selected
articles
from

The *Lenkurt* Demodulator

Copyright© 1966

by

LENKURT ELECTRIC CO., INC.

Articles contained in this volume

Copyright© 1955, 1956, 1957, 1959, 1961, 1962, 1963, 1964, 1965

Previous Edition

Copyright© 1959 by Lenkurt Electric Co., Inc.

Printed in the United States of America

0-1015 REV 1

World Radio History

selected articles from

The

Lenkurt®

Demodulator

LENKURT ELECTRIC San Carlos, California, U.S.A.
SUBSIDIARY OF
GENERAL TELEPHONE & ELECTRONICS GT&E

World Radio History

PREFACE

The Lenkurt Demodulator is an informative technical periodical published monthly and circulated without charge to technicians, engineers, and managers of companies or government agencies who operate communications systems, and to educational institutions. Each issue features an interesting and instructive article dealing with the subject of telecommunications.

This volume is a collection of 74 of the best and most popular articles selected from past issues of **The Lenkurt Demodulator.** The articles are grouped topically into five sections—(I) MULTIPLEX TECHNOLOGY, (II) MICROWAVE RADIO, (III) DIGITAL DATA TRANSMISSION, (IV) GENERAL COMMUNICATIONS, and (V) SEMICONDUCTOR DEVICES. An expanded Table of Contents is included for quick and easy reference to the subjects covered in each article.

It should be noted that these articles have been reprinted exactly in their original form. Occasional references may be found to articles not appearing in this volume. Also, some articles, dealing with such subjects as satellite communications, may be slightly outdated because of rapid technological changes, but are included for their historical and tutorial value.

EDITOR

The Lenkurt Demodulator

World Radio History

CONTENTS

SECTION I

MULTIPLEX TECHNOLOGY

World Radio History

SECTION II
MICROWAVE RADIO

World Radio History

World Radio History

SECTION III

DIGITAL DATA TRANSMISSION

World Radio History

SECTION IV

GENERAL COMMUNICATIONS

World Radio History

World Radio History

SECTION V
SEMICONDUCTOR DEVICES

World Radio History

SECTION I
MULTIPLEX TECHNOLOGY

World Radio History

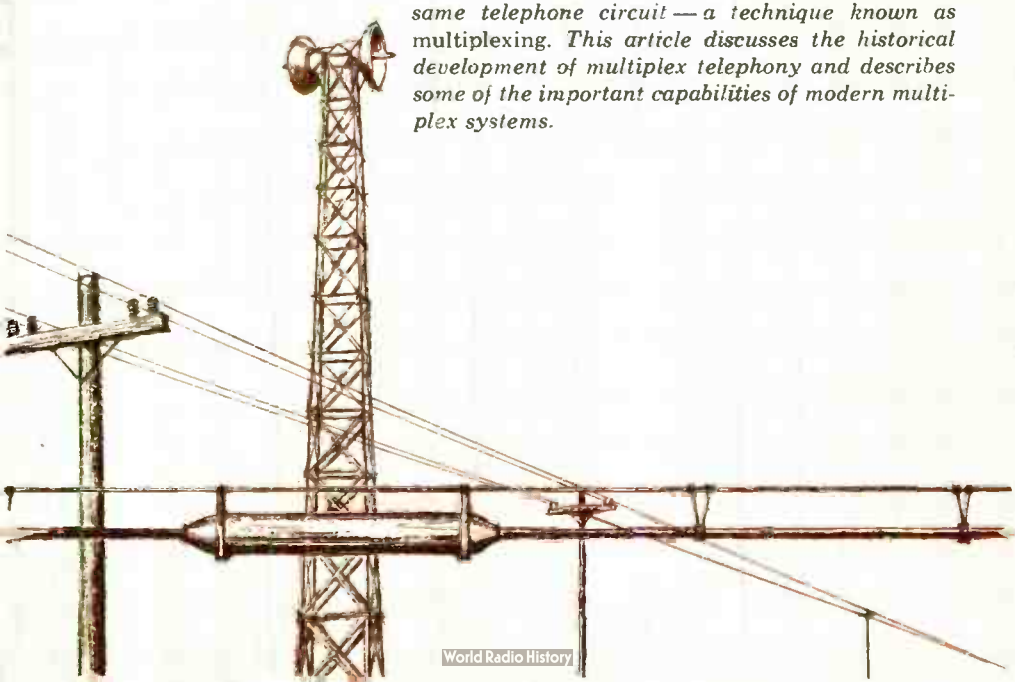
World Radio History

the Lenkurt Demodulator

VOL. 14, NO. 12 DECEMBER, 1965

The Development of MULTIPLEX TELEPHONY

The advancement of telephone communications from a laboratory experiment to a highly sophisticated technology records some rather extraordinary achievements. Prominent among these achievements is a means whereby two or more speech signals can be transmitted simultaneously over the same telephone circuit — a technique known as multiplexing. *This article discusses the historical development of multiplex telephony and describes some of the important capabilities of modern multiplex systems.*



World Radio History

The ability to transmit the spoken word over long distances by means of electrical circuits was indeed a revolutionary step in man's progress. Since Alexander Graham Bell's remarkable invention in 1876, the telephone has become a vital and indispensable element in promoting economic and social progress. Today, with millions of miles of telephone circuits, there seems to be no bounds on man's desire to communicate with distant places.

The first practical telephone circuits consisted of a single grounded wire with a telephone connected at each end. With this arrangement, each telephone could be connected only with the telephone at the opposite end of the circuit—and not to any others. Such a simple arrangement was very limited.

It didn't take long to realize that there was a need for some practical means of interconnecting all the telephones in a local area. This need was satisfied by establishing a central point where all local telephone circuits would come together and where any two telephones could be interconnected, upon request, through a switchboard. This common point became known as the telephone switching, central, or exchange office. The first such commercial telephone office, opened on January 28, 1878, in New Haven, Connecticut, served twenty-one telephones over eight open-wire lines called *subscriber loops*. Service was soon extended by interconnecting the switchboards in the telephone offices with additional wire lines which became known as *trunks*. Trunks interconnecting local offices were designated *exchange* trunks, while those interconnecting long distance offices were designated *toll* trunks.

During its early years, telephony was involved with transmitting only voice-frequency electrical signals over a single grounded wire line. The techniques first used to develop the outside wire plant for telephone circuits were borrowed from the older telegraph industry. Soon hundreds of wire lines, carried on crossarms which were mounted on wooden poles planted along streets and roadways began to appear in the towns and cities and along routes interconnecting metropolitan areas.

It was soon discovered that the single grounded wire line was not completely suitable for telephone communications because of such things as excessive electrical disturbances that were annoying to the users. This problem was solved with the development of the two-wire or *metallic* circuit. This type of circuit consisted of two closely paralleled wires, with one of the wires providing the current return path instead of returning the current through the earth.

Although the metallic circuit solved an interference problem, it presented the enormous problems of reconstructing practically the entire telephone plant and also doubling the already burdensome and oftentimes unsightly mass of wire lines. This seemingly overwhelming task was performed by the telephone industry during the period between 1890 and 1900.

Putting the wire pairs into cables served to remove some of the wire lines from view, but the problem of continually having to enlarge the outside wire plant to satisfy the increasing demand for more circuits still remained. A method of increasing the number of telephone circuits without having to add thousands of miles of more wire was sorely needed.

Multiplexing

Early in the development of telephone communications, it was found that a frequency range from about 300 to 2800 cps would convey speech with sufficient fidelity and clarity for commercial telephone service. (Modern telephone systems transmit speech sig-

World Radio History

nals ranging from about 300 to 3400 cps.) However, since it was possible to transmit electrical waves of hundreds of thousands of cycles per second over wire lines, all the capabilities of the existing telephone circuits were not being used. This fact resulted in a search for a means of transmitting more than one telephone conversation simultaneously over a single pair of wires, a process known as *multiplexing*.

The underlying principles of multiplexing actually predate the invention of the telephone. Bell himself was experimenting with a type of multiplexing for telegraph systems at the time he conceived the idea of the telephone. Since electronic devices were not available to the early experimenters, they had to generate and select the alternating current *carriers* required for multiplexing by mechanical means. Initially, vibrating reeds, each with a different resonant frequency, were used for this purpose.

In these so-called *harmonic* telegraph systems, it was necessary to produce alternating current carriers in the order of only a few hundred cycles since the original signals were dc pulses. However, to multiplex telephone signals the carrier frequencies would have to be much higher. Carriers in the order of tens of thousands of cycles were considered to be necessary to preserve the characteristics of speech signals and to properly separate them electrically. But to produce frequencies of this higher order of magnitude, new and different techniques of generating alternating current were needed. Tuning forks, high-frequency commutator generators, and dc arc interruptors were among the first carrier producing devices considered for multiplex telephony.

Unfortunately, the early experimental types of telephone multiplexing were of little practical use. The development of successful multiplex systems had to await the arrival of the wireless or radio technology which occurred shortly before 1900. During the decade following 1900, a number of advances occurred which later helped to develop practical commercial telephone multiplex systems. Among these were the invention of the vacuum tube by Edison, the addition of a grid to the vacuum tube by DeForest, and the improvements in electrical wave filters.

In 1910, Major G. O. Squier, a United States Signal Corps officer, developed an experimental multiplex system which was operated over a short length of cable. This experiment restimulated interest in commercial multiplex telephony and led to extensive developmental effort by the Bell System

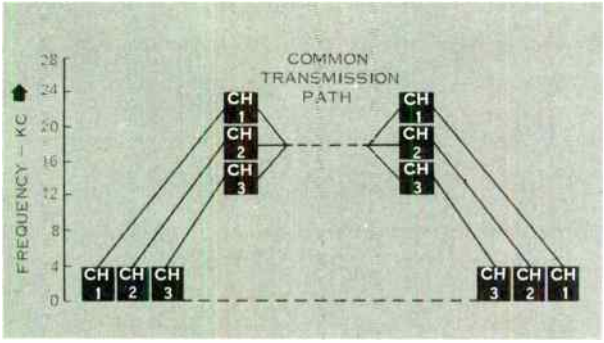


Figure 1. In frequency division multiplexing, low-frequency speech signals, ranging between 0 and 4 kc, are shifted to separate positions in a higher frequency range.

World Radio History

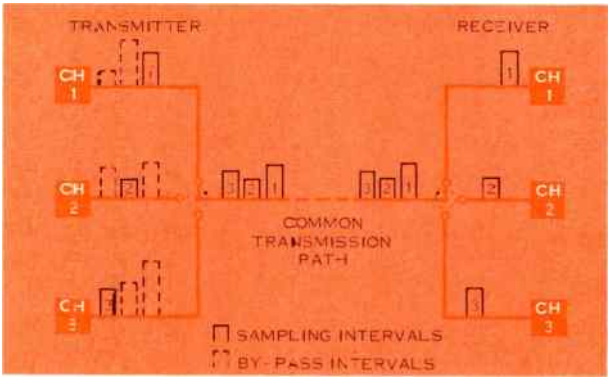


Figure 2. In time division multiplexing, speech signals are separated by briefly sampling each channel in a regular sequence.

—first by Western Electric and later by the Bell Telephone Laboratories.

It is interesting to examine some of the names that have been associated with the technique of multiplexing. The terms carrier current telephony and multiplex telephony appear often in early references, with other terms such as high frequency telephone, wired radio, and line radio receiving some notice. The term carrier telephony prevailed afterwards, but has been applied ordinarily only to wire systems, while the term multiplex telephony has been applied to radio systems. More recently, the term multiplex telephony has been used exclusively, regardless of the type of system to which it refers.

How It Works

In order to transmit two or more telephone signals simultaneously over the same circuit, the signals must be separated so that they do not interfere with each other. This can be done by separating them either in *frequency* or in *time*. Separating signals in frequency is known as frequency division multiplexing, whereas separating them in time is called time division multiplexing. The concepts of these two types of multiplexing are shown in Figures 1 and 2. Both methods were experimented with in the early stages of multiplex telegraphy. The vibrating reed technique, which used a different frequency for each telegraph channel, was an example of frequency division multiplexing. The most prevalent type of multiplexing used in the telephone industry has been frequency division.

Either frequency modulation or amplitude modulation may be used to transform speech signals to separate frequency bands, as required for frequency division multiplexing. Amplitude modulation is most commonly used. In this type of modulation, the resulting modulated wave consists of a carrier wave, an upper sideband wave, and a lower sideband wave. The two sideband waves are separated from the carrier wave by a frequency equal to the modulating speech signal. Each sideband wave includes all of the frequency components of the modulating speech signal. It soon became evident that only one sideband wave had to be transmitted. Therefore, the carrier wave and the other sideband wave could be suppressed, provided an equivalent carrier was available at the receiving end to demodulate the signal.

By using only one sideband, the energy required to transmit the signal is reduced considerably and the fre-

World Radio History

quency band used is essentially half of that required if both sidebands and the carrier are transmitted. Thus, twice as many telephone channels can be obtained in the same multiplex frequency band. The technique of transmitting only one sideband, known as *single-sideband suppressed-carrier,* is used in most of the multiplex systems that have been developed for toll circuit use.

The basic components of a frequency division multiplex system, using single-sideband suppressed-carrier, are the modulators, demodulators, filters, and a source of carrier frequencies. Additional circuits are necessary to provide such things as power, signaling, and regulation. Figure 3 illustrates the use of the basic components in a simplified two-channel multiplex system.

In this system, speech signals received from the telephone transmitters associated with each channel pass through a low-pass filter which limits the upper end of the frequency range to about 4000 cps to conserve the frequency spectrum. Next, the signals are applied to a balanced modulator where they combine with a carrier received from an oscillator. (Note that the two carriers are different.) The carrier is suppressed in the modulator; therefore the output of the modulator contains only the upper and lower sidebands. The upper sideband is then attenuated in the bandpass filter leaving only the lower sideband for transmission. At this point, signals in channel one range between 6 and 10 kc, while signals in channel two range between 11 and 15 kc. Signals in both channels are now combined, amplified, and transmitted over the same transmission line.

At the receive terminal the combined signal appears at the input of two bandpass filters. The channel one bandpass filter passes only the 6 to 10 kc signals and rejects the 11 to 15 kc signals, while the channel two bandpass filter passes only the 11 to 15 kc signals while rejecting the 6 to 10 kc signals. Next, the signals in each channel are applied to a demodulator where they combine with a carrier of the same frequency as that used in the transmit terminal. The output of the demodulator is passed through the 4-kc low pass filter which attenuates the upper sidebands, leaving only the original speech signals to be transmitted to the telephone receiver.

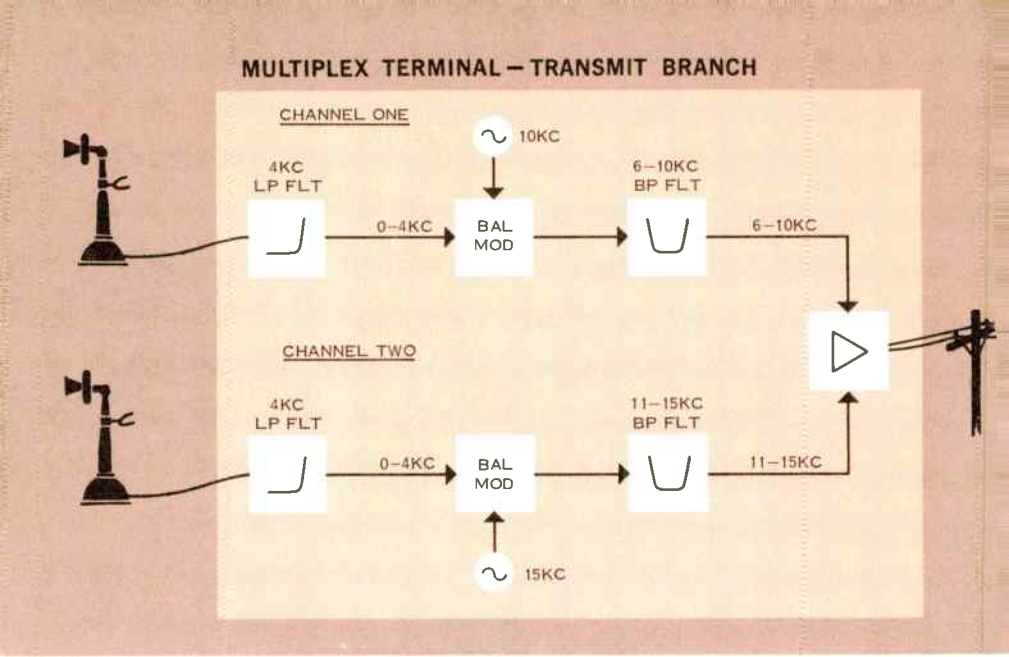
Hence, this simplified multiplex circuit is able to electrically separate two speech signals so that they can be transmitted simultaneously over the same transmission line, without interfering with each other, and properly detected at the receiving end.

Early Commercial Systems

In 1914, a laboratory model of a complete multiplex telephone system was tested by the American Telephone and Telegraph Company on an open-wire line that extended from South Bend, Indiana to Toledo, Ohio. The results of this test proved the feasibility of multiplex telephony in commercial applications. Later, in 1918, the first commercial multiplex system began operating between Baltimore, Maryland and Pittsburgh, Pennsylvania.

The Bell System designated their original multiplex system as type A, thus beginning a succession of different systems with alphabet designations. The type A system provided four two-way channels for use over a single open-wire line, with the same carrier frequencies being used for each direction of transmission. This required the use of hybrid coils to separate the transmit and receive signals, a feature which later proved to be objectionable. In the next type of multiplex system, installed in 1920 and designated type B, three two-way channels were provided using different frequencies for each di-

World Radio History



rection of transmission. This permitted the use of filters to separate the channels and the hybrid coils were no longer necessary. This technique of using different frequencies for each direction provided what is known as an *equivalent 4-wire* system.

Both the type A and B systems used amplitude modulation to superimpose the telephone signals onto the carriers. In the type A system, the carrier and one sideband were suppressed with only the other sideband being transmitted. However, in the type B system, one sideband was suppressed while the other sideband and the carrier were transmitted. When the type C system was developed in about 1925, it incorporated the best features of the two earlier systems. This system provided three channels using different frequencies for each direction of transmission, and transmitted only one sideband.

The initial developments in multiplexing were directed toward telegraphy, and the frequencies used for this service occupied the range below 10 kc. The frequency band from 10 to 30 kc, therefore, was used in the early telephone multiplex systems, limiting them to 3 or 4 voice channels. Although higher frequencies could have been used, at the time it was not considered practical because of the higher attenuation and crosstalk, and other factors associated with open-wire lines.

The type C multiplex system developed rapidly and was used extensively throughout the Bell System's long distance toll routes. By 1928, several transcontinental 3-channel multiplex systems were in operation along with many shorter systems between such points as Chicago and Pittsburgh, and San Francisco and Los Angeles.

Before the first commercial multiplex systems could be successfully used, it was necessary to measure and analyze the characteristics of the transmission medium to be used. Tests showed that the attenuation of open-wire and cable was a function of fre-

World Radio History

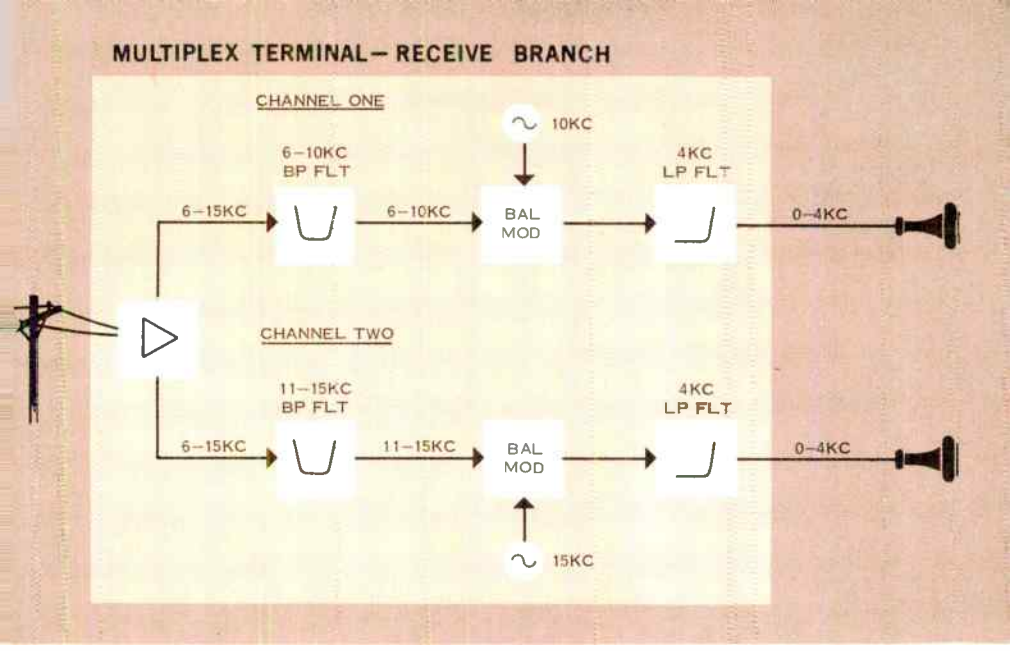


Figure 3. Simplified two-channel telephone multiplex system.

quency, and increased as the frequency was increased. The attenuation-frequency characteristics of open-wire pairs differed for different wire sizes and for different spacings of the wire pairs. The attenuation per unit length was much lower for open-wire pairs than for cable pairs under most conditions, but the open-wire pairs were more severely affected by changes in temperature and humidity, and by icing.

In the multiplex frequency band, the characteristic impedance of open-wire pairs varies somewhat with frequency, different wire gauges and spacing, but is generally considered to be about 600 ohms. However, for non-loaded toll cable the characteristic impedance is approximately 130 ohms.

Differences such as these greatly affected the design requirements of early multiplex equipment. In general, if a multiplex system was designed for one transmission medium, such as open-wire, it could not be readily adapted for use on a cable.

Soon, however, technical advances proceeded far enough to permit the development of twelve-channel multiplex systems designed to operate over non-loaded cables and open-wire lines. In the design of these systems, many new types of components were used and standardized so that they could be interchanged, thereby making multiplex systems more economical.

The first of these systems, designated type J, was designed for open-wire lines and had a line frequency range of about 36 to 140 kc. The frequency band from 36 to 84 kc was used for transmission in one direction and the frequency band from 92 to 140 kc was used for transmission in the opposite direction.

In addition, this system could be used with a type C system, operating in the 6 to 30 kc range, and a v-f cir-

World Radio History

cuit to provide up to 16 telephone channels over one open-wire line.

The next 12-channel system, designated type K, was used on a transcontinental cable system and was put into service in 1938. This system used the frequency band from about 12 to 60 kc for transmission in both directions. This was done by using a different wire pair for each direction, thus establishing a *physical 4-wire* system.

The lower line frequency of the type K system was achieved using a new technique called *group modulation*. In the earlier systems, the multiplex line frequencies were accomplished by a single direct-modulation step. Group modulation, however, consists of using two or more steps of modulation to establish the line frequencies.

One of the most significant advantages of group modulation was that it provided a simplified means of interconnecting standard sub-groups of channels at line frequencies, a technique employed extensively in later multiplex systems.

Later Developments

Prior to World War II, multiplex systems were designed for operation over medium and long distance telephone routes. The development and installation of both the Western Electric 12-channel J open-wire and the 12-channel K cable systems established multiplexing as the method for deriving toll circuits over long distance routes. Not only was it possible to increase circuit capacity more economically, but the grade of circuit was improved.

Following the war, there was an unprecedented demand for more short, medium and long distance telephone circuits. At this same time, there was a shortage of materials for outside plant construction, and the costs of labor and materials were increasing.

To supply the telephone circuit requirements, new approaches to multiplexing were necessary, since its use prior to this time was considered economical only for long-haul circuits. For short- and medium-haul use, various multiplex systems were developed which were simple to install, operate and maintain, and which were competitive with voice-frequency circuits, even over very short distances.

These systems included the type N system and Lenkurt's type LN system, which provided 12 channels over two cable pairs, and the type O system and Lenkurt's type 45C system which provided four 4-channel groups or sixteen channels over an open-wire transmission line.

In a number of instances, the full channel capacity of a 12- or even a 3-channel system was not required. For this reason, *stackability* was desirable in multiplex equipment, and systems, such as the Lenkurt type 33A, were developed with a minimum of so-called common equipment. Thus, individual channel equipment could be added (stacked) later to meet future circuit needs.

The extreme flexibility afforded by stackable multiplex equipment, and its short-distance *prove-in* cost, permitted economical expansion and was a contributing factor to the use of multiplex systems for short- and medium-haul toll telephone circuits.

