta\omega_1}{\pi}$ omitted. The peak amplitude is attained only after an infinite time, since the ideal low-pass characteristic is approximated only if $t_d \to \infty$. The impulse characteristic is zero when $\omega_1 t_o = \pm n\pi$, or $t_o = \pm t_1, \pm 2t_1, \ldots \pm nt_1$, where

$$t_1 = \frac{\pi}{\omega_1} = \frac{1}{2f_1} \tag{27-4}$$

Impulses can be transmitted at intervals of t_1 without mutual interference between the peaks of the received pulses. As noted previously this is basic to the establishment of the maximum transmission capacity of ideal systems.

In addition to the elaborate phase equalization and long delay required to approximate this characteristic, several other practical disadvantages arise. They are

1. The gain equalization required will be difficult to realize, and, furthermore, the gain stability of such a characteristic is difficult to maintain. It is almost axiomatic that the band edge shape of a transmission system will "swing in the breeze" since both the active and passive elements of the system vary most widely at the extremeties of the band. Any variation in ω_1 will displace the zeroes of the impulse response from the desired spacing and result in intersymbol interference.
2. Variations in synchronization (i.e., timing errors) will also introduce rather severe intersymbol interference with this type of characteristic due to the relatively sharp slope of the characteristic in the neighborhood of the zeroes. This will be examined in greater detail in a succeeding section.

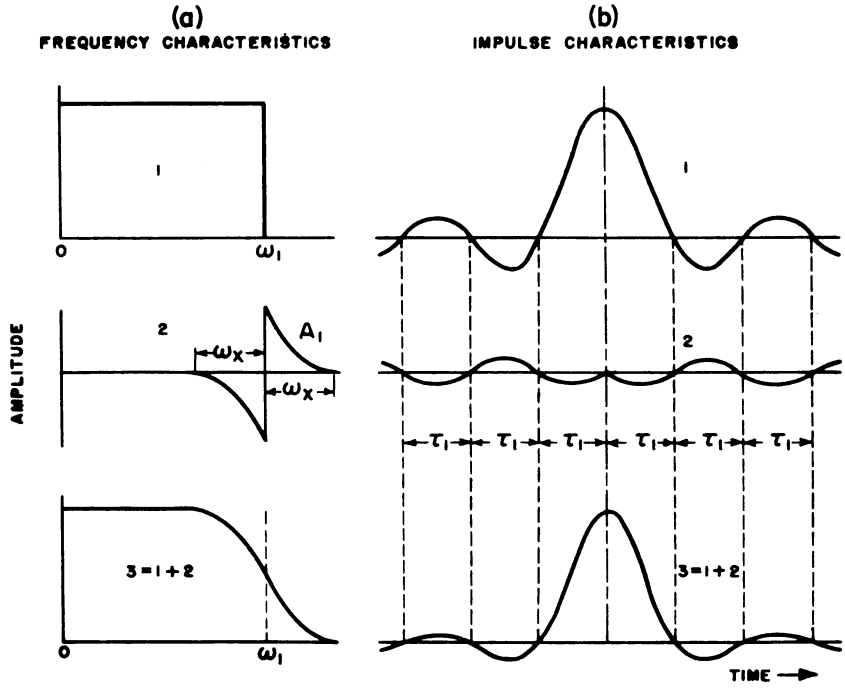
Impulse Response of a Gradual Roll-off Characteristic

The review of the amplitude-phase relationships for minimum phase type systems revealed that the infinite discontinuity in the phase characteristics associated with sharp cut-off amplitude characteristics can be eliminated by providing gradual cut-off of the amplitude characteristic. Furthermore, as will be discussed here, providing a gradual

rather than a sharp cut-off, as shown in Figure 8, not only reduces the non-linearity of the phase characteristic but also reduces the amplitude of the build-up and decay oscillations of the impulse response.

In the work with the Fourier Transform, the impulse response for a gradual roll-off transmission characteristic was derived by assuming an ideal characteristic with sharp cut-off supplemented by an amplitude characteristic which has odd symmetry about the cut-off frequency ω_1. This is illustrated by Figure 8. In addition, the phase was assumed to be linear and given by $\theta(\omega) = \omega t_d$. The resulting impulse response is given by

$$y(t) = \frac{\delta\omega_1}{\pi} \frac{\sin \omega_1 t_o}{\omega_1 t_o} \frac{\cos \omega_x t_o}{1-(2\omega_x t_o/\pi)^2} \qquad (27\text{-}5)$$



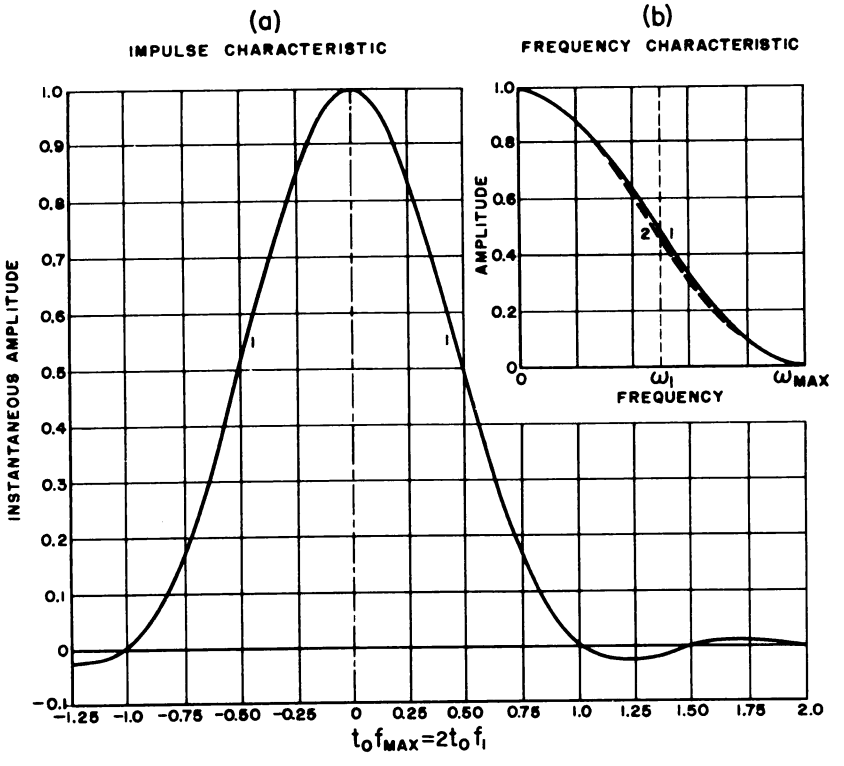
Impulse Response of Gradual Cut-Off Characteristic, Obtained by Superposition of: 1, Ideal Low-Pass Filter, and 2, Band-Pass Filter with Odd Symmetry About ω_1. Linear Phase Shift Assumed.

Figure 27-8

For the particular case $\omega_x = \omega_1/2$ the zeroes of the impulse response are identical with those obtained with the ideal sharp cut-off characteristic. For other values of ω_x, the original zeroes will be preserved, but additional zeroes in the pulse response will be introduced.

As long as ω_x is of the order of one-half ω_1 or less, it is necessary to employ phase equalization to obtain a linear phase characteristic. Furthermore, oscillations of appreciable amplitude remain in the impulse characteristic. A virtually linear phase characteristic together with a reduction of these oscillations can be attained by a further extension of the gradual roll-off, such that $\omega_x = \omega_1$. An amplitude characteristic of this type, together with the corresponding impulse response, is shown in Figure 9. The amplitude characteristic is given by

$$G(\omega) = \frac{1}{2}\left[1 + \cos\frac{\pi\omega}{2\omega_1}\right] = \cos^2\frac{\pi\omega}{4\omega_1} \qquad (27\text{-}6)$$



Impulse Response of Gradual Cut-Off Characteristic 1. Characteristic 2 is Same as Shown by Solid Lines in Figure 5.

Figure 27-9

between $\omega = 0$ and $\omega = 2\omega_1$. The impulse response is found by letting $\omega_x = \omega_1$ in Equation 5, which gives

$$y(t) = \frac{\delta\omega_1}{\pi} \frac{\sin 2\omega_1 t_o}{2\omega_1 t_o[1 - (2\omega_1 t_o/\pi)^2]} \tag{27-7}$$

Here ω_1 is now the bandwidth to the half-amplitude point on the transmission frequency characteristic and $2\omega_1$ is the bandwidth to the point of zero transmission.

Figure 9b also shows (dashed line) an amplitude characteristic which is known to have a linear phase characteristic in the transmission band, i.e., from $\omega = 0$ to $\omega = 2\omega_1$. Because of the close approximation of the solid line to this amplitude characteristic, which is ideal as regards phase linearity, the phase characteristic associated with the solid line (i.e., Equation 6) may for practical purposes be regarded as linear.

Impulse Response of a Gaussian Characteristic

Another type of amplitude characteristic resembling that shown in Figure 9 and frequently considered in connection with pulse transmission is a Gaussian characteristic given by

$$G(\omega) = \varepsilon^{-\Delta\omega^2} \tag{27-8}$$

where Δ is a constant. The corresponding impulse response is also Gaussian and can be shown to be

$$y(t) = \frac{\delta}{2(\pi\Delta)^{1/2}} \varepsilon^{-t_o^2/4\Delta} \tag{27-9}$$

Again it has been assumed that $\theta(\omega)$ is equal to ωt_d. If we require that the amplitude of the impulse response be reduced to 1 per cent of the peak value after an interval $t_o = \pi/\omega_1$, corresponding to the first zero point of an ideal impulse characteristic, it is necessary that $4\Delta/t_o^2 = 4.6$, or $\Delta = 0.54/\omega_1^2$. The corresponding amplitude and impulse characteristics are

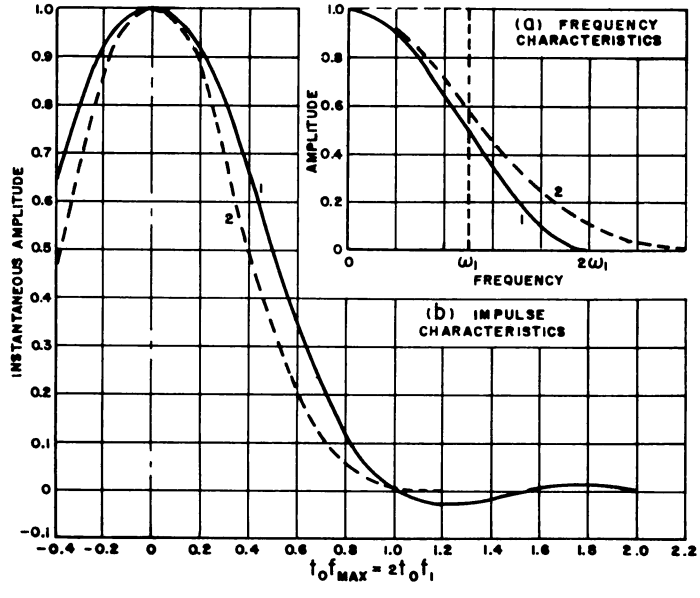
$$G(\omega) = \varepsilon^{-0.54(\frac{\omega}{\omega_1})^2} \tag{27-10}$$

and

$$y(t) = \frac{\delta\omega_1}{0.83\pi}\,\varepsilon^{-0.46(\omega_1 t_o)^2} \qquad (27\text{-}11)$$

In Figure 10, a comparison is made of the two amplitude characteristics (6) and (10) considered above, and the corresponding impulse responses (7) and (11). The comparison shows that for the same pulse transmission rate, the impulse oscillation, and, therefore, the intersymbol interference, is negligible with the Gaussian characteristic, but that a somewhat wider band must be provided for this shape. This is a disadvantage, particularly when the band is restricted within prescribed limits by considerations of interference in adjacent transmission bands.

A further note is in order at this point. If the information for retiming is extracted from the incoming pulse train, it will be necessary to provide an amplitude characteristic which is not too far down at $2\omega_1$ (the pulse repetition frequency) in order to obtain sufficient clock power. This consideration will make a characteristic with increased bandwidth, such as the Gaussian characteristic, necessary. Further elaboration on this point will be found in the chapter on timing.



Impulse Response of Two Gradual Cut-Off Characteristics

Characteristic 1: $G(\omega) = \frac{1}{2}\left(1+\cos\frac{\pi\omega}{2\omega_1}\right)$

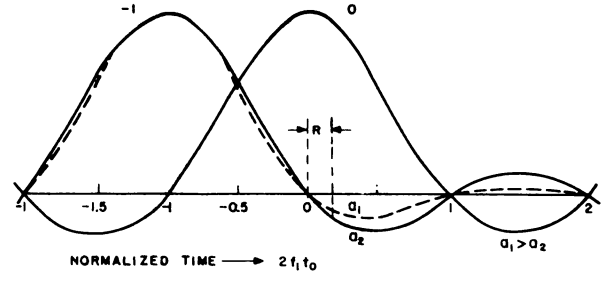
Characteristic 2: $G(\omega) = \varepsilon^{-0.54\left(\frac{\omega}{\omega_1}\right)^2}$

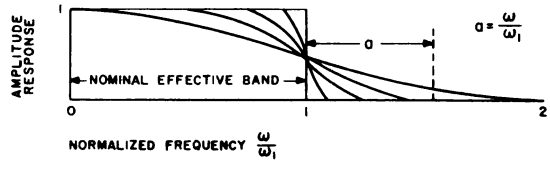
Figure 27-10

In the next section some quantitative results are given which document the advantages of the roll-off type of characteristic when timing errors are made in sampling the equalized pulse train.

Influence of Roll-off on the Effect of Timing Errors

A convenient measure of the intersymbol interference between pulses is the reduction which it requires in external noise to maintain a specified frequency of error. As noted previously, this intersymbol interference can arise from deviations in the gain and phase characteristics from the desired transmission characteristic, or can arise from errors in the time at which the received pulse is sampled. In this section we will consider the latter effect to determine the efficacy of the roll-off characteristic in reducing the noise impairment associated with timing errors.





Timing Error Parameters: Frequency Characteristics and Time Responses

Figure 27-11

Figure 11 defines both the time response and the associated amplitude response under consideration. As in previous discussions, linear phase is assumed. The parameter "a" is a measure of the amount of roll-off that is used. For example, a=0 yields the ideal sharp cut-off low pass characteristic, while a = 1 corresponds to complete roll-off

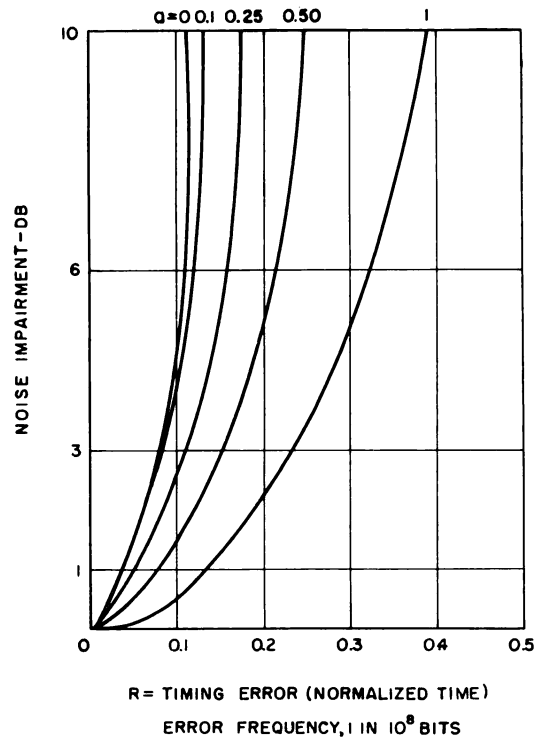
as shown in Figure 9. R in Figure 11 indicates the timing error in sampling the pulse that normally achieves its maximum value at $2f_1 t_o = 0$, where it will be recalled that $t_o = t - t_d$.

There are essentially two sources of noise penalty due to timing errors. The first arises from the reduction in pulse amplitude due to sampling away from the peak. This means that a smaller negative noise pulse may cause misrecognition of the information-bearing pulse. Secondly, pulses remote from the sampled pulse have non-zero values at the mistimed sampling point. The latter is intersymbol interference and can be either positive or negative depending upon the particular pulse pattern transmitted. The procedure to be followed in evaluating the effects of timing errors is summarized below and the results of the analysis are presented in Figures 12 through 14.

Since the original pulse pattern is statistical in nature, it is necessary to determine a distribution of the amplitudes of the intersymbol interference. This can be accomplished by considering all possible combinations of 1's and 0's (or marks and spaces of equal duration) and determing the disturbance caused by each combination at a point displaced by R units in time from the point of peak pulse amplitude. The procedure employed is to consider each distribution of amplitudes as a separate normal distribution, compute the noise impairment, and average the impairment contributed by each distribution. It is assumed that the external noise is random, with a Gaussian amplitude distribution. A detailed analysis of the steps involved in arriving at the results is beyond the scope of this text.

With the basis of the study thus set forth, the results may now be presented and summarized. They can be grouped into two sets. In the first set the noise impairment can be plotted as a function of the timing error of the sampler, with the roll-off ratio "a" as a parameter. A separate plot is required for each error frequency considered. Figure 12 shows the results for an error rate of 1 in 10^8 bits.

The meaning of Figure 12 may be made clearer if we discuss one point on one of the curves in terms of a specific numerical example. Suppose we have a transmission characteristic with a roll-off such that $a = 0.5$. We might ask, for example, what is the effect of a timing error of about 0.2 of a time slot? Figure 12 tells us that it results in a



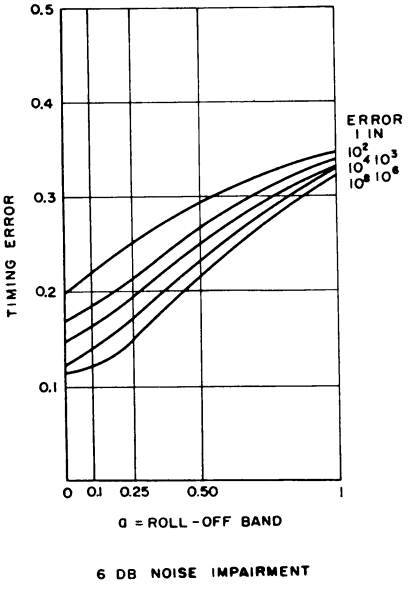
Noise Impairment vs. Timing Error for Various Roll-Offs

Figure 27-12

noise impairment of 5 db if our objective is to have only one error per 100,000,000 pulses. Table II of Chapter 25 tells us that, ideally, a peak signal to average noise power of about 21.0 db will result in an error rate of 10^{-8}. Therefore, the effect of the 5 db noise impairment caused by the 0.2 timing error is to require that the signal-to-noise ratio be increased to 26.0 db to maintain the same probability of error.

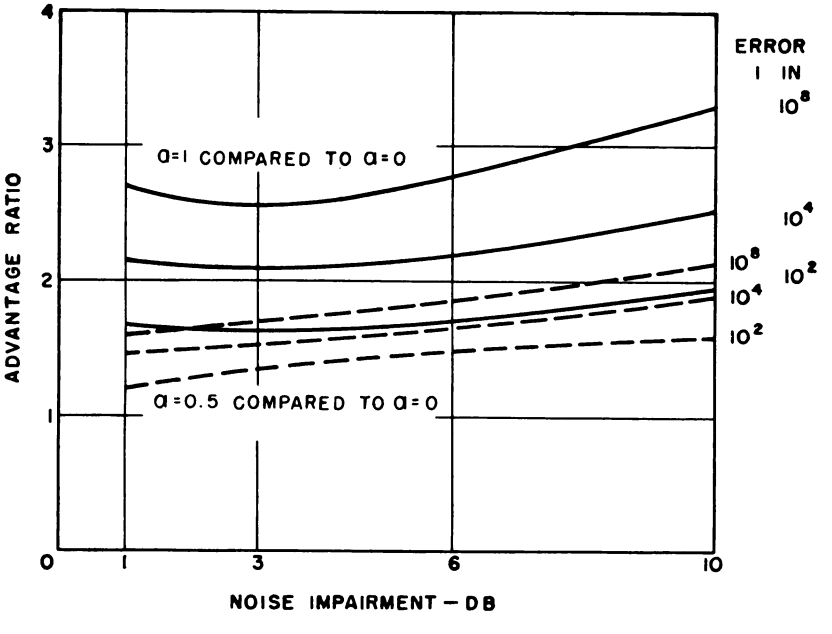
The second presentation of the results shows the timing error tolerance as a function of relative roll-off bandwidth, with error tolerance as a parameter. One impairment (6 db) is considered here, as shown in Figure 13.

A single set of summarizing curves is shown in Figure 14. The solid lines in this plot show the advantage obtained in timing error by extending the roll-off band to be equal to the normal effective band (a = 1). The "advantage" is defined as the ratio of the tolerance of timing error with the use of the roll-off band to that for the case of no roll-off. The advantage is plotted as a function of the db noise impairment allotted to intersymbol interference, and curves are



Timing Error vs. Roll-Off for 6 db Noise Impairment

Figure 27-13



Advantage Ratio vs. Noise Impairment for Various Roll-Offs

Figure 27-14

presented for each of three error tolerances. As an example, if our objective is an error rate of 10^{-8} and we are willing to suffer a 3 db noise impairment, we might ask: By what factor can we increase the timing error if we go from no roll-off to a roll-off of a = 1.0? Figure 14 tells us that the answer is "by about 2.6 times" (advantage ratio = 2.6).

The dotted curves in Figure 14 represent the advantage obtained by extending the roll-off band half as far (a = 1/2).

The conclusions may be numerically summarized briefly as follows:

1. The advantage obtained from the use of the maximum roll-off band considered, as compared with no roll-off band, runs from the ratio of 1.65 to 1, to 3.3 to 1.
2. The advantage is greater for the lower than for the higher error probabilities; i.e., the advantage runs from 2.55 to 1, to 3.3 to 1, for an error frequency of 1 in 10^8 bits, as compared to 1.65 to 1 up to 1.95 to 1 for an error frequency of 1 in 10^2 bits.
3. The advantage is generally greater for the larger than for the smaller noise impairments allocated to intersymbol interference. Thus at an error frequency of 1 in 10^4 bits the advantage is 2.5 to 1 at 10 db noise impairment. This drops to 2.1 to 1 at 3 db impairment (but rises slightly again to 2.16 to 1 at 1 db impairment).
4. The advantage that can be obtained with a half roll-off band runs from .6 to .85 of that obtained with the maximum roll-off band. The low figure generally comes at the low error frequencies (1 in 10^8) and the higher figure at the higher error frequencies.

Low Frequency Suppression - Nature of the Problem

Low frequency cutoff in the transmission frequency characteristic of wire systems results from transformer coupling of terminals and repeaters to the cable medium. Transformer coupling may be dictated in order to:

1. Secure good impedance terminations.
2. Isolate the operating voltages in a repeater.
3. Couple a balanced line to an unbalanced repeater.
4. Provide a phantom path for transmission of dc over the line to power the repeaters.

Another advantage associated with attenuating the low frequency portion of the transmission band arises when we transmit pulses over wires in the same cable bundle with other pairs used exclusively for low frequency channels. In this case the mutual interference problem is practically eliminated.

The price of this inability to transmit low frequency components is an increase in intersymbol crosstalk. With dc suppression, each received pulse must have equal areas above and below the zero line. This results in a transient tail which can extend over many time slots and thereby interfere with subsequent pulses. In fact, poor low frequency response threatens the eventual collapse of the output wave from a long string of like pulses as shown in Figure 15d. In this figure it is assumed that the equalized pulse without low frequency suppression is triangular and the pulses are transmitted at the rate of $\frac{1}{T}$ pulses per second without intersymbol interference. Figure 15a shows the undistorted pulse, while Figure 15b depicts the envelope of a pulse train consisting of eight consecutive pulses. The sampling instants are indicated by arrows.

With no low frequency suppression, the combination of sampling in time and slicing in amplitude permits us to recognize the existence of the eight pulses in the trapezoidal envelope and, therefore, regenerate them. However, with a poor low frequency response, this will not be the case. If the low frequency cut-off is approximated by a single R-L or R-C high-pass section, as shown in the figure, then the response of either of these circuits to the triangular pulse and triangular pulse train is as shown in parts (c) and (d) of Figure 15. The RC time constant has been chosen such that C loses half its charge in time 2T. This corresponds to RC=2.885T and to 3 db loss in the high-pass network at the frequency $\frac{0.0552}{T}$. At time T the response to an isolated triangular pulse of amplitude E reaches

$$E_T = \frac{CR}{T}\left(1-\varepsilon^{-T/RC}\right) E \quad \frac{CR}{T} \qquad (27\text{-}12)$$

For the time constant assumed, E_T will be equal to 84.5% of the undistorted pulse height E. If the slicing level is set at half the peak pulse amplitude of the undistorted pulse, a negative-going noise peak of 34.5% of the undistorted pulse peak, occurring at the sampling

instant, will result in an erroneous decision; i.e., the pulse will be considered as a space. The reduction in pulse amplitude represents a 3 db loss of margin to noise.

At time 2T, the next recognition instant, the pulse has reversed sign and is given by

$$E_{2T} = -\frac{CR}{T}\left(1-\varepsilon^{-T/RC}\right)^2 E \qquad (27\text{-}13)$$

In this case, $E_{2T} = -25\%$ of the undistorted peak. Therefore, when two pulses occur in succession, the composite output wave drops at time 2T to about 60% of the impressed peak. Negative noise peaks of only 10% of the peak of the impressed wave will result in an erroneous decision in the repeater. As more pulses are joined in succession, the output sags steadily towards zero. If the slicing level had been fixed at 50% of the undistorted peak, only the first two pulses would have been recognized in the absence of other interference. The remaining six pulses would be absent from the regenerated pulse train. After the eighth successive pulse, the response is down to 7.5% of the value without low frequency suppression. This occurs because the capacitor has become charged to nearly the full peak voltage by the first pulse, so that only a small charging current flows through the load to produce an output voltage on succeeding pulses. When the string of eight pulses is terminated, the output voltage goes negative to a value of -79% of the input peak and the output voltage begins to decay towards zero. Setting a fixed slicing level to distinguish pulses and spaces in the face of this distortion is extremely difficult. It is true that with a random pulse train, where pulses and spaces are equally likely, the probability of eight successive pulses is $(\frac{1}{2})^8$, so that complete collapse of the pulse train is rather unlikely. This, of course, depends on what the system objectives are in terms of the desired probability of error. With a random pulse train, the dc level of the received pulses will, nonetheless, vary sufficiently to justify a search for corrective action. The following paragraphs examine several means for combating the effects of low frequency cut-off.

Methods of Combating the Effects of Low Frequency Suppression

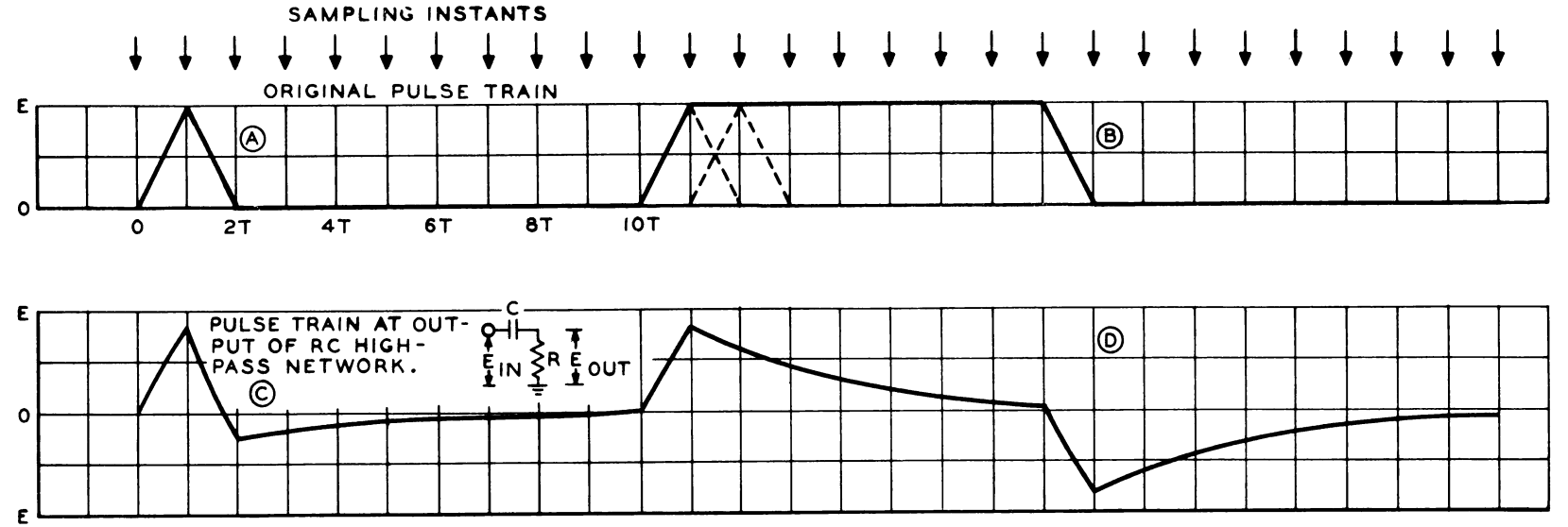
Various methods for correcting the effects of low frequency cut-off have been proposed. These include: (1) quantized feedback,

(2) linear equalization, (3) dc restoration, (4) clamping techniques, and (5) ac pulse transmission. On a purely theoretical basis it appears that quantized feedback has decisive advantages over all the other methods. Let us, therefore, consider this method first.

If the relative amount of bandwidth trimmed off by the low frequency cut-off is small, the first pulse in a train is not severely affected and can be recognized at the prescribed instant in time. In the method of quantized feedback, as soon as recognition occurs a new outgoing pulse is generated and a small portion of the new pulse is fed back to cancel the remaining transient from the first input pulse. Assuming that the cancellation is complete at the instant at which a second recognition must be made, the next pulse can be treated as if it were the first pulse of the train. In theory, any number of consecutive pulses could then be recognized as well as the first one.

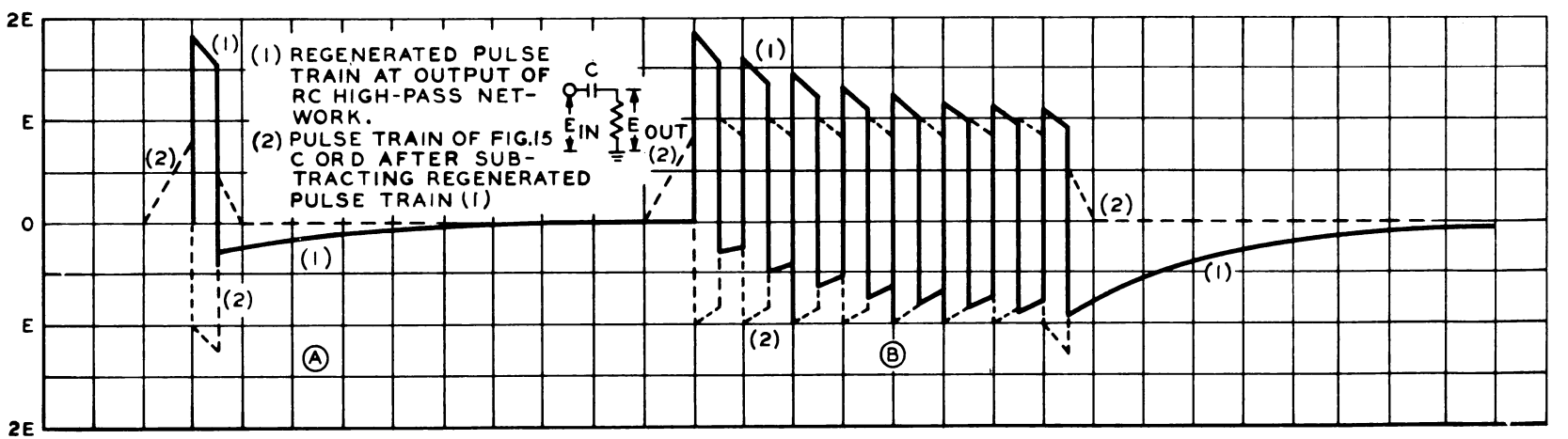
The simplest case to examine is that of an exponential form of transient decay. Perfect cancellation can be attained in this case, since two exponentials with the same damping constant can always be matched by multiplying one of them by a constant. For example, consider an input pulse which terminates at time t_o. Let its transient decay be of the form $E_o\varepsilon^{-a(t-t_o)}$. Then, suppose that we generate locally a pulse which terminates at time t_1, with a transient $E_1\varepsilon^{-a(t-t_1)}$. We can make the two transients cancel after time t_1, if we multiply the locally generated pulse by a constant, $k = \frac{E_o}{E_1}\varepsilon^{-a(t_o-t_1)}$. This cancellation can be accomplished by passing the regenerated pulse through a low-pass network and then adding it to the incoming pulse train. Alternately, we can pass the regenerated pulse through a network identical with that which distorted the received pulse and subtract it from the incoming pulse.

In order to illustrate these points, let us return to the previous case in which we assumed a triangular pulse in the absence of low-frequency suppression and show how quantized feedback eliminates the intersymbol interference. This is shown in Figure 16a and b. We have assumed that the regenerated pulse is rectangular, of amplitude V, and duration T/2, and that it is applied to a high-pass R-C network. The solid line in Figure 16 shows the response of such a network to the regenerated rectangular pulse. Let us determine the magnitude of V required for perfect cancellation of the transient. Beginning at time 2T, the response of the R-C network to a single isolated triangular pulse is given by



Effects of Low Frequency Suppression

Figure 27-15



Correction of Low Frequency Suppression by Quantized Feedback

Figure 27-16

$$E_p(t) = E_{2T}\, \varepsilon^{-(t - 2T)/RC} \qquad (27\text{-}14)$$

The response of the same network to the regenerated pulse is, after time $t = 3T/2$,

$$E_s(t) = -V(1-\varepsilon^{-T/2RC})\, \varepsilon^{-(t - 3T/2)/RC} \qquad (27\text{-}15)$$

For perfect correction

$$E_p(t) - E_s(f) = 0$$

Therefore, equating Equations 14 and 15, and substituting Equation 13 for E_{2T}, we obtain

$$V = \frac{CRE}{T}(1-\varepsilon^{-T/RC})(1+\varepsilon^{-T/2RC})\, \varepsilon^{T/2RC} \qquad (27\text{-}16)$$

By the principle of superposition, the correction may be extended to any sequence of pulses and spaces. Figure 16b shows how the correction operates for eight consecutive pulses. The oscillations appear to be violent, but the timing wave allows triggering only as we approach the right-hand extremities of the solid lines above the axis.