Once multiplexing had been firmly established as the standard method of deriving toll circuits, there was increasing pressure to reduce the physical size of the equipment. During and following World War II, a continuing effort was made to reduce the size of electronic components, and to add new devices—such as germanium and silicon diodes and transistors — which were small and required much less power than vacuum tubes. Lenkurt's type

World Radio History

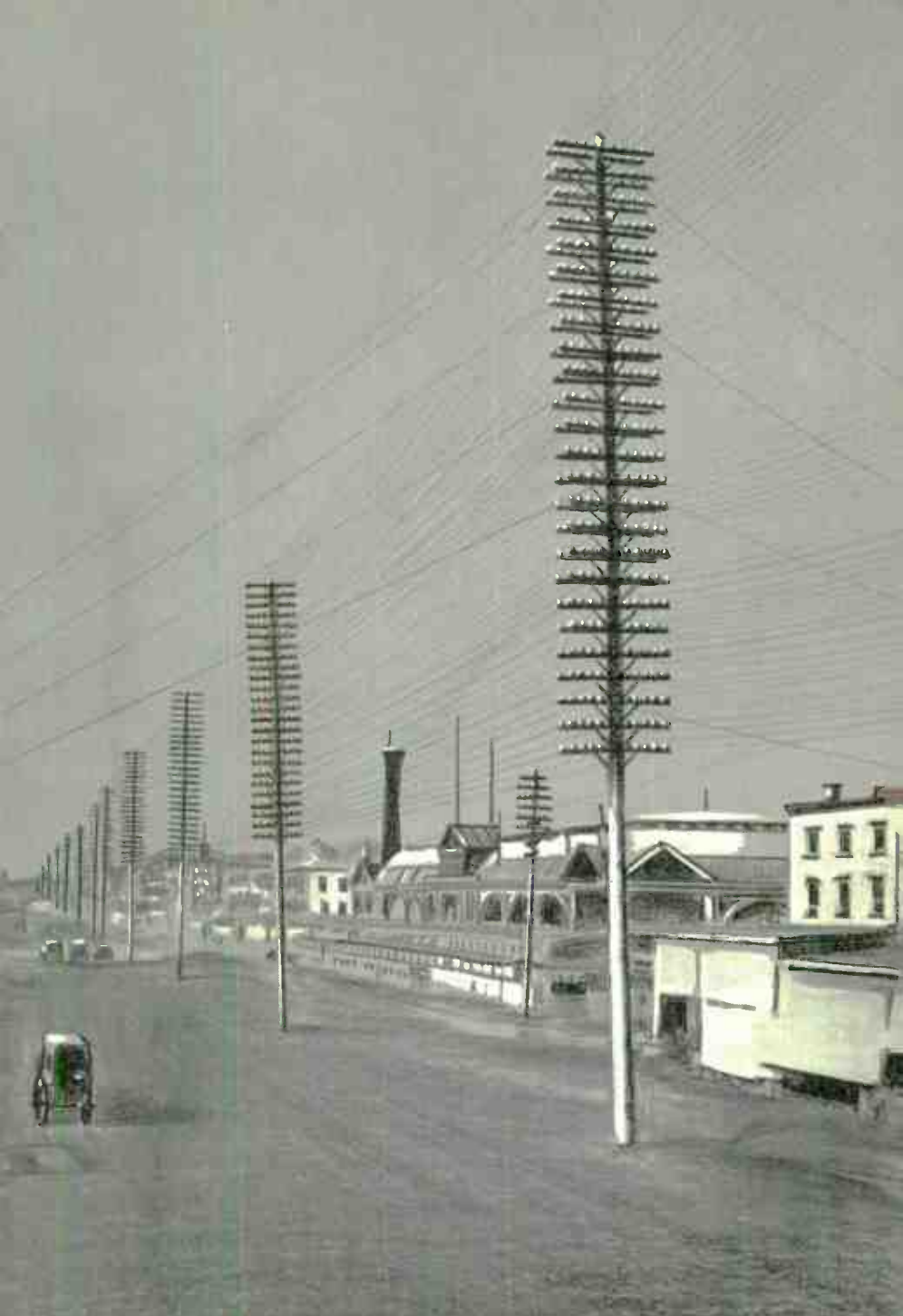


Figure 4. In the early days of telephony, a pair of wires was needed for each trunk circuit. Often, hundreds of wire pairs had to be strung up on a single pole line to provide the necessary number of trunk circuits between metropolitan areas.

World Radio History

45A 12-channel system for open-wire lines, developed in 1952, was the first multiplex system designed to take advantage of miniaturization techniques and modular plug-in components. This system led to the development of a complete family of miniaturized multiplex systems for all transmission mediums.

Today's modern solid-state frequency division multiplex systems, such as the type N3 and Lenkurt's type 46B, designed to operate over standard types of toll cable, provide up to 24 channels with a line frequency ranging from about 36 to 264 kc. The line frequencies are different for each direction of transmission. These modern transistorized systems provide economical service for intertoll and toll-connecting trunks, and other medium- and short-haul applications.

The development of multiplex systems for short-haul applications led to its use in exchange trunks and subscriber circuits. These short-haul systerms provided a simple and economical means of providing up to 10 or 20 channels over a single open wire or cable pair. Subscriber multiplex proved to be particularly suitable in rural or sparsely-populated areas where one rural line could provide adequate service to as many as 50 homes. Exchange multiplex systems, such as the Lenkurt type 81A and type X, provided economical, high quality trunks for Extended Area Service and toll-connecting applications.

To compete with the cost of loaded cable pairs used for exchange trunks shorter than 10 miles, the Bell System developed a multiplex system significantly different from the conventional frequency division systems used in the past. This system, designated type T1, is a time-division multiplex system which uses pulse-code modulation to provide 24 telephone channels over an exchange trunk cable.

High Capacity Systems

Most of the development of multiplex systems prior to World War II was directed toward increasing the efficiency of open-wire and multipair cable facilities which provided almost all of the telephone circuits. However, the useable bandwidth of open-wire and multipair cable circuits limits the multiplexed channels to a rather small number. The capacity of open wire systems is limited to about 16 channels, while the capacity of multipair cable systems is about 24 channels. However, multiplex systems could be made to operate with much greater bandwidths than those provided by conventional wire systems. All that was needed was some wideband transmission medium.

Modern wideband transmission mediums are provided by coaxial cable and microwave radio facilities which are capable of handling hundreds of channels. In about 1948, the Bell System completed a transcontinental coaxial cable transmission facility, designated type L1, to be used for television as well as to provide a large number of telephone channels. The L1 facility is capable of handling up to 600 single-sideband suppressed-carrier frequency division multiplex telephone channels, or one television channel. Later, a higher capacity coaxial cable, designated type L3, was developed. This facility is capable of handling 1860 multiplex telephone channels, or one television channel and 600 multiplex telephone channels. The Bell System is presently working on the type L4 coaxial cable which will have a capacity of about 32,000 multiplex telephone channels.

In addition to the L-type coaxial cable transmission facilities, the Bell System has developed two long-haul microwave radio relay systems, designated type TD-2 and type TH. The

World Radio History

TD-2 system, operating in the 4000 mc common-carrier band, has a capacity of 6000 multiplex channels, with a later version, the type TD-3, capable of handling up to 12,000 channels. The TH system, operating in the 6000 mc common-carrier band has a capacity of over 11,000 multiplex channels.

The single-sideband suppressed-carrier frequency division multiplex systems developed by the Bell System for use with the coaxial cable and microwave facilities have been designated type L. These systems are presently capable of providing up to 600 or up to 1860 multiplex telephone channels.

The type L multiplex systems combine the voice-frequency telephone circuits into 12-channel groups. A series of group modulation steps are then used to form up to fifty 12-channel groups (600 channels) into a baseband with a frequency range of 60 to 2788 kc, or up to 155 12-channel groups (1860 channels) into a baseband with a frequency range of 312 to 8284 kc. The Lenkurt type 46A multiplex system, developed for use with microwave radio, uses a modulation scheme compatible with that of the type L to multiplex up to 1200 channels.

These high channel capacity multiplex systems have provided an efficient and economical means of expanding the long-haul telephone plant to meet the great demand for more communications services. In addition, the development of these systems has resulted in considerable standardization, especially in group modulation schemes, thus permitting sub-groups of multiplex channels, derived from different systems, to be interconnected at multiplex frequency levels rather than at voice-frequency levels. This feature provides large savings in equipment costs and results in higher quality transmission. Also, these systems have greatly improved the performance standards and reliability of multiplex systems and have been instrumental in reducing the size of the equipment and in lowering the per-channel costs.

Conclusion

The telephone industry has grown at a significant pace since its beginning in 1876. This growth has been greatly effected by the development of multiplex systems which permit many speech signals to be transmitted simultaneously over a single open-wire, cable, or radio transmission facility. In recent years other industries have taken advantage of multiplexing to solve the problems of expanding their communications facilities. Among these other users are the railroads, pipeline companies, utilities, airlines, various government agencies, and the Armed Forces.

Today's modern multiplex systems are capable of handling not only speech signals, but many types of digital signals such as low and high speed data, which have added new dimensions to communications technology. In the near future, multiplexing will be providing channels for *waveguide* transmission facilities capable of handling perhaps 200,000 speech signals simultaneously. Perhaps further in the future, laser beams will be piped between major population centers carrying over one million multiplexed signals simultaneously.

Multiplexing has played a key role in promoting worldwide telephone communications, which has become essential to the efficiency and success of our society. The great communications systems that have emerged through the technological advances of multiplexing have indeed become a national asset.

World Radio History

World Radio History

the Lenkurt Demodulator

VOL. 7 NO. 12 DECEMBER, 1958

MULTIPLEXING AND MODULATION

In Carrier Telephone Systems
Part I

Multiplexing is the means by which a number of circuits may be combined for transmission over a common transmission medium. Modulation is the process by which multiplexing may be effected. Although a number of different types of modulation are possible, only two basic multiplexing methods—frequency division and time division—are in common use. For the two different multiplexing methods, any one of several types of modulation may be used.

In this article, frequency-division multiplexing is defined and its use with the various standard methods of amplitude modulation is discussed. Frequency modulation (another form of modulation which may be used in frequency-division multiplexing), time-division multiplexing and modulation methods used with time division are considered in a subsequent article.

Even at the time of the invention of the telephone, the advantages of transmitting several information circuits over a common medium were recognized. Since then much effort has been expended in developing multiplexing techniques. In present-day practical systems one of two basically different multiplexing methods is employed—frequency division or time division.

Frequency Division

Frequency division is so called because each multiplexed circuit is preassigned a specific frequency band (channel) for transmission of its information. In this way, individual circuits may be combined on a facility and simultaneously transmitted over a common transmitting medium. The concept of multiplexing is shown diagrammatically in Figure 1; and the technique for frequency-division multiplexing (FDM) is shown in Figure 2.

In combining a number of circuits for frequency-division multiplexing, the principal requirement is that the technique include a method of translating the original circuit frequencies into the frequency band assigned for transmission. Any one of a number of modulation processes may be used for this

World Radio History

purpose. Forms of either amplitude modulation or angle modulation are often used, with amplitude modulation being the most common.

Modulation

In telecommunications, modulation is used in many different applications. For example, the conversion of sound energy into electrical energy by a telephone transmitter is a form of modulation. However, in frequency-division multiplexing, modulation is the word used to describe the process in which an electrical wave acts to change a characteristic—for example amplitude or frequency—of a second wave. The resulting wave then contains properties of both original waves.

In the process of modulation, there are three basic waves—modulating wave, carrier wave and modulated wave. The modulating wave may be made up of any type of information which has been converted into electrical impulses. In carrier telephony, the modulating wave is the complex electrical wave form of speech obtained from the telephone transmitter.

The carrier wave is normally a single-frequency electrical signal that has been derived from a frequency generator (oscillator). The frequency of the carrier wave establishes the circuit position in the available frequency spectrum.

The modulated wave is the resultant output wave of the modulation process;

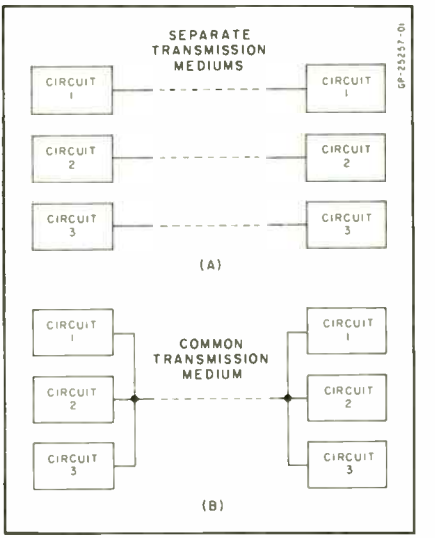


Fig. 1. Illustrating the more effective utilization of a transmission medium by multiplexing. (a) Without multiplexing separate facilities are required for each circuit. (b) With multiplexing the three circuits may use one transmission path.

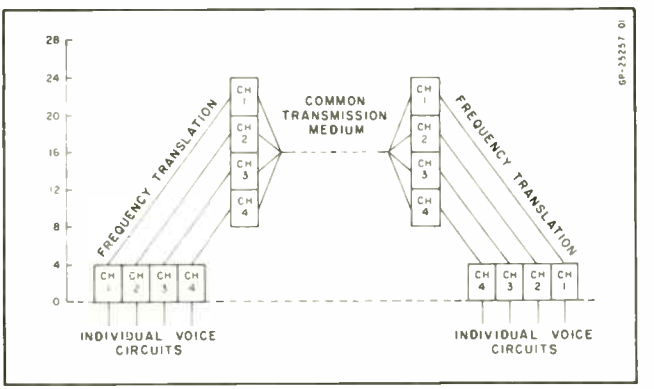


Fig. 2. In frequency-division multiplex systems, each circuit (channel) is translated to its own position in the frequency spectrum before being applied to a common transmission medium.

World Radio History

it contains the carrier wave modified by the action of the modulating wave.

Amplitude Modulation

In amplitude modulation, the amplitude of the carrier wave is controlled by the modulating wave. As shown in Figure 3, the resultant modulated wave has the same frequency as the carrier, but the carrier wave amplitude varies in direct relation to the modulating wave. In fact, the curves through both the positive and negative peaks of the modulated wave are identical to the modulating wave, and they are called the wave envelopes. The modulation factor, m, is a measure of the degree of modulation. For a sinusoidal variation as shown in Figure 3, the modulation factor—normally called the *modulation index* and sometimes the *degree of modulation*—is equal to the peak amplitude of the envelope minus the amplitude of the unmodulated carrier divided by the amplitude of the unmodulated carrier. For more complex signals, the modulation index is more difficult to determine since it will vary from one instant to another. In any case, the modulation index is the fractional extent by which the modulation varies the amplitude of the carrier, and is often expressed as percent modulation by multiplying the modulation index by 100.

From Figure 3 it is apparent that the maximum amount that the carrier amplitude can be varied, without loss of signal, is equal to the carrier amplitude. When this occurs, 100 percent modulation is obtained.

An analysis of the modulated wave resulting from amplitude modulation shows that the modulated wave consists essentially of the carrier wave and frequencies above and below the carrier wave. These side frequencies are separated from the carrier by a frequency equal to that of the modulating wave. Where a complex modulating wave—such as speech or music—is used, the side frequencies above and below the carrier each consist of a band (sideband) of frequencies. A sideband includes all of the frequency components of the modulating wave.

Three important factors in the use of amplitude modulation are derived from an analysis of the modulated wave: (1) the sidebands obtained from a complex wave each have the same bandwidth as the original modulating

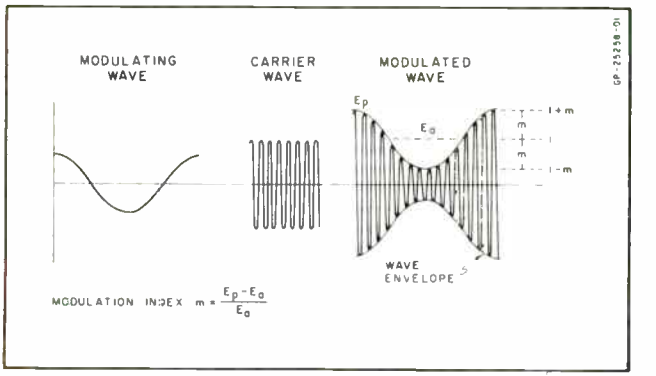


Fig. 3. Amplitude modulation. The amplitude of the carrier is varied by the modulating wave. The frequency of the modulated wave envelope is the same as the modulating wave.

World Radio History

wave; (2) the same intelligence is contained in each sideband; (3) the frequencies in the upper sideband have the same relative relationship as the modulating wave, but those in the lower sideband have an inverse relationship. Not so apparent is the fact that the power distribution in the sidebands is directly related to the distribution of power in the modulating wave.

Double Sideband

As a result of amplitude modulation, the frequency band of the modulating wave is translated to a different position in the frequency spectrum. It is this ability of translation during the modulation process which is used in combining circuits for frequency-division multiplexing. If both sidebands and the carrier are used, the multiplexing technique is called double-sideband amplitude modulation (DSB-AM, or normally AM).

Although AM is quite commonly used in radio broadcasting, a disadvantage of this method in carrier telephone multiplexing is the magnitude of power in the carrier in relation to the sideband (information) power. Even with 100 percent modulation, the power in each sideband is only ¼ of that in the carrier. Since the power in the sidebands is proportional to the square of the modulation index, for low modulation levels the sideband power will become only a fraction of the carrier power.

In multiplexing, this is quite important because of the number of channels that are involved. The various parts of the system common to more than one channel must be designed to be capable of handling a large amount of power that is not useful in the transmission of information. In addition, the sideband-to-noise power is relatively low.

For these reasons, it is common in double-sideband multiplexing either to

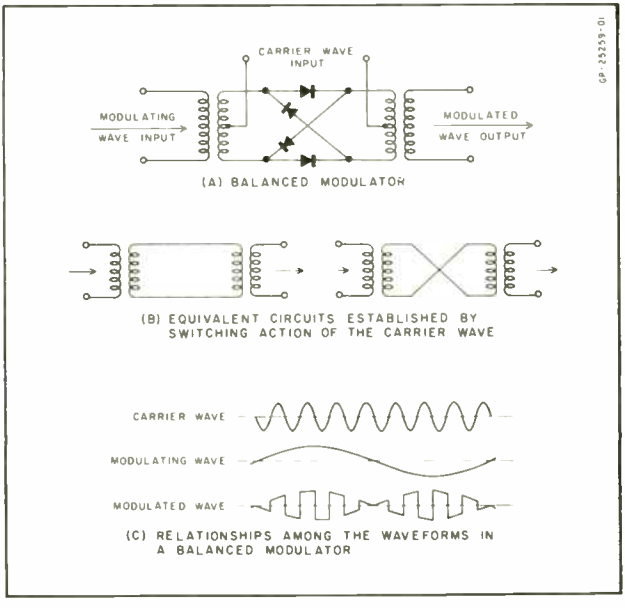


Fig. 4. A balanced modulator may be used where it is desired to suppress the carrier. The modulated wave then contains the upper and lower sidebands.

World Radio History

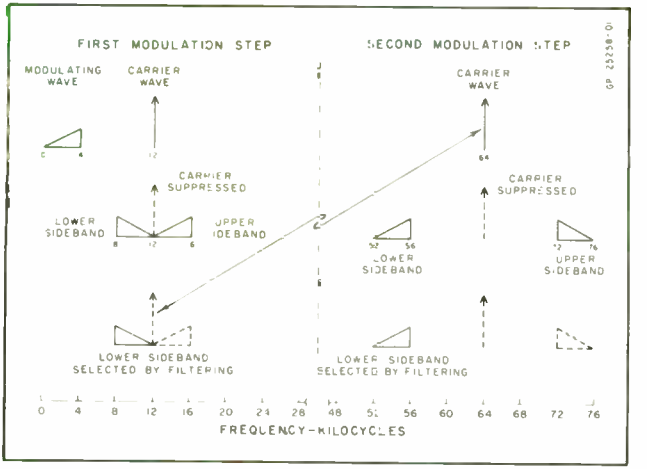


Fig. 5. A number of modulation steps are sometimes required to position a channel in the carrier frequency spectrum. Where SSB-SC is used, the carrier and unwanted sideband are removed at each step of modulation.

suppress the carrier (DSB-SC) or to transmit the carrier at a relatively low level. Where the carrier is suppressed, some method of deriving a carrier frequency at the receiving terminal is necessary. This may be done either by separately generating the carrier frequency, deriving the carrier frequency from the transmitted sidebands, or by transmitting a separate tone from which the demodulating frequencies may be derived.

After the carrier is suppressed, it is possible to increase the sideband power and still keep the power handling capacity of common equipment units below that which would have been required if the carrier were transmitted. This increases the operating range of the equipment considerably.

Some of the advantages of DSB-SC multiplexing are: (1) the relative simplicity of the modulation process; (2) relatively lenient filter requirements, particularly in the transmitting direction; and, (3) the relative immunity to distortion which may occur in transmission. A particular disadvantage is the bandwidth used for the information transmitted.

Single Sideband

In amplitude modulation, the two sidebands produced each carry the same information. If only one sideband were transmitted, the required bandwidth could be reduced at least in half.

The frequency separation between the carrier and each sideband is equal to the frequency of the modulating wave. Where a complex modulating wave such as speech or music is used, the sideband frequencies may differ from the carrier frequency by a few cycles per second up to several kilocycles. Unless the low frequencies are restricted to above about 200 cycles, the problem of suppressing the carrier and the unwanted sideband, without undue distortion of the wanted sideband, becomes very difficult. The loss of the low frequencies has very little effect on the quality or fidelity of speech, and is usually tolerated to a very high degree in music. For this reason single-sideband suppressed-carrier mul-

World Radio History

tiplexing is suitable for carrier telephony, and in fact has been the most commonly used method.

The principal advantage of SSB-SC is the efficient bandwidth utilization in the transmission of information. In toll telephone applications, the increased channel capacity that can be obtained in a limited bandwidth far outweighs the necessary complexity of equipment design. In addition to bandwidth conservation, the reduced bandwidth improves the signal-to-noise ratio as compared to a DSB-SC system. However, SSB-SC systems suffer from an inherent delay limit which is a result of filtering out one sideband and the carrier; and, an SSB-SC system cannot be used directly for pulse transmission because of the low frequency limit.

World Radio History

the Lenkurt® Demodulator

VOL. 8 NO. 1 JANUARY, 1959

MULTIPLEXING AND MODULATION

In Carrier Telephone Systems

Part II

The use of amplitude modulation in frequency-division multiplexing was discussed in the first of this two-part article. Because of its widespread usage in toll applications, single-sideband suppressed carrier has become almost synonymous with frequency-division multiplexing. However, current applications of carrier—such as rural subscriber and exchange systems—have made other types of modulation and multiplexing attractive.

In this article, frequency modulation and its use in frequency-division multiplexing is considered first. Then, time-division multiplexing and some of the modulation methods used in this type of multiplexing are discussed.

Although frequency-division multiplexing is often associated with amplitude modulation, the various types of angle modulation may also be used in frequency-division multiplexing applications. Angle modulation is the general term used to describe any form of modulation in which the frequency or phase of a sinusoidal carrier wave is controlled by the modulating wave. Frequency and phase modulation are the two types of angle modulation most commonly used. While these types of angle modulation are somewhat different, they are also closely interrelated. In fact, both are often used in a single modulation system. Because the basic considerations are similar, the following discussion will be restricted to frequency modulation.

Frequency Modulation

Frequency modulation (FM) is the process in which amplitude changes of the modulating wave are used to vary the instantaneous frequency of the carrier wave from its unmodulated value. An example of the action of the modulating wave on the carrier wave to produce a frequency-modulated output wave is shown in Figure 1.

The magnitude of frequency change for a given amplitude of the modulating signal is called *frequency shift*, fre-

World Radio History

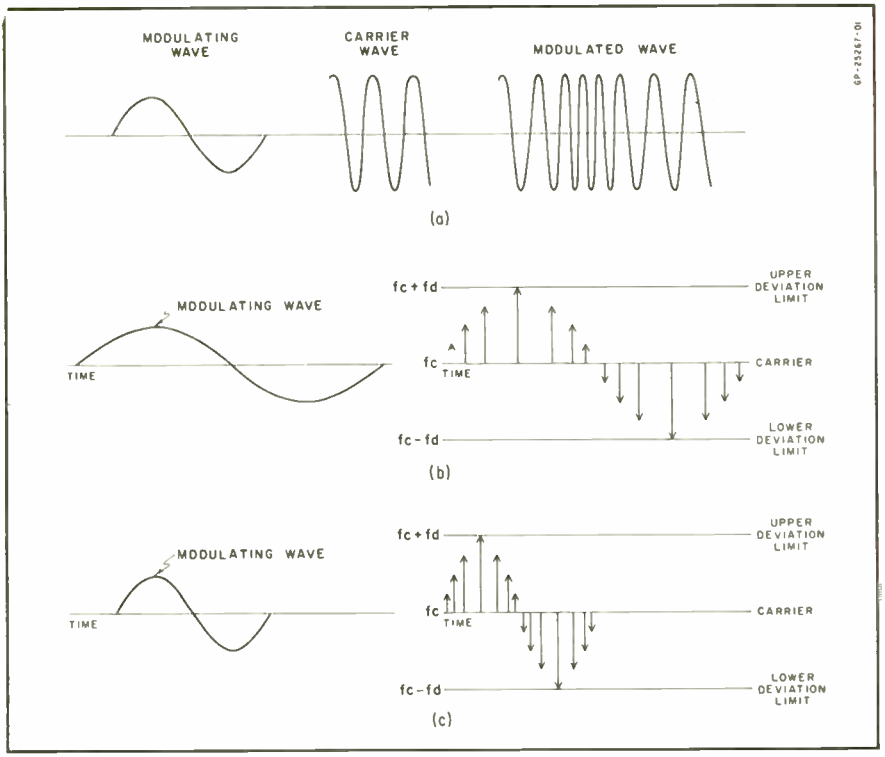


Fig. 1. Frequency Modulation. Both (a) and (b) show how the carrier frequency varies with the amplitude of the modulating wave. A comparison of (b) and (c) shows how the deviation rate is controlled by the frequency of the modulating wave.

quency swing or frequency deviation. However, the last two terms are also often used to define other properties of the change in frequency. *Frequency swing* is normally reserved to denote the maximum frequency shift that occurs when a sinusoidal modulating wave is employed; *frequency deviation* is the maximum value of frequency shift permitted by equipment design, and is also called *peak deviation*. The term frequency deviation is also often used to denote the instantaneous difference between the instantaneous frequency of the modulated wave and the carrier frequency.

While frequency shift is controlled by the amplitude of the modulating wave, the rate at which the carrier frequency is shifted, called *deviation rate*, is controlled by the frequency of the modulating wave. If a carrier of 1 megacycle is modulated by a 1000-cycle sinusoidal modulating signal, the frequency swing that will occur will depend upon the amplitude of the modulating wave. If the frequency swing is 500 cycles for a given amplitude, the carrier will swing between 1,000,500 and 999,500 cycles at a rate of 1000 cycles per second. If a signal of a different frequency but same amplitude is

World Radio History

used, the frequency swing will still be ±500 cycles from the carrier, but the deviation rate will be equal to the frequency of the modulating wave. This is shown diagrammatically in Figure 1.