Important factors worthy of stress are:

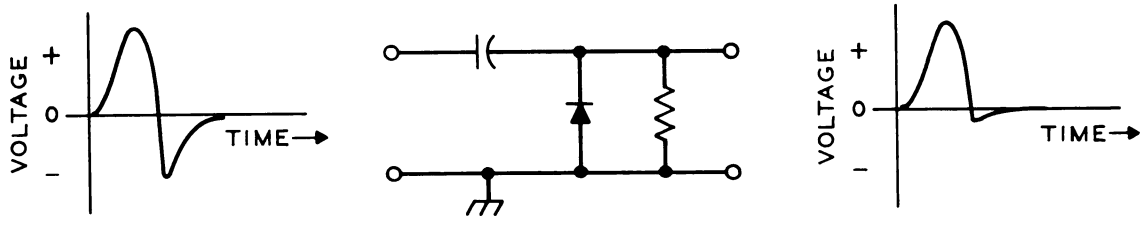
1. If a cancelling wave can be locally generated in a system which becomes independent of the signal wave after triggering, noise accompanying the signal does not get through to the cancelling wave. This means that intersymbol interference is corrected without introducing noise from the imperfectly transmitted part of the signal band: that is, we do not match the level of weak, low frequency signal components contaminated by noise of comparable magnitude, but rather supply new and clean pulses for transient cancellation.

2. By correcting with a transient from the high-pass circuit instead of from the low-pass type we avoid the necessity for dc coupling of the feedback loop.

Now let us discuss some of the alternative methods for low frequency compensation, comparing them to quantized feedback on the as-

sumption that quantized feedback is perfect. We will later qualify these comparisons by taking note of the difficulties of instrumenting a quantized feedback system.

The application of a linear equalizer is subject, of course, to the limitation that if no dc is transmitted none can be received. It would, however, be possible in practical cases to fix an upper bound to the number of successive pulses of like sign which need be recognized accurately. Transmission could then be equalized down to a sufficiently low frequency to adequately preserve the response to this particular train of pulses. The necessary equalization becomes more and more drastic as we require longer pulse sequences, and hence approach closer to zero frequency (dc) transmission. If we equalize by increasing the low frequency gain at the receiver, we increase the noise contributed by the low frequency band, and if we equalize at the transmitter, we increase the power required on the line. These penalties approach infinity as we approach more and more closely the conditions required for a large number of successive pulses with no dc transmission. The penalty imposed on the quantized feedback case is only a slight reduction in pulse peak because of the relatively small narrowing of the band associated with the low frequency cut-off. It thus appears that quantized feedback has an advantage over linear equalization.



DC Restorer

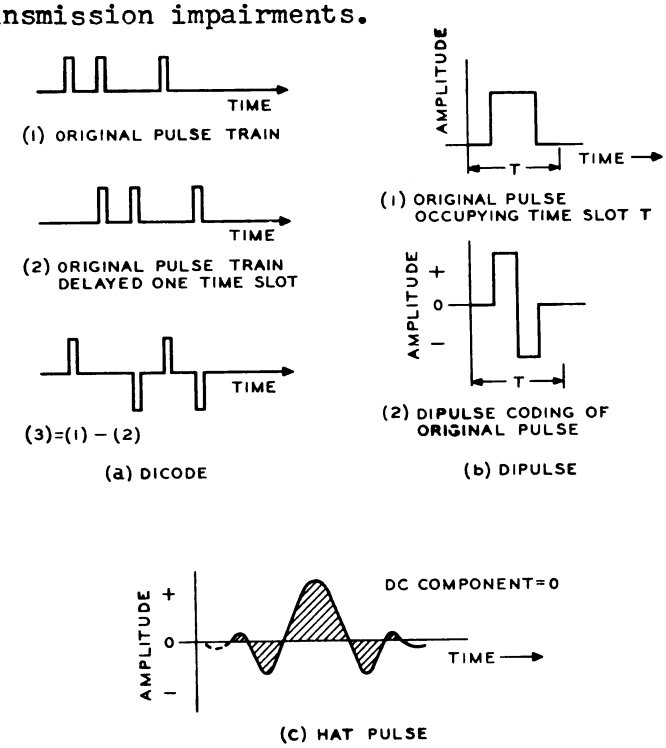
Figure 27-17

In a dc restorer a rectifying element is shunted between the load side of a capacitor and a fixed reference voltage, as illustrated in Figure 17. The voltage at this point is thus forced to stay on one side of the reference voltage. One limit of excursion of the received pulse train is therefore fixed. In the case of the pulse train of Figure 15 the reference voltage might be set at zero. This would prevent the negative excursions, but it would not prevent the positive peak response from sagging when a succession of pulses occurs. By itself, it is clearly inferior to the quantized feedback method.

By clamping we mean closing a switch between the load side of a series capacitor and a fixed reference potential at regularly spaced instants of time. In the usual clamping operation, a pulse of fixed height is sent at regular intervals and the clamping voltage is made equal to the voltage of this pulse. Aside from the fact that the reference is thereby affected by the noise, it is clear that in a drastic cut-off condition, such as Figure 15 illustrates, we would lose most of the benefit of the reference value within one or two pulse intervals. The fixed pulses injected for clamping purposes would, therefore, have to be so numerous as to seriously reduce the rate at which information bearing pulses could be sent. The most effective way to use clamping would be to clamp in every pulse interval without sacrificing any information pulses. This is not only difficult to accomplish accurately, but brings in an inherent noise penalty because the voltages clamped must include noise samples.

The last method mentioned for working through a poor low-frequency response characteristic makes use of a pulse having no important low frequency components. Examples of such pulses include dicode, dipulse, and the "hat" pulse, all of which are illustrated in Figure 18, as well as various kinds of carrier pulse transmission. In dicode, Figure 18a, the binary pulse train is delayed one pulse interval and then subtracted from the undelayed train. The resultant pulse train has no dc component but assumes three possible magnitudes, instead of two, at the recognition times. It suffers a six db loss of margin over noise and requires a more complicated pulse repeater. Its advantage is that it is not sensitive to the exact shape of the low frequency transmission characteristic. Dipulse transmission, Figure 18b, employs a combination of matching plus and minus segments for the typical pulse. It preserves two-level signaling since the negative segments can be ignored

at the receiver, but it requires a widening of the band at the high frequency end to transmit the two segments in one pulse interval. The hat pulse, Figure 18c, is an ac pulse of approximately double bandwidth, which can be interleaved in time with another like pulse in such a way that a zero of one coincides with a peak of the other. The bandwidth penalty is removed at the expense of increased precision requirements. Carrier pulse transmission schemes include: (1) straight AM, which is simple and practical but requires twice the baseband width; (2) quadrature AM, which exchanges the bandwidth penalty for a severe requirement on phase stability by sending independent pulses on carriers ninety degrees apart; and (3) vestigial sideband pulse transmission, in which the transmitted bandwidth can be made to approach the baseband width at the expense of increased complication of apparatus and greater susceptability of the signal to transmission impairments.



Dicode, Dipulse, and "Hat" Pulse

Figure 27-18

As pointed out earlier, on a purely theoretical basis it would seem that quantized feedback has decisive advantages over all other proposed methods for reducing the effects of low frequency suppression. In practice this advantage may dissolve because of difficulties of instrumentation. Though not specifically stated, a wide local bandwidth

was implied in our previous discussion of quantized feedback. Parasitic elements associated with the feedback loop will distort the shape of the feedback signal so that exact cancellation of the transient tail will not take place. This problem becomes increasingly severe as the pulse repetition frequency increases. In addition, exact timing was assumed for both the signal sampling and the feeding back of the regenerated signal for elimination of the tail of the incoming signal. This too is a mathematical fiction, for the presence of timing errors will prevent exact cancellation of the unwanted transient. Finally, if timing considerations dictate extracting information from the incoming pulse train, quantized feedback cannot be used conveniently. This point will be expanded and clarified in the next chapter.

A combination of low frequency equalization, limitation of the length of a consecutive string of pulses,* and a dc restorer appears to solve the low frequency suppression problem, at least for pulse repitition frequencies less than about 2 mc. At higher frequencies, sluggish response in the dc restorer becomes a limiting factor in achieving restoration of the dc level during the time when no pulse is present.

Transmission Deviations: Pulse Echoes from Phase Distortion

The preceding sections have been devoted to a discussion of transmission-frequency characteristics suitable for pulse transmission. A linear phase characteristic has been associated with each amplitude characteristic considered. This section discusses the effect of deviations from phase linearity within the transmission band and the associated distortion in pulse response. We shall confine ourselves to deviations of a sinusoidal character. In principle, the impulse response for any transmission-frequency characteristic can be determined from the Fourier integral. If the transfer function of the medium can be adequately approximated by a rational function, the impulse response and, therefore, the effect of the deviations, can be readily evaluated. In many cases this will not be convenient and the Fourier integral will have to be evaluated numerically from the measured or computed frequency

- -

*Specifically, the number 127, which requires seven pulses in a row in one channel slot, is never transmitted. This limits the maximum number of pulses that can occur consecutively to thirteen: the number 63 and a signalling pulse in one channel followed by the number 126 in the next channel.

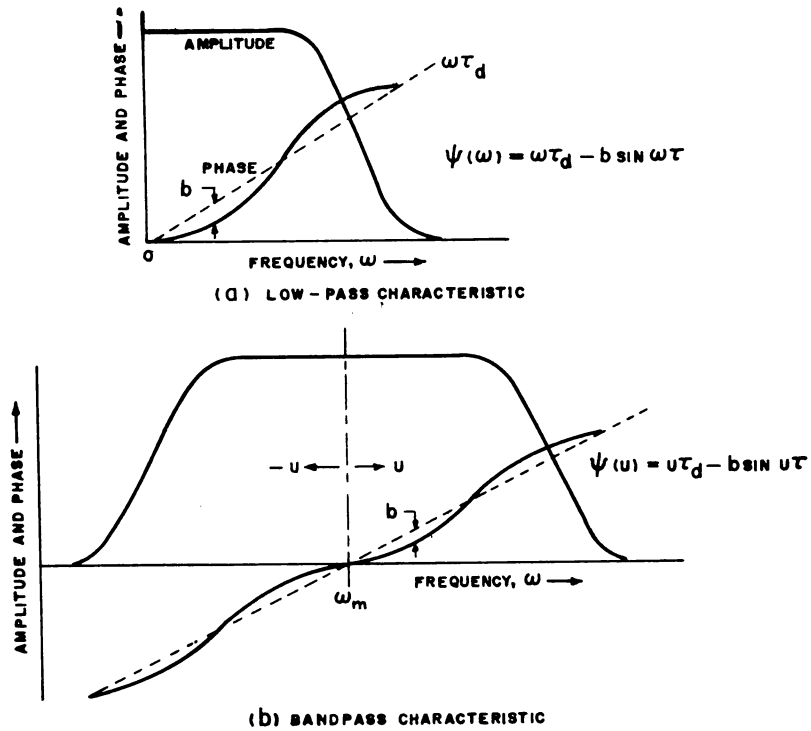
characteristic.* An alternate approach, often suitable for engineering accuracy, is to use the theoretical results given previously (assuming no ripples or deviations) as a point of departure or first approximation. This is followed by a second approximation which considers the deviations from the ideal amplitude and phase as a Fourier cosine and Fourier sine series, respectively. This is the approach adopted previously in dealing with transmission deviations in television systems. When the phase deviation from linearity is small, resulting approximately in a single pair of echoes, the method of paired echoes is sufficiently accurate. In data transmission over voice frequency channels, band-pass filters introduce appreciable delay distortion, resulting in the production of more than a single pair of echoes. When this is true, a large number of echoes of considerable amplitude must be considered. The procedure to be used in this situation is, nonetheless, basically simple. The reader is referred to the bibliography for more detail. Here we shall merely discuss the concepts involved and the results obtained.

A given amplitude characteristic within the transmission band may be associated with various phase characteristics, depending upon the outband amplitude characteristic and on whether or not a minimum phase shift system is involved. It is, therefore, permissible to consider phase departures from a given phase characteristic independent of the amplitude characteristic within the band. Thus, consider the type of phase deviation illustrated by Figure 19, in which the actual phase is seen to ripple sinusoidally about the ideal linear phase characteristic.

It can be shown that the impulse response of a transmission system having such a phase characteristic consists of a train of pulses made up of the main pulse and a series of leading and lagging echoes. The time separation of each pulse in the train is equal to τ, which is the reciprocal of the ripple frequency of Figure 19. The amplitude of the received pulses is a function of the ripple amplitude b. These principles are shown in more detail in Figure 20, where the addition of the echoes to obtain the resultant pulse is illustrated. Figure 20a

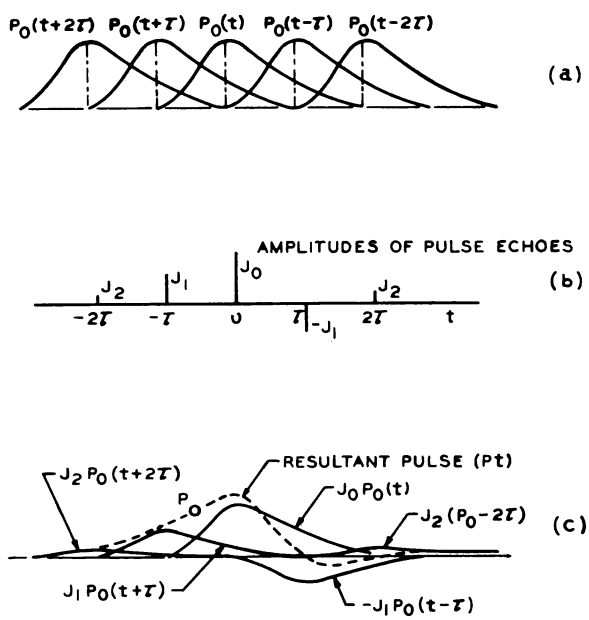
---

*Both procedures can be coded for IBM computers. Presented with a rational function, the computer can be programmed to print out the impulse response. When the frequency response is known only at a discrete set of points, programming can be set up to evaluate responses to commonly encountered driving functions.



Low-Pass and Band-Pass Characteristics with Sinusoidal Phase Distortion

Figure 27-19



Determination of Resultant Pulse by Addition of Pulse Echoes

Figure 27-20

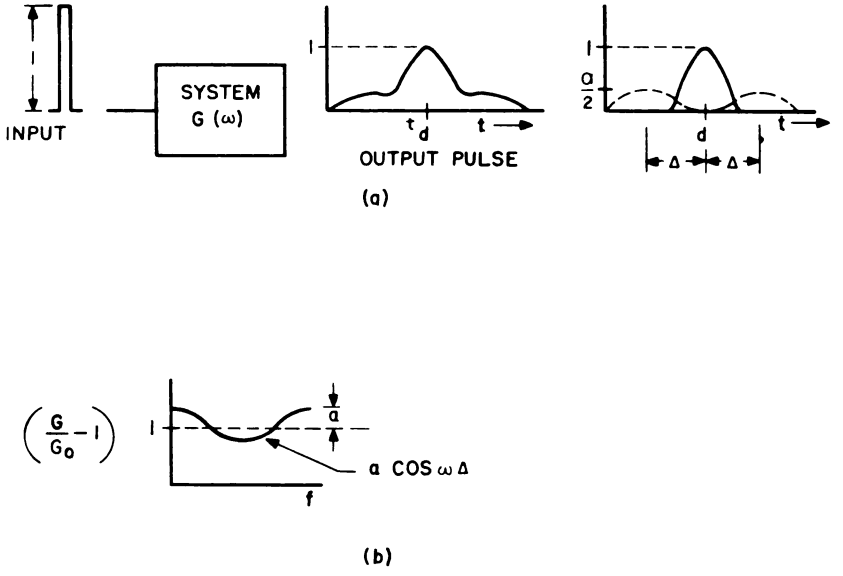
shows a train of pulses, each pulse being intended to represent the impulse response of a transmission system which has the sinusoidal phase deviation illustrated by Figure 19 together with some arbitrary amplitude characteristic. Note that the separation of pulses is τ. Figure 20b shows amplitudes of Bessel functions, $J_n(b)$, which multiply the corresponding pulse shapes in 20a to give the overall pattern of main pulse with leading and lagging echoes shown in 20c. The resultant pulse is shown by the dashed line in 20c.

For ripple amplitudes $b \ll 1$, the Bessel function coefficients become negligible except for J_0, J_1, and J_{-1}. Therefore, to a first approximation, the effect of a small sinusoidal deviation is to produce a single pair of equal amplitude echoes asymmetrically located about the main signal. As the amplitude of b is increased, a greater number of echoes must be considered. In some cases the phase distortion must be represented by a number of sinusoidal components.

The general method referred to here for determining overall pulse distortion becomes quite laborious unless each deviation is small in amplitude. When the amplitudes are small, each deviation can be treated independently, and the resultant pulse shape is found from the superposition of a number of pairs of leading and lagging echoes, each deviation giving rise to one pair.

Transmission Deviations: Pulse Echoes from Amplitude Distortion

The pulse echoes due to cosinusoidal departures in the amplitude characteristic from a given shape can be treated in an analogous manner to that discussed for sinusoidal phase variations, as shown in Monograph 2284. An alternate approach will be given in this section. Instead of proceeding from cause (cosine variation in amplitude characteristic) to effect (echoes in the time domain) we will postulate the problem in a different manner. Assume that we have a transmission system which has a nominal amplitude characteristic $G_o(\omega)$ and a linear phase characteristic ωt_d. Associated with this frequency characteristic is an impulse response defined by $y_o(t)$. Suppose the system is tested by applying a short unit amplitude rectangular pulse to the input and that a measurement of the received pulse is made. It is found that the received pulse departs from the desired output, but that the measured output can be described in terms of the desired output and a pair of leading and lagging echoes, as shown in Figure 21. How has the transmission system changed to result in this distorted response?



Pulse Echoes Due to Amplitude Distortion

Figure 27-21

This can be found by determining the transfer function of the system under the conditions of test. Had the test pulse been an impulse, the output waveform would be expressed by

$$y(t) = y_o\,(t-t_d) + \frac{a}{2}\,y_o(t-t_d+\Delta) + \frac{a}{2}\,y_o(t-t_d-\Delta) \qquad (27\text{-}17)$$

The corresponding Fourier Transform is given by

$$G(\omega)\varepsilon^{-j\theta(\omega)} = G_o(\omega)\varepsilon^{-j\omega t_d} + \frac{a}{2}\,G_o(\omega)\varepsilon^{-j\omega(t_d-\Delta)} + \frac{a}{2}G_o(\omega)\varepsilon^{-j\omega(t_d+\Delta)} \qquad (27\text{-}18)$$

or

$$G(\omega)\varepsilon^{-j\theta(\omega)} = G_o(\omega)\varepsilon^{-j\omega t_d}\,(1 + \frac{a}{2}\,\varepsilon^{j\omega\Delta} + \frac{a}{2}\,\varepsilon^{-j\omega\Delta}) \qquad (27\text{-}19)$$

which is equivalent to

$$G(\omega) = G_o(\omega)\,(1 + a\cos\omega\Delta) \qquad (27\text{-}20)$$

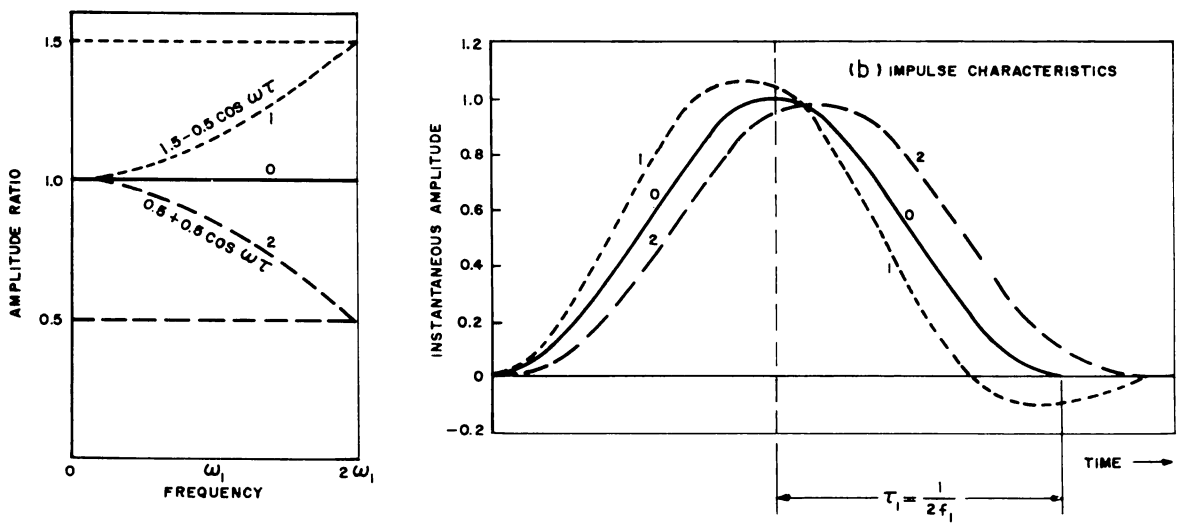
and $\theta = \omega t_d$

It will now be assumed that the rectangular test pulse is sufficiently short that its spectrum is flat over the frequency band of interest, so that it can be treated as an impulse. In this case, the transmission-frequency characteristics of the system are given by $G(\omega)$ and $\theta(\omega)$ in Equation 20 above. The question originally posed can now be answered. The amplitude characteristic of the system has been modified by the term $(1+a \cos \omega\Delta)$. The change in system performance can be defined by

$$\frac{G(\omega) - G_o(\omega)}{G_o(\omega)} = \frac{G(\omega)}{G_o(\omega)} - 1 = a \cos \omega\Delta \qquad (27\text{-}21)$$

This shows that the pair of equal echoes symmetrically located about the main signal results from a cosine ripple in the amplitude characteristic.

In a similar manner, other relationships can be derived which specify the imperfections in a transmission system which give rise to specific echo patterns. For example, a small cosine modification in the amplitude characteristic, accompanied by a corresponding sine deviation in the phase characteristic, gives rise to an impulse response characterized by the desired pulse plus a lagging echo. In Figure 22 is shown the effect of positive and negative cosine variations when the amplitude at zero frequency is held constant, a condition which may be approached in wire systems as a result of variation of attenuation with temperature.



Effect of Slow Cosine Variation in Amplitude Characteristic When Amplitude at Zero Frequency is Held Constant

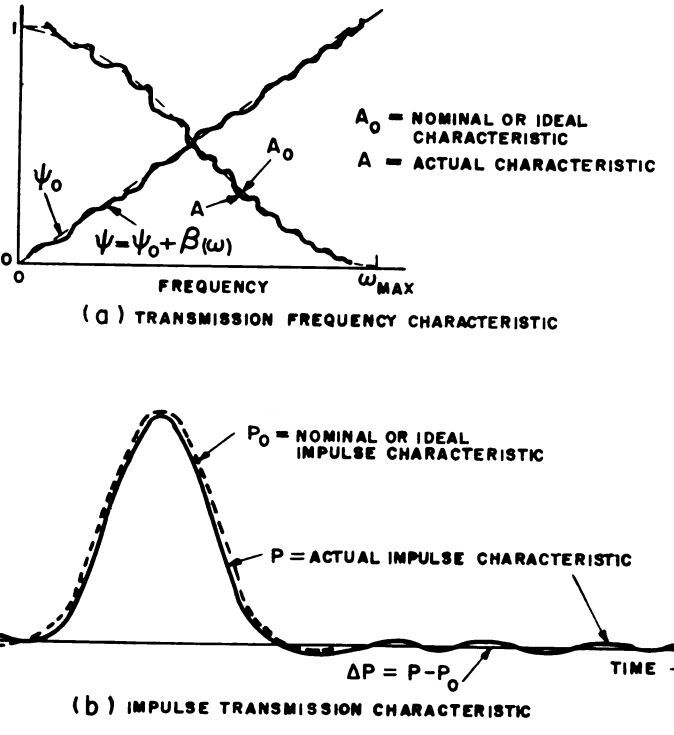
Figure 27-22

Curve 1 would correspond to a 3.5 db smaller loss at the maximum frequency $2\omega_1$ than at zero frequency, and curve 2 to a 6 db greater loss at the maximum frequency. It will be noticed that pulse distortion as well as the variation in the peak amplitude of the pulses is greater under the first condition, i.e. curve 1, which corresponds to a decrease in temperature of the cable medium. In both cases the pulse spreads out into the next pulse position.

Fine Structure Imperfections in Transmission Characteristics

As a result of imperfections in the transmission medium and in equalization, there may be fine structure departures from a nominal transmission characteristic, as illustrated in Figure 23. They are often caused by echoes resulting from impedance mismatches in very long lines. Fine structure deviations from a specified amplitude characteristic may, in principle, be represented by a cosine Fourier series, since the amplitude function is an even function of ω. Thus, if the specified amplitude characteristic is $A_0(\omega)$, the actual amplitude characteristic $A(\omega)$ may be represented by an infinite cosine Fourier series as

$$A(\omega) = A_0(\omega)(1 + a_1\cos\omega t + a_2\cos 2\omega t + \ldots + a_m\cos m\omega t + \ldots). \qquad (27\text{-}22)$$



Fine Structure Imperfections in Transmission Frequency Characteristic and Resultant Prolongation of Impulse Characteristics

Figure 27-23

The coefficients a_1, $a_2 \ldots a_m \ldots$ are determined in the usual manner by Fourier series analysis. If $A_o(\omega)$ closely approaches $A(\omega)$ the fine structure departures in the transmission characteristic and hence the coefficients a_1, $a_2 \ldots a_m \ldots$ will be small.

A corresponding analysis can be made for fine structure imperfections in the phase characteristic. The deviation $b(\omega) = \theta(\omega) - \theta_o(\omega)$ from a prescribed phase characteristic $\theta_o(\omega)$ may in this case be represented by a sine Fourier series since the phase characteristic is an odd function of ω:

$$b(\omega) = b_1 \sin \omega t + b_2 \sin 2\omega t + \ldots + b_m \sin m\omega t + \ldots \qquad (27\text{-}23)$$

In the case of minimum phase relationships, it can be shown that a small cosine deviation of peak amplitude a_m in the amplitude characteristic will be accompanied by a sinusoidal phase deviation having a peak deviation value b_m. If the amplitude deviation is given in nepers and the phase deviation in radians, it is found that $b_m = a_m$.

As shown in Figure 23, the effect of fine structure imperfections in the band is to stretch out the impulse characteristic. It follows that the intersymbol interference is increased as a result of these imperfections. Generally, the effect of the fine structure imperfections can be expressed in terms of the resulting rms intersymbol interference.* Thus, it can be shown that the deviations in the amplitude characteristic give rise to an rms intersymbol interference which is a function of the sum of the squares of the amplitude coefficients a_m. Similarly, the phase deviations produce an rms intersymbol interference which is proportional to the sum of the squares of the phase coefficients b_m. When both deviations are present the resultant rms intersymbol interference is equal to the square root of the sum of the square of the amplitude and phase interferences. In the next section the requirements placed on various transmission imperfections, including the fine structure deviations, are discussed.

- -

*In a PCM system, peak intersymbol interference is of principle interest. So long as the fine structure is random, i.e., cannot be predicted, peak intersymbol interference is estimated from rms interference by applying a peak factor of about 4.

Allocation of Requirements on Transmission Deviations

With the aid of relationships such as those discussed above, and using mathematical techniques covered in the bibliography, we can develop a technique for placing requirements on various portions of the system to achieve the desired transmission objective. In a pulse transmission system the desired performance is specified in terms of the error rate, i.e., the expected number of pulses in error per unit time. As in other transmission systems, the tolerable error rate is arrived at by subjective tests. For example, a probability of error of about one in 10^5 pulses has been found to be tolerable for a 24 channel digit telephone transmission system. Initial system planning and design is concerned with allocating portions of the system margin to various sources of imperfection. Under the assumption that we have developed equations which can be used to describe the rms intersymbol interference due to transmission deviations, and assuming that these effects combine on a power basis with those due to timing errors, pulse width variations, pulse height variations, noise, and other effects one can apportion the margin against error among the sources. In effect, we assume that all sources of noise impairment can be treated like random noise.

As an example, consider the problem of allocating the margins against error when we have random noise present and where all intersymbol interference is due to fine structure transmission deviations. In Chapter 25, it was shown that a signal-to-noise ratio of about 18 db results in an error rate of the order of one pulse in every 10^5. It will be assumed that this is the system transmission objective. The $\frac{S}{N}$ ratio in db is defined by

$$\frac{S}{N} \text{ db} = 20 \log_{10} \frac{V}{\sigma} \qquad (27\text{-}24)$$

where V = Peak pulse amplitude.
σ = rms random noise voltage.

An 18 db $\frac{S}{N}$ ratio is equivalent to the ratio 8 to 1. This means that the root sum of mean square intersymbol interferences plus mean square noise should be less than or equal to $\frac{V}{8}$ to meet our objectives. Furthermore, we might assume that we will give half of our margin to intersymbol interference caused by fine structure amplitude and phase deviations and give the other half to random noise. On this basis, the required rms amplitude deviation must not exceed $\frac{1}{16}$ of a neper, or .5 db.

The associated fine structure phase deviation is $\frac{1}{16}$ of a radian, or about 4°. To accomodate these transmission deviations, the signal to random noise ratio in the medium must be 21 db to meet the overall 18 db objective. From this example, it can be seen that intersymbol interference due to fine structure deviations must be severely limited in a PCM system just as in a conventional AM television transmission system. PCM will, however, retain a significant transmission advantage due to regenerative repeatering.

Expressions for rms interference due to the other systems imperfections mentioned can be developed and combined as above to determine how tolerances can be allocated to each. This will not be pursued further here. Instead, the limitation behind this method of adding systems deviations will be briefly discussed.

It has been assumed that the probability distribution for each source of intersymbol interference is normal, with zero mean and a specified standard deviation. Verification of this assumption for each contributor is a very difficult chore and is rarely possible at the outset of the system analysis. In many cases the use of power addition results in tolerances on system components which are almost an order of magnitude more stringent than actual field trials indicate. Therefore, this method of addition should be used with caution since the reliability of absolute results is open to question. Comparative results, on the other hand, can be used with some confidence to indicate areas where deviations from desired performance are most harmful. The approach should be viewed as an engineering expedient for getting a preliminary feel for the problem of setting tentative allocations of margin. As more detailed data become available, the distribution of margin can be altered. This further pinpoints the need for early PCM system field trials.

Conclusions

The three R's -- Reshaping, Retiming, and Regeneration -- must be performed in a regenerative repeater in order to satisfactorily reconstitute a signal that has been distorted during transmission through a bandwidth-limited medium. Several transmission-frequency characteristics of the combined medium and equalizer have been considered. It has been shown that a gradual roll-off instead of a sharp cut-off amplitude characteristic is desirable for reducing intersymbol interference. In addition, some widening of the band, such as accompanies the gradual roll-off characteristic, may be necessary to provide sufficient power at the pulse repetition frequency to operate the timing circuitry of

the repeaters. Since dc is not transmitted, several methods of combating the effects of low frequency suppression were discussed. It is found that a dc restorer and a limitation on the maximum number of pulses that can occur in succession provides an adequate solution to this problem, at least for the 24 channel system under development. Methods for handling deviations from the desired transmission characteristic have been described. It is shown that ripples in the amplitude and/or phase characteristic produce echoes about the main pulse which result in intersymbol interference. The only solution to this problem is to engineer and maintain the system so as to keep the deviations small in amplitude. In the next chapter, the problems of retiming and regeneration will be considered more thoroughly.

Bibliography

Gain - Phase Relationships

1. Y. W. Lee, Synthesis of Electric Networks by Means of the Fourier Transforms of Laguerre's Functions, J. Math. and Phys., June, 1932.
2. H. W. Bode, Network Analysis and Negative Feedback Amplifier Design, D. Van Nostrand Book Company, 1945.

Transmission Characteristics

3. H. Nyquist, Certain Topics in Telegraph Transmission Theory, A.I.E.E. Trans., April, 1928.
4. Theoretical Fundamentals of Pulse Transmission, E. D. Sunde - B.S.T.J., Vol. 33, 721-788, May 1954 and Vol. 33, 987-1010, July 1954. BTL Monograph 2284. The bulk of the material of this chapter was taken from this reference.
5. Time Division Multiplex Systems - W. R. Bennett, B.S.T.J., Vol. 20, 199-221, April 1941, BTL Monograph B-1291.

Paired Echo Theory

6. H. A. Wheeler, The Interpretation of Amplitude and Phase Distortion in Terms of Paired Echoes, Proc. I.R.E., June, 1939.
7. C. R. Burrows, Discussion of 5 above, Proc. I.R.E., June, 1939.
8. G. N. Watson, Theory of Bessel Functions, Cambridge University Press, 1944. - Math. reference.
9. E. Jahnke and F. Emde, Funktionentafeln, 1928, page 149. - Math. reference.

Low Frequency Suppression

10. Synthesis of Active Networks - W. R. Bennett, Proc. of the Symposium on Modern Network Synthesis, New York, N.Y. - April 1955. Most of the material on low frequency suppression was taken from this excellent paper. Contains list of patents related to Quantized Feedback.

Clampers and Diode Restorers

11. Clampers in Video Transmission - S. Doba and J. W. Rieke, Trans. AIEE, Vol. 69, Part 1, 477-487, 1950.

Chapter 28

REGENERATION AND TIMING

The processes of regeneration and retiming can be described as either partial or complete depending upon the degree to which variations in pulse amplitude and position are removed. Partial regeneration and retiming can be obtained with relatively simple circuitry. The cost to the system, however, is an increase of about 6 db in the requirement on the signal-to-noise ratio at the input to each repeater in order to maintain the same error rate as a system using ideal complete regeneration and retiming. The retiming wave is generally obtained from a tuned circuit driven from either the regenerator input (forward acting retiming) or output (backward acting retiming). The Q of the tuned circuit must be a compromise between desired suppression of pulse position deviations and frequency stability of the tank. An accumulation of timing errors, even in a system using complete retiming, may necessitate the use of mop-up repeaters.

Introduction

The main advantage of a binary transmission system resides in the possibility of periodically reconstructing the sequence of pulses after amplitude distortion and time dispersion occur in transmission. A necessary prelude to reconstruction, particularly in the case of wire circuits, has been shown to be the reshaping of the band limited signal. After reshaping has been effected, a decision can be made as to the presence or absence of a pulse. The reconstruction process which follows involves removal of amplitude distortion of a pulse by regeneration, and retiming for realignment of pulse positions that have been "jittered" during transmission.

These operations are essential to the operation of a PCM system. In this chapter various methods which are under consideration to accomplish this job are examined and their shortcomings pointed out. The reader should remember that here we are not dealing with operating hardware but are looking at the early stages of a new art.