Bandwidth

Since the frequency shift is dependent upon the amplitude of the modulating wave, it would appear that the bandwidth required for FM transmission could be made considerably less than that of the modulating wave. However, the individual cycles of a modulated wave obtained from an FM modulator are not sinusoidal because of the instantaneous variations in frequency which occur during modulation. An analysis of the modulated wave shows that this complex wave contains a large number of sidebands rather than the two normally associated with amplitude modulation.

Where a single sinusoidal modulating wave is used, the spectrum of the modulated wave is symmetrical with respect to the carrier frequency. In this case, the sideband frequencies are displaced from the carrier by integral multiples of the modulating frequency. For example, a 1000-cycle modulating wave would produce a first-order pair of sidebands that differ from the carrier frequency by 1000 cycles, a pair of second-order sidebands located at 2000 cycles on either side of the carrier as well as higher order sidebands. This is illustrated graphically in Figure 2.

If a more complex modulating wave —such as more than one sinusoidal signal or speech— is used, the frequency spectrum of the modulated wave becomes very complicated. The sideband frequencies present include not only those that would be obtained with each modulation frequency acting separately, but also various combination frequencies. However, although complex modulation greatly increases the number of frequency components present in the frequency-modulated wave, it does not widen the bandwidth occupied by the high energy portion of the wave.

Although the total bandwidth of a frequency-modulated wave is quite large, the higher-order sidebands often contain only a small portion of the total wave energy. In these cases, the actual bandwidth can therefore be reduced considerably without introducing un-

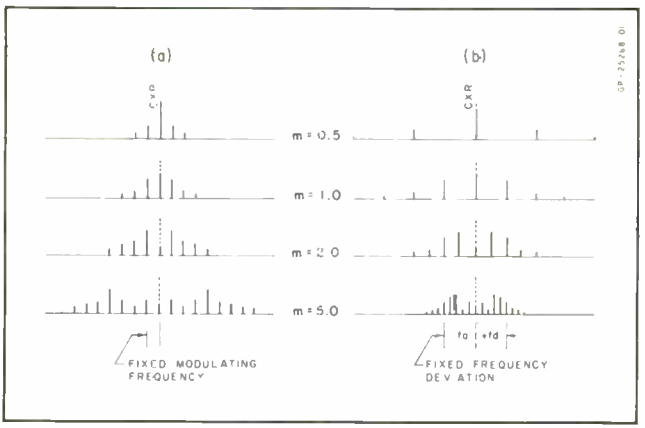


Fig. 2. Energy distribution in an FM wave with a sinusoidal modulating signal: (a) distribution as frequency deviation is increased with modulating frequency constant; (b) distribution as modulating frequency is decreased with frequency deviation constant.

World Radio History

due distortion. The energy distribution depends upon the amplitude of the different frequency components. Component amplitudes, in turn, are related to the modulation index, and may be calculated with the aid of a table of Bessel's functions. The results of a number of such calculations are shown graphically in Figure 3, from which the bandwidth for a variety of conditions may be determined. Bandwidths determined from Figure 3 will contain all but about 1 per cent of the energy in the modulated wave. The loss of 1 percent of the energy will cause distortion. Although the distortion introduced can be tolerated in some applications, in others the distortion requirements are more severe, and a greater proportion of the energy must be included. For these cases, a detailed analysis is made, and Figure 3 is not used.

With reference to Figure 3, some useful rules can be derived for determining the approximate bandwidth required for transmission of a frequency-modulated wave. When the modulation index is greater than 1, the bandwidth is equal to twice the sum of the frequency deviation plus the modulating frequency. Although not shown in Figure 3, as the modulation index decreases to values of 0.5 and below, the bandwidth becomes essentially equal to twice the modulating frequency. In this last case, the bandwidth is the same as for an amplitude modulated wave.

Noise Advantage

Because both methods are used in frequency division multiplexing, the advantages and disadvantages of frequency modulation are normally expressed in terms of similar characteristics in amplitude modulation. On this basis, the chief advantage of FM is in its ability to exchange bandwidth occupancy in the transmission medium for improved noise performance.

The noise-power reduction advantage of FM over AM for random noise is often given as $R = 3 f_d^2/B^2$, where f_d is the frequency deviation and B is the output bandwidth of the receiver. (B is equal to the highest modulating frequency.) This advantage is also expressed in terms of FM advantage *(db of quieting)* and may be written as $db = 10 \text{ Log } 3f_d^2/B^2$. For example, for a frequency deviation of 3 kc and a 3-kc modulating signal, a random-noise reduction advantage of about 4.8 db is obtained. In both equations, 100 percent amplitude modulation is assumed and the comparison is made for average output power.

On the basis of the same peak power at the transmitter, SSB-SC has a noise advantage over AM of about 8:1 (approximately 9 db). However, when peak power is used as the basis of comparison, FM has an additional 4:1 advantage (total of approximately 15 db) over AM than that given by the above equation. Therefore, on a peak power basis, the noise advantage of FM as compared to SSB-SC is $R = 3f_d^2/2B^2$.

In addition to noise advantage, FM exhibits a characteristic often called *capture effect*. Where two signals in the same frequency band are available at the receiver, the one appearing at the higher level is accepted to the near exclusion of the other.

Improvement Threshold

The noise advantage of FM is obtained for normal signal and noise lev-

World Radio History

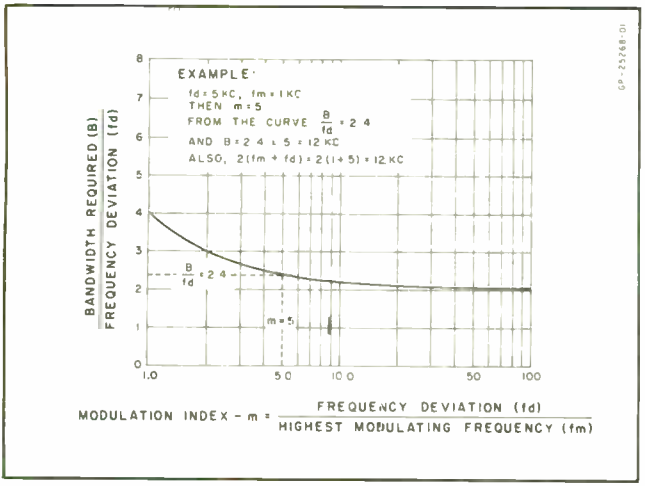


Fig. 3. From the above curve the bandwidth for different modulation indexes of an FM system can be determined. All sideband components except those having less than 1 percent of the wave energy are included.

els at the input to the receiver. And this advantage increases as the frequency deviation (modulation index) is increased. However, as the peak signal level is decreased to that of the peak noise level, there is a rather sharp transition between good and poor signal-to-noise ratios. The point at which this transition occurs is often called the *improvement threshold* (sometimes shortened to threshold).

Although the noise advantage increases as the modulation index is increased, the corresponding increase in bandwidth increases the noise. For large bandwidths, the threshold becomes more critical and is reached at higher signal levels. The optimum value of modulation index is thus a compromise between service range and noise advantage.

Other Features

Other features of FM make its application to frequency-division multiplexing attractive where bandwidth is not a critically limiting factor. These are: (1) separate regulation is not necessary since in a properly designed FM receiver, the output signal level is insensitive to input signal level variations above threshold level; and, (2) synchronization is not a problem because of the detection method.

Time-Division Multiplexing

Although less well known because of its relatively limited usage, time-division multiplexing is the most direct and is basically the simplest multiplexing method. In this method, the circuits to be transmitted over a common transmission medium need not be connected together as in the frequency division case, but are arranged so that each circuit is successively connected to the transmission medium for a short time interval. An example of a technique for accomplishing this purpose is shown in Figure 4.

A commutator or other type of switching device may be used, and in the earlier systems motor-driven commutators were employed. Most present-day multiplexing equipment uses electronic switching and gating techniques.

World Radio History

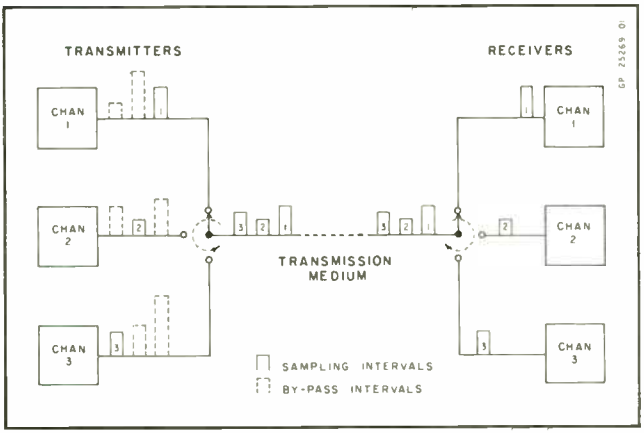


Fig. 4. Time-Division Multiplex. Samples of the input wave are transmitted by successively connecting each circuit to the transmission medium.

Regardless of the switching method, it is apparent that for each circuit the input wave is broken up into a series of pulses called samples. This is an example of pulse-amplitude modulation, since the amplitude of each sample is directly proportional to the instantaneous amplitude of the modulating wave.

Sampling Rate

Because in time division the modulating wave is repetitively sampled, the questions of how large must the samples be, and how often samples must be taken to retain the intelligence naturally arise. It has been found that the time duration of the sample is not critical, and can be reduced as much as required without appreciably degrading the information. However, the samples must be taken at a rate that is at least twice the highest frequency appearing in the modulating wave. For telephone voice channels, the sampling rate commonly used is 8,000 cycles per second. Or, each circuit is sampled once every 125 microseconds. The 125-microsecond interval between successive samples of one channel may be used by other channels of the system. Theoretically the number of channels that may be obtained is very high.

Bandwidth

However, as the number of channels is increased, the bandwidth required for transmission increases rapidly. The reason for this is that the pulse train spectrum is made up of a fundamental frequency that is equal to the sampling rate (8 kc) and its harmonics, all of which have an upper and lower sideband produced by the modulation process.

The actual bandwidth depends upon a number of factors which must be determined during the design of the system. As an example of the bandwidth required for present day systems, a pulse-amplitude modulated (see Figure 5b) system of 30 channels requires a bandwidth of approximately 1 megacycle. This is about 10 times the bandwidth of the intelligence (4 kc x 30 = 120 kc), and severely restricts the application of this type of modulation.

Much work has been expended on the reduction of the bandwidth required in time-division systems, and by

World Radio History

using pulse-coding techniques and appropriate filtering of the transmitted signal the bandwidth can be reduced to approach that required for the intelligence. However, the equipment necessary to accomplish this complicates the terminal design, and some of the advantage of simplicity of the multiplexing method is lost.

Synchronization

From the elementary diagram of Figure 4, it is apparent that synchronization is relatively critical in time-division multiplexing, and in time division systems a means of maintaining synchronization is normally provided. A common method of doing this is to transmit synchronizing pulses —often called marker pulses— at the beginning of each frame. A frame is the interval occupied by one complete set of pulses, and in this case the frame is made up of a marker pulse and one pulse from each channel. This is shown diagrammatically in Figure 5.

PTM

Pulse-time modulation (PTM) is modulation in which values of the instantaneous samples of the modulating wave are used to vary the time of occurrence of some parameter of a pulse carrier. Pulse-duration modulation and pulse-position modulation are particular forms of pulse-time modulation.

Pulse-duration modulation (PDM), sometimes designated pulse-length modulation or pulse-width modulation, is modulation of a pulse carrier in which the value of each instantaneous sample of a modulating wave is used to vary the duration of a particular pulse. This is shown in Figure 5. The modulating wave may vary the time of occurrence of the leading edge, the trailing edge or both edges of the pulse.

In pulse-position modulation, (PPM) the value of each instantaneous sample of a modulating wave is used to vary the position in time of a pulse relative to its unmodulated position. An example of PPM is shown in Figure 5.

Like FM, PTM has the property that, by using extra bandwidth, some of the overall performance characteristics of a PTM system can be improved, provided that the peak interference is less than the peak signal. A feature of PPM is that, for fixed average power out of the transmitter, the peak power can be increased as the pulse duration is reduced. However, much of the advantage is lost where a large number of channels are considered.

PCM

Pulse-code modulation (PCM) is a relatively new innovation in pulse systems. As in the earlier pulse systems, each modulating wave is sampled periodically at a rate somewhat in excess of twice the highest frequency component in the modulating wave. Unlike the continuously variable samples used in PAM, PDM, and PPM, in PCM the samples are quantized into discrete steps. The individual steps may be alike or they may vary depending upon the characteristic properties of the modulating wave.

In addition, each quantized sample is assigned a particular code pattern, the code pattern assigned being uniquely related to the magnitude of the quantized sample. This gives rise to various possible patterns of coded pulses. For

World Radio History

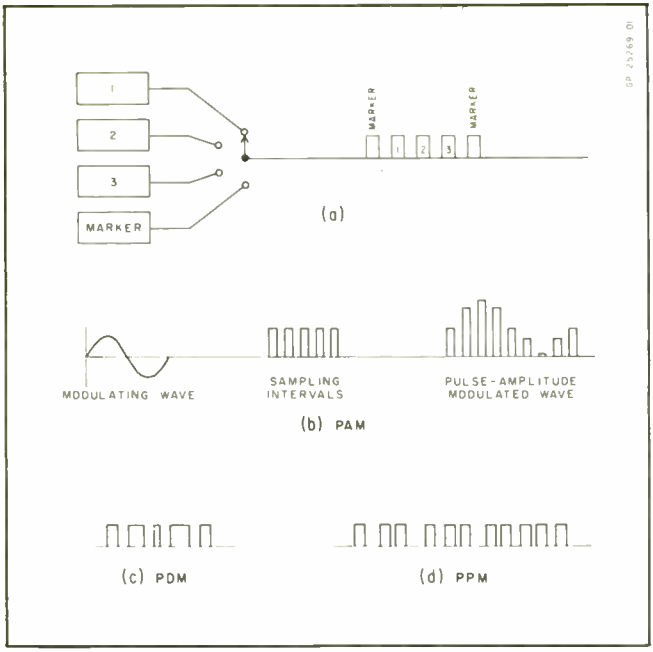


Fig. 5. Pulse Modulation. (a) Showing the position of the marker pulses used in time-division multiplex systems. (b) Pulse-Amplitude Modulation. (c) Pulse-Duration Modulation. (d) Pulse-Position Modulation.

example, if amplitude is the parameter quantized, there may be patterns of on or off pulses; or patterns of three-value code elements, namely, values of +S, O, and –S, as in the familiar case of submarine cable telegraphy; or, in general, code patterns which contain a number of code elements, and in which each code element will assume one of several distinct amplitude values. At the receiving end, each code pattern is identified, decoded, and used to produce a voltage proportional to the original quantized sample. From a succession of such samples, the original wave is approximated. By making each quantum step sufficiently small, theoretically the original wave may be approximated as closely as desired.

Pulse-code modulation has two outstanding properties: first, it affords marked freedom from noise and interference; and, second, it permits repeating the signals again and again without significant distortion. For example, consider the code patterns formed by on or off pulses. At each regenerative repeater, as long as each incoming pulse can be correctly identified in the presence of accumulated noise, interference, and distortion, a new and correct code pattern can be generated and started out afresh to the next repeater.

Conclusion

The inherent circuit capacity of a transmission medium can be filled using either frequency or time division techniques. The choice of multiplexing method and the type of modulation employed depends upon such factors as: compatibility with existing systems; loss over which the system may be effectively operated; bandwidth, signal-

World Radio History

to-noise performance; interference-rejection capabilities; distortion characteristics; and, the required stability of the frequency generators. Each of the types of multiplexing and modulation discussed have inherent advantages and disadvantages, and in a number of cases any one of several methods might appear to have about equal capabilities for a given application. The final choice might then depend upon the relative complexity of the system.

In practice, toll carrier systems have been primarily made up of SSB-SC frequency-division multiplex systems. For these applications, this method is economically compatible and the total bandwidth required for the intelligence transmitted is held to a minimum in the present state of the art. Other types of modulation in frequency-division as well as time-division multiplex appear attractive for other carrier telephone applications.

World Radio History

World Radio History

the Lenkurt Demodulator

VOL. II, NO. 11 NOVEMBER, 1962

TIME-DIVISION MULTIPLEX—

New Promise for Old Technique?

Time-division multiplex, oldest and simplest method of deriving additional communications channels from a transmission medium, has never been really practicable before now. Frequency-division multiplex, or "carrier," although more complex, was originally easier to achieve than time-division because the particular techniques required were further developed than those required for time-division multiplex. Progress in semiconductor technology and computer techniques has now largely removed this disparity in knowledge so that time-division systems may now be economically feasible in certain applications. This is the first of several articles which discuss the characteristics of such systems.

Conventional frequency-division multiplex techniques have developed from our ability to produce electrical filters which can separate and, in a sense, create *bands* of frequencies. By using each frequency band as a separate channel for transmitting information, it is possible to transmit many individual channels simultaneously over a single transmission medium. The bandwidth required is the sum of the individual channel bandwidths, plus a small amount between channels.

Much the same result is obtained with time-division systems. Many individual channels share the transmission medium by "taking turns," each being connected to the line very briefly, then replaced by the next. This is repeated again and again so swiftly that there is no loss of message intelligence in any of the channels. If the time during which each channel is connected to the line is kept very short, many channels can share the transmission facility.

At the receiver, the same process occurs in reverse. Some kind of signal distributor or commutator arrangement is required to "sort" the samples as they arrive in sequence, and distribute them to the appropriate lines at the

World Radio History

proper instants. This requires precise synchronism between transmitter and receiver. If synchronism fails, all channels are garbled and lost. Significantly, it was the lack of ability to maintain synchronization that prevented the general adoption of time-division multiplexing during the last century, shortly after the invention of telegraphy, and later, telephony. This left the field open for the frequency-division techniques which have become nearly universal.

By now, most technical obstacles to time-division multiplexing have been overcome. The development of cheap but efficient solid-state devices such as transistors, plus highly refined electronic switching techniques, has eliminated the relative economic disadvantages that time-division previously suffered. These, plus certain other advances have stimulated new interest in time-division multiplexing for communications systems.

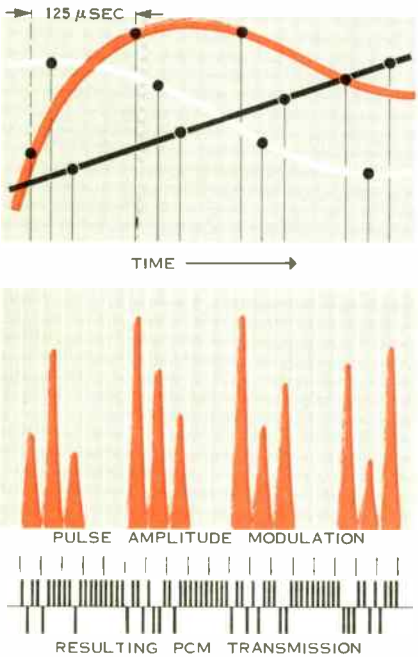


Figure 1. Two samples per cycle of highest frequency adequately define each channel. If samples are brief, many channels can share the common line.

The Role of PCM

An important factor behind the growing interest in time-division multiplex is its natural suitability for use with electronic time-division switching and pulse code modulation (PCM). Electronic switching is based on the same digital techniques used in modern electronic computers. PCM is also a digital system, able to transmit any sort of signal—even television—in the form of coded binary pulses. This at once yields several advantages. Unlike conventional analog transmission methods (in which a current or carrier wave is varied in a manner analogous to the original message), the message is sampled periodically and the values observed are *represented* by a coded arrangement of several pulses. Each separate signal value has a unique arrangement of pulses. Thus, only the presence or absence of pulses — not their shape —determine the received message and its quality.

This is very similar to conventional teletypewriter communication, in which the transmitted characters are represented by various combinations of a five-pulse code. Regardless of how badly the code pulses may be distorted or degraded in transmission, the sharpness or clarity of the *characters* reproduced at the receiving printer are obviously not changed or altered in any way. Distortion of the transmitted pulses merely increases the chances of a mistake in interpreting the code and the printing of a wrong character.

World Radio History

In addition to providing improved transmission, the combination of these digital techniques can result in substantial economies. By integrating the equipment used for switching and multiplexing, equipment cost may be much lower than if the two techniques were used separately.

It should be pointed out that there are many types of pulse modulation. Almost without exception, however, PCM or related approaches (which we will class as "PCM" in this issue) are the only ones now seriously considered for time-division multiplexing systems. This stems from the great efficiency of PCM in overcoming interference and noise. Even more important, PCM permits the use of regenerative repeaters.

Regenerative repeaters detect the presence or absence of "oid," distorted pulses and replace them with perfect new ones. The regenerative repeaters must be spaced closely enough to correctly identify the incoming distorted pulses, and be able to send out new pulses which precisely match the original pulse stream. When done perfectly, it is theoretically possible to transmit messages for unlimited distances without degradation. This would allow uniformly high transmission quality regardless of distance; a call across a continent would be as clear and distinct as a call next door. In practice, small timing errors tend to add up as the number of repeaters increases, eventually placing a limit on system length.

Because of the very close association between time-division multiplexing and PCM in system planning, and because many of the problems of such systems stem from the PCM, subsequent references in this article may be only to "PCM." It should be understood that time-division multiplexing is assumed, even though not always stated.

Bandwidth versus Noise

Pulse code modulation requires more bandwidth than amplitude modulation, but it uses this bandwidth far more efficiently in overcoming noise and interference than almost any other modulation method. For instance, frequency modulation (FM) also trades bandwidth for noise improvement, but rather slowly. With FM, the signal-to-noise ratio improvement (in db) is proportional only to the *logarithm* of the increase in bandwidth. Thus, to improve the signal-to-noise ratio by 10 db requires bandwidth ten times the original. By contrast, the PCM improvement in db is in *direct* proportion to the increase in bandwidth; a PCM signal requiring ten times the bandwidth of the original signal will yield a signal-to-noise ratio of 70 db. Note: this signal quality is inherent in the transmitted signal and is not the result of noise encountered in transmission, so long as the signal remains above the noise threshold of the system.

In order to reconstruct the original wave, at least two amplitude samples must be transmitted for every cycle of the highest frequency in the original waveform. Therefore, good reproduction of 4000-cycle telephone channels requires that 8000 samples per second be transmitted. A continuous waveform has an infinite number of discrete amplitude values. In order to represent the waveform by groups of pulses, it is necessary to limit the number of points which represent the wave to a convenient number which can be handled by the code. This is called "quantizing" the wave and is a form of approximation. The random differences between the actual waveform and the quantized approximation results in "quantizing noise" being introduced into the transmission. With error-free

World Radio History

transmission, this is the only source of noise in the system, and yields the signal quality defined above. Quantizing noise may be minimized by using as many steps as possible, since this reduces the difference between the original signal and its coded approximation (see Figure 3).

This, however, increases the number of code pulses which must be transmitted. The number of steps is determined by the number of binary pulses or digits in each code group: number of steps $= 2^n$, where n is the number of digits in the code. Thus a five-digit code yields 32 amplitude steps, while a seven-digit code provides 128 steps. The latter reduces the quantizing error considerably. The bandwidth required is directly proportional to the number of digits; a seven-digit code requires seven times as much bandwidth as a binary (single-digit) code, but improves the signal-to-noise ratio 36 db. (The signal-to-quantizing noise ratio in db can be calculated from the expression $S/N = 10.8 + 6n$).

In view of the requirement for so much greater bandwidth than amplitude-modulated or single-sideband systems, one might conclude that PCM is totally unsuited for practical communications system inasmuch as transmission bandwidth is very costly. In many applications this is true. However, several studies have shown that the noise advantage of PCM can, under some circumstances, be used to "buy back" frequency spectrum. For instance, it might be possible to overlap PCM transmissions in frequency. Although this would be expected to result in a certain amount of interference, the systems could exhibit a very high tolerance to their mutual interference, depending on the amount of overlap.

New sources of bandwidth are being made available by various technological advances. Rapid progress in the development of lasers (optical masers) confirms the high hopes and optimistic predictions about the possibility of using light as a wide-band communications medium (see DEMODULATOR, June, 1961). Of more immediate importance, a certain propagation mode in circular waveguide actually reverses the normal loss characteristics of waveguide, resulting in *less loss as frequency increases*. This also could make almost unlimited bandwidth available for techniques such as PCM which require great bandwidth. Not all of the practical difficulties in using this type of transmission have been overcome, however. Bends, junctions, and imperfections in the waveguide have a tendency to cause the energy to be transformed from the desired low-loss mode to other modes, any of which cause increasing loss at higher frequencies.

The Practical Situation

Assuming that transmission difficulties can be overcome, is the introduction of a new approach worthwhile? In the United States alone, nearly $50 billion is invested in communications, and a similar investment exists elsewhere. Only very substantial benefits could justify radical, large-scale changes in the basic system.

The desired objective is to achieve a nation-wide or even global communications network in which any telephone or station can instantly dial any other; transmission quality would be uniformly excellent, and the influence of distance would seem to vanish. Since the network would employ digital pulses for most transmission, data could be accommodated as easily as speech. None of these objectives are completely realizable or economically practicable with the conventional tech-

World Radio History

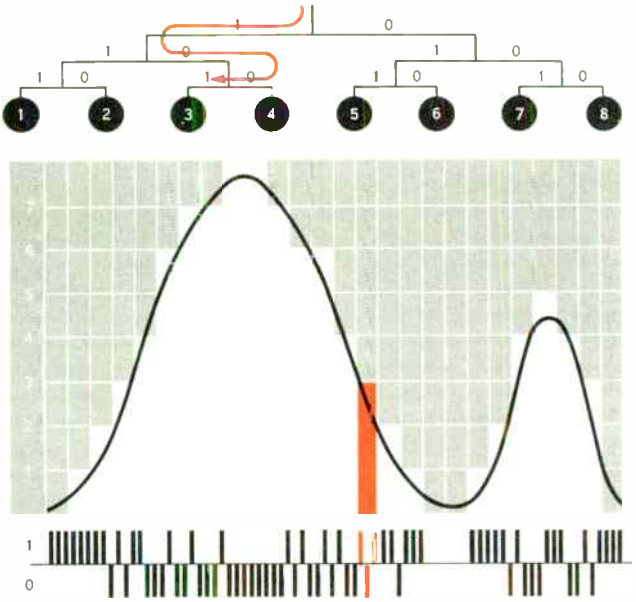


Figure 2. How a waveform can be quantized and converted into code pulses. Bandwidth restrictions limit the number of digits per sample, thus establishing maximum number of quantizing levels.

niques presently in use. However, it remains to be proved whether PCM transmission systems can successfully eliminate some of the technical and economic difficulties of present methods.

Even if a radical change is found to be desirable, it obviously cannot be accomplished overnight. Improvements have historically been evolutionary rather than revolutionary in the communications industry. New techniques have had to be compatible with the old, and this continues to apply.

There is no question but that adequate techniques for very limited PCM systems are now available. More than four years ago a laboratory model of a system which integrated time-division multiplexing, PCM, and electronic switching was successfully demonstrated. However, except for a narrow range of applications, a great gulf exists between laboratory devices and practical equipment suitable for field use.

Satisfactory PCM requires that all the code pulses be regenerated with great precision as often as required to suppress noise and distortion. The problem may be quite difficult when transmission is over facilities such as wire and cable which are severely restricted in bandwidth and which introduce large amounts of noise and crosstalk. Not only must the pulses be accurately identified and regenerated, but they must preserve their original timing exactly, since only by their exact timing can the individual channels be located and recovered. The problem of pulse regeneration and re-timing is so important that it determines the major limit on the length and performance of the new semi-experimental systems now being tried.

PCM Economics

The real benefits of PCM appear to be realizable only in very large networks spanning great distances. This poses a problem, since it isn't desirable

World Radio History

or practicable to innovate on such a scale, particularly with techniques which are still only a little beyond the experimental stage. Accordingly, it is necessary to find some other application in which PCM can be introduced and in which practical experience can be obtained for perfecting and refining the equipment.

At this time, PCM seems to be most suitable for linking relatively close terminals linked by wire or cable. In modern telephone plant, this is fulfilled best by relatively short trunks between exchanges. Transmission by radio could reduce the need for frequent regenerative repeaters. However, relatively few channels could be accommodated due to the great bandwidth required compared to conventional methods. This would also result in very high cost per channel, since suitable radio equipment costs the same whether it is used to transmit a few channels or many. The same economic philosophy applies to other transmission media such as cable or open wire. Where multiplexing methods that use less bandwidth can be used economically, they will be preferred, since more channels can be accommodated in a given bandwidth. This eliminates almost all present toll circuits and tends to restrict PCM to competing with those transmission methods which are even more wasteful of capacity—notably voice frequency circuits. Thus, the principal area of application—today at least—is in linking centers which are so close together that cable pairs have been cheaper than carrier channels.

If enough channels are involved, PCM *terminals* are inherently very inexpensive. Very little equipment is re-

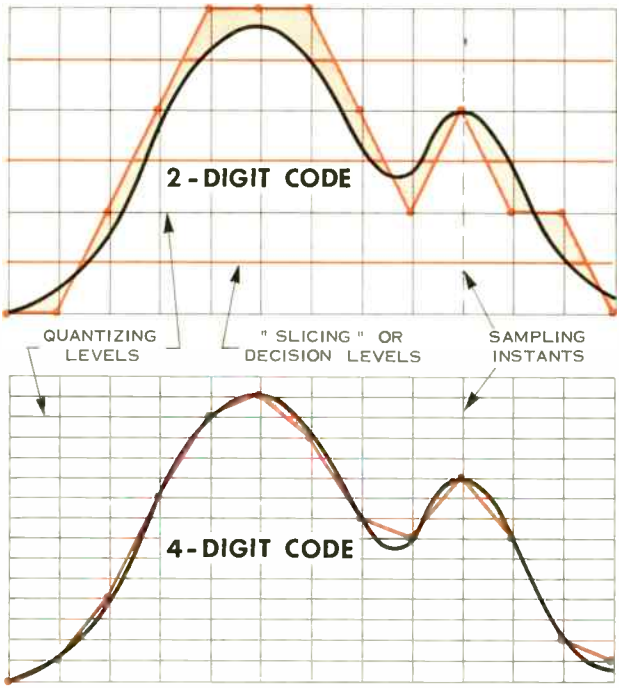


Figure 3. Signal quality is determined by how closely the quantized signal approximates original signal, a function of the number of code digits. Differences show up as quantizing noise. Each additional digit improves signal-to-quantizing noise by six db. With ideal regeneration and retiming, this quality is maintained regardless of number of repeaters or system length.

World Radio History

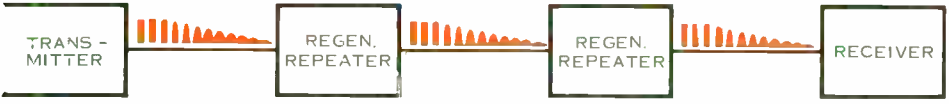


Figure 4. Pulses undergo constant attenuation and degradation in cable. Most crucial function in system is the accurate retiming and regeneration of distorted pulses.

quired by individual channels and all share the common equipment, which primarily consists of switching and encoding circuits. The number of channels transmitted is limited only by the ability of repeaters to identify and restore pulses, and this is controlled by the pulse rate and the transmission characteristics.

Most present systems use a seven-digit code, with an eighth digit for signaling. Since each channel should be sampled 8000 times per second, 64,000 pulses per second are required for each channel. The greater the number of channels which must share the transmission medium, the higher the pulse rate that is required. In order to transmit 24 channels by this means, more than 1.5 million pulses per second must be transmitted.

Obviously, this imposes a severe transmission requirement on ordinary exchange cable pairs, which are normally considered to be rather limited in bandwidth. This is a secondary consideration with PCM, however, since bandwidth limitations only determine the degree of distortion suffered by pulses as they travel over the line. The higher the pulse rate, the more the pulses are degraded in transit, and the harder they become to identify after traveling over the line. The practical significance is that regenerative repeaters merely have to be closer together along the cable. If fewer channels are transmitted, or if better cable is used, repeaters could be more widely spaced, other factors being equal. Since existing cable must be used in the majority of cases, it is desirable to design the system for repeater spacing which matches that of existing loading coils—usually about a mile. Although loading coils improve the line characteristics for speech transmission, they drastically impair pulse transmission, particularly at high speeds. By matching repeater and loading coil spacing; it becomes particularly convenient to remove the coils and substitute repeaters.

If existing exchange cable characteristics largely determine repeater spacing, repeater quality strongly influences the number of channels which may be accommodated, and the over-all economics of the system. An important conflict exists between the cost and capability of repeaters. Since repeaters are required every mile they should be relatively inexpensive in order to be competitive. The stringent requirements for timing accuracy are easily met only with relatively complex or refined techniques, which tend to be expensive.

This, and such other technical problems as crosstalk considerations and the reduction of quantizing noise in PCM systems will be discussed in the next article in this series. It will appear in the January, 1963 issue. •

World Radio History

World Radio History

the Lenkurt Demodulator

VOL. 12, NO.1 JANUARY, 1963

The Transmission of PCM over Cable

For reasons outlined in a recent issue (November, 1962), new multiplexing systems which use time division and pulse code modulation are now beginning to appear in commercial service. Initially, at least, these new systems are restricted to the relatively short trunks between telephone exchanges. The relative immunity of PCM to interference enables it to use bandwidth in exchange cable that cannot be used by conventional carrier systems. It isn't easy, however; the problems of transmitting multichannel PCM signals over cable are formidable — and not yet perfectly solved. This article reviews the more basic problems and some current approaches to solving them.

Three major problems must be overcome by time-multiplexed PCM systems. The conflict between available bandwidth and transmission quality must be resolved. The effects of crosstalk between two or more systems sharing a cable must be minimized. Most important, each of the pulses which comprise the signal—more than 1.5 million per second—must be restored and held to extremely close timing standards, in the face of distortion and interference which tend to introduce uncertainty as to the presence of a pulse or the exact instant when it should most easily be detected.

A conflict between transmission quality and available bandwidth stems from the basic nature of PCM. The PCM signal consists of a stream of code pulses which *represent* samples of the original waveform, and from which the waveform can be reconstructed. These pulses are much easier to squeeze through an unwilling and noisy transmission medium than the analog signal which they represent.

Inevitably, however, the reconstructed signal can be only an approximation of the original waveform, since only a limited number of amplitude "steps" can be represented by the code groups. The differences between the original message waveform and the code-derived facsimile show up as "quantizing" noise. Assuming that all

World Radio History

pulses are accurately regenerated in transmission, the noise acquired by the pulse train in transmission is essentially eliminated in the regeneration process. So long as the PCM remains above threshold, so that the pulse codes can be identified accurately, transmission quality remains uniform. This is in contrast to the performance of conventional methods, where transmission quality varies with signal level, as shown in Figure 1.

The use of additional digits in the code improves signal quality and reduces quantizing noise, but this also increases the bandwidth required for transmission. Bandwidth is directly proportional to the number of digits in the pulse code. This bandwidth requirement places a practical limit on the number of digits that may be used in PCM transmission. This is particularly true when multi-channel PCM signals are transmitted over exchange trunk cables, since these have a very high rate of attenuation at the required frequencies. If signal quality is improved by the use of additional code digits, it reduces the number of channels that may be accommodated or makes necessary more frequent or more expensive regenerative repeaters.

Instantaneous Companding

One way of reducing the total quantizing noise present in a PCM transmission without increasing the number of code digits is to vary the size of the quantizing steps to take advantage of the nature of speech. Statistically, low speech amplitudes (that is, the softer sounds) are much more probable than the very great amplitudes. Figure 2 shows a typical distribution of speech signal voltages relative to the rms or effective level. Note that there is only a 15% probability that the voltage will exceed the rms value, and that fully 50% of the time speech voltage will be less than one-fourth the rms value. Where uniform quantizing is used, many of the quantizing steps are "wasted" and others are "overworked." By altering the quantizing characteristic to favor the weak signals at the expense of the very high amplitudes, more of the speech energy is subjected to relatively "fine-grained" quantizing, thus lowering the total amount of noise. Of course, high signal amplitudes (which are relatively rare) will suffer more degradation than with uniform quantizing.

This technique, called *instantaneous companding*, can be achieved in either of two principal ways: compress the amplitude range of waveform samples before they are quantized, then use a

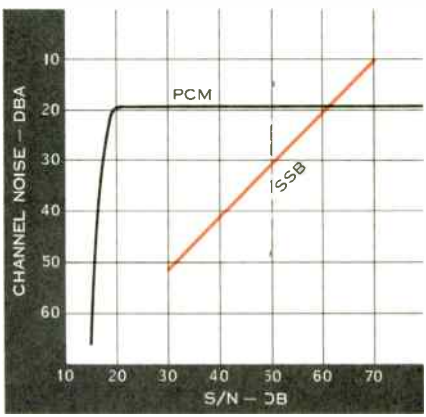


Figure 1. Arbitrary comparison of signal quality versus transmission S/N ratio. SSB signal quality improves in direct proportion to increase in signal power. PCM signal quality is based on quantizing noise caused by coding. Relative positions of the two curves may vary widely, since different factors control each. PCM threshold is largely determined by intersymbol interference, timing accuracy, and differential delay. SSB signal quality is controlled by the absolute value of line noise.

World Radio History

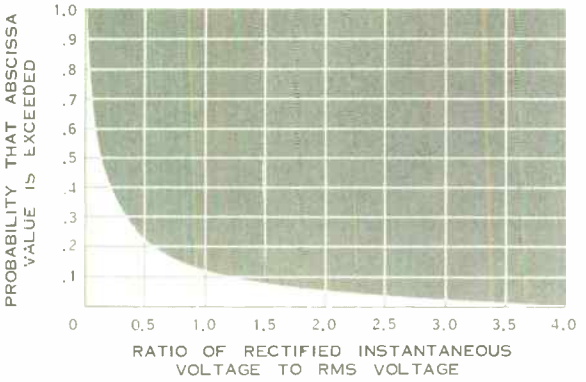


Figure 2. Statistical distribution of single-talker speech amplitudes relative to nominal rms amplitude value. Very low speech volumes predominate, and high amplitude values are relatively rare.

linear quantizer (one with steps of equal size), or vary the sizes of the quantizing steps themselves, so that the steps are smaller and more numerous for low amplitudes. With either method, it is necessary to reverse the process at the receiver in order to restore the original range of amplitude values.

Instantaneous companding produces a significant improvement in signal quality. So long as the compressor and expandor complement each other, a wide range of compression-expansion characteristics may be used. The optimum characteristic depends on the nature of the talkers, instrument weighting and similar factors. Using a certain typical compression-expansion characteristic that varies logarithmically with signal amplitude, an improvement of better than 26 db in the signal-to-quantizing noise ratio is obtained. This is the equivalent of about four additional digits in the pulse code. In systems which use a seven-digit code, instantaneous companding provides quality equivalent to an eleven-digit code.

Regardless of the method by which instantaneous companding is achieved, it is extremely important to match the expansion and compression curves of transmitters and receivers quite precisely. Since compression is a form of "pre-distortion," it is vital that the expandor "track" the compressor as closely as possible in order to cancel out the deliberately introduced signal distortion. Mistracking will cause changes in the transmission net loss as signal level changes, and will restore some of the noise that companding seeks to suppress.

Tracking error between compressor and expandor is sufficiently important that great care must be taken to stabilize the drift of diodes or other non-linear components used to achieve the compression and expansion characteristics. Such techniques as temperature control may be required, adding considerably to the cost and complexity of PCM terminals.

One solution to this problem is provided by a *non-linear coder* which, instead of using the inherent non-linearity of such components as diodes to achieve the companding characteristic, employs linear components (such as resistors) to form a network in which component values yield the desired characteristic. Even if the network components change value with age or temperature, they all tend to change together, thus avoiding alterations in the compandor characteristic curve. If maximum and minimum values on the curve are controlled,

World Radio History

all the points between tend to remain the same.

The companding curve obtained with such a network may not quite achieve the full quantizing noise improvement possible if the speech amplitude distribution is exactly matched, since the network only approximates the desired companding characteristic — which, itself, is only an approximation—with a series of short chords, rather than with a continuously smooth curve. However, quantizing error noise is still reduced on the order of 25 db, and terminal equipment can be more reliable and less costly

The Timing Problem

Rigorous control of the timing of the transmitted pulses is obviously necessary in a multiplex system based on time separation. When the system employs PCM, timing accuracy requirements become even more stringent. High-speed pulses undergo severe attenuation and distortion when they are transmitted through cable. Energy is robbed from each pulse by attenuation and cross talk. Variations in the speed of propagation and delay of the various frequency components of the pulses spread some of each pulse's energy into the time slots occupied by other pulses. This fusing together and mutual intrusion on each other by adjacent pulses is called *intersymbol interference*. At the high frequencies necessary in a multi-channel PCM system, attenuation becomes very high. Crosstalk coupling into adjacent pairs increases with frequency. At the 1.5 megabits-per-second pulse rate, coupling loss between pairs in the same cable may, in some cases, be less than the transmission loss in the pairs themselves. Accordingly, a severe near-end crosstalk problem may exist when two or more systems share a cable.

The ability of PCM to overcome crosstalk and other interference is based on the ability to recognize the presence or absence of the code pulses with great accuracy. Transmission errors result in clicks and snaps, since a false amplitude —one that has little or no relation to the correct signal waveform being reconstructed—is produced by the decoder. Such transmission errors become increasingly likely as the pulse rate increases. The transmission limitations

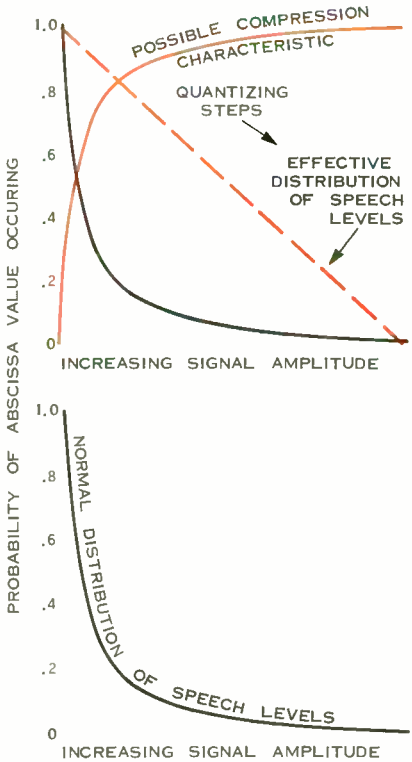


Figure 3. Linear quantizing, shown in A, uses relatively few quantizing steps for a majority of speech amplitudes. By using a compression characteristic (solid red curve) which amplifies the plentiful low amplitudes more than the high values, more uniform quantizing is obtained. As shown in B, this increases the number of quantizing steps for low amplitudes, results in less quantizing noise.

World Radio History

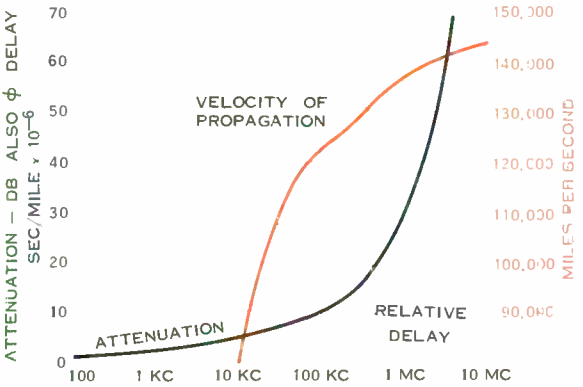


Figure 4. Measured values of attenuation, velocity of propagation, and relative delay of various signal frequencies in typical 22 gauge exchange cable.

of the cable, diagrammed in Figure 4, result in large amounts of intersymbol interference, with the result that pulses become harder and harder to identify. Under these conditions, noise and crosstalk have a greater effect in increasing transmission errors.

In order to regenerate and re-time pulses accurately despite the effects of severe intersymbol interference, it is necessary to sample the pulse train periodically, at just the instant that a pulse, if present, would achieve its peak value. This requires very accurate timing.

In addition to masking the signal pulses, crosstalk or other interference tends to shift the position of individual pulses, creating "timing jitter" which adds to the uncertainty about the presence of pulses. In order to avoid an accumulation of timing error from repeater to repeater (which would sharply restrict system length), each repeater must regenerate pulses which have the correct duration and spacing. This also requires that precise timing information be available at each repeater. There are a number of ways that this can be done. One is to use extremely accurate and stable oscillators at each repeater. This would be successful only if the oscillators did not deviate more than about 30° in phase from the incoming signal. Although this is attainable, it is expensive.

Another approach might be to transmit a pilot tone. This presents the difficulty of providing a very high quality channel for the pilot so that noise of various sorts would not introduce jitter or very short term variations in the regularity of the pilot signal itself. The need for this type of auxiliary channel defeats the purpose of PCM transmission, since one of the objectives is to reduce the vulnerability to interference and to eliminate the need for relatively "delicate" transmission methods. Since the pilot would require complete freedom from crosstalk, it would almost certainly have to be sent over a separate but parallel transmission path, possibly another cable.

An accepted solution is to derive the timing information from the data stream itself. Since the signal consists of a series of periodically recurring pulses, these can be used as a reference for controlling an oscillator, or for exciting a resonant circuit to produce a timing wave. In both cases, reasonably high oscillator stability or resonant circuit Q is desired, but not so high that the resulting timing wave is too "stiff" to be corrected by the incoming pulse

World Radio History

stream. Above all, economic considerations demand that the simplest method that will provide satisfactory timing be used, since practical PCM repeaters are still, at present, more complex and expensive than repeaters for conventional exchange trunk carrier systems. For this reason, contemporary systems use a simple resonant circuit tuned to the pulse repetition frequency. The sine wave obtained from this circuit is amplified and used to derive a narrow sampling pulse. Inevitably, economically realizable circuits fall short of ideal performance, so a small amount of timing jitter is unavoidable, and this shows up in small perturbations of the reconstructed pulse train. Since these irregularities must affect pulse retiming at the next repeater, timing error accumulates and eventually limits the overall length of the system. If a precision "clock" or timing generator is included at intervals, this source of error can be greatly reduced.

Transmission Techniques

The very rapid increase in cable attenuation as frequency becomes higher

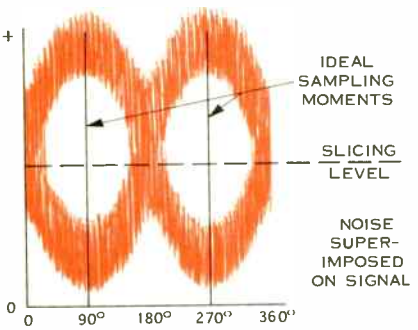


Figure 5. Shaped pulses assume sine-wave form. Noise superimposed on the waveform tends to obscure actual signal value. Timing errors can result in non-optimum sampling, increase vulnerability to crosstalk and noise.

poses several serious problems. Part of the loss is due to crosstalk coupling into adjacent pairs in the cable. Not only is transmission loss great at the 1.5 mc pulse repetition frequency, but crosstalk is formidable. There are several ways of coping with these difficulties to obtain better transmission. One is to "shape" the transmitted pulses so that they have the minimum bandwidth consistent with their repetition rate. Another is to "encode" the signal into a form that shifts the energy spectrum of the pulse train so that more of it is concentrated at lower frequencies. The binary pulses from the transmitter or the pulse regenerator are shaped by passing them through a filter that selectively attenuates some of the very high-frequency harmonics that are inherent in square-wave pulses. The pulses can be shaped at very low level, then amplified for transmission. The filter characteristic is chosen to assure minimum attenuation in the transmission medium. Since less of the transmitted energy is absorbed or coupled into adjacent pairs, (due to a lowering of the extraneous high frequency components), crosstalk is reduced, and there is greater likelihood of the pulse retaining its identity. This technique is not new, having been used in telegraphy for decades, but it assumes much greater importance at the very high pulse repetition rates used in PCM.

Although it is convenient and conventional to diagram a pulse train as consisting of a series of unipolar pulses such as would be obtained by making and breaking a circuit through which a direct current flowed, there are objections to unipolar pulses in practical PCM systems. It becomes necessary to use dc amplifiers instead of the much less expensive ac amplifiers, and transformers cannot be used for signal coupling at the terminals and repeaters. Furthermore, a train of unipolar pulses

World Radio History

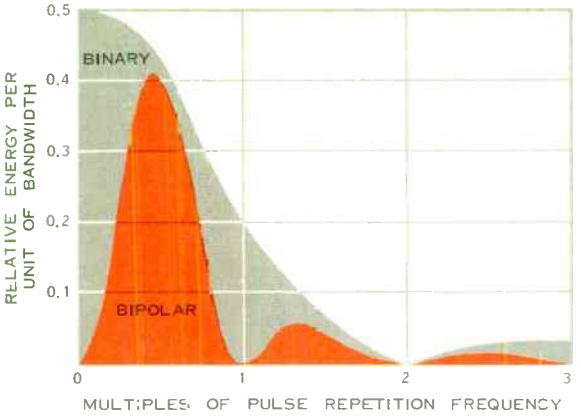


Figure 6. Energy distribution of binary and bipolar signals for 50% duty cycle pulses. Binary signal has strong discrete component at the pulse repetition frequency. Bipolar signal has no frequency components at dc or at the pulse repetition frequency; most energy falls near half the pulse rate frequency.

requires more bandwidth than other signals which have the same information capacity.

These objections are overcome by using a "pseudo-ternary" or bipolar type of signal. These are signals in which consecutive marks are of opposite polarity, and space is represented by zero voltage or neutral. This type of transmission has the additional advantage of requiring only ¼ the power-handling capacity for a given overall voltage range. As shown in Figure 6, most of the energy of bipolar signals is concentrated near frequencies of about half the pulse frequency. Accordingly, there is much less energy coupled into other cable pairs because of the reduced line loss and crosstalk coupling. By forcing the signal to alternate between positive and negative values, the need to transmit direct current is eliminated and transformer coupling becomes permissible. Although there is no energy component at the actual pulse rate with bipolar signals, it is easy to obtain the required 1.5 mc clock signal with a simple full-wave rectifier or frequency doubler. Since a bipolar signal may assume any of three levels, it is somewhat more vulnerable to interference than a true binary signal of the same amplitude. However, this can be overcome by either increasing the signal amplitude, increasing the stability of the timing signal at each repeater, or by shortening the span between repeaters.

An even more promising method of processing PCM signals for transmission has recently been invented at Lenkurt. This technique, known as Duobinary Coding, doubles the number of pulses that may be transmitted through a given channel, compared to bipolar or binary signals. This technique will be described in the February, 1963 issue of the DEMODULATOR. •

BIBLIOGRAPHY

1. H. E. Vaughan, "Research Model for Time-Separation Integrated Communication," *The Bell System Technical Journal;* July, 1959.
2. B. Smith, "Instantaneous Companding of Quantized Signals," *The Bell System Technical Journal;* May, 1957.
3. M. R. Aaron, "PCM Transmission in the Exchange Plant," *The Bell System Technical Journal;* January, 1962.
4. R. H. Shennum and J. R. Gray, "Performance Limitations of a Practical PCM Terminal," *The Bell System Technical Journal;* January, 1962.

World Radio History

World Radio History

the Lenkurt® Demodulator

VOL. 9 NO. 8 AUGUST, 1960

TIME SHARING

Growing Trend in Communications

The constant search for better communications at lower cost has led to new interest in various forms of time sharing. Although the advantages of time sharing have been long appreciated, only recently have techniques appeared which make it attractive for entire communications networks. This article discusses several applications of time sharing and how it may improve the communications systems of the future.

The concept of time sharing is certainly not new. Actually, time sharing is nothing more than several users of a common facility "taking turns." Traffic lights at a busy intersection provide a form of time sharing. Airplanes approaching a busy airport may be "stacked" to allow a single runway to be used by each in turn. This is certainly more economical and practical than building as many runways as needed to let all aircraft land simultaneously.

Telephone subscribers take turns using toll and trunk circuits. The amount of switching equipment in a telephone office and the number of circuits between offices is carefully restricted to meet the requirements of the normally-expected maximum traffic, rather than the maximum possible traffic. If this idea of "taking turns" is extended to carrier equipment and to the method used in central office switching, even greater economies may be achieved.