Regeneration

Amplitude regeneration may be classified as either complete or partial depending on the potency of its effect on variations of pulse amplitudes. Complete regeneration requires that the output-input characteristic of the repeater be discontinuous so that values of input below one-half the normal pulse amplitude will produce no output (i.e., a space occurs), and inputs above one-half the normal value will produce

standard output pulses. Curve A of Figure 1 depicts the output-input characteristic of a completely regenerative, two code value (1,0), digit repeater.* Curve B represents a linear response and is included for comparison. Curve C indicates the form of repeater response for partial regeneration. This curve follows the general law

$$\text{Output} = 2^{(N-1)} \ (\text{input})^N \qquad (28\text{-}1)$$

for those values of input below half the peak pulse amplitude. Above this level the curve is represented by

$$\text{Output} = 1-[2^{N-1} \ (1-\text{input})^{\ N}] \qquad (28\text{-}2)$$

Curve C has been plotted for N = 2. Curve D, also shown on Figure 1, shows the regenerative effect of five partially regenerative repeaters in tandem, each having the response of Curve C. In obtaining Curve D it has been assumed that the deviation in input pulse amplitude has occurred only at the input to the first repeater (i.e., the deviation is not systematic). This indicates that a succession of partially regenerative repeaters has an extremely potent effect on completely random variations in pulse amplitude.

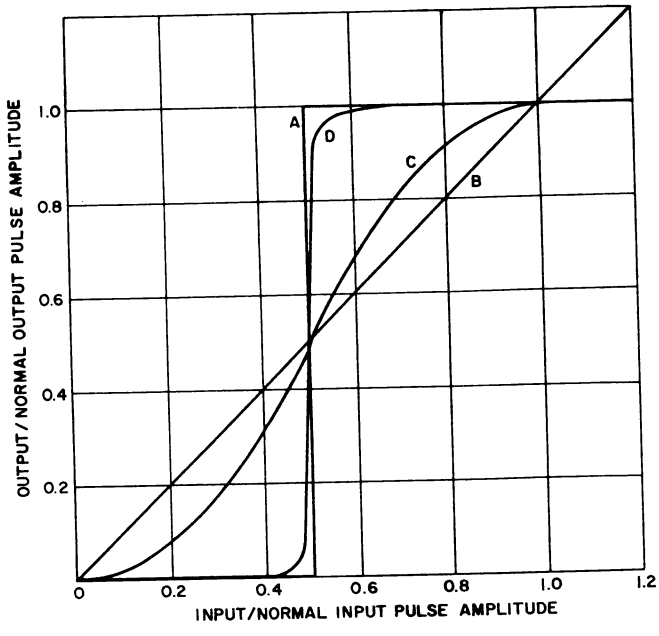
On the other hand, if we have a systematic deviation in pulse amplitude and, for example, a square law regenerator, we will slowly work our way down Curve C until the pulse falls below half-amplitude and is, therefore, lost. To illustrate, suppose the gain of each repeater has been set too low (possibly because of a faulty test set) so that the output pulse is only 0.8 the normal output amplitude of the repeater. In this case, starting with a normal amplitude pulse at the input to the first repeater, the input to the eighth repeater will be below one-half the normal amplitude and will be lost at this point. Although this situation probably represents an extreme deviation from normal performance, it nevertheless indicates that when systematic errors occur there is a distinct advantage to using a regenerator having more nearly the ideal characteristic (Curve A), which can be approximated by using higher values of N.

In a similar manner, a systematic noise source of sufficient amplitude, when introduced into a string of partially regenerative

---

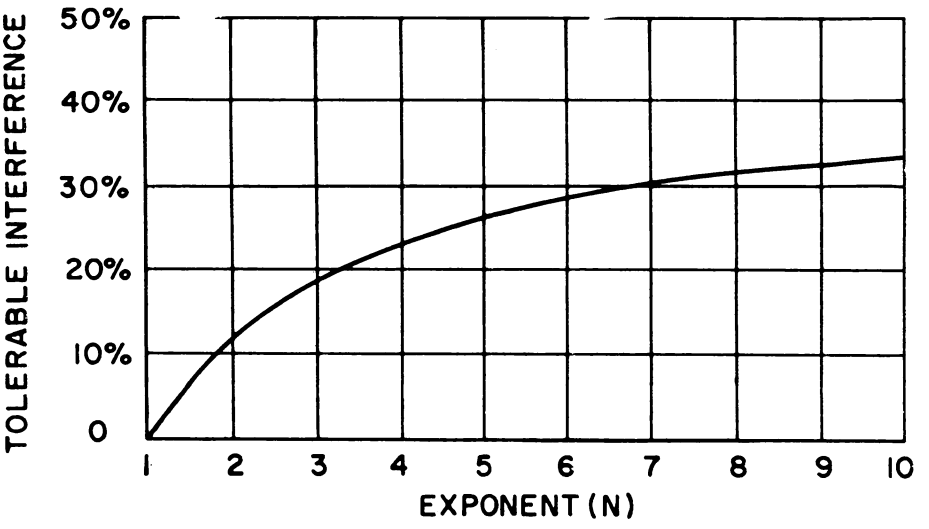
*A repeater capable of handling, for example, three code values +1, 0 and -1 would have an output-input characteristic with odd symmetry about the origin. Thus, the recognition and regeneration circuitry becomes more complex as the number of amplitude levels increases.

repeaters, can result in a noise build-up which may ultimately produce an error. For example, consider a square law regenerator and a situation in which a non-random noise (or pulse), with an amplitude of 0.2 the normal pulse amplitude, appears at the input of each repeater. As seen from Figure 1, the noise at the output of the first repeater will have a relative amplitude of only 0.08. However, because a systematic noise source was assumed, this output adds to the noise at the input to the second repeater to produce a relative amplitude of 0.28 at this point. After going through six repeaters in this fashion, the total noise at the input to the seventh repeater will be greater than the normal pulse amplitude. Obviously, this kind of performance for a systematic noise source is highly undesirable. Fortunately, this situation does not prevail for all amplitudes of systematic disturbance. It can be shown that if the systematic noise at the input to each repeater is sufficiently low in amplitude, the system noise will build up so as to converge upon some constant amount after passing through a large number of repeaters. Figure 2 shows that amount of additive or systematic input interference which, in an infinite number of repeaters, will build up to less than one-half the normal input pulse amplitude. Values of systematic interference below the solid line will converge upon an amount less than one-half the normal pulse amplitude, while systematic noise above the line will build up relatively rapidly to equal the pulse amplitude.



Curve A - Complete Amplitude Regenerator
Curve B - Linear Response
Curve C - Square law Regenerator
Curve D - 5 Square Law Regenerators in Tandem

Figure 28-1



Tolerable Additive Interference, in Percent of Normal Input Pulse Amplitude, vs Exponential "N", when Output = $2^{(N-1)}$ $(\text{Input})^N$

Figure 28-2

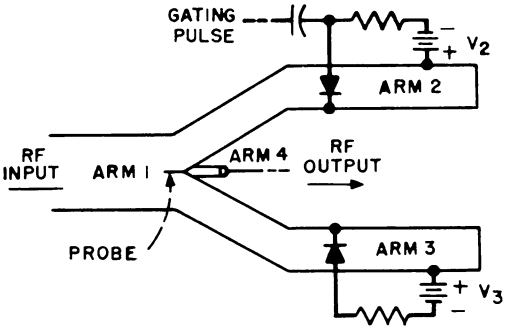
Example of a Partial Regenerator*

Transmission of television signals by PCM requires considerable bandwidth. A seven digit system, for example, would require the transmission of seventy million pulses per second. This need for wide bands makes the microwave region an attractive one in which to work. S. E. Miller** has pointed out that a binary system employing regeneration might prove to be especially advantageous in waveguide transmission. Simplicity dictates regeneration of pulses directly at microwave frequencies. Simple equipment also implies the use of partial rather than complete regeneration. A partial regenerator for such an application at 4 kmc is shown in Figure 3. It consists simply of a waveguide hybrid junction with a silicon diode followed by a shorting plate (i.e., the microwave equivalent of a short circuit) in each side arm. The purpose of the shorting plates is to provide a reflection point for signals entering the side arms. The operation of the device can best be explained by tracing the path of a signal in the circuit shown in Figure 3a. A signal which enters arm 1 splits, one-half the power going into side arm 2 and the other half into side arm 3. If the

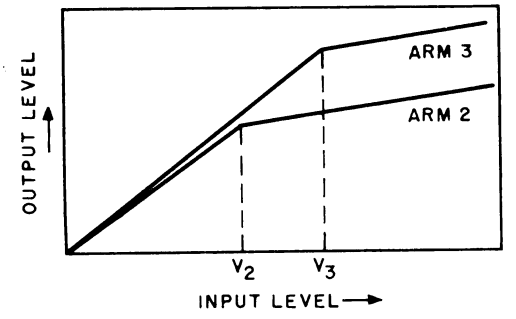
---

*Experiments on the Regeneration of Binary Microwave Pulses - O.E. Delange BSTJ, January 1956. Much of the material in this section was abstracted from this paper.

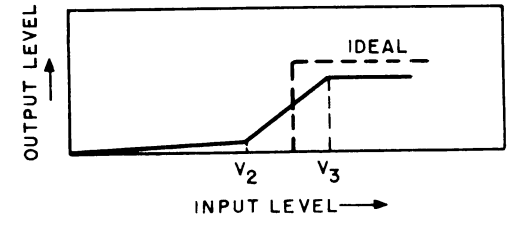
**Waveguide as a Communication Medium - S. E. Miller - BSTJ, Nov. 1954.



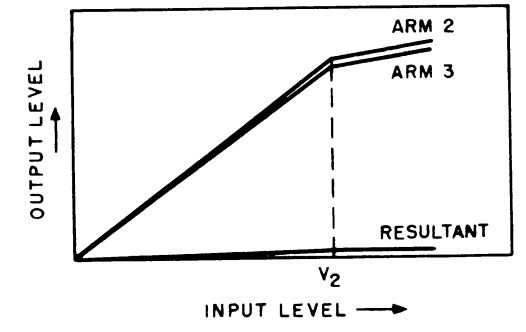
(a) MICROWAVE REGENERATOR EMPLOYING A SINGLE HYBRID JUNCTION.



(b) CHARACTERISTICS OF THE SEPARATE ARMS WITH DIFFERENTIAL BIAS.



(c) RESULTANT OUTPUT WITH DIFFERENTIAL BIAS



(d) CHARACTERISTICS OF THE SEPARATE ARMS AND RESULTANT OUTPUT WITH EQUAL BIASES.

Microwave Regenerator

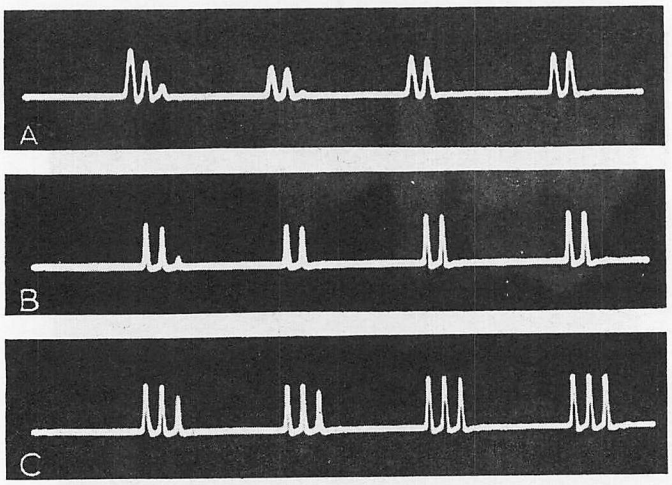
Figure 28-3

diodes are back-biased sufficiently to be non-conducting in the presence of the signal, their effect on the signal will be small and can be neglected. Therefore, the signal in each side arm remains unchanged until the shorting plate is reached, at which point total reflection occurs. Upon returning to the junction of the four arms, one-half the power from each side arm (or one-fourth the original input power) is transmitted into arm 4 and the other half goes back into arm 1. For low signal magnitudes, such that the diodes do not conduct, the input-output characteristic for signal paths from arms 1 to 2 to 4, and from arms 1 to 3 to 4, are shown in Figure 3b by the portion of the curves having unity slope. As soon as the signal voltage in a side arm reaches a value equal to that of the back-bias, however, the diode will start to conduct, thus absorbing power and decreasing the slope of the characteristic. As shown in Figure 3b, the reflection from arm 2 starts to flatten off when the input to that arm reaches the value V_2, while the reflection from arm 3 does not flatten until the input to that arm reaches the value V_3. The combined output, which by proper phasing of the reflected signals can be made equal to the differences of the two arm reflections, is then that shown by the solid line of Figure 3c and is seen to have a transition region between a low output and a high output level. If the two branches are accurately balanced and if the signal voltage is large compared to the differential bias $V_2 - V_1$, the transition becomes sharp and the device is a good slicer (regenerator).

If the two diodes are equally biased, as shown on Figure 3d, the reflections from the two branches should be nearly equal regardless of input. In this case the total output, which is the difference between the two branch reflections, will always be small.

The differential bias referred to above can be provided by a timing pulse. In this way the characteristic of the regenerator is shifted from that shown on 3d to that shown on 3c during the time the information bearing pulse is present. The regenerator is, therefore, also made to act as a gate, though not an ideal one.

A more detailed description of this partial regenerator is contained in the paper by Delange from which the material in this section was obtained. In addition, it contains some very instructive photographs depicting the performance of the regenerator in a circulating loop. This method of testing reconstructive repeaters permits the simulation of a chain of repeaters without the necessity of building a



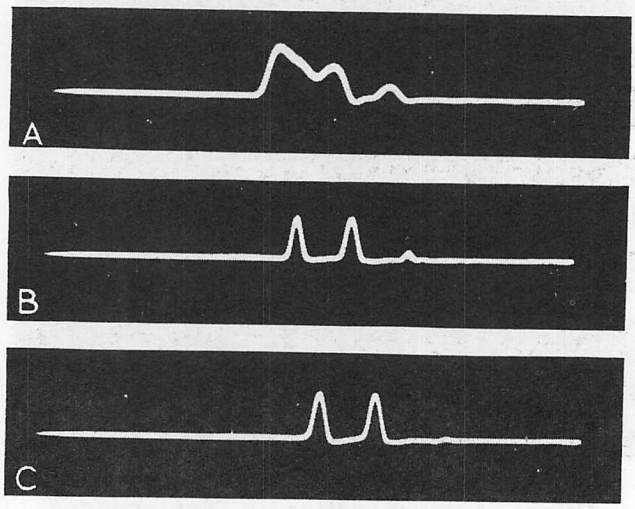
Effect of Regeneration on Disturbances
Which Occur at Only One Repeater

Curve A - Input to regenerator, first four groups
Curve B - Output of regenerator, first four groups
Curve C - Output of regenerator, increased input level

Figure 28-4

large number of repeaters. Figure 4 displays the effects of successive regeneration on disturbances that occur at only one repeater in a string. The left hand pulse group in Figure 4a represents the initial input to the regenerator. Notice the small disturbance that appears as a weak third pulse in the group. The next three pulse groups in this picture show the succeeding inputs to the regenerator after passage through the regenerator and circulating loop 1, 2, and 3 times, respectively, viewing from left to right. Figure 4b shows the corresponding output from the regenerator. After only two passes through the regenerator, all visible effects of the disturbing third pulse have been removed. The effect of increasing the amplitude of the disturbing pulse so that it exceeds one-half the amplitude of the normal pulse is shown in Figure 4c. After the third passage through the regenerator the disturbance has built up to full pulse amplitude, thereby creating a permanent error in this code group. These pictures clearly show that the discrimination of the regenerator to unwanted noise fails when the disturbance occurring in a time slot exceeds the slicing level.

In Figure 5a, a disturbing pulse has been inserted between the desired pulses to simulate intersymbol interference. Figure 5b



Effect of Regeneration on Disturbances
Which Occur at Only One Repeater

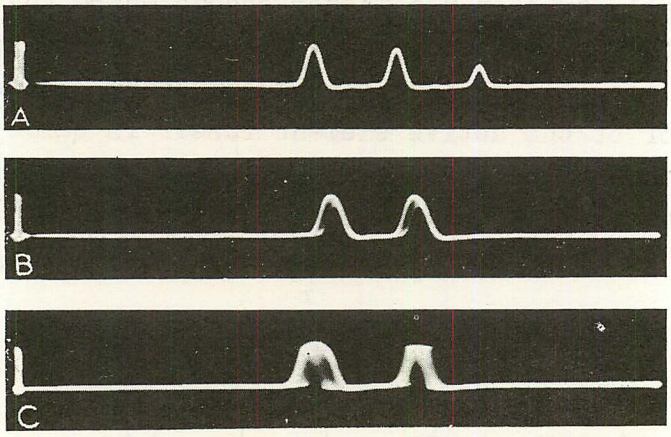
Curve A - Input to regenerator, original signal
Curve B - Output of regenerator, first trip
Curve C - Output of regenerator, 24th trip

Figure 28-5

shows that after only one trip through the regenerator the effect of the added pulse is very small. After a few more trips the effect is completely eliminated, leaving a perfect pulse group as shown in Figure 5c. Discrimination against the intersymbol pulse has been achieved by proper timing, since the interference occurs at a time when no gating pulse is present and, therefore, no output is produced from the regenerator.