Time-Division Multiplexing

Many separate channels may be transmitted over a single wire or radio path if some form of multiplexing is used. Frequency-division multiplexing, or carrier, is the most widely used method for accomplishing this. Another method, used far less, is *time-division multiplexing*. Theoretically, both methods produce equal results, neither one requiring more bandwidth than the other, nor having greater vulnerability to noise, in an ideal system.

World Radio History

Until now, frequency multiplexing has been far more practical because simpler techniques were involved.

Frequency-division multiplexing is accomplished by modulating a carrier frequency with the desired signal. A number of these modulated carriers can be applied to a wire or cable pair, or can, in turn, be used to modulate a wide-band radio carrier of much higher frequency.

Time-division multiplexing requires that all channels "take turns" using the common line. Imagine a situation such as diagrammed in Figure 1, where two talkers require connection to their respective listeners, but have only one line available. If the line is rapidly switched from talker *A* and listener *A* to talker *B* and listener *B*, then back again, both share the line.

If the line is switched between the users at too slow a rate, words or syllables in each conversation may be lost. If the rate is increased slightly, both parties may receive all the words and syllables, but still experience distortion because of the loss of some transient sounds and frequencies during switching. If the switching rate is increased so that each circuit is connected several times during each cycle of the highest frequency that the telephone circuit transmits, neither listener will be able to detect any interruption or distortion. Furthermore, there need be no difference in performance between each of the time-multiplexed signals and a direct, uninterrupted connection.

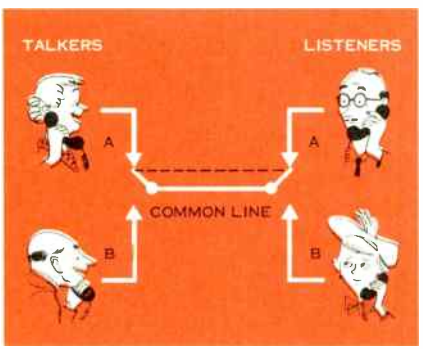


Figure 1. Two or more parties may share common line without loss if line is switched to each user at a rate twice the highest frequency that might be transmitted.

What are the limits to such multiplexing? Researchers have discovered that the necessary sampling rate depends on the highest frequency which must be transmitted. It turns out that a waveform may be perfectly reconstructed if it is sampled at a rate at least twice the highest frequency in the waveform. Thus, a signal may be reconstructed perfectly, without distortion, from samples taken at the rate of 6000 per second. Two samples per cycle allow perfect reproduction because the waveform is inherently unable to assume any surprising or unpredictable values due to the circuit bandwidth limits.

Although telephone channels rarely carry frequencies higher than 3000 cycles per second, the standard telephone channel occupies 4000 cycles. The extra thousand cycles provides a guard band for isolation between channels, and may be used for signaling or other special functions. Therefore, a normal telephone channel would require a sampling rate of 8000 samples per second to completely reproduce all the frequencies it might carry.

A sampling rate of 8000 samples per second requires one sample every 125 microseconds. The shorter the sampling period, the greater the amount of time between samples that might be used for other channels. Shorter pulses, by necessity, have

World Radio History

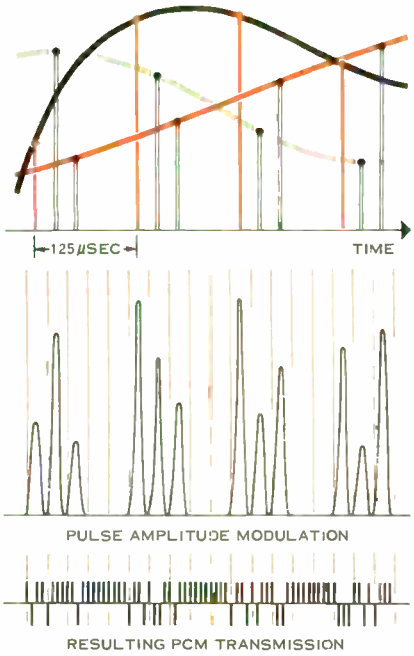


Figure 2. Time between channel samples may be used for carrying samples from other channels. Pulse code transmission increases bandwidth but permits distorted pulses to be completely restored at each repeater.

steeper sides than pulses of longer duration, and this steepness requires more bandwidth for transmission. Insufficient bandwidth tends to "melt" tightly-packed pulses together, thus costing them their identity.

It works out that both time multiplexing and frequency multiplexing require about the same bandwidth. Since the filters of frequency-division systems cannot be made perfect, a guard band must be left between channels to prevent interference. Noise and delay distortion interfere with the transmission of very short pulses, so that they too require operating margin, perhaps a "guard" time interval. Even so, 120 channels might be time-multiplexed by using samples not much briefer than one microsecond, provided that good synchronization is maintained between both ends of the channel.

Pulse Code Modulation

Transmission of such brief pulses presents no particular difficulty over radio, but is considerably more difficult over cable. The very high frequencies involved cause high crosstalk coupling between adjacent cable pairs. Since each pulse represents an amplitude sample, crosstalk tends to change the amplitude of the samples, thus distorting or destroying the message.

A happy solution to this problem is provided by pulse code modulation, a method of converting variable-amplitude signals to digital form. (See DEMODULATOR, *June*, 1959). In this method of modulation, the amplitude range of a signal is divided into a discrete number of steps ("quantized"), and a code combination is assigned to each step. If the code is a binary code, that is, a code having only two states such as mark and space, the transmission is extremely resistant to noise or crosstalk interference, much more so even than frequency-modulated signals. The price paid for this noise advantage is increased bandwidth.

The greater the number of amplitude values or steps to be recognized and transmitted, the greater the number of code elements or digits that must be transmitted for each sample. Since a code combination is used to represent the value of each sample, the number of individual values must be limited, or else the code becomes too long and cumbersome. Accordingly, each code combination is assigned a *range* of values. Naturally, the greater the number of these steps, the smaller

World Radio History

the amplitude range that each must cover, and the truer the reproduction of the original signal.

If a binary code is used, the number of amplitude values which may be distinguished is 2^n, where n is the number of digits required for each sample. Thus, a five-digit code can represent 2^5 or 32 amplitude levels, and a six-digit code permits 64 levels. In addition to achieving greater fidelity in reproducing the original signal, each additional digit provides a 6-db noise advantage, but increases bandwidth in proportion to the total number of digits. For instance, a four-digit pulse code could convey sixteen different amplitude values of a signal, but would require four times the bandwidth of an amplitude-modulated transmission. However, it could tolerate 24 db more noise for a given freedom from error. Similarly, a seven-digit code would tolerate 42 db more noise than an AM signal, could transmit 128 different levels, but would require seven times the bandwidth. By contrast, it is interesting to note that an FM transmission must *double* bandwidth for each 3-db noise improvement. Therefore, to obtain a 6-db improvement, an FM transmission requires four times the bandwidth of an AM transmission. To match the 30-db noise advantage of a five-digit pulse-code transmission, FM would require $2^{10}/5$ times the bandwidth of PCM, an increase of more than 200 times!

The tremendous ability of PCM to overcome noise and crosstalk interference has suggested the possibility that a PCM system could be designed that would actually reduce frequency occupancy. Channels could be stacked together without the guard band so necessary in most other modulation methods. Due to the ability of PCM to overcome interference, channels might even overlap to a certain extent. PCM's ability to overcome noise is dramatically illustrated in Figure 3.

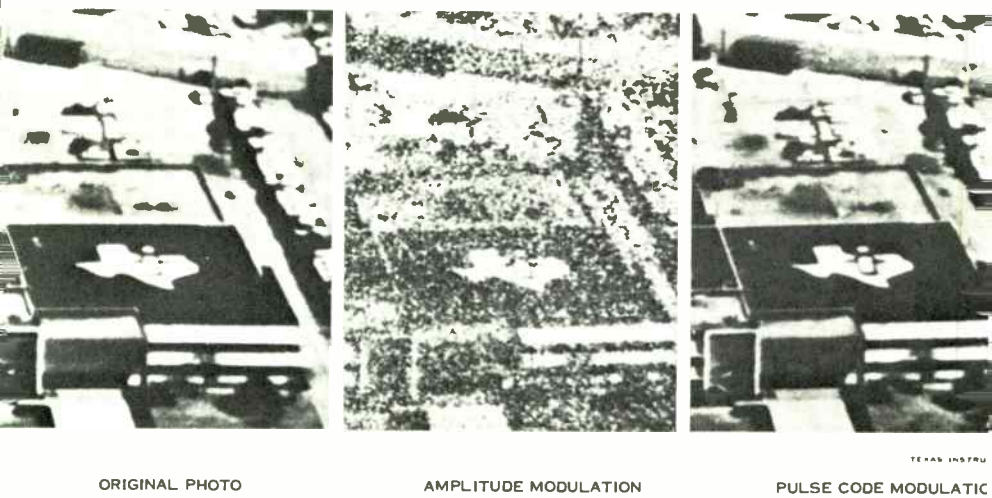


Figure 3. Comparison of transmission by amplitude modulation and pulse code modulation. Both transmissions were made under identical conditions of noise and transmitting power (4 db signal-to-noise ratio). Improved transmission by PCM is obtained at expense of bandwidth.

World Radio History

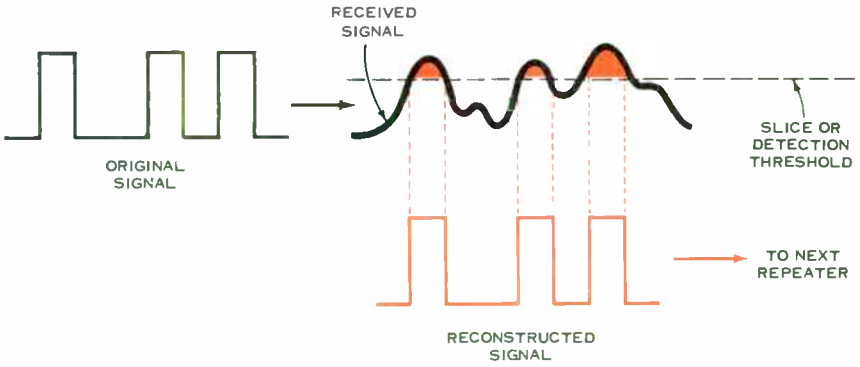


Figure 4. PCM transmission conveys information by presence or absence of pulses, not by pulse shape. Distortion and noise is eliminated by regenerative repeater if it can accurately detect presence or absence of pulses. Reconstructed pulses are synchronously generated to avoid timing errors.

New Pulses for Old

One of the most appealing qualities of PCM is that by using repeaters which regenerate the code pulses, messages can be sent any distance, through any number of repeaters, without acquiring any additional distortion or noise whatsoever. A 3000-mile call would be just as clear and noise-free as a call to a neighbor. All calls would have a standard quality and the distinction between "toll quality" and local or trunk quality equipment would disappear.

Since message information is carried by the presence or absence of pulses, rather than by their shape, regenerative repeaters eliminate the relentless addition of noise and distortion which characterizes all other transmission methods. A regenerative repeater detects only the presence or absence of the distorted pulse and replaces it with a new pulse having the same shape and timing as the one originally transmitted. Repeaters must be spaced close enough that no pulses are so obscured by noise or distortion that they cannot be detected accurately and regenerated.

Electronic Switching

One of the most important factors stimulating interest in time sharing and PCM is their extreme compatibility with electronic switching. The development of semiconductor components having very high reliability and consuming only tiny amounts of current, show the way toward far better and faster service, even while reducing the size and cost of switching equipment. Military communications systems for tactical use already employ electronic switching centers, and several prototype electronic exchanges are going into commercial service.

Most electronic switching centers use the principle of time sharing. Conventional switching methods require that an interconnection be provided for every possible combination of subscriber's lines and trunk circuits, as diagrammed in Figure 5. Note that 45 crosspoints are required to permit any

World Radio History

one of the nine telephones to be connected to any one of the five trunk circuits.

An electronic switching network may employ a "memory" to avoid the need for the crosspoints. Instead of 45 crosspoints, only fourteen "gates" are required. In effect, the memory repeatedly scans the gates representing the nine telephones and the five trunk circuits. Each telephone is assigned a specific "time" in each scan. To make a connection between telephone "4" and trunk "B," for instance, the memory has only to open the gate representing trunk B at the regular time that gate "4" opens. It is this recurrent switching that makes the combination of time-division multiplex and electronic switching so compatible. Since certain identical functions, such as synchronization and pulse generation, are required by both, an integrated system would provide substantial savings in equipment and performance over similar systems in which multiplex and switching were provided separately.

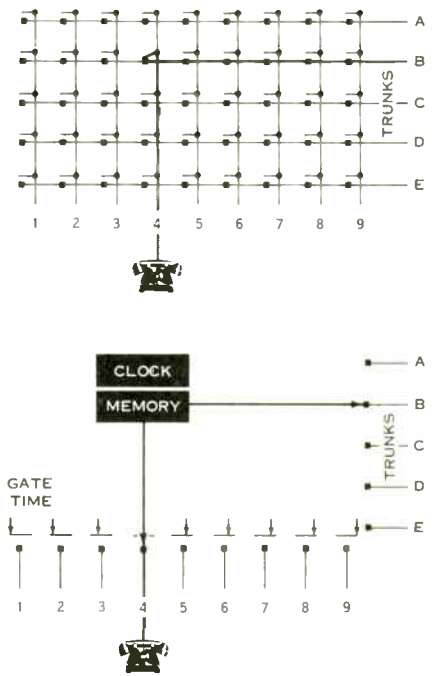


Figure 5. Physical switching requires 45 switches in this example. Electronic system uses only 14 "gates", opened and closed at correct instant by electronic "memory."

The *Bell Laboratories Record* speculates that we might "... some day have a 'super' transmission plan based largely on PCM techniques. Fairly small numbers of PCM channels might be assembled into larger and larger groups in coaxial cables, and finally combined for transmission over 'backbone' waveguide routes at information rates perhaps as high as a trillion bits per second. This large capacity would include thousands of telephone conversations, many channels for data, teletypewriter, and other special services."

Despite the special advantage of these new methods, it is not likely that there will be any sudden change in the switching and multiplexing methods now used in telephone systems. New methods will be introduced gradually, in such a way as to be fully compatible with existing plant and transmission methods.

An excellent example of how new methods may be introduced on a fully compatible basis is the time-division signaling used in the new Type 81A Exchange Carrier equipment. Time-division signaling is used in order to reduce the complexity of channelizing equipment, thus increasing reliability and reducing the space required for each system. Each of the 24 channels is assigned a specific "time slot" for signaling. When an individual channel has a signal to transmit, such as a dial

World Radio History

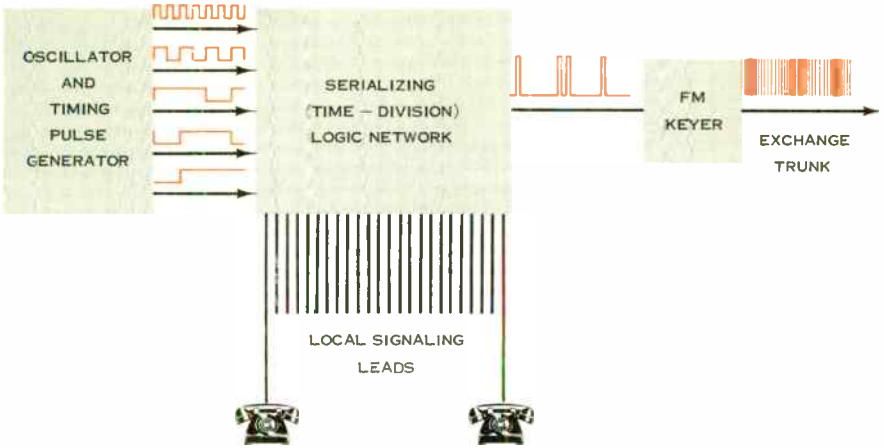


Figure 6. Type 81A Exchange Carrier system employs time-division for signaling. Each local line is assigned its own "time slot." Local signals from all 24 telephones are carried in special binary FM signaling channel.

pulse or ringback signal, a signaling pulse is transmitted only during the time assigned that channel. All 24 channels are scanned 500 times a second, and the resultant signal pulses are transmitted over a single FM binary signaling channel on a frequency slightly above the baseband. Since the fastest dial pulsing rate that might be transmitted over this system is about 20 pulses per second, each dial pulse is scanned at least 25 times, and a signaling pulse is transmitted on each scan. Should a noise "hit" on the line destroy two or three scanning pulses at the beginning or end of a dial pulse, the worst that happens is that dial pulse bias is changed momentarily. If the noise hit occurs in the middle of a dial pulse, nothing at all happens to the signal.

This approach considerably reduces the amount of equipment devoted to signaling in the sytsem. Although the common equipment is somewhat more elaborate, each channel requires fewer components and less space. The use of simple logic circuits and highly reliable semiconductor components, in the fashion of modern computers, provides a high degree of reliability which is in keeping with the best telephone industry traditions. •

BIBLIOGRAPHY

1. W. D. Lewis, "The Idea of Time Sharing," *Bell Laboratories Record*, July, 1959.
2. C. B Feldman and W. R. Bennett, "Band Width and Transmission Performance," *The Bell System Technical Journal*, Vol. 28, No. 3; July, 1949.
3. H. E. Vaughan, "Research Model for Time-Separation Integrated Communication," *The Bell System Technical Journal*, Vol. 38, No. 4; July, 1959.
4. Donald K. Melvin, "A Fully Electronic Automatic Telephone Exchange Using Time-Division Multiplex Technique," *General Telephone Technical Journal*, Vol. 6, No. 4; April, 1959.

World Radio History

World Radio History

the Lenkurt Demodulator

VOL. 9 NO. 10 OCTOBER, 1960

The Difficult Problem of Exchange Trunk Carrier

In most fields of endeavor, progress or improvement is measured in terms of increasing values and larger numbers. Aircraft fly faster, missiles soar farther; radio transmitters increase in power and antennas become ever larger.

One group of engineers, at least, is working hard to reverse this trend toward larger numbers. Their project is to produce a carrier system for transmitting messages shorter distances than any other carrier system has been designed for. The problem is greater than it sounds.

Carrier was invented to save money. The first telephone circuits were simple loops, joined at a common office. As the telephone came into greater use, local offices were connected together by special circuits called *trunks*, and which were not associated with any particular customer's telephone. The main difference between loops and trunks is that loops are always associated with a specific customer (or small group of customers), while a trunk serves any or all the customers, in turn.

The first telephone and telegraph messages were transmitted over metallic wires, each wire or pair of wires carrying one message circuit or channel. It wasn't long before ways were discovered for obtaining additional channels on the same wire by so-called phantom and simplex arrangements. However, only a very limited number of additional channels could be obtained in this way. More circuits required additional wires, and these became quite expensive as distances increased.

The problem was relieved by the development of carrier systems which permitted many different message channels to be transmitted over a single pair of wires. Because carrier equipment may be quite complex, it was economical only

World Radio History

on long circuits, where the cost of additional wire or cable exceeded the cost of the carrier equipment. In addition to providing more channels over existing paths, carrier was found to provide a superior transmission quality.

Carrier Economics

For very long transmissions, many repeaters are required, and these have the characteristic of amplifying distortion and exaggerating small variations in level, as well as amplifying the message. Thus, the longer the system the greater the care required in regulating levels and reducing distortion. For this reason, carrier equipment designed for toll or "long haul" service tends to be quite elaborate and relatively costly.

For shorter distances, the same complex terminal equipment may be too expensive. Furthermore, shorter systems don't need all the features found in a long system. Regulation need not be so strict, and somewhat more noise contributed by the carrier equipment can be tolerated. This allows some cost reduction and permits the system to be competitive with wire or cable over shorter spans than possible with toll carrier equipment.

Even if carrier equipment is stripped of some of its refinements for short haul use, there is a limit as to how far this can be extended. Under no circumstances can the quality or reliability of the equipment be reduced. Components and mechanical construction must be no less dependable than in the finest toll system. Thus, the cost of carrier systems cannot be reduced in direct proportion to their length. There is always a certain minimum cost for terminal equipment, no matter how short the system may be. By contrast, the cost of wire or cable circuits is almost directly proportional to their length; a two-mile cable circuit costs almost exactly half as much as a four-mile circuit.

As recently as 1957, the Chief Engineer of the American Telephone and Telegraph Company stated that it was difficult to "prove in" carrier systems at distances below about 15 miles, although he expected that the rising cost of outside plant and the ingenuity of carrier system designers would eventually reduce this distance.

There is considerable incentive for reducing the economic prove-in distance below 15 miles. As shown in Figure 1, 80% of all trunk circuits in the Bell system are less than 15 miles long. In the independent telephone companies and the General Telephone System, 94% of all trunks are less than 15 miles long! New concentrations of business and commerce in large cities, and the

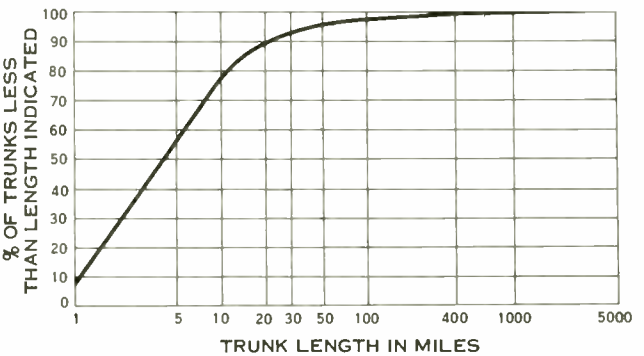


Figure 1. Only 20% of Bell System trunks exceed 15 miles, even with long toll circuits included. About 94% of all other trunks in U.S. are shorter than 15 miles.

World Radio History

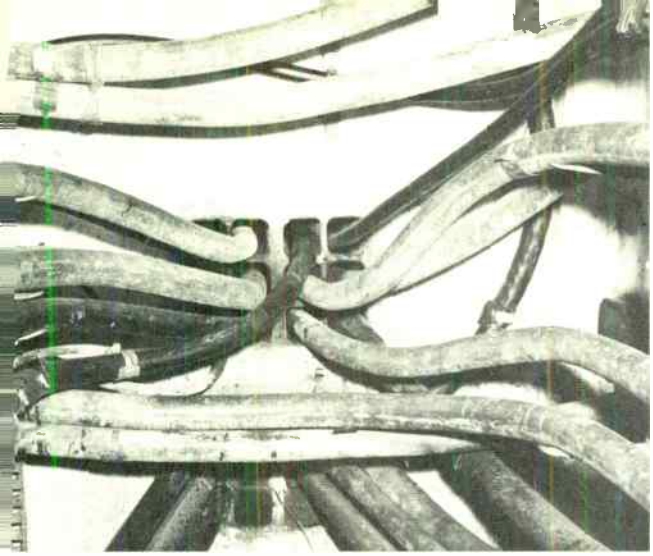


Figure 2. Typical manhole in metropolitan area. When all cable ducts become filled, additional circuits may be obtained by adding new conduit, replacing old cable with finer gauge, or by using carrier.

continued shift of the population from the city to suburban areas is expected to maintain—or even increase—the preponderance of shorter trunk circuits.

Better Transmission

Aside from considerations of cost, carrier offers many transmission advantages over voice-frequency circuits. On spans which are too short to require repeaters, loss will vary with circuit length and the gauge of cable or wire that is used. As a result, there may be considerable level variation between circuits. This is particularly undesirable if these circuits interconnect with toll offices. Although voice-frequency repeaters provide a means of controlling transmission levels, this may be their only justification on short trunks. Not only does carrier provide this same control over loss, permitting transmission level to be independent of the length or nature of the circuit, but it also provides many additional high-quality circuits at low cost.

A special advantage of modern exchange carrier is its flexibility in expanding plant to meet new demand. As cable circuits become fully utilized, carrier channels can be added one by one to meet demand. This provides a means for orderly expansion even in areas where demand is much greater than expected. The cost of expansion can be spread out over a long period of time, since channelizing equipment may be purchased only when needed. By contrast, cable expansion requires the entire expenditure to be made at one time.

Since growth of facilities is largely controlled by available capital, comparative costs usually determine the method used for obtaining additional circuits. An exception may occur in large cities where telephone cable and other utilities must be buried underground in conduits. As demand for service has increased, conduits may have been filled to capacity. When this occurs, streets may have to be opened, and new conduit added. Some cities sharply restrict how often this may be done, and other measures must be found. Often, large-gauge cable is replaced with smaller-gauge cable, thus increasing the number of circuits available in the limited space. While this affords a temporary solution, it cannot be continued indefinitely be-

World Radio History

cause the smaller cable degrades transmission. Eventually, more conduit and cable must be installed, or additional circuits must be obtained by using carrier. Under these circumstances, even toll carrier may provide a *relatively* economical solution. It is not an optimum solution, however, since the refinements usually built into toll carrier may not be removable, yet they require space, power, and increased investment.

More typical is the situation in suburban areas and between small towns. Aerial cable is almost universally used for trunks between exchanges and switching offices. Unlike underground cable which must be installed in conduit, there is little problem in adding circuits as demand grows, particularly with the newer, light-weight cables. Even though the cost of cable and voice-frequency repeaters varies in almost exact proportion to the distance, some sort of trunk terminating equipment is required at each end of the cable circuits, and this cost is independent of trunk length. Carrier systems are usually competitive only when they can provide the additional channels at a cost lower than the cost of new cable and its accessory equipment, other factors being equal.

Economic Design

Since cost considerations dominate the use of exchange carrier, the design engineer must find some way of resolving the conflict between cost and keeping the carrier equipment fully adequate. Several approaches to solving this problem are always open.

One is to eliminate the need for additional expenditure in central office equipment. Following this approach, well-designed exchange trunk carrier systems eliminate the need for additional trunk circuit repeaters which extract signaling and supervisory information from the two-wire transmission path. This function should always be incorporated into exchange trunk carrier equipment.