The need for retiming is shown in Figure 6. Figure 6a shows the input signal to the regenerator on the first trip. Figure 6b shows that after ten trips without retiming, a noticeable time jitter caused by residual system noise has developed, while in Figure 6c, after 23 trips, the jitter has become severe although the pulses are still recognizable. In this experiment the signal-to-noise ratio was much better than that which would probably be encountered in an operating system. In an actual system, therefore, the time jitter would be expected to build up much more rapidly than indicated here.

The reader is referred to the paper by Delange for further illustrations of the operation of the partial regenerator under various



Effect of Regenerating in Amplitude Without Retiming

Curve A - Output of regenerator, no timing, first trip
Curve B - Output of regenerator, no timing, 10th trip.
Output of regenerator, no timing, 23rd trip.

Figure 28-6

input signal conditions. The purpose of this discussion has been to show that a relatively simple device can be constructed to regenerate pulses, in this case directly at microwave frequencies. Furthermore, the partial regenerator has been shown to have a rather potent effect on disturbances associated with a pulse group. To prevent the adverse accumulation of noise in a partial regenerative repeater system, however, it is found that the signal-to-noise ratio at each repeater should be about 6 db above that required for a system using ideal complete regenerators.

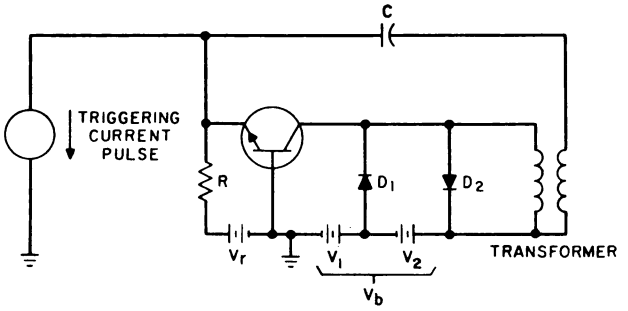
Example of a Complete Regenerator

The discontinuous input-output characteristic of the complete regenerator of Figure 1 is, of course, fantasy in real life. It can be approached, however, with circuitry that is conceptually simple. In most applications to date, a blocking oscillator has been used to approach the desired complete regenerator. It is a relatively simple circuit to construct, but far from simple to analyze and design intelligently.*

---

*Linvill and Mattson (Monograph 2487) have developed an analytical method for designing junction transistor blocking oscillators for pulses in the microsecond region. For faster speeds a similar treatment is too restrictive and the assumptions underlying the approach are subject to considerable question. A somewhat crude, but useful, method which combines the Linvill-Mattson approach with some cut and try is commonly used in the design of blocking oscillators.

This is particularly true at the faster speeds where the inherent energy storage effects (memory) introduced by the stray capacities, leakage inductances, and magnetizing inductances cannot be neglected. Furthermore, the behavior of the active element itself is quite complicated. For purposes of these notes, however, we will only describe the operation of a typical circuit and indicate its limitations and the implications of these limitations on repeater and system design.



A Typical Transistor Blocking Oscillator

Figure 28-7

In terms of Figure 7, the sequence of operations of the blocking oscillator for a cycle is as follows. At the end of the quiescent period, just as the trigger pulse starts, the emitter is back-biased and the collector voltage is V_b. The trigger pulse causes current to flow in the emitter, and at this point the transistor becomes active. After a short transit time, this emitter current is substantially all collected by the collector, and, by virtue of the transformer, a much larger emitter current is drawn in the same direction as that initiated by the trigger pulse. It is desirable to select a transformer with suitable turns ratio to lead to the fastest possible build-up of current in the transistor. The diffusion delay, collector capacitance, and leakage inductance, along with the other transistor parameters, all need to be considered. As the current build-up proceeds, the collector voltage drops from V_b until it reaches V_1, when diode D_1 begins conduction. At this point the circuit regeneration is stopped and the sequence associated with switching ON is complete. The collector voltage remains at V_1 and the diode D_1 merely carries an amount of current to satisfy Kirchhoff's current law at the collector node. However, one observes that the magnetizing inductance of the transformer is in parallel with D_1 and at this stage has a steady voltage V_2 across it. Therefore, the current

through this magnetizing inductance simply builds up linearly, and there is a corresponding decrease in the current carried by D_1. The end of the ON period of the pulse comes when the current in D_1 has dropped to zero. At this point the remainder of the circuit cannot support the growing current required by the magnetizing inductance at the voltage V_2, so this voltage starts to decrease. D_1 opens and the collector voltage starts to rise. There follows in reverse the sequence of events that came with the switching action just described, and the result is that the transistor is switched off. When the voltage at the collector is back again to V_b, diode D_2 prevents any further rise and the magnetizing inductance is discharged through it. By virtue of the charge stored on the capacitor C during the pulse, the emitter is again biased off even in the free-running case and a recharging from V_r through R must occur before the cycle is repeated. In the case of the driven blocking oscillator, the value of V_r is such that the transistor will not begin to conduct again until a new triggering pulse is applied.

From the above discussion, it can be seen that the circuit goes through essentially three distinct operations during a cycle. In each of these portions - switching on, ON period, switching off - simplifications can be made which permit the derivation of equations for rise-time, on-time, and fall-time in terms of the circuit and transistor parameters. In the analysis, the diodes are treated ideally, and it is assumed that the initial energy stored in the blocking oscillator at the beginning of each cycle is zero.*

Neither of these assumptions is exactly true in practice. First, the diodes have finite transition times dependent upon their past history. Secondly, diode D_1, which prevents the transistor from going into saturation, cannot be conveniently employed in some circuits. Without this diode the transistor goes into saturation and the simple equivalent linear circuits which are approximately valid during the on-time and fall-time are no longer adequate descriptions of the true situation.

Energy storage elements associated with a realizable regenerator make its action dependent on the past history of the pulse train being processed. This energy storage results in variations of height and width of the regenerated pulses. Furthermore, changes in the reactive elements of the regenerator with temperature yield additional changes in the pulse shape. The variations in amplitude,

---

*This is in essence the Linvill-Mattson approach.

rise time, and duration of the regenerated pulses result in variations of the pulse shape at the decision making point in the next regenerator, with an attendant S/N degradation. The obvious remedial measure is to use some external control independent of the regenerator to both turn on and turn off the regenerator. This external control can be exercised by the timing wave and will be discussed in the following sections on retiming.

Retiming

Variations in the slicing level of regenerators as well as noise can cause the spacing between pulses to vary from the desired value. To prevent the eventual obliteration of pulses it is mandatory to provide some means for realigning the pulse train. In addition, in order to maintain the best S/N ratio, each pulse should be sampled in the neighborhood of its peak. The problem of retiming has several facets:

1. Derivation of the timing wave.
2. The actual retiming process.
3. Addition of timing errors in a string of binary repeaters.

Derivation of the Timing Wave

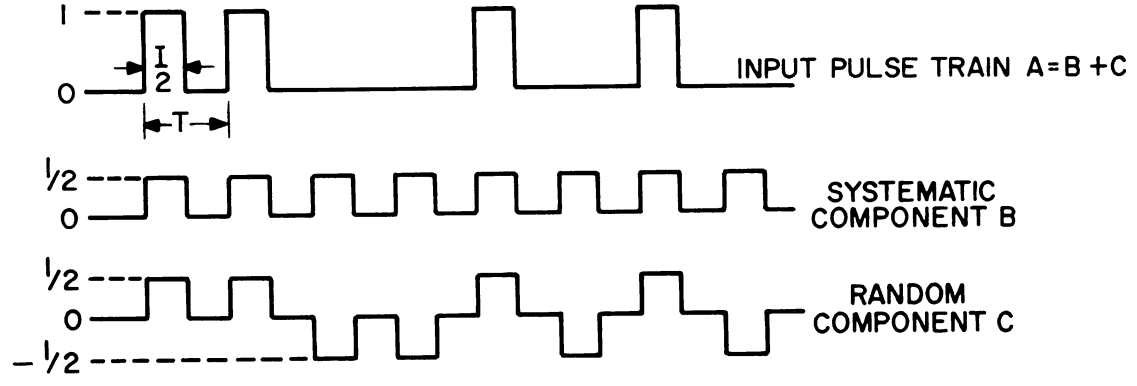
If a transmitted pulse train containing equally likely pulses and spaces is decomposed into a systematic and a random train of pulses as shown in Figure 8, it can be seen that there is a discrete component in the spectrum at the pulse repetition frequency. The simplest means for extracting this component is to pass either the equalized or regenerated pulse train through a series resonant circuit. The output of the resonant circuit will then be a sinusoidal signal of the pulse repetition frequency with both random amplitude and phase modulation.* Random modulation arises due to noise, the statistical nature of the signal, variations in the pulse positions from the required spacing, and tuning error in the resonant circuit. This random modulation is dependent upon the bandwidth of the tuned circuit. The higher the Q, the narrower the bandwidth of the resonant circuit, and, therefore, the greater the suppression of unwanted position modulation. As noted above, phase errors can also arise due to mistuning of the resonant

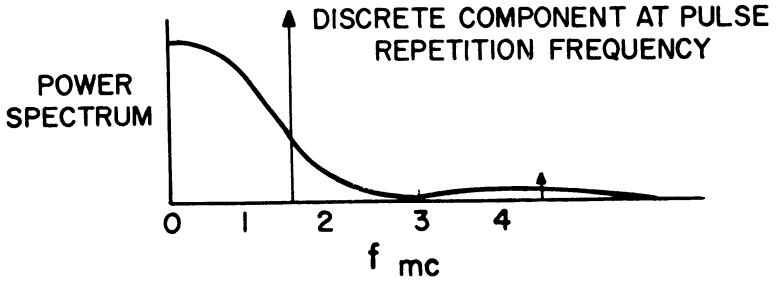
---

*MM 52-170-29 - Derivation of a Timing Wave from a Binary Pulse Train - W. R. Bennett.

circuit. For small amounts of mistuning, this phase shift is proportional to the product of the change in resonant frequency times Q. Therefore, the allowable resonant frequency error for a given amount of timing wave phase shift is inversely proportional to Q. This means that the Q chosen must be a compromise between the amount of retiming and the stability requirements of the tank. A Q of about 100 with presently available inductors and capacitors appears to meet both requirements at least for short systems.

Another approach to retiming uses an extra pair of wires for separate transmission of a sinusoidal signal at the pulse repetition frequency. Carrier repeaters are spaced throughout this pair to provide sufficient clock power to time several systems. Besides using an extra pair of wires, this scheme suffers from other defects, the principal one being that differences in electrical path length due to actual and temperature induced length differences shift the phase of the timing wave with respect to that of the pulse train.



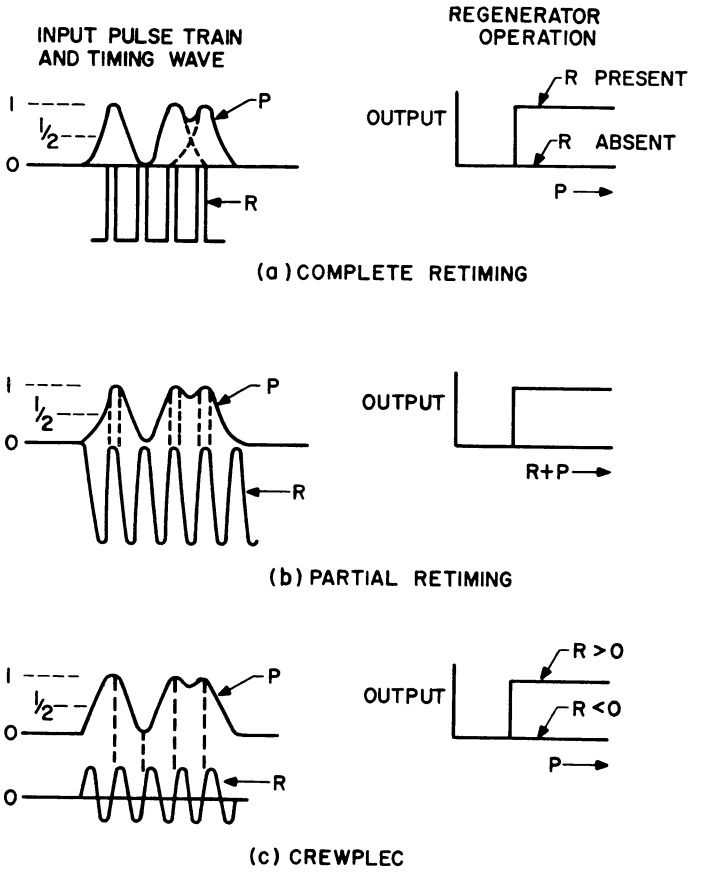


Timing Extraction

Figure 28-8

Methods of Retiming

There are several ways in which the timing wave can be used to reduce time "jitter" in pulse transmission. Three basic approaches are shown in Figure 9.



Retiming and Regeneration Methods

Figure 28-9

Complete Retiming

The first method consists of operating on the extracted timing wave to generate narrow pulses at the pulse repetition frequency so as to sharply sample the input pulse in the neighborhood of its peak. The regenerator invokes a logical "and" operation and regenerates the pulses when the clock pulse is present and the signal amplitude exceeds the slicing level. This is known as complete retiming because the timing of

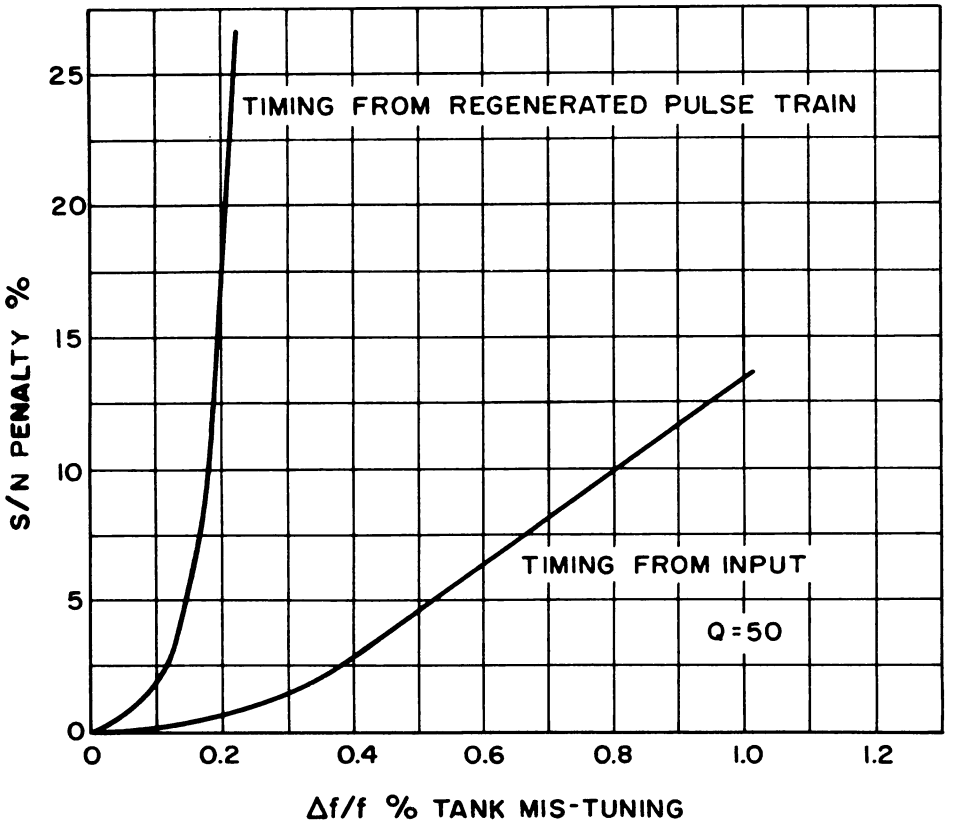
the leading edge of the regenerated pulse is completely controlled by the timing wave and is not affected by noise or timing errors associated with the input pulse. There are two disadvantages to this approach. First, additional circuitry is required to derive the narrow timing pulses and secondly, the regenerator is left to turn itself off. The latter problem is the most critical and we shall return to it later.

Partial Retiming

The second option shown uses a simple algebraic "and" arrangement whereby the timing wave extracted from the tank has its positive peaks clamped to ground and acts as a pedestal. When the algebraic sum of the timing wave plus the signal exceeds the slicing level, the regenerator becomes operative. With such a system, noise and timing errors on the input pulse affect the timing of the leading edge of the regenerated pulse. However, it can be seen that this scheme approaches complete retiming as the clock amplitude increases.

There are two ways that the timing wave can be derived in this partial retiming arrangement. In the backward acting retiming method, the timing wave is extracted by driving a tuned circuit on the output of the regenerator. The alternative procedure, known as forward acting retiming, derives the timing wave from a tuned circuit driven by the incoming equalized pulse train. These two variations in partial retiming can be compared quantitatively by determining the reduction in noise tolerance when the natural resonant frequency of the tank deviates from the pulse repetition frequency. When this occurs, the essentially sinusoidal output is shifted in phase and the information bearing pulses are "sampled" away from their peaks. A smaller negative noise amplitude can now prevent recognition of the pulse. The S/N penalty associated with tuning error is shown in Figure 10, in which it has been assumed that the information bearing pulse is a raised cosine and that the Q of the resonant circuit is 50. For the case of forward acting timing, the noise penalty increases relatively slowly with tuning error. In backward acting timing the reduction in tolerance to noise increases quite rapidly. With the inductors currently available for use in the tuned circuit, this latter scheme is highly unattractive.

In addition to the increased susceptibility to noise, the partial retiming scheme has the same critical disadvantage as complete retiming. The regenerator must still turn itself off.