Another approach for resolving the conflict requires new techniques, ad-

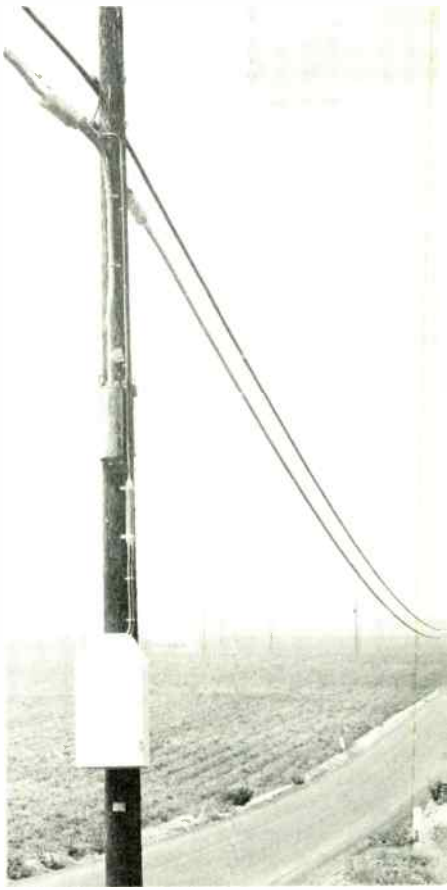


Figure 3. Many exchange trunks use aerial cable. Here, 81A carrier repeaters share pole with older loading coil pots. Most short-haul carrier systems are designed for 6000-foot ("H") spacing of repeaters, same as most voice-frequency loading.

World Radio History

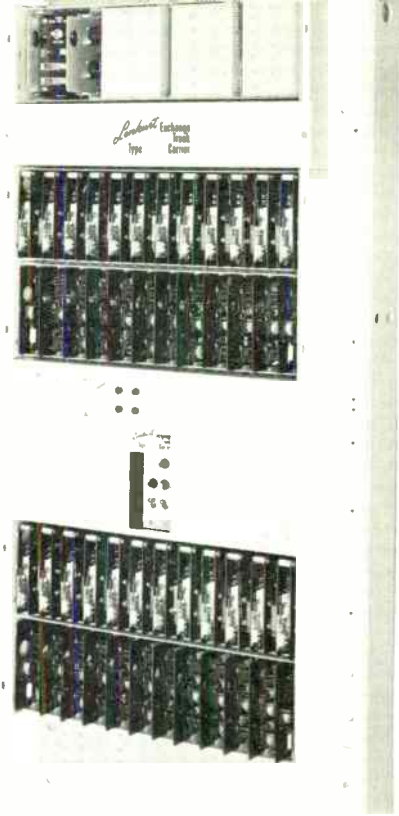


Figure 4. Improved mechanical design concept and complete transistorization permits unusual space economy in the 81A system. Single 11½ foot rack holds 96 channels, complete with carrier supply, power supply, and all trunk signaling.

vances in the state of the art. Since the end of World War II, there have been an abundance of these. The invention and development of the transistor and its related family of semiconductors is one of the most significant. The new carrier system designs now appearing avoid the use of electron tubes, and by so doing achieve reliability of a sort rarely available with electron tubes.

New methods of assembly and mechanical construction have appeared which are particularly suitable for transistors, and which not only decrease costs, but actually improve reliability and serviceability of equipment (See DEMODULATOR, JUNE, 1960. Such state-of-the-art improvements, however, are not restricted to short-haul carrier, but improve all carrier systems.

Specialized Design

Another approach, and one that is particularly appropriate for short-haul carrier, is that of *specialization*. Under this concept, quality is not compromised, but every design decision is based on achieving the desired performance under the special conditions for which the system is intended. Instead of seeking versatility and broad capabilities, the designer concentrates on doing the limited task particularly well — and economically.

Many engineering factors can be varied to achieve savings without losing quality. The type of modulation used is one such factor. Three types of modulation are used almost exclusively in frequency-division carrier: (1) single-sideband, suppressed-carrier, (2) double-sideband, suppressed-carrier, and (3) double-sideband, transmitted-carrier. Toll carrier almost invariably uses single-sideband, suppressed-carrier modulation in order to carry as many channels as possible in the available bandwidth. When the carrier is suppressed, common amplifiers can accommodate more channels before overloading and increased non-linear distortion become likely. This is of dubious advantage in short-haul systems because these systems don't need to transmit very many channels over one facility. A carrier channel unit must be more complex and costly if the carrier is suppressed, and this is more of a handicap

World Radio History

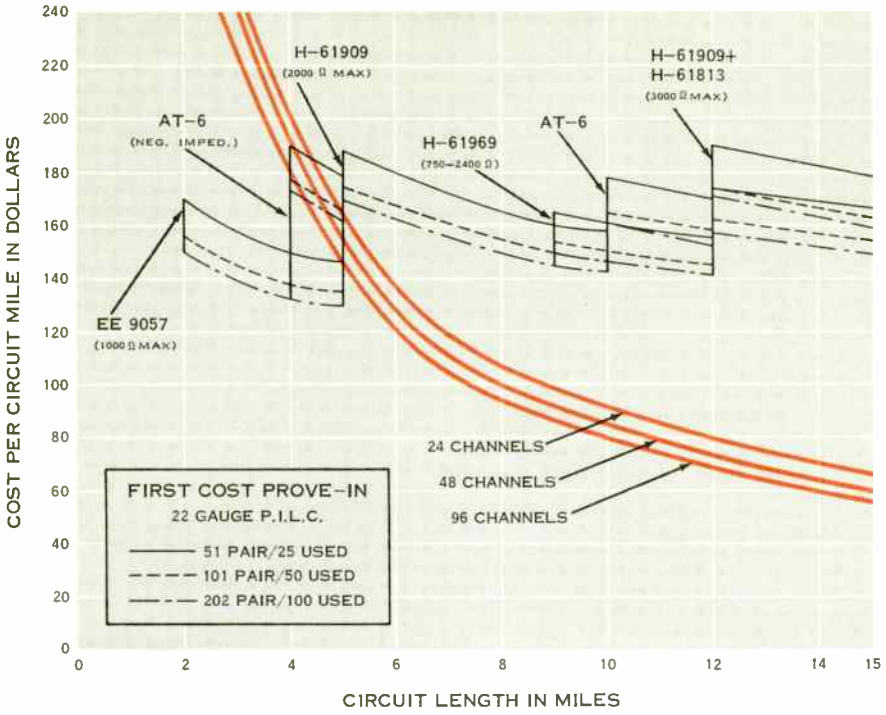


Figure 5. Economic prove-in curves for 81A Exchange Trunk Carrier System as compared with several commonly used types of cable. Cable costs show total installation and material, based on average value of 130% of material, and include negative impedance repeaters or trunk repeaters suitable for distance indicated. Carrier costs include all equipment, suitable repeaters, and installation based on 140% of equipment cost.

to a short-haul system than to a toll system.

An additional difficulty which affects cost and complexity, is that when the carrier is suppressed at the transmitting end, it must be re-inserted at the receiving end. This requires very close control of carrier frequencies at both ends. In the case of single-sideband, all carrier frequencies should be within a few cycles per second of each other. The problem is much greater for double-sideband, suppressed carrier. The transmitting carrier (which is suppressed) and the carrier which is reinserted at the receiver must be locked in phase as well as frequency. Should these get out of synchronism by as much as 90° (¼ cycle), severe distortion results.

Although there are various ways of coping with these problems, the solutions require a more complex system, one which may be more difficult to maintain. Because of the additional complexity, there are more chances of failure, since more parts and more functions are required to do the same job.

Both single-sideband, suppressed-carrier, and such modulation methods as phase or frequency modulation are use-

World Radio History

ful in *extending* or lengthening systems, but are less practical for short-haul systems, because their advantage in overcoming noise or interference is less needed. Again, the additional complexity adds unnecessary cost both in the initial investment and in future maintenance. In general, the simplest design that will provide the desired performance is to be preferred.

Another engineering factor that can be manipulated in designing carrier systems is the distribution of functions between common equipment and channel equipment. If the system is to have more than just a few channels, it is more economical to reduce the channelizing equipment and add to the common equipment. The traditional argument against this is that the operation of *all* channels depends on the reliability of the common equipment. However, in very short-haul applications, it seems preferable to perform as many functions as possible with common equipment in order to reduce complexity and costs of channel units. Failure of the common equipment is of relatively less consequence on short trunks than it would be on long-distance trunks where more costly transmission mileage would be tied up and a much greater percentage of the total communications network involved.

Of course, as much care as possible must be taken in designing common equipment, in order to eliminate as many chances of failure as possible. For instance, *active* devices, such as electron tubes, are more likely to fail than such *passive* devices as diodes, capacitors, and the like. Thus, common equipment (or any equipment) will tend to be more reliable if it accomplishes its duties with passive devices rather than active devices. The allocation of duties between common equipment and restricted equipment such as channel units, and between active components and passive components, requires some of the most sophisticated and knowledgeable engineering design to be found in carrier communications.

Conclusions

The art of specialization of equipment design is really a form of refinement. The resulting equipment achieves great efficiency in its assigned task, but at the cost of reducing the variety of tasks for which it is suited. In this respect, it may be likened to the relationship between a general-purpose carving knife and a surgeon's scalpel. Although the scalpel will never replace the carving knife, it is much to be preferred for its special task!

The recent appearance of modern short-haul carrier systems, such as the Lenkurt Type 81A, which proves-in against aerial cable at distances less than five miles (see Figure 5), is a step forward in improving transmission quality and providing a tool with which telephone companies can provide better service at lower cost. •

BIBLIOGRAPHY

1. H. R. Huntley, "Where We Are and Where We Are Going in Telephone Transmission," *Communication and Electronics:* March, 1957.
2. A. B. Clark, "Some Recent Developments in Long Distance Cables in the United States of America," *Bell System Technical Journal*, Vol. 9, No. 3; July, 1930.
3. L. B. Bogan and K. E. Young, "Simplified Transmission Engineering in Exchange Cable Plant Design," *Communication and Electronics;* November, 1954.
4. H. R. Huntley, "Transmission Design of Intertoll Telephone Trunks," *Bell System Technical Journal*, Vol. 32, No. 5; September, 1953.

World Radio History

World Radio History

the Lenkurt Demodulator

VOL. 11, NO. 9 SEPTEMBER, 1962

COORDINATION BETWEEN CARRIER SYSTEMS

When two or more carrier systems must share the same cable or open wire transmission path, they may interfere with each other, causing increased crosstalk and noise in both. Although this may be reduced by suitable design of the carrier equipment, increased crowding of transmission facilities may bring together systems having characteristics which increase interference. This article discusses how interference between systems occurs, the nature of the interference, and ways of reducing it.

An ideal communication circuit would be so very clear and free from noise or interference that it would seem as though the talker were in the next room—or even face-to-face. In the case of data circuits, this freedom from interference would eliminate all errors at any transmission speed.

Unfortunately, the ideal case never prevails, and communications circuits are always subject to interference from many sources. One of the more serious problems is the mutual intereference which may occur between carrier systems sharing the same transmission facility.

When signals of any type are transmitted over wire or cable, some of the energy from the signal is coupled into adjacent pairs, where it may appear as crosstalk or noise. The degree of interference which occurs is directly related to the coupling between the pairs and certain transmission characteristics of the two interfering systems. These include relative transmission levels, frequency plans, and the type of modulation used.

Although various measures such as open-wire transpositions and the use of variable-pitch twist in cable* pairs reduce the coupling between pairs, they do not eliminate it altogether. One way of controlling the mutual interference is by using auxiliary devices such as compandors which reduce the apparent *effect* of interference without actually diminishing the transfer of signal energy from one system to the other. Since this approach is largely psychological, it is ineffective in telegraph or data transmission.

*For convenience, most subsequent references in this article will be to transmission over cable. However, it should be understood that the discussion applies just as well to open wire.

World Radio History

Coordination of Levels

Interference between two systems is directly proportional to the difference in the operating levels of the two. If the signal level in the disturbing circuit is high, the crosstalk will tend to be high, on a db-for-db basis. For instance, if circuit A operates at a level 10 db higher than circuit B, crosstalk from A will appear in B 10 db higher than if the operating levels were equal. However, intereference from B will be 10 db lower.

It is important, therefore, that systems operating between the same two points maintain the same nominal transmission levels, and this level coordination should be maintained at all points along the line. If a system joins another at an intermediate point between terminals, the transmission level of the entering system should be adjusted to correspond to the levels of the other systems in the cable.

Where cables with mixed wire gauges are employed, the difference in attenuation characteristics may require a special "compromise" in operating levels in order to minimize level differences. For instance, one signal is carried on a 22-gauge pair and another on a 19-gauge pair, the signals transmitted over the 22-gauge pair will be attenuated more rapidly than those on the 19-gauge pair. In such cases, it is generally necessary to reduce the transmitting level of the system on the 19-gauge pairs by *half* the attenuation *difference* between the two.

Frequency Coordination

The above techniques—reduction of coupling between pairs, use of compandors, and coordination of operating levels—can be used to minimize crosstalk and interference in most types of systems, including those which operate only at voice frequencies. When *carrier* systems share the same cable, however, interference can be further controlled by proper selection of frequency allocations and type of modulation. Since each carrier channel occupies its own small frequency band, it is vulnerable only to interference which falls within that band; other interference is rejected by the carrier filters which separate one channel from another.

Even the weighting characteristics of the instruments used in the communications system may have an important effect in reducing crosstalk. As shown in Figure 3, weighting characteristics and channel filters effectively "reshape" the frequency distribution of energy entering the system. The desired messages transmitted over the system are also altered, but because they are received at a much higher level than interfering energy, this shaping has less effect on clarity and intelligibility. Note that the most important frequencies for intelligibility lie between 800 and 1500 cycles per second, while most of the energy present in speech is concentrated between about 200 and 500 cps. Part D of Figure 3 represents the equivalent signal that is transferred from the "disturbing" system into the "disturbed" system.

Figure 4(A) illustrates the transmission or frequency response characteristics of the "disturbed" carrier channel. If the frequency allocations of

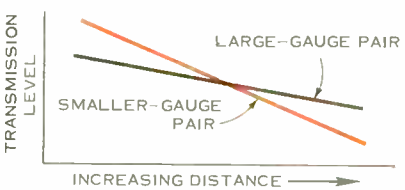


Figure 1. Where two systems experience different attenuation rates, "compromise" reduces maximum level difference. Equal transmission levels would result in much greater difference at receiving end.

World Radio History

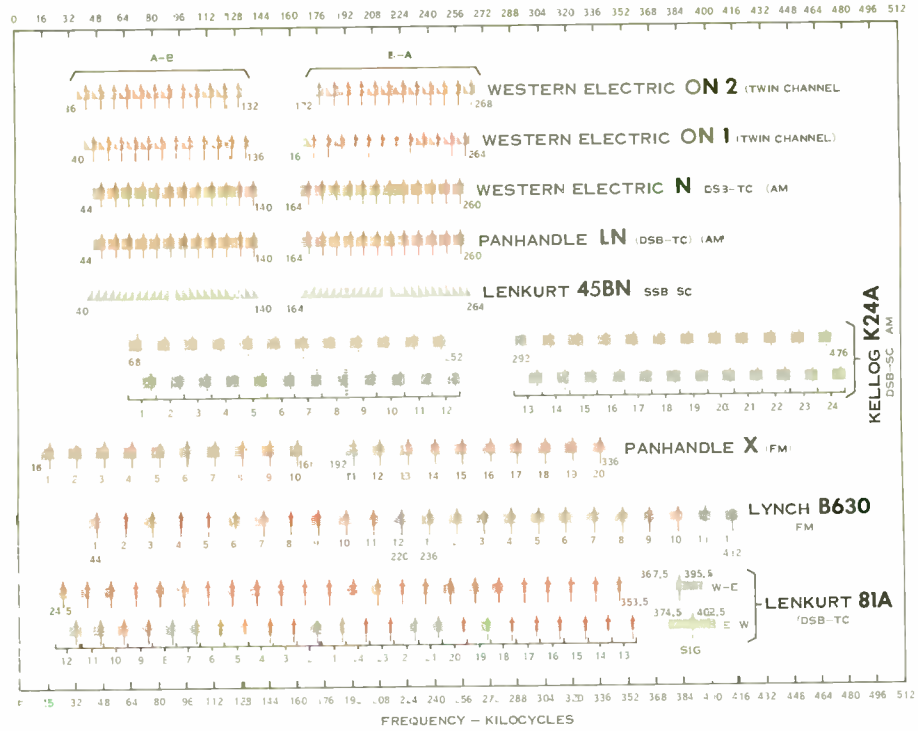


Figure 2. Frequency allocations of several typical cable carrier systems. Note that frequency allocations differ from system to system. Interference can occur when channel frequencies coincide. Systems at upper left use carrier frequencies and operating levels designed to minimize interference, and Panhandle LN and W.E. "N" operate end-to-end.

the two interfering carrier systems are the same, the disturbing energy will appear in the disturbed channel with the relative magnitude shown in Figure 4 (B), after having undergone attenuation by channel filters and weighting characteristics of the telephone equipment.

Frequency Inversion

An important technique for reducing interference between systems is the use of frequency inversion between systems operating in the same basic channel frequencies. This is possible when single-sideband modulation is used, if one system transmits the upper sideband of each channel carrier, while the other system transmits only the lower sidebands. In addition to affecting intelligibility of the crosstalk, this reduces the energy coupled into the disturbed system significantly, by shifting energy peaks of the interfering signal to new locations on the transmission characteristic where there is more attenuation. Figure 4(C) shows the resulting energy spectrum when the disturbing signal diagrammed in Figure 3(D) is inverted in frequency and passed through the channel having the characteristics of Figure 4(A). Since most of the

World Radio History

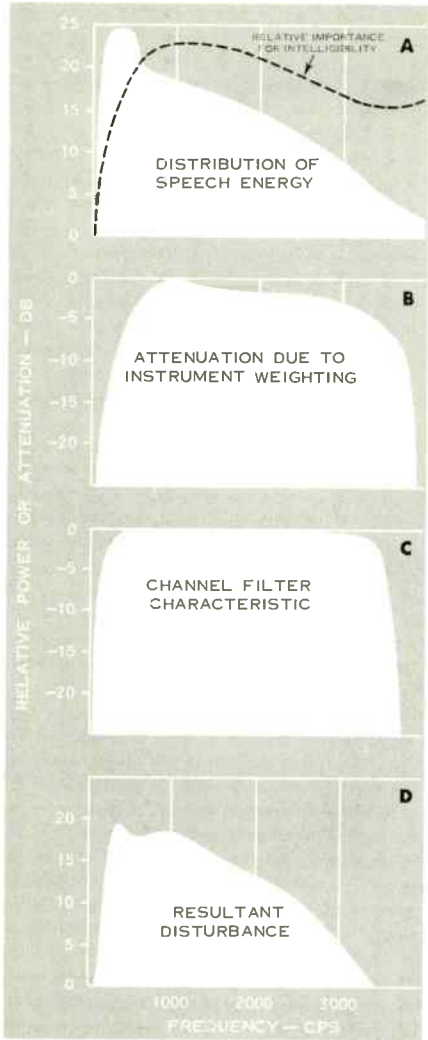


Figure 3. Typical power spectrum of speech and how it is affected by channel filter and instrument characteristics. Part (D) shows resulting spectrum of disturbing signal.

disturbing energy lies in the vicinity of 1000 cps, it is sharply reduced by the relatively high attenuation introduced by the transmission characteristic of the disturbed channel. In practical systems which use this method to reduce interference (Lenkurt 33A and 45A, and Western Electric C and J open-wire carrier systems, for instance), a reduction in interfering energy of about 3 db is realized.

Frequency Staggering

Another technique that successfully reduces interference between two systems is to shift the carrier frequencies relative to each other. Like frequency inversion, the interfering energy may be made to fall outside the pass-band of the disturbed channels. Figure 5(A) shows the energy distribution of interference when the disturbing channel carrier is shifted 1000 cps higher in frequency than the disturbed channel. When the disturbing channel frequencies are shifted 1000 cps lower than the disturbed channel, the interference appears as shown in Figure 5(B). Further shifts would result in even more improvement, but for the presence of adjacent channels. As channel carrier frequencies are shifted further, the disturbed channel begins to pick up energy from *two* channels, and interference increases with further shifting.

An additional benefit is gained by channels used for conversion. When channel frequencies are inverted or shifted in frequency 1000 cycles or more, the crosstalk becomes unintelligible. Although the same amount of interference energy may be present, it is less disturbing to talkers. Subjective tests reveal that unintelligible crosstalk can be as much as 3 db higher in level than unintelligible crosstalk to produce the same disturbing effect. Thus, frequency inversion can yield a total improvement of about 6 db, while staggering of carrier frequencies can produce even more improvement, depending on the nature of the disturbing signal and the transmission characteristics of the disturbed channels.

World Radio History

In the case of pulse transmission such as digital data or telegraph, "intelligibility" is not a factor, and interference is strictly a function of the amount of energy coupled into the disturbed circuit, and its frequency. If the data is transmitted over a single frequency assignment within a voice channel (as in certain types of "switched data" transmissions), the use of frequency inversion or staggering may be sufficient to cause the disturbing tones to fall outside the pass-band of the data receiver channel filter. However, where several data channels are transmitted in a single voice channel, frequency inversion and staggering may merely transfer the interference from one data channel to another. In such a case, a reduction of interference results only when the interfering tones fall outside the pass-band of the voice channel filter.

It is important to make sure that the frequency allocations of two systems which share a cable are not staggered in such a fashion that at one terminal the high level transmission of one system coincides in frequency with the much weaker incoming signal of the other system. This creates the maximum possible difference in levels, since the transmitted signal is at its strongest and the received signal is at its weakest. Assuming that both cable pairs are of the same gauge and that both systems operate at the same nominal transmitting level, near-end crosstalk will be increased by the transmission loss of the path.

Single-sideband Versus Double-sideband

In order to simplify and reduce the cost of terminal equipment, some carrier systems employ double-sideband amplitude modulation, in which the carrier and both sidebands are transmitted. Most of the power in such a signal is in the transmitted carrier. Even at 100% modulation, the carrier has twice the power of both sidebands together. Carrier power remains constant regardless of modulation, while sideband power will vary directly with the degree of modulation.

If such a signal is staggered in frequency from another, so that the carrier frequency falls within the pass-band of another channel, it will create interference in the form of a tone, the frequency of which will be determined by its relative location in the disturbed channel. For this reason, systems that are designed to operate in the same cable with other systems which trans-

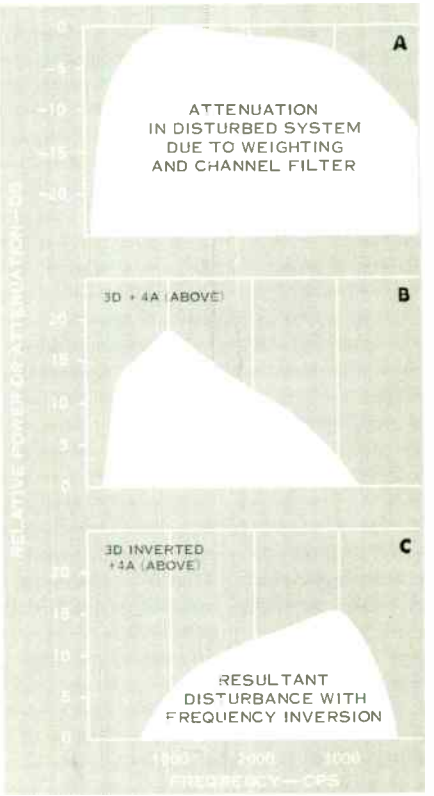


Figure 4. Characteristics of disturbed channel and how it reduces interference. (C) shows improvement due to frequency inversion.

World Radio History

mit the carrier almost invariably employ the same carrier frequencies.

In most single-sideband systems, one sideband and the carrier are eliminated, so that the transmitted signal power will depend entirely on the modulation of each channel. Only when one or more channels are modulated is power transmitted over the line. Thus, the interference caused by such a system will vary with the modulation characteristics and the operating levels used, and this will vary with the individual types of equipment.

When double-sideband systems are used in the same cable with single-sideband systems, the double-sideband (DSB) systems are inherently less susceptible to interference than the single-sideband (SSB) systems, other factors being equal. Assuming that there is no frequency staggering or inversion, only one sideband of the double-sideband transmission will be coupled into the SSB channel, so that the resulting crosstalk will be directly proportional to the difference in the levels of the two—the same as between two SSB systems.

When a DSB signal is demodulated, the voltage of the two sidebands add in phase to yield an effective increase in level of 6 db. Thus an interfering SSB signal produces interference in a DSB channel 6 db weaker than in another SSB channel.

Even when *two* SSB channels interfere with a single DSB signal, one into each sideband, the two interfering channels have no coherent phase relationship, and so add on a power basis rather than on a voltage basis. This results in a 3-db increase in interfering energy, which is still effectively 3 db lower than the DSB signal.

Other Modulation Methods

Although single-sideband and double-sideband modulation are the most widely used means of transmitting multiple channels, other methods are sometimes used. For instance, frequency modulation may be used to reduce the need for level regulation, or to seek an improvement in noise performance. Ideally, FM is insensitive to amplitude variations and should achieve better noise performance than amplitude-modulated signals such as SSB and DSB. This improvement is actually achieved only with a relatively high modulation index or deviation ratio; that is, when the frequency deviation is

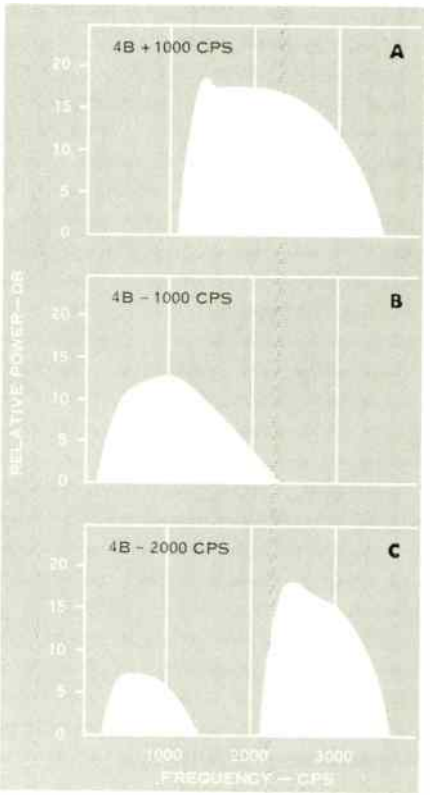


Figure 5. Interference between two SSB channels when (A), disturbing signal is 1000 cps above disturbed channel; (B), disturbing channel is 1000 cps below disturbed channel; (C), disturbing channel is 2000 cps below disturbed channel.