S/N Penalty vs Tank Mis-tuning

Figure 28-10

Crewplec

The last retiming scheme shown in Figure 9 depicts "Complete Retiming with Pulse Length Control", which is abbreviated to Crewplec. In this approach, the regenerator turn-on is controlled by both the timing wave and the input pulse. When the input signal is greater than the slicing level (ideally half the peak pulse amplitude) and the clock wave is passing through zero in the positive voltage direction, the generator is turned on. After turn-on the regenerative action is independent of the input signal. The turn-off of the regenerator is controlled by the action of the clock. When the timing wave goes negative, the regenerator is turned off by diverting or extracting the feedback current from the blocking oscillator. In

principle, the regenerator is cleared after each pulse is regenerated, and the length of the regenerated pulse is independent of the basic blocking oscillator circuit and its component variations.

The use of a Crewplec regenerator in a reconstructive repeater is a pivotal decision which affects the methods of realizing the remaining repeater functions. It can be demonstrated that backward acting timing combined with a completely retimed regenerator requires extreme stability of the tank frequency. Hence, to obtain reasonable tank requirements, complete retiming dictates the choice of forward acting rather than backward acting timing.

The forward acting timing decision then leads to the choice of a DC restorer rather than quantized feedback for reducing the effects of low frequency suppression on the pulse train. Quantized feedback, as previously discussed, involves the cancelling of input pulse tails with output pulse tails. Unfortunately, it also feeds back a large timing component from output to input which would violate the previous decision to use forward acting timing. Finally, it is apparent that wide band input equalization is required to give the strong timing component required in a forward acting timing arrangement.

Accumulation of Timing Errors

The concluding remarks on retiming relate to the accumulation of timing errors in a long string of reconstructive repeaters. This problem* has and is receiving considerable attention. Most of the analyses to date have been concerned with mathematical models that involve a considerable idealization of the true physical situation. They represent a first look at the problem and are a useful starting point to guide further analysis and experimentation. For purposes of this text, we will concern outselves with a brief summary of the present knowledge.

If we consider variations in the positions of the pulses as a position modulation or timing noise, under certain idealizations we can trace this timing noise as it propagates through the system along with the signal pulses. When we have timing noise introduced at only the first repeater, it is reduced considerably as we go down the chain,

due to the reduction in bandwidth afforded by the chain of tuned circuits. On the other hand, when broadband timing noise is added at the input to each repeater, the output timing noise accumulates as we proceed down the chain. For a Q of about 100 in each tank, Rowe's analysis indicates that this effect is negligible in the short systems presently being developed. Before we can make this conclusion with confidence we should go back and consider some of the idealizations that were made in making the mathematical model of the real world. First, impulses were assumed as the excitation to the tank; secondly, the tanks were assumed all to be on tune; thirdly the regeneration characteristics were idealized. When these assumptions are relaxed, it can be shown that additional phase shifts are introduced into the timing wave in each repeater. These phase shifts vary with the pulse pattern, as is borne out by experiments.

It should be emphasized that all of the above effects can be combated by increasing the complexity of the repeater or by introducing special repeaters at the terminal or other points in the system to mop-up timing deviations. This problem is very similar to the previous comparison of partial and complete regeneration. The decision to employ more complex circuitry for retiming or regeneration or both is intimately tied up with the type of transmission medium to be used and the interference (impulse noise, crosstalk, thermal noise) that is expected. The economics of a particular situation might show that despite the increased circuit complexity of repeaters which approach the ideal, reliability is actually improved and maintenance costs are reduced by using such circuits. Each situation must be examined in the light of its own peculiar constraints.

Conclusions

The four chapters on PCM have served to indicate the fundamental differences between frequency division and time division transmission systems. The real payoff in PCM is the ability to periodically reconstruct the binary pulse train. The increased bandwidth associated with PCM is a small price to pay for the relative freedom from noise which results. The reconstructing process entails reshaping, retiming, and regeneration. All of these functions can be performed to various degrees of perfection. As in TV transmission, fine structure gain and phase deviations must be minimized in PCM systems.

With partial regeneration, additive interference at each repeater can accumulate adversely in a long chain of repeaters. An increase in signal power of about 6 db suffices to make a partially regenerative repeater string have an error rate as good as that of a string composed of complete regenerative repeaters. This is also true of repeaters with partial retiming. Finally, even with complete retiming and regeneration, timing errors can accumulate in a repeater chain. This accumulation is slow and can be limited by various means. In a long chain of repeaters this effect may necessitate the use of mop-up repeaters.

The inherent ruggedness of PCM permits the operation of a system in a relatively noisy environment. Very little has been said, however, about the effects of such things as impulse noise, crosstalk between PCM systems, compatability with AM carrier systems in the same environment, and temperature effects on PCM error rates. These effects are presently under study and means for combating them are being developed.

Chapter 29

SIGNAL PROCESSING

Introduction

The purpose of transmission systems is to transport information-bearing signals from one point to another. Logically, the design of these systems should be tailored to the properties of the signals to be encountered. The modulation systems described previously can accept any signal, within certain bandwidth limitations, from a white gaussian noise to a simple sinusoidal wave. Many transmission facilities, however, are called upon to handle exclusively very specific signals, for instance, voice or television. Considerable economies of bandwidth and signal-to-noise requirement can be achieved by studying the particular nature of such signals and in the light of such considerations, processing the signal to make the most efficient use of the available transmission medium.

General Philosophy

The purpose of this chapter is to indicate the sort of examination of signal properties which is necessary to achieve this end. Several schemes to embody these ideas which are presently under study are briefly described. These are only examples and the future should have in store many more applications of the kind of thinking shown here.

Typically, a message signal such as speech contains a great deal of superfluous structure or detail, commonly called redundancy. It is redundant in the sense that a considerably simplified transmission signal can be obtained by eliminating this detail, while at the receiver the original message can be restored by utilizing sets of predetermined relations. The first step in processing the signal, therefore, is to remove this redundancy. For instance, it is common in transmitting literal text by telegraph to use numerous abbreviations, which are familiar to all telegraph operators. By such coding, much of the redundancy of written English is removed, resulting in more efficient (in this case, more rapid) transmission. The text can be reconstructed at the receiving point since the abbreviations are tacitly agreed upon in advance.

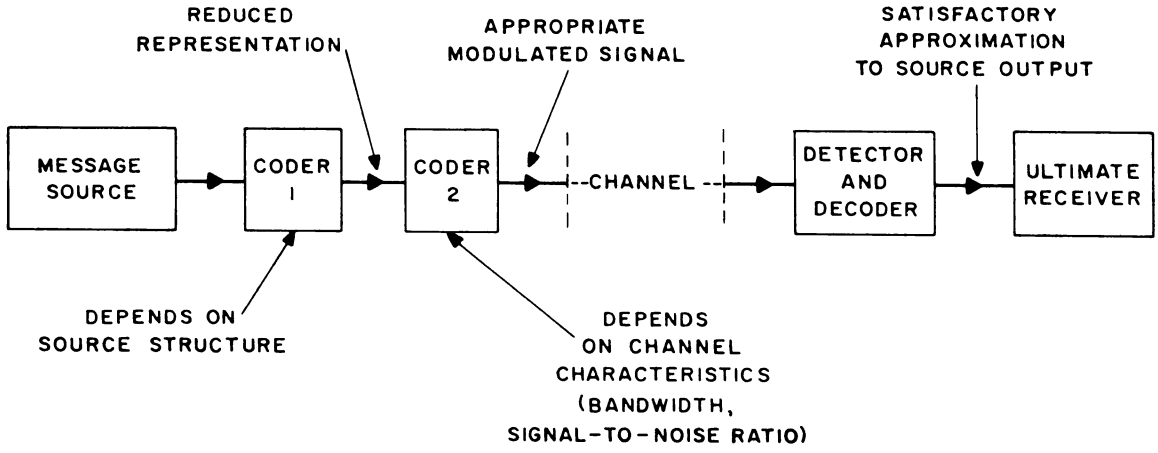
Of course, in such an abbreviated or "reduced" message, errors in transmission can cause serious confusion, since there is little context to fall back on. Thus the next step in processing a signal for transmission is to "match" it by suitable coding to the available transmission medium. The purpose of this coding is to keep the error

rate within bounds so that few confusions result at the receiver. For instance, if the medium is inherently narrow-band and has high signal-to-noise ratio, single sideband or vestigial sideband AM is appropriate. This is often the case in coaxial cable or radio systems. On the other hand, a wideband, noisy channel calls for wideband FM or, even better, PCM. This "matching" consists, then, merely of trading bandwidth and signal-to-noise requirement so that the transmitted signals have minimum bandwidth consistent with the resistance to noise required of them by the channel characteristics.

These two coding processes; namely the reduction of redundancy or superfluous structure and tailoring the transmission signal to the available medium, can yield a highly efficient communication link.

Ideally, then, coding for transmission involves the two steps. Such coding steps are shown in block diagram form in Figure 1. As was pointed out in Chapter 25, the structure which lends resistance to noise, or ruggedness, to a signal is a very specific sort, namely the kind that enables the receiver to separate the signal from the noise. The structure appearing in the original message is a characteristic of the message source and, in general, may not contribute noise resistance to the signal. Thus a good coding scheme trades this kind of detail for a sort which protects the signal sufficiently during transmission through a medium. This is not to say that the original message structure gives no protection from any kind of disturbance. For instance, speech at the hearing point is often subjected to numerous interferences: other sounds in the environment, delayed reflections of itself, and sometimes the listener's own speech. Also, there are mispronounciations, dialects, and accents to confuse the listener. Notwithstanding, there is little trouble in understanding even under extreme conditions. Generally, however, these are not the kind of interferences encountered in transmission, so that a different kind of structure is then called for. The advantages to be had by exchanging one structure for another must be examined for the case at hand; but, in all instances known thus far, a net gain can be had by a judicious combination of codings.

The transmission characteristics of various modulation schemes, i.e., AM, FM, and PCM, have been discussed in the chapters preceeding this one. Here we will concern ourselves with the preliminary coding of a particular message, namely speech, for efficient transmission. Other signals such as television and numerical data have been the subjects of similar codings, but space precludes discussing them here.



System Coding

Figure 29-1

The Nature of Speech

When several people repeat the same words, the human ear and the brain has no difficulty, under ordinary conditions, in recognizing their utterances as representing the same words. On the other hand, an acoustic analysis of the corresponding speech samples shows wide differences. Thus there must be certain properties which all the samples have in common. The listener must use these properties in interpreting the speech. Thus, the key to speech coding for reduced redundancy is the ability to extract from the signal those specific features which convey meaning. The coding transformation must preserve these characteristics. The formulation of the speech coding schemes presently in use is based upon a knowledge of the production and perception of speech itself.

Speech is produced by the coordinated action of the vocal cords and the vocal tract. By muscular manipulations of the oral components, using them in suitable combinations, a speaker can make all of the sounds used in his speech. The gestures of the vocal tract occur slowly compared to the frequencies contained in the speech itself, implying that the latter has a "carrier" nature. In a vowel sound, for

instance, the vocal cords produce a triangular acoustic waveform at a repetition rate between 100 and 300 cps. The frequency spectrum of this wave contains many harmonics of the repetition frequency and has a rather uniform or flat spectrum shape.

The vocal tract, which is composed of the lips, teeth, tongue, palate (roof of the mouth), and throat, can be considered as an acoustic tube about 18 centimeters long and of variable cross sectional area, depending upon the positions of the vocal organs. For any particular cross sectional configuration, this tube has certain resonant frequencies or poles, crudely analogous to the resonances of an organ pipe.

As the vocal cord energy traverses the vocal tract, its spectrum is shaped by the tract resonances which emphasize the vocal energy near these frequencies. The emergent speech, then, has a complicated spectrum, reflecting the instaneous dimensions of the vocal tract. This shaping of the speech spectrum may be looked upon as a low-frequency modulation of the vocal cord carrier. Consonant sounds differ from vowel sounds only in that a noise source producing "voiceless" energy may replace the vocal cords as the tract excitation. As in other modulation systems, the low-frequency modulation function, whose basic rate is 8 to 15 cps for normal speech, contains the transmitted information.

We find, then, that although speech ordinarily proceeds in a steady time sequence, there are periods as long as 1/10 to 1/4 second during which the waveform is very nearly periodic. These segments of speech are the vowels and semi-vowels. They are interspersed with shorter sounds, the consonants, which often have a more random waveform. Because of this structure, linguists and phoneticians have found that speech can be accurately represented by a series of discrete symbols, one of which is identified with each self-consistent section. Thus, the utterance can be broken down into a series of basic sounds which are known as phonemes. For instance, all the vowel phonemes of general American speech are contained in the words heed, hid, head, had, heard, hud (as in Hudson), who'd, hood, hard, hawed, and hod. In English, surprisingly few phonemes, about 40 in all, are necessary to represent accurately the spoken word.

By considering speech as a phonemic sequence, an estimate of the information content can be made. This will provide a hint as to the degree of channel compression which can be reasonably expected from a good coding scheme. Assume for the moment that all phonemes are equally

likely and that each is independent. Then there are about 5 bits/phoneme*; if these occur at an average rate of 10 per second, the information rate is 50 bits/second. On toll quality circuits, signal-to-noise ratios of about 30 to 40 decibels are common. If optimum coding is achieved on such a circuit, the bandwidth necessary for transmission is

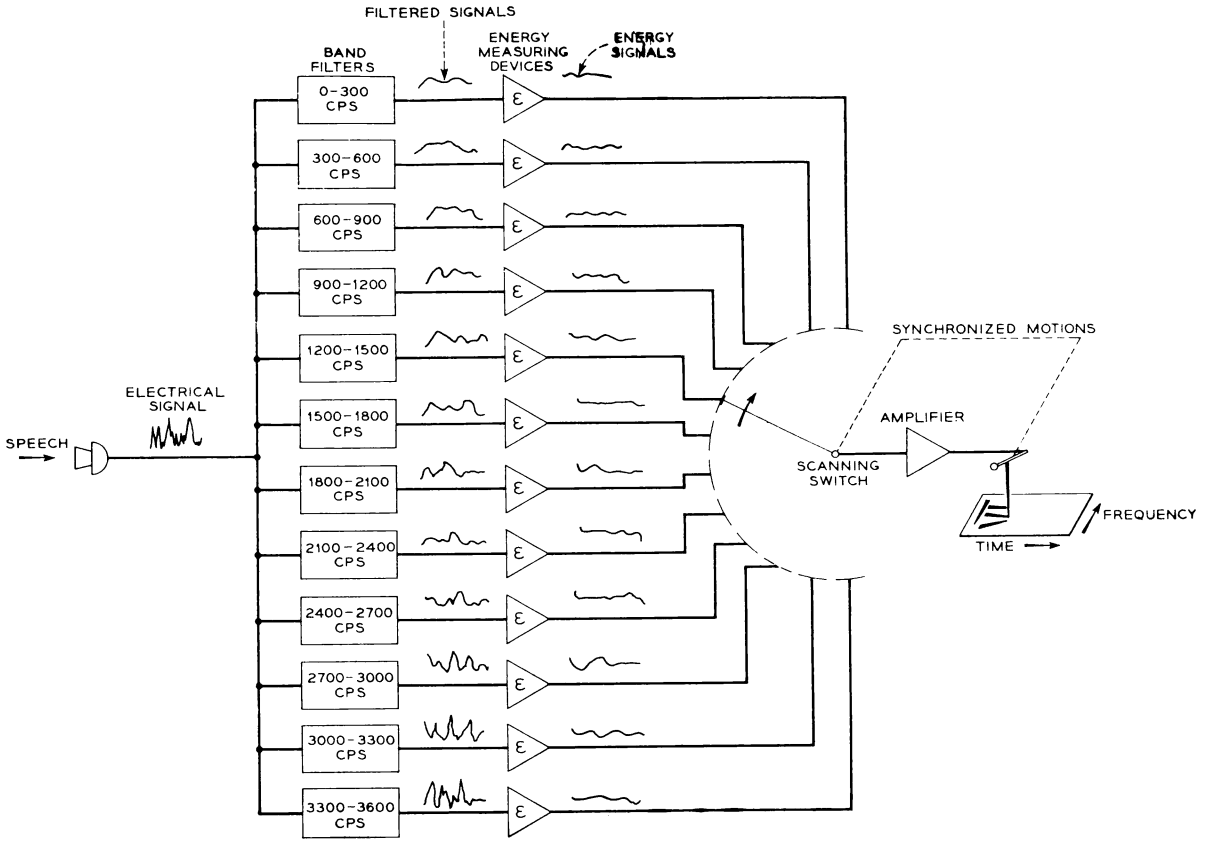
$$B = \frac{C}{\log (1+S/N)} = \frac{50}{7} \approx 7 \text{ cps}$$

If the dependence of successive phonemes in the sequence is taken into account, the information per phoneme, and hence the bandwidth, would be decreased even further. This calculation is not intended to suggest that such a large compression can actually be attained in practice. It is presented to demonstrate the high degree of redundancy present in vocal communication. Also, it should be stated that the information content as determined above includes only phonemic data and does not account for properties of voice quality, such as voice pitch and inflection (pitch variations), which on occasion can themselves carry appreciable information and, of course, give a voice "natural" quality.