World Radio History

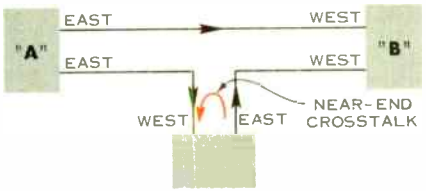


Figure 6. "Directional" coordination is important in reducing interference. Identical allocations should be used at same location whenever possible, as at A *and* B. *Unavoidable use of East and West terminals at* C *may require special treatment, such as using separate cables to junction.*

several times as great as the highest modulating frequency. Because wide deviation rapidly "uses up" the frequency spectrum available for transmitting the signal (thus sharply limiting the number of channels that can be accommodated), most FM systems restrict the modulation index to little more than unity. The result is that noise performance is essentially equal to that of a conventional DSB system. Although most noise and interference are amplitude-varying phenomena which the FM receiver should eliminate, practical deficiencies of the receiver limiters allow some interference to pass.

When no modulation is present, the carrier of the FM channel may appear as an interfering tone in other channels which take in the FM carrier frequency. This source of interference is reduced when the channel is modulated, because of the distribution of carrier power into the sidebands.

Interference from FM sidebands is very much like that from frequency-inverted SSB channels, unintelligible but proportional to the amount of energy present within the pass-band of the disturbed channel. Unlike AM sidebands, however, the energy distribution across the band will vary from instant to instant in a way that is not proportional to the energy distribution of the modulating signal, but changes as carrier energy is distributed into farther sidebands with increase in modulation level.

PCM or pulse code modulation multiplex systems introduce special problems of coordination with other systems. Practical PCM systems transmit pulses at a very high rate — about 1½ million per second. The minimum bandwidth required for transmission at this rate is nearly 1.0 mc, thus imposing a very severe transmission requirement on the cable pair. At the higher frequencies which must be transmitted in PCM, coupling between pairs becomes very much greater than at the lower frequencies which characterize SSB, DSB, or FM transmission. Although the basic nature of PCM usually permits adequate transmission under these conditions, interference with other systems in the same cable becomes intolerable, unless they are also of the PCM type. Although PCM systems are inherently able to withstand 20 to 50 db more interference than voice or carrier systems, the transmission of several systems over a single cable increases mutual interference so badly that a separate cable may be required for the return direction, in order to avoid near-end "crosstalk" or interference which causes pulse transmission errors.

Although equipment characteristics are extremely important in minimizing interference between systems, other factors, such as the way they are applied, the nature of the transmission medium, and the quality of maintenance can spell the difference between acceptable and intolerable performance. Such techniques as pair selection, and even the separation of transmission facilities, might have to be resorted to in difficult cases.

World Radio History

World Radio History

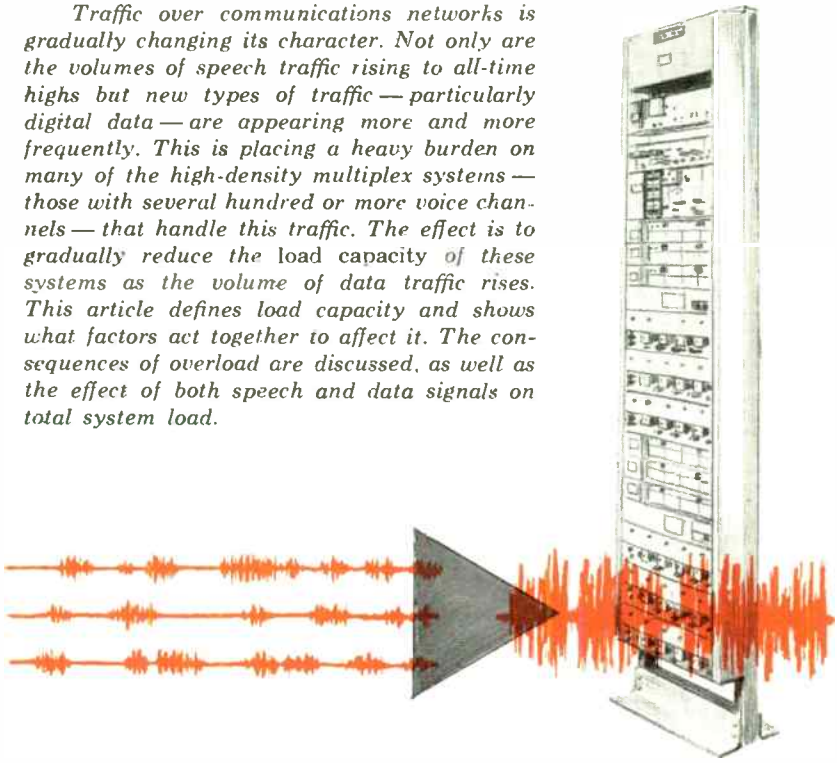
the Lenkurt Demodulator

VOL. 14, NO. 3 MARCH, 1965

LOAD CAPACITY

of high-density multiplex systems

Traffic over communications networks is gradually changing its character. Not only are the volumes of speech traffic rising to all-time highs but new types of traffic — particularly digital data — are appearing more and more frequently. This is placing a heavy burden on many of the high-density multiplex systems — those with several hundred or more voice channels — that handle this traffic. The effect is to gradually reduce the load capacity *of these systems as the volume of data traffic rises. This article defines load capacity and shows what factors act together to affect it. The consequences of overload are discussed, as well as the effect of both speech and data signals on total system load.*



World Radio History

WHAT is a 600-channel multiplex system? If this question were asked of people only lightly engaged in communications work, the answers would probably be as varied as they were numerous. A typical answer might be: "A 600-channel multiplex system is one that can carry 600 voice-frequency signals simultaneously — one signal in each voice channel." This seems logical. Such a system does have 600 channels and each channel is capable of carrying a signal. Surprisingly enough, this answer is almost never true. Few high-density systems can approach carrying signals simultaneously on all channels without being severely overloaded. Such an extreme capability is not required in most of these systems, since traffic will rarely be present in all channels at once. The traffic-handling capability, or *load capacity,* of a high-density multiplex system, therefore, is based on the probable signal load at the time of heaviest traffic, rather than the maximum load that could occur.

Load capacity can be defined as the volume of traffic a system can handle without undue distortion or noise. The exact meaning of "undue" varies, depending on the quality of service required, but generally it is the point at which interference seriously affects either the accuracy or the intelligibility of the transmitted information.

Many factors act together to determine load capacity. The physical makeup of the major circuit elements — amplifiers, modulators, and demodulators

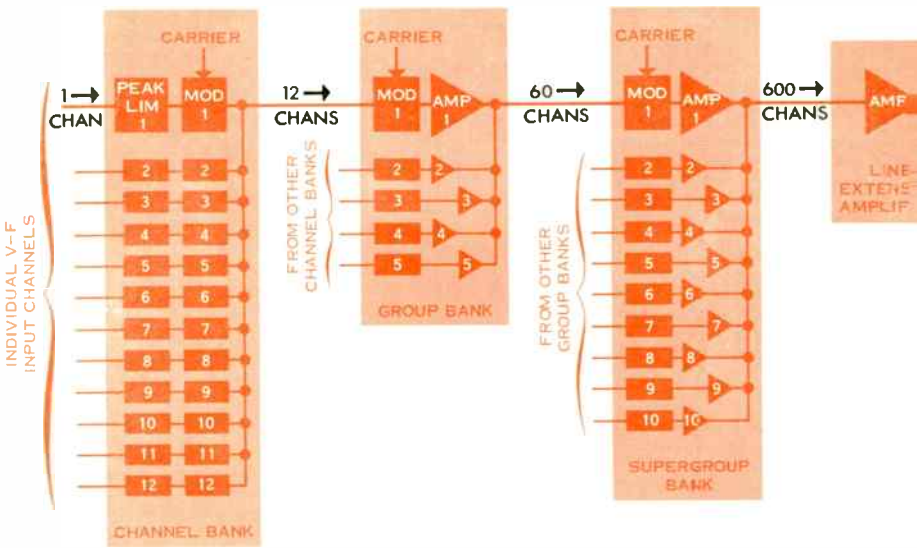


Figure 1. Arrangement of the major active circuit elements in the transmit portion of a typical 600-channel multiplex system. Many elements are common to more than one voice-frequency channel.

World Radio History

— fixes the maximum signal load a system can handle. This factor is constant in any system. The types and quantities of signals carried by a system determine the total signal load. Some types of signals impose much heavier loads on multichannel systems than do others. Data and telegraph signals, for example, normally present a continuous load, while speech, which is quite sporadic, does not. If the number of channels carrying data or telegraph signals exceeds that which a system was designed to carry, then other voice channels may have to be disconnected from service to prevent overloading. This, of course, reduces the system's load capacity.

Transmission requirements of a particular network also affect load capacity. These requirements set the levels of input signals to a multiplex terminal and establish the maximum permissible noise level at the terminal output. If signals have to be applied at high levels, or if noise requirements are unusually stringent, then the load capacity may be effectively reduced.

Speech is the predominant type of traffic in most multiplex systems. Consequently, most systems are designed around the characteristics of speech signals and the statistics of telephone talkers. Allowance is usually made for signaling tones, pilot signals, carrier leak, and relatively small amounts of telegraph traffic. Probability has played an important part in arriving at a suitable load capacity for these systems. As stated previously, these systems are designed to handle the probable signal load at the time of heaviest traffic, rather than the maximum load that could occur. Of course, there will be instances when the total load exceeds the design limitations. These instances, called the *periods of overload*, must not occur except for a small percentage of the time if a system is to provide adequate service.

Effects of Overload

A multiplex system allows several signals to be transmitted simultaneously over a single transmission medium. Most high-density systems use a frequency-division form of multiplexing, where individual input signals are translated to separate positions in the frequency spectrum by means of amplitude modulation. The lower sideband of the frequency-translated signal is usually transmitted, while the carrier and upper sideband are suppressed.

In these systems, many voice channels are handled by common amplifiers, modulators, and other active circuit elements. If a system were perfect, each of these elements would be completely linear, and a signal at the output of any element would be a faithful reproduction of the input signal, with nothing added or taken away. For practical reasons these are ideals that are approached but never fully realized. All active elements are non-linear to some extent and induce a certain amount of distortion into the signal. In a well-designed transmission system, distortion is simply held to within limits that are acceptable to the user.

Of all the active circuit elements in a multiplex system, multichannel amplifiers contribute the most distortion and are most likely to overload when input signals reach high levels. In a typical system these amplifiers may be common to anywhere from a few voice channels to several hundred, and have a frequency bandwidth of from several thousand to well over four million cycles per second. Over such a wide band of frequencies, amplitude response and phase shift will not be uniform. Most amplifiers use negative feedback to improve linearity, but prac-

World Radio History

tical considerations cause amplifier performance to be somewhat less than ideal.

When signals of different frequencies are applied to an amplifier, the output contains not only the input frequencies but an almost infinite number of other frequencies. Harmonics of the input frequencies appear in the output. In addition, an amplifier generates a great many intermodulation products including not only the sums and differences of the original input frequencies but also the sums and differences of their various harmonics. Since these products were not present in the original input signals, they are the distortion caused by the amplifier's nonlinearity. The more nonlinear the amplifier the greater the power of the distortion products.

As long as the amplifier is operated within its dynamic range, power levels of the distortion (or intermodulation) products normally will be too low to interfere greatly with the original input signals. Second-order products (A + B, A — B, 2A, and so on) will have the greatest amplitude, followed successively by third-order products and higher orders. However, with every 1-db increase in rms ouput power, second-order products increase in power by 2 db, third-order products by 3 db, and so forth. All products follow a power series increase until the power of the input signal reaches a certain critical point—the *breakpoint* of the amplifier.

Above the breakpoint, second- and third-order intermodulation products quickly rise in power. But the biggest power rise is in higher order products. These products jump abruptly from very low levels to levels very near those of the .second- and third-order products. If the amplifier has a bandwidth of one octave or less, most even-order

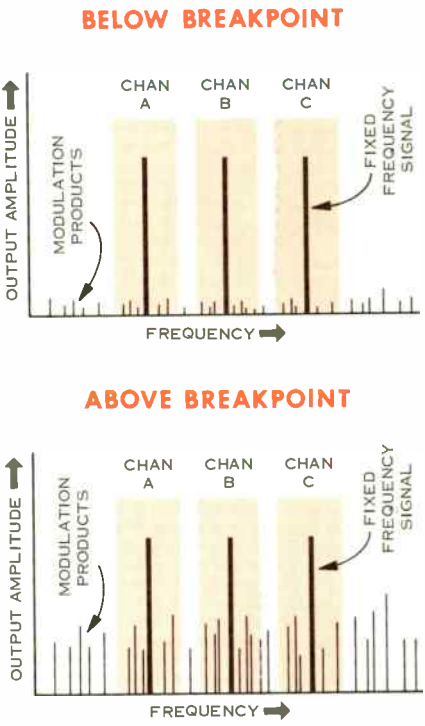


Figure 2. Signal distortion in a multichannel amplifier. Below the amplifier breakpoint, modulation products falling within a channel bandwidth are at relatively low levels and do not significantly interfere with the input signals. But above the breakpoint the magnitude of these products increases dramatically, causing excessive noise in the channels.

products will fall outside the passband, and only the odd-order products falling within the passband greatly interfere with the impressed signals. In a wideband amplifier both even- and odd-order products may cause interference. Sizable disturbances may occur in nearly all channels handled by the amplifier. They appear not as distortion of the impressed signals, but rather as a type of

World Radio History

noise whose level depends on the load in all channels.

The breakpoint, therefore, defines the instantaneous load capacity of an amplifier. The CCITT (International Telegraph and Telephone Consultative Committee) defines the breakpoint as the power at the output of an amplifier, at which a 1-db increase in input signal power causes a 20 db or more increase in power of the third harmonic. There are, however, a number of other definitions. The breakpoint of a system depends on the quality of service required. With speech, experience has shown that adequate service will be achieved if during the *busy traffic hour* the sum of all periods of overload does not exceed more than 1 percent (36 seconds).

Speech Loading

The total load applied to a multichannel amplifier is simply the sum of the loads in the individual channels. It might appear that the total speech load in an amplifier handling N number of voice channels is simply N times the load in a single channel. This, however, is not the case. There are certain peculiarities about speech transmission that tend to reduce and stabilize the total load as the number of channels increases. The net result is that multichannel amplifiers, particularly those which handle a great many channels, need not be designed to handle an extremely wide range of amplitudes — as broad a range, for example, as would be required for a single-channel amplifier. This permits using amplifiers that are considerably less complex, thereby reducing their cost.

With speech, three factors act together to determine the total load:

- the number and distribution of channels actively transmitting speech
- the volumes of speech in the individual channels
- the distribution of speech signal peaks

The first two factors cause rather slow variations in the total load and combine to equal what is known as the *maximum rms load* on the amplifier. The third factor causes instantaneous variations and results when peaks in the speech of several talkers occur at the same time.

Numerous studies have shown that in high-density multiplex systems all voice channels will rarely, if ever, be actively transmitting speech at a given instant. Some channels will be completely *idle*, meaning that they are available for an operator to complete a call. Other channels will be *busy*, in that a connection will have been made between two parties; but there will be no speech in the channel at that instant. Still others will be *active*, meaning that they are busy and speech is present in the channel. It is only the active channels that, at a given instant, contribute to the total load on a multichannel amplifier.

Both the total number of simultaneously active channels and their distribution throughout a system are questions of probability. In an N-channel system, it is possible for either zero or N channels to be simultaneously active, or for all active channels to be grouped in such a way that the total load is concentrated at a few amplifiers or within a narrow frequency band — but it is not very probable. Such extremes become even less probable in systems with greater numbers of channels. In high-density systems, the number and distribution of active channels will vary, but in most cases only between narrow limits.

Several years ago measurements were made on large numbers of telephone channels to determine exactly what per-

World Radio History

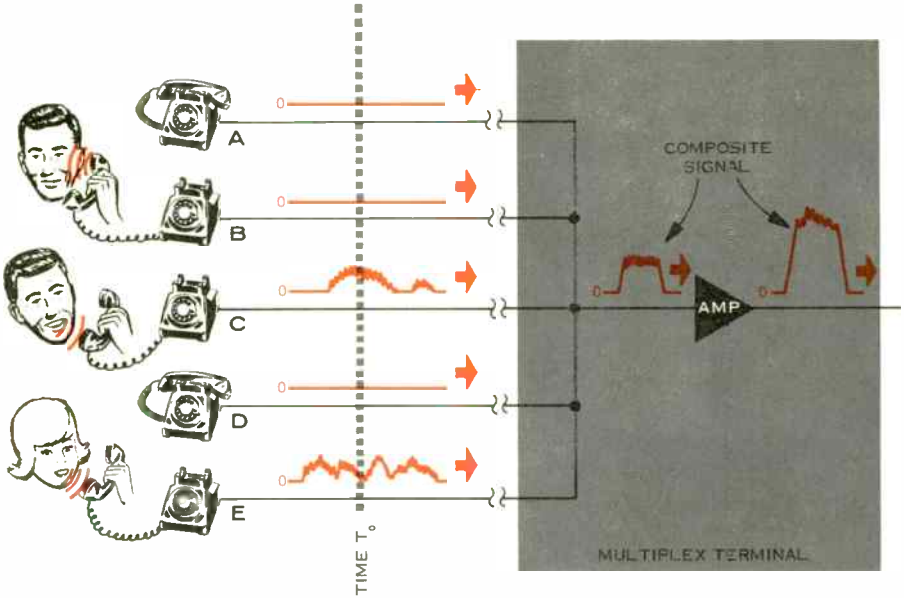


Figure 3. At a given instant in time, speech will not be present in all multiplex voice channels. Some channels will be completely idle (A and D), some will be busy but there will be no speech in the channel at that instant (B), and some will be actively transmitting speech (C and E). It is the active channels only that contribute to the total load on a multichannel amplifier.

centage of the busiest traffic hour a channel might be active. This percentage is called the *activity factor*. Results showed that in high-density systems the largest percentage of the busy hour that this might occur is about 25 percent. For smaller groups of channels this percentage may be larger, but it is highly unlikely that any increase in group size would change it appreciably. This activity factor is considered standard for multiplex systems that carry predominantly speech traffic.

Volumes of speech in active voice channels also affect the total load. Speech volumes are not constant in a channel, but vary considerably, depending on the characteristics of the talker's speech and the loudness of his voice. Obviously a system can tolerate a great many more soft talkers than it can loud talkers. An unusually high percentage of loud talkers can overload a system just as easily as an excessive number of active channels.

Speech volume is an approximate measure of the average speech energy introduced into the voice channel. This energy varies with the words and syllables spoken, but is generally concentrated in the lower voice frequencies between about 250 and 1000 cps. The amount of energy in speech is quite small compared to energy from other sources. An incandescent lamp, for example, expends almost *three million*

World Radio History

times as much energy each second as a person speaking a simple six- or seven-word sentence. Indeed it would take 500 people talking continuously for one year to produce enough energy to heat a cup of tea!

Speech energy is commonly rated in terms of the intensity level of a speaker's voice measured one meter from his mouth. The American Standards Association has adopted a reference intensity of 10^{-16} watts per square centimeter for such measurement. Numerous experiments have shown that the *average speech intensity* for all people is about 66 db above the reference intensity, with men having a slightly higher average level than women. When a person talks as loudly as possible, this level can be raised to about 86 db; and when talking as softly as possible, it can be lowered to nearly 46 db; so that from a soft whisper to a loud shout, there can be a range of 40 db.

When several channels are combined into a group, such wide variations in average power tend to average out. In groups of 64 or more channels, the total rms load will vary quite slowly, despite rather wide power variations in the individual channels. Only when an unusually high percentage of either loud or soft talkers is present will the total rms load reach extreme levels.

Changes in active channels and speech volumes are concerned only with the maximum rms load on a multichannel amplifier, causing more and more gradual variations as the number of channels increases. But it is the total input voltage applied to an amplifier, and not just the rms portion, that determines whether or not the amplifier will overload. This total voltage is the vector sum of the instantaneous voltages in the separate channels and thus is a function of both the phase and amplitude of each speech signal.

Instantaneous voltage in an active voice channel fluctuates widely, even when the volume of speech in the channel is constant. The fine structure

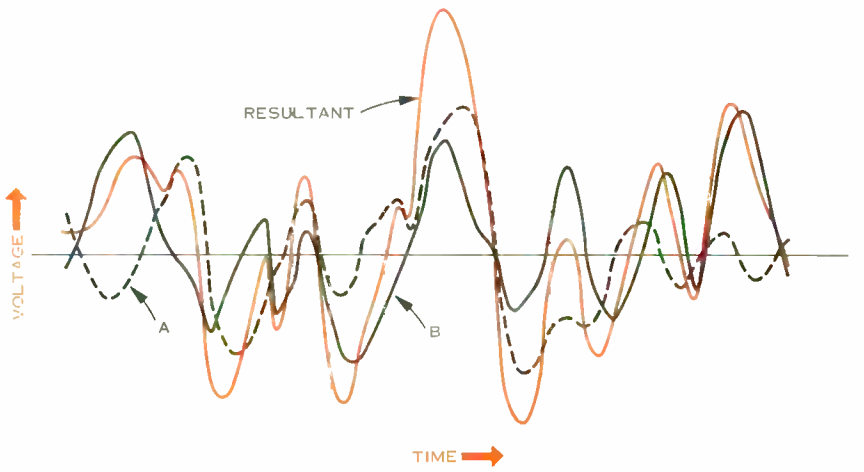


Figure 4. Phase addition of two random speech signals (A and B). High voltage peaks appear in the resultant whenever the individual signals peak together.

World Radio History

of individual speech sounds, and the differences between successive syllables and between vowel sounds and consonants, cause instantaneous variations in speech voltage. Extremes reached by speech signal peaks are very dramatic, often reaching as high as 100 times the average levels. If these peak variations are coupled with the 40-db range in average speech volumes, the peak power can conceivably range as much as 70 db — corresponding to a power ratio of *10 million to 1*.

When several channels are combined into a group, peaks tend to average out as do variations in average power. However, this averaging does not occur in all instances Since many different frequencies are transmitted, the phase relation between these frequencies varies randomly. Sometimes several frequencies will reach a peak together, causing a momentary rise in total voltage. At other times, the various frequencies may combine to lower the total voltage well below average. It can be shown mathematically that as the number of channels increases, the possibility that several frequencies will peak together decreases. Nevertheless, it always exists and must be considered in the design of multichannel amplifiers.

To prevent a few loud talkers from upsetting the balance between probable maximum instantaneous load and the load capacity of the system, individual voice channels are usually provided with peak limiters. With these devices, excessive voltage peaks are prevented from entering the system.

Data Loading

In recent years the volume of digital and analog data transmitted over communications networks has shown a substantial rise. Businesses are transmitting more data over commercial and private networks than ever before. Facsimile, teletype, and graphics are being carried more and more frequently. Consequently, a multiplex system may be called upon to handle more data than it was designed to carry, thereby decreasing its load capacity.

As mentioned earlier, most multiplex systems are designed to handle predominantly speech traffic, so design criteria are based on statistical probability. Human speech is extremely random. This characteristic, together with the fact that only a small percentage of the voice channels are active at a given instant, allows speech signals to average out in high-density systems, thereby reducing and stabilizing the load variations on multichannel amplifiers. These criteria, while in most cases suitable for a limited amount of data transmission, are far from being optimum when larger volumes of data become involved.

For one thing, data does not flow randomly into a voice channel, as does speech. The flow is either continuous, as with frequency-shift telegraph transmission, or interrupted, as with on-off type data signals. Data signals also are more or less of constant amplitude, in contrast to the wide level variations of speech. Thus, the average level of a data signal is considerably higher than that of speech, and imposes greater loads on amplifiers common to many channels.

To compensate for an excessive number of data signals in a multiplex system designed primarily for speech, one of two steps can be taken: either the levels of the data signals must be reduced; or some of the voice channels must be *dropped* or disconnected. The first alternative is not too desirable, since the data signals may suffer a decrease in signal-to-noise ratio and, consequently, an increase in error probability. The second alternative simply decreases the number of channels available for connection.

World Radio History

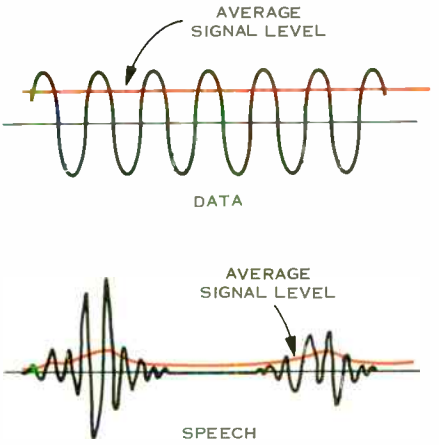


Figure 5. Data signals flow continuously and evenly into a voice channel, while the flow of speech is sporadic. Thus data has a much higher average level than speech and imposes more severe loads on multichannel amplifiers.

Determining Load Capacity

The load capacity of a high-density multiplex system depends on the types and quantities of signals the system must carry. In designing a system, extensive traffic studies are made to determine the quantities of speech, telegraph, and data signals to be handled. Based on these studies, a suitable loading formula is derived, and the power-handling capability of the system is calculated.