It is clear, therefore, that the speech information must be concerned with properties which change at the phonemic rate. These properties can be measured by means of apparatus such as the sound spectrograph, a block diagram of which is shown in Figure 2. Speech is first split into a number of bands by filters which cover the appropriate frequency range. The output of each filter is rectified and averaged, thus indicating the energy traversing the filter during the averaging time. When these signals are marked on a paper so that an energy concentration is indicated by a dark area, the resulting picture is known as a sound spectrogram or a "visible speech" presentation. It indicates the distribution of speech energy as time procedes. A typical spectrogram is shown in Figure 3.

For instance, at the time marked 0.2 seconds, most of speech energy is concentrated near four frequencies, 200 cps., 1200 cps., 2000 cps., and 3000 cps. Indeed, it can be seen that the vowel and vowel-like sounds are characterized by concentrations of energy (known as formants) which traverse the picture along the time dimension and change their positions in frequency at the phonemic rate. These concentrations

*The amount of information represented by a single choice from N equally - likely alternatives is $\log_2 N$. Thus a single phoneme represents $\log_2 40 \approx 5$ bits of information.

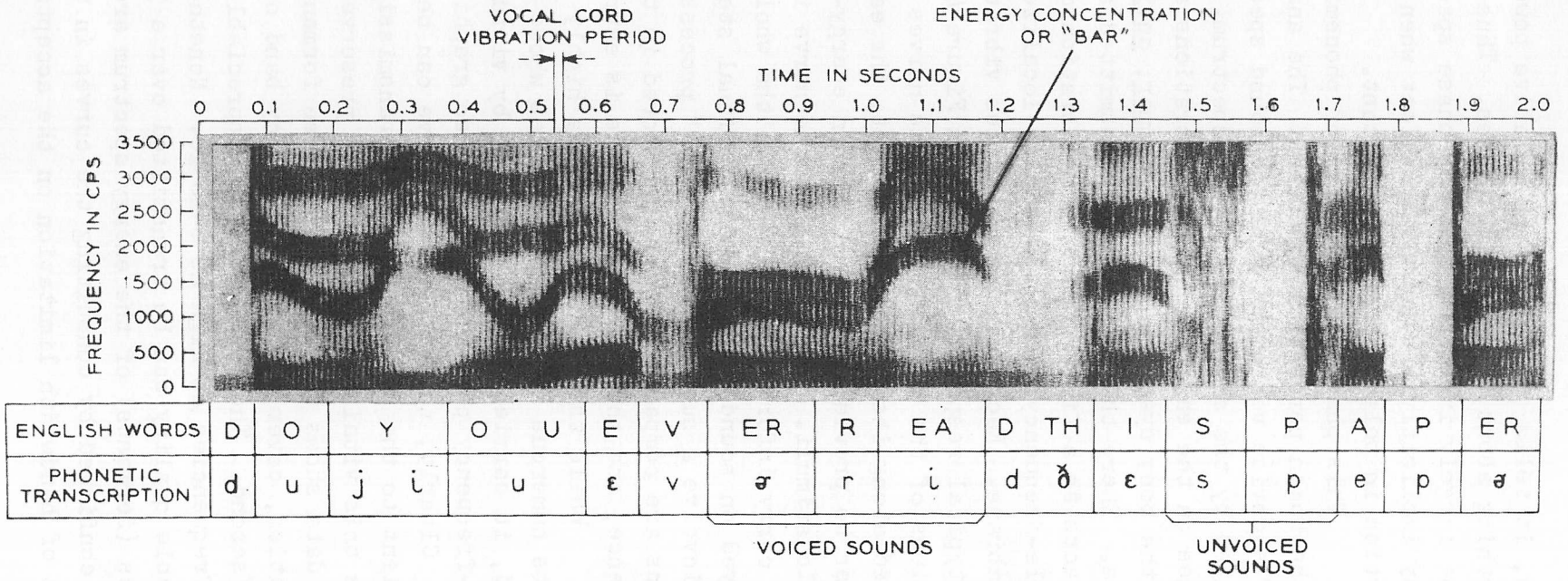


Block Schematic of Sound Spectrograph

Figure 29-2

fall at nearly the same frequencies as the vocal tract resonances. Consonant sounds are seen to be shorter in duration and are typified by smears of energy across large parts of the frequency band.

Usually there are at least three distinct formants in a vowel or vowel-like sound, and it has been shown empirically that their relative positions are sufficient to determine the phonemic content of a sound to a first-order approximation. For instance, a sound can be considered as a point in a rectangular space whose coordinates are the formant frequencies themselves. The points for the same phoneme spoken by various people tend to fall in a volume which is distinct from the similar volumes for other phonemes. The consonant sounds can be sorted by a similar procedure which depends upon the shape of their energy spectra. These facts say that speech sounds can be identified using only the information contained on a sound spectrogram. Since this presentation reflects the "running" or "short-time" power spectrum of



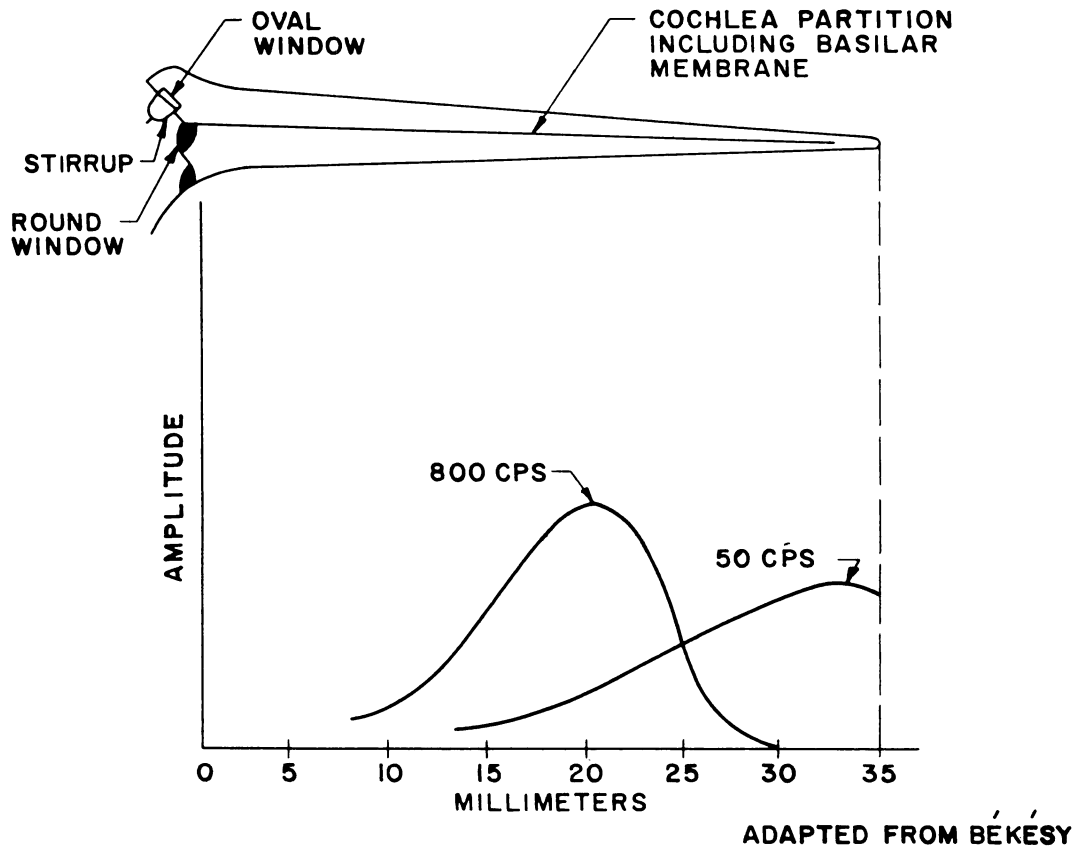
Example of Sound Spectrograph

Figure 29-3

speech, it tells only about the relative power at various frequencies and nothing about the relative phases. Thus, the phonemic information must be largely independent of the phase spectrum of speech. We have already implicitly recognized this fact when we observed that phase distortion in telephony was unimportant.

This phase insensitivity of phonemic data is not inconsistent with the human perceptual apparatus. The anatomy of the ear is in itself a device very much like the sound spectrograph in that it too is activated by the short-time energy spectrum. The incoming speech impinges on the eardrum, and the vibrations are conveyed by a mechanical path (the bony system containing hammer, anvil, and stirrup) to the cochlea. There the motions are transmitted along the basilar membrane which acts as a filter bank. Its vibration displacement responding to a single-frequency wave is spatially localized to a restricted area of the membrane. Each frequency produces vibrations in a somewhat different area; typical responses are shown in Figure 4 for 50 and 800 cps. waves. Vibrations of the membrane excite the nerves whose endings are distributed along its length. In effect, the ear acts as a transducer and analyzer to provide the brain with an energy-frequency picture of the acoustic stimuli. It may be that the nerve impulses arriving in the cortex carry information which is on the whole quite similar to that displayed on sound spectrograms. A final step in the hearing process might involve a subjective "matching" process whereby the incoming patterns are compared to patterns stored in the brain during previous experience, and thus the intelligence is extracted from the speech.

While the speech wave has a highly complex waveform arising from its numerous harmonic frequencies whose phases are not simply related, it carries information only by virtue of its slowly changing energy-frequency pattern, displayed so graphically on sound spectrograms. Clearly, then, the speech wave can be transformed in any manner convenient to the purpose at hand (transmission to a distant point) so long as this vital characteristic is preserved. For instance, experimental data shows that the first three formants, which carry the phonetic information, cover a maximum frequency band of only about 100 to 3600 cycles/second. There is, however, appreciable energy in the speech wave up to frequencies as high as 8-10 kc. Nonetheless, speech of quite acceptable quality can be transmitted over a 3.6 kc band since the main features (formants) of the energy spectrum are preserved. This fact can be confirmed by consulting the curves in Chapter 2, which show the effects of bandwidth limitation on the acceptability of telephone service.

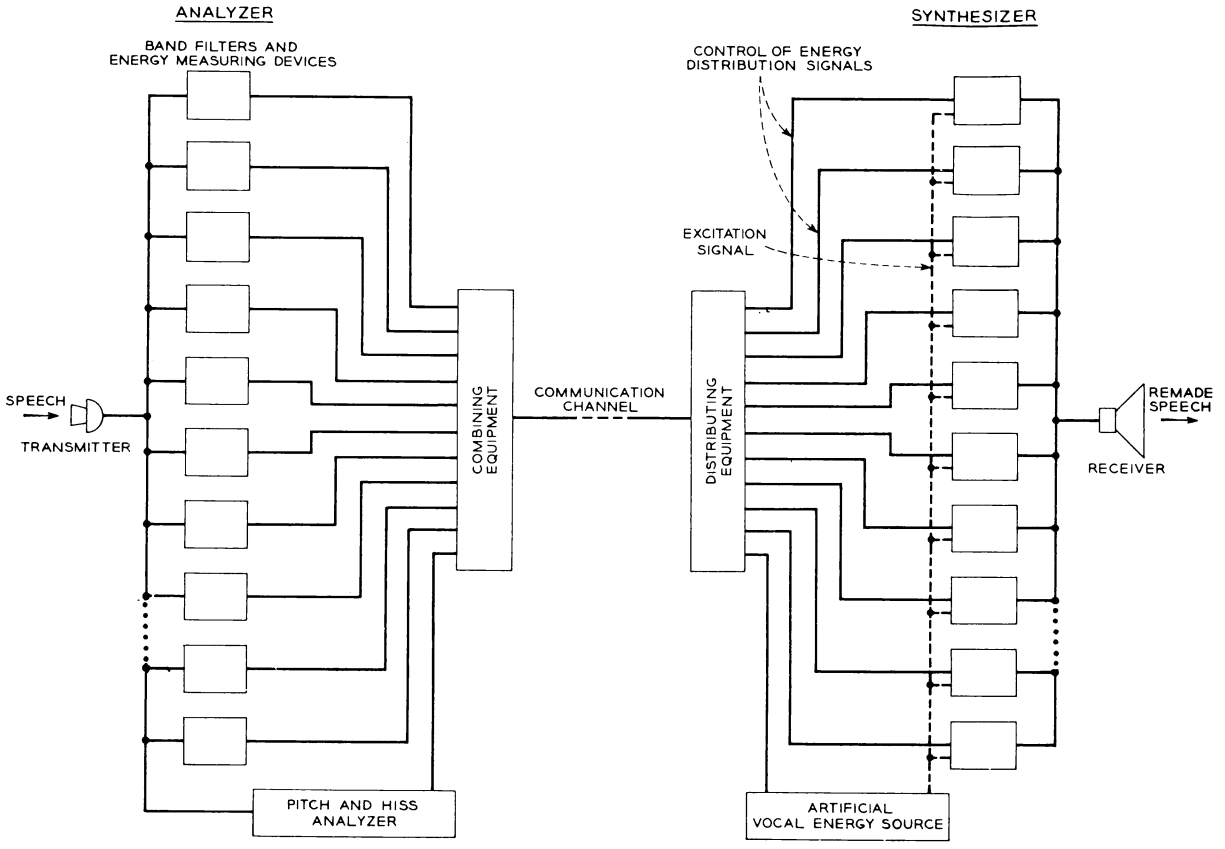


Spatial Response of Basilar Membrane

Figure 29-4

There are other more subtle ways to take advantage of this property of speech for transmission purposes. One of these is H. B. Dudley's invention, the Vocoder. The essentials of this device are shown in Figure 5. The sending terminal is much like the sound spectrograph; the speech is passed through a bank of sixteen contiguous bandpass filters, rectifiers, and low-pass filters, which produce corresponding signals indicating the energy fluctuations in each pass-band. The energy signals vary at the phonemic rate, and hence each can be limited to a band of about 20 cps. Utilizing a group of single-sideband channels, these signals can be transmitted over a total band of approximately 320 cps for the 16-channel vocoder.

The signal-to-noise ratio required in the transmission channel for satisfactory reproduction is of the order of 25 decibels. Since this is about the same as would be required for the uncoded speech, the



Vocoder Block Diagram

Figure 29-5

structure which the vocoder removes from the speech signal is not part of an important noise-resistant property.

At the receiving terminal, the output of an impulse generator, simulating the vocal cords, or a noise source, simulating the voiceless energy in sounds such as "sh", is passed through a bank of filters identical to those at the sending terminal. The amplitude of the output signal from each filter is controlled by the low-frequency energy signal derived from the corresponding transmitting filter. The controlled outputs from the filters are added, thereby creating a signal whose energy spectrum is a close approximation to the original speech. A switching voltage selects the appropriate artificial vocal source (impulse or noise generator). In order that the synthetic speech not sound too artificial, it is necessary to synchronize the vocal cord simulator with the pitch frequency of the original speech, thereby retaining the talker's inflection. This is accomplished by deriving a

low-frequency signal whose amplitude is proportional to the voice pitch. Both the switching voltage and the pitch control voltage can be sent over one additional 20 cps. channel, making 17 in all for the "16-channel" vocoder.

A 16-channel vocoder reproduces speech of rather good quality when the filters are arranged properly across the band. It has been found that the best results are achieved when the filters are equally spaced on the Koenig Aural Scale. This scale is linear with frequency below 1000 cps. and logarithmic above, and corresponds approximately to the spatial response of the basilar membrane in the ear. Thus, the filter bank is a crude analog of the ear itself. Some years ago Dudley proposed that a computer which could identify the 40 English phonemes with reasonable accuracy would be the key to an even more efficient bandwidth compression system. The computer would operate on the speech in much the same way as the Vocoder. In this case, however, the energy signals would be applied as inputs to a group of coincidence detectors which would be prearranged to identify the characteristic energy-frequency pattern of the phonemes. When a certain phoneme was spoken, its energy pattern would actuate one of the detectors. Speech would thereby be broken down into a series of phonemes at the sending terminal. Keying signals indicating which phonemes were spoken would be transmitted to the receiver. There, a facsimile of the original speech would be reproduced by means of an artificial vocal tract. Thus the necessity of actually transmitting the energy signals themselves is eliminated. Instead, a code signal is sent which is used to control an artificial voice in the receiving terminal.

Recently, a system applying these principles has been built, using, however, a classification system which is based on only 16 energy-frequency patterns. Even with this degree of discrimination, a quite acceptable speech output is obtained. The band necessary to transmit the keying signals is about 30 cps. In its most elementary form, described above, the words reproduced by the system would convey the same intelligence as the original speech, but such properties as pitch, inflection, loudness, and stress would be lost. The addition of a pitch circuit in the present model, however, eliminates much of the artificiality but requires a transmission band as great as that for the phonetic data itself. The channel signal-to-noise ratio necessary for satisfactory transmission has not been measured as of the present writing.

The two compression systems just described achieve a high degree of bandwidth reduction. They do, however, require an extensive amount of terminal equipment in order to operate satisfactorily. There are other systems which achieve a somewhat more modest compression and which require correspondingly less equipment. A description of these will not be detailed here, but generally they are based on the same phonemic-acoustic principles.

None of these systems are yet utilized in the telephone system because speech processing of the kind necessary invariably degrades, to some extent, the quality of the speech transmitted. Present research in the speech coding field is directed toward reducing such degradations so that acceptable speech quality can be achieved. Thus, there is the possibility that at some future time vocoders and similar devices will prove to be highly useful in enabling the telephone system to better utilize the bandwidth in existing transmission channels.

INDEX

Where discussion of an item extends over several pages, reference is to the initial page.

A Partial List of Symbols Used in This Text

The following list is restricted to those symbols which have been given a special meaning and used in a number of places in the text. No attempt is made to list such common places as ω for frequency, ω_c for carrier frequency, etc.

Where several references are given, the first is to the definition; subsequent references list particularly important uses of the symbol.