In actual loading calculations the input level for each type of signal must be specified. Experience has shown that the average speech power at the input to a voice channel is about −16 dbm0 (16 dbm below the power at the so-called *zero transmission level point*). Although peaks in speech power may reach high levels, tests indicate that they will not exceed −3 dbm0 more than 1 percent of the time. The average levels of telegraph and data signals are normally higher than the average level of speech. Telegraph signals are just about standardized at a level of −8 dbm0 at a channel input. Standard levels for data signals, however, are not yet firmly established. To use existing communications facilities, the input level of a data signal must be compatible with the power-handling capabilities of the average voice channel. Also, the level must be as high as possible for the signal to be relatively free of impulse noise. The exact level for data signals depends on the particular type of data transmission equipment used and the quality of service required. Normally the level selected will be between −5 dbm0 and −15 dbm0.

The above signal levels apply to the loading of a single voice channel. To determine the total load on circuit elements common to many channels it is necessary to calculate the sum of the individual channel powers. When speech is the principal type of traffic, loading formulas recommended by the CCITT may be used to determine the multichannel load. These formulas give the mean absolute power (P_m) of the distributed speech signals that the system must be capable of carrying. The signal power, as measured at the zero transmission level point, depends on the number of channels involved, and is calculated from one of two formulas:

$$P_m = -15 + 10 \log N \text{ (for } N \text{ greater than 240 channels)}$$

$$P_m = -1 + 4 \log N \text{ (for } N \text{ between 12 and 240 channels)}$$

where N is the total number of channels in the system. The formulas include a small margin for loads caused by signaling tones, pilot signals, and carrier leak, and are valid for a limited amount of telegraph transmission.

World Radio History

When a substantial amount of telegraph or data is involved, the CCITT formulas may be insufficient for determining the total system load, although they may be used to determine that portion of the load imposed by speech signals. The loads imposed by telegraph and data signals must be computed from other formulas. One formula that may be used to compute both loads is

$$P_m = P_c + 10 \log N$$

where

P_m = *rms power of the multichannel signal*

P_c = *rms power of the input data or telegraph signal, referenced to the zero transmission level point*

N = *the number of channels carrying data or telegraph signals*

The method of determining load capacity in Lenkurt's 46A multiplex system provides an example of how loading calculations are actually made. The 46A, because of the expected increase in data traffic, was designed to carry substantially more data (or telegraph) than allowed for in the loading formulas recommended by the CCITT. It was decided that the 46A should be capable of carrying speech in 75 percent of its channels, telegraph in 17 percent of the channels, and data in 8 percent.

Two important criteria were first established: the levels of telegraph and data signals at the input to each voice channel; and the maximum noise level at the output of the system. The standard level of −8 dbm0 for telegraph signals was selected, while a level of −5 dbm0 was selected for data. Actually the −5 dbm0 level is probably the highest level at which a data signal would be applied. For the total noise contribution, a design level of 23 dba0 (F1A weighted) was selected. This is the level recommended by the CCITT as the total noise contribution of a pair of terminals, when these terminals are part of a long-haul transmission system. The reference level, or 0 dba0, is equivalent to a 1000-cps tone with a power of −85 dbm.

For a single channel, speech signals impose the greatest load. Elements in a channel are not designed to carry only the average speech level of −16 dbm0, since the elements would be overloaded about 50 percent of the time. However, as stated previously, these elements will not overload more than 1 percent of the time if the channel is designed for an input level of −3 dbm0. This level was selected for the 46A. Telegraph and

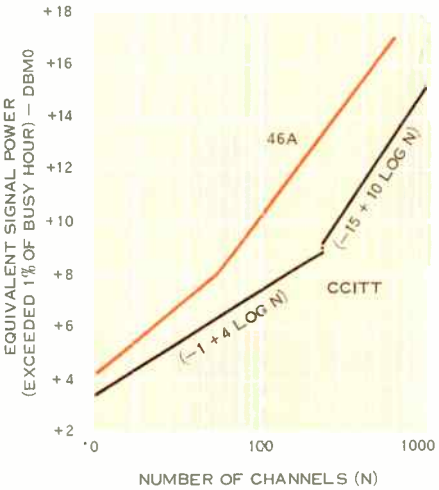


Figure 6. A comparison of the load-handling characteristics recommended by the CCITT and the characteristics of the Lenkurt 46A. The 46A will carry substantially more data traffic than allowed for in CCITT recommendations.

World Radio History

data signals would be applied at levels lower than −3 dbm0 and thus did not influence individual channel design.

In the 46A, considerable advantage was taken of the statistical distribution of speech signals in computing the total speech load on circuit elements common to many voice channels. CCITT loading formulas were used for this purpose. The total load imposed by data and telegraph signals was computed using the general loading formula for these signals. With data, P_c was made to equal −5 dbm0, while with telegraph a value of −8 dbm0 was assumed. Adding the powers of these services for the total number of channels in the system gives the total system load. When more than 240 channels were involved, the following formula was used:

$$P_m = -11 + 10 \log N$$

As stated previously the 46A was designed to carry 75 percent speech, 17 percent telegraph, and 8 percent data. However, this was only a design objective. It assumed that the input levels are −5 dbm0 for data and −8 dbm0 for telegraph, and that 23 dba0 of noise is the maximum allowed at the output. There are many other combinations of signal levels that result in the same total load on the system. For example, 67 percent of the 46A channels will carry data signals at a −5 dbm0 level, if the remaining 33 percent of the channels are disconnected from service. And if the input level of each data signal is reduced from −5 dbm0 to −10 dbm0, the 46A will carry data on 100 percent of its channels, and still have an output noise level that does not exceed 23 dba0.

If better noise performance is needed, the loading of the 46A may be adjusted accordingly. For example, if all channels are used only for speech traffic, then the output will contain approximately 19 dba0 of noise. This same noise level is also possible with data on *all* 46A channels, providing each data signal is applied at approximately −15 dbm0. However, problems may arise if data signals are applied at too low a level. Impulse noise may be too great and cause excessive errors to appear in the transmitted data.

Future Needs

In coming years, the volume of data transmitted over commercial telephone circuits should rise dramatically. As the cost of computers and data processing machines decreases, putting them within the reach of more and more businesses, data traffic will become more prevalent in telephone circuits. More widespread use of dial-operated data transmission sets, which permit fast and economical data transmission, will accelerate the pace even faster. Prominent leaders in the communications industry recently estimated that within ten years the volume of data transmitted over telephone circuits will equal, and possibly surpass, the volume of speech traffic. One thing is certain — there is an ever-growing demand for multiplex systems having a load capacity greater than today's recognized standards, and this demand is sure to increase in future years.

World Radio History

World Radio History

the Lenkurt Demodulator

VOL. 12 NO. 9 SEPTEMBER, 1963

A Discussion of "LEVELS" AND "POWERS" IN A CARRIER SYSTEM

Two of the most troublesome terms in communications language are "level" and "power." Although they are often used interchangeably in daily conversation, the two terms are not synonymous. This article discusses some of the reasons for the very common erroneous usage of the two terms, and reviews the actual meaning and proper usage of each term to show why they are necessary in communications work.

Basically, *level* is an expression of *relative* signal strength at various points in a communication circuit. *Power*, on the other hand, is an expression of *absolute* signal strength at a specific point in a circuit.

Generally speaking, the word "level" is used to indicate the value of a signal relative to an established reference signal. The reference signal is known as the "zero reference level." This general conception of level has many applications. For example, in the aircraft industry the speed of supersonic aircraft is measured with respect to the speed of sound rather than in terms of distance per unit time. The speed of sound is arbitrarily called Mach 1, and the speed of any aircraft can then be stated as a Mach Number to express that speed with respect to the speed of sound.

In telephone work the term "level" is used in a similar manner to express the relative amount of power at various

About This Article

This is a slightly condensed version of an article first published in the December, 1952 DEMODULATOR. Widely acclaimed for clarifying certain confusing terminology, the article has been reprinted by several magazines. Today the DEMODULATOR reaches some 33,000 readers in 95 countries (five times the 1952 circulation), many of them relatively new to communications. Although the article is written in terms of American telephone practice, we think it has value for anyone concerned with the operation and maintenance of a communication system.

World Radio History

points in a circuit. Just as the speed of an aircraft is expressed as a multiple of the speed of sound, the amount of power at the output of a telephone repeater can be expressed as one-half, two, or three times the power at the zero reference level.

In practice, relative levels in a telephone circuit are expressed in db (decibels) rather than in arithmetic ratios. This is done because of the convenience of using this logarithmic expression for the relatively large ratios involved. They are sometimes as great as 100 million to one (80 db).

Unless some other reference is stated, the zero reference level for a signal in a telephone circuit is the power which the signal has when measured at the two-wire input to the toll circuit.

The concept of level is illustrated in Figure 1. Since the level at the input to the toll switchboard has been defined as zero reference level, the level at the talking subscriber's subset is somewhat higher, depending on loop loss.

From the toll switchboard the transmitted speech passes to a carrier terminal which provides a gain of 17 db. The line attenuation between carrier terminals and the repeater is 42 db. Therefore, the level of the received signals at the repeater is −25 db. Since the gain of the repeater is 42 db, the transmitted level from the repeater is +17 db. With 42 db line attenuation between the repeater and the receiving carrier terminal, signals will be received at the terminal at a level of −25 db. The receiving branch of the

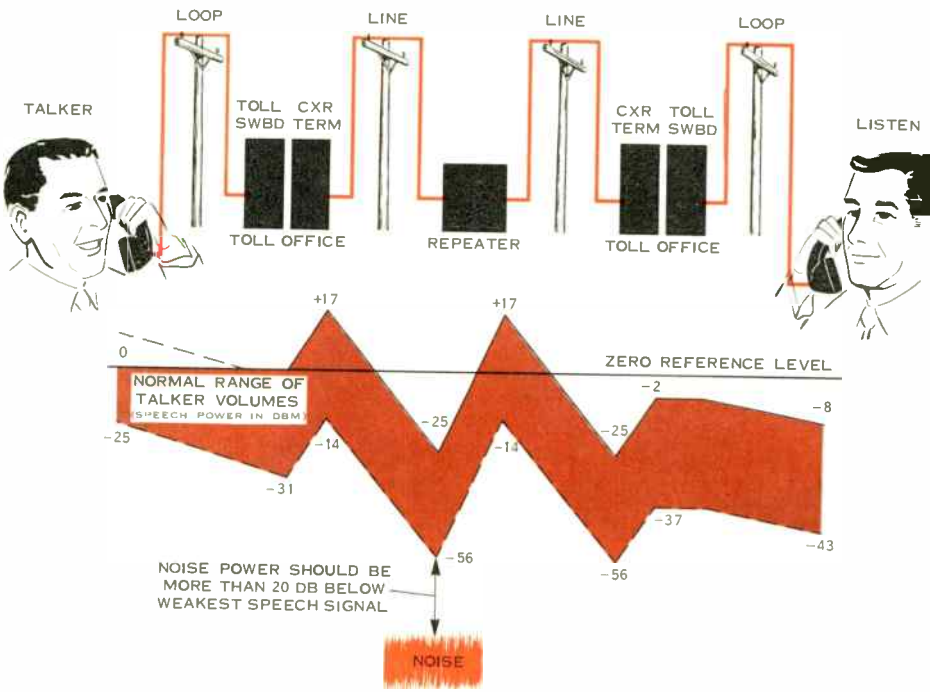


Figure 1. Relative power of transmitted signal varies widely during passage through circuit, but is measured with respect to zero reference level. Red area indicates the normal range of talker volumes.

World Radio History

carrier terminal provides a 23 db gain so the signals will be delivered to the toll switchboard at a level of −2 db. Then, since the loop attenuation is 6 db, the received level at the listener's subset will be −8 db. Thus, the range of speech signals from the talker will be heard by the listening subscriber at a level 8 to 12 db below the signal strength leaving the transmitter.

At all level points the strength of the transmitted speech has been clearly stated as having a definite ratio to the strength of the speech at the zero reference level. The statement of level at each point indicates only how much gain or loss the transmitted signals have received between the various points along the transmission path.

Level, therefore, is purely a relative term. Whenever level is expressed, the zero reference level is understood to be at the point where the circuit being considered becomes a toll circuit.

What is meant by "Power"

While level is always a ratio, "power" always designates a definite quantity. This quantity is defined in electrical terms as the rate at which electric energy is taken from or supplied to a device. The most common unit for expressing power is the "watt."

In addition to the watt, a number of other defined units are commonly used for expressing the amount of power in telephone equipment. Among these are "dbm," and "dba." Both of these units are based upon using the decibel to express the amount of power above or below a convenient amount of reference power.

Because of the use of the decibel and of a reference power in defining these units, powers expressed in decibels are sometimes erroneously called levels. They are not—because in every case a value stated in dbm or dba can be readily converted to a value in watts.

The difference between power and level can be shown more clearly by considering the use of "dbm." This unit is perhaps the most common of the three mentioned. Stating that the power at a certain point is $\pm X$ dbm simply means that the power is X db greater or less than one milliwatt.

A 1000-cps test tone with a power of one milliwatt is ordinarily available at toll switchboards. When this test tone is transmitted over a telephone circuit the test tone power in dbm at any point in the circuit is numerically equal to the level in db at that point. When the test tone power is any other amount, or when power is measured in any unit other than dbm, the numerical value of *level* is not the same as that of the power at that point. It is this similarity which can cause confusion between proper usage of level and power.

The distinction between level and power can also be illustrated by considering the two terms with respect to a fixed-gain amplifier, as shown in Figure 2. Two conditions are shown. In the first, the input to the amplifier is 0.001 watt. In the second the input to the amplifier is 0.0005 watt. In both conditions the amplifier has a fixed gain of 30 db.

In this example the input in both cases is arbitrarily considered to be zero level. Therefore, the output level in both cases is +30 db and it cannot change unless the amplifier gain changes.

The power input and the power output change in both cases, however. In the first, the input signal of 0.001 watt is amplified 1000 times to produce an output of 1 watt. In the second, the input signal of 0.0005 watt is amplified by the same amount since the gain of the amplifier is fixed at 30 db (or a factor of 1000). The output power is therefore 0.5 watt.

It is obvious that the power output

World Radio History

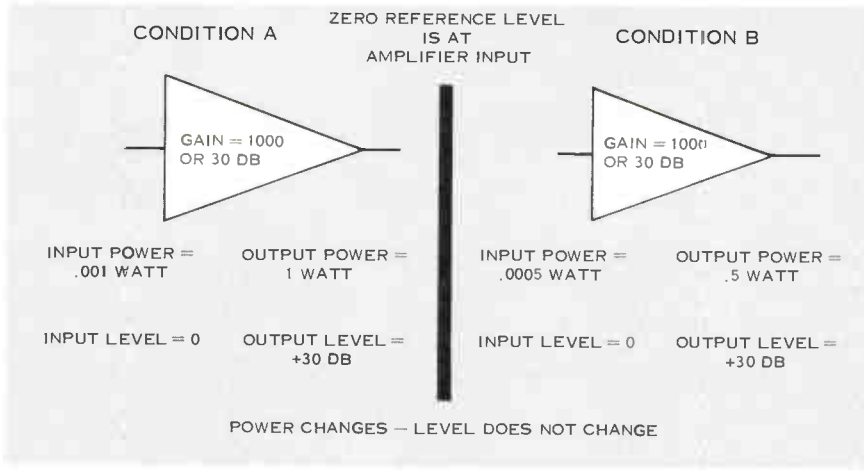


Figure 2. Halving the input power to a fixed-gain amplifier also halves the output power, but does not change the relative *output level because the zero reference level is at the input.*

of a fixed-gain amplifier will change when the input power changes. *But the level remains the same* so long as the gain or loss (in decibels) between zero or reference level and the output of the amplifier remains the same.

Why "Level" is Used

Transmitted speech consists of a large range of frequencies and powers which vary widely for different speakers. For this reason it is impossible to determine exactly what power will exist at any point in a circuit when the circuit is in use. However, regardless of the specific power at any point, the level, or in other words, loss or gain between the point in question and other points in the circuit, can be determined either by calculation or by measurement of the test tone which is transmitted at the reference level.

When laying out telephone circuits it is necessary to know the net loss which the circuit imposes on speech currents passing through it. It is neither necessary nor practical in this type of planning to know exactly what the *actual power* will be at any point, particularly since the power will vary over wide limits depending upon the talker and the words spoken.

Since the gain or loss of a circuit is independent of power (within the power handling capacity of the equipment) it is convenient to have the concept of relative level to express the relative strength of a signal at any point and to determine net loss of the circuit between any two points.

Why "Power" is Used

Although level has definite value in circuit planning, it is necessary to consider the actual power involved when designing and operating the electronic equipment used at voice and carrier frequency terminals and repeaters.

Operation of electronic equipment depends upon the minimum and maximum powers which can be supplied to the input of the equipment and delivered from the output. Equipment sensitivity and the amount of noise and other disturbances present in the circuit usually determine the lowest prac-

World Radio History

tical input power. Maximum output power (and consequently the maximum input) depends upon the power handling capacity of the equipment.

Specifications for carrier equipment normally give the test tone power at the inputs and outputs of each channel. In some cases it is desirable or necessary to know the total peak power that may be delivered to common equipment or to the line by several channels.

Per-channel power is normally stated in dbm. Because dbm is a logarithmic value, two powers expressed in dbm cannot be added to obtain the total power. Instead, each channel power must be converted to watts, added, and the total then reconverted to dbm. Alternatively, the powers can be added graphically with the aid of Figure 3.

If the per-channel power of all channels is the same, doubling the number of channels increases the total power by 3 db. Thus, if a system has 8 channels, each with a signal output power of +10 dbm, the total power delivered to the line is +19 dbm.

In speech communication circuits it is unlikely that all channels of a carrier system will be transmitting signals of the maximum value simultaneously. Therefore, the common equipment is usually designed to handle the total expected power rather than the total possible power.

Although levels are more important than actual power to the engineer laying out a telephone circuit, the transmission engineer interested in the installation or operation of a carrier system must usually know the actual power at the various points. Otherwise, there is a possibility of operating a circuit with either less input power or more output power than the equipment is designed to handle.

Any consideration of power in a carrier system can be divided into two sections—the amount of power at the connections to the carrier equipment, and the amount of power at various points inside the carrier equipment. Power values inside the carrier equipment are of interest primarily to the design engineer.

The signal power which appears at the line and drop terminations of the transmitting and receiving branches is of great importance to the operator of carrier equipment. These power values determine where carrier systems may be operated, how repeaters must be spaced, and how coordination may be achieved.

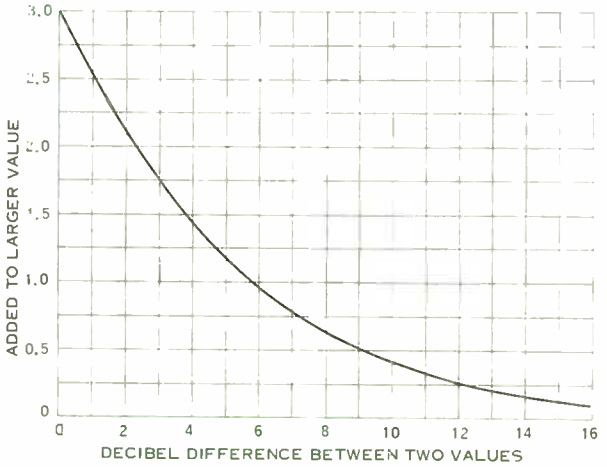


Figure 3. Graph for adding noise or signal power expressed in decibels. If signals differ by more than 16 db, smaller signal makes no significant contribution to total.

World Radio History

A typical two-wire carrier circuit with a four-wire termination at one drop and a two-wire hybrid termination at the other is shown in Figure 4. Typical test tone power values at the equipment connections are indicated.

Voice-Frequency Power

The voice-frequency (v-f) power required at the input to the transmitting branch of a carrier system is primarily based on the normal amounts of power delivered to the line from the toll switchboard. Because many telephone offices are arranged for patching circuits on a four-wire basis at a level of −16 db, and the usual test tone power at the transmitting toll switchboard is 0 dbm, the input stages of carrier systems are often adjusted to receive a test tone power of −16 dbm on a four-wire basis. This really means that the input to the carrier system is at a −16 level or at a circuit point 16 db removed from the two-wire v-f level at the transmitting toll testboard.

The amount of v-f power obtained from the receiving branch of a carrier system is also determined primarily by switching requirements. If all of the other values indicated in Figure 3 are within proper limits, the v-f output power for each channel with 0 dbm test tone at zero level would normally be about +7 dbm on a four-wire basis.

Although the traditional test tone power has long been 0 dbm, modern equipment may require lower values. As the quality of communications has improved over the years, speech volumes have tended to become lower. When this is reflected in the design of multiplex equipment, the 0 dbm test tone is excessive. Accordingly, lower test tone levels are often specified. For example, the Lenkurt Type *81A* and Panhandle Type *X* exchange trunk carrier systems specify a test tone of −10 dbm.

Transmitted and Received Power

A number of factors influence the amount of power which can be transmitted or must be received from the line. Since the difference between these

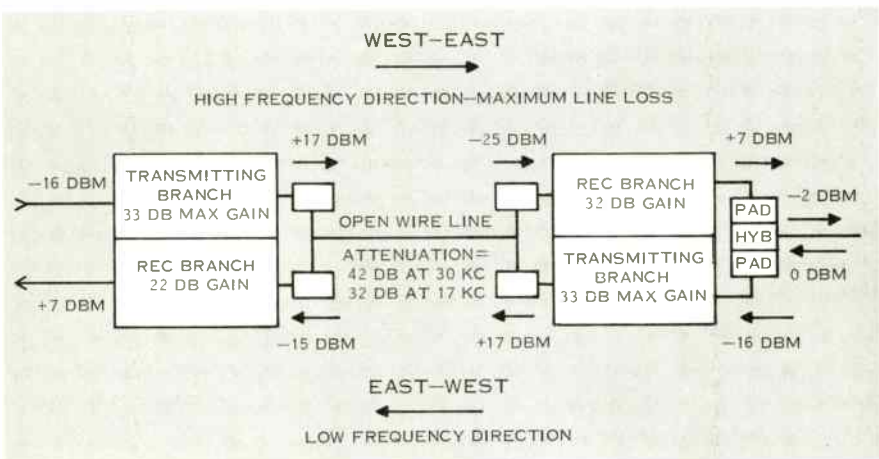


Figure 4. Simplified block diagram of a two-wire carrier system illustrating typical test tone power values at the connections to the carrier equipment.

World Radio History

two values is the maximum span attenuation, these factors also influence repeater spacing for a carrier system. Among these factors are the noise level of the line, the line attenuation, the system operating frequencies, the characteristics of the directional filters, and crosstalk considerations.

Basically, the amount of power transmitted must be high enough that sufficient power will reach the receiving terminal to permit recovery of the transmitted intelligence unimpaired by excessive noise. The power received must be sufficient so that the proper v-f output power can be delivered with the receiving branch gain available, and so that the received power will be sufficiently higher than the line noise to maintain the proper signal to noise ratio.

The amounts of power commonly transmitted were established as a result of attenuation studies conducted during many years of experience with telephone lines used for carrier circuits. These studies provided engineers with information concerning line characteristics and their effect on carrier systems under a variety of conditions. By using this information, standardized transmitted power values and levels were adopted for various carrier applications.

Among the factors which determine the minimum amount of power which should be received from the line at a carrier terminal are the receiving branch gain and the noise level of the line.

The ultimate objective of a carrier circuit is to deliver a certain amount of signal power to the telephone "drop." Therefore, the minimum received power must be such that, after being amplified an adequate amount, the signal will have the proper amount of power at the drop.

Higher receiving branch gain will not necessarily permit lower minimum amounts of power to be received since the received signal must be sufficiently greater than the noise level of the line to maintain the desired signal to noise ratio. Since noise is amplified as much as the desired intelligence, the signal-to-noise ratio at the output of the receiving branch cannot be any better than at the input.

Loop Gain and Level Coordination

Loop gain is defined as the sum of the gains experienced by a signal of a particular frequency in passing around a closed loop. The loop can be a carrier terminal, a repeater, or a complete carrier circuit.

Excessively high gains in the transmitting or receiving branches of a carrier system terminal or repeater can cause "singing." This occurs if the gain around the loop for any frequency is greater than the losses around the same loop at that frequency.

Loop gain is affected by a number of complex factors. Among them are the suppression supplied by directional filters, the losses due to hybrid balance, and the effect of the other frequency-selective elements in a carrier system. All of these factors are considered by design engineers when they determine operating levels and the amounts of power which will be transmitted and received by a carrier system when operating under various conditions.

A further limiting factor which must be considered when determining the amounts of power which will be transmitted or received by a carrier system is coordination of the levels and powers between two or more systems operating at the same frequencies on the same lead. If all systems transmit the same amount of power, they are not subjected to power differentials along the line. Any crosstalk between adjacent line conductors is then not further increased by a difference in power. •

World Radio History

World Radio History

the Lenkurt Demodulator

VOL. 8 NO. 7 JULY, 1959

TELEPHONE SIGNALING

Telephone signaling is of vital importance to any telephone system. This importance grows as the size of the telephone system increases. Only through the medium of signaling can the various parts of a telephone network act together to enable efficient service. Although basic signaling principles are simple, signaling is often regarded as being complex and formidable. This probably stems from the variety of signaling methods that are used.

This article reviews some signaling fundamentals, describes some of the differences between subscriber and inter-office signaling methods, and illustrates common techniques of handling telephone signaling over carrier systems.

A telephone circuit, in addition to transmitting telephone messages, must handle certain *telephone signaling* functions necessary for the operation of the telephone system. Telephone signaling is to the telephone system what the sympathetic nervous system is to the body. The sympathetic nervous system regulates many internal functions related to the proper operation of the body, such as blood pressure, digestion, and others. In telephone practice, signaling provides the internal management and supervision of the system. Unlike the body, which provides special paths for its internal signaling, telephone companies must make one set of "nerves" do all the work—the same lines that handle the customer's messages must also carry signaling functions.

Even the most primitive telephone circuit requires some form of signaling to announce that one terminal desires to talk to the other. In more complex systems, a variety of signaling functions may be required. In general, these may be classified as signals used for estab-

World Radio History

lishing a connection, and signals that pass information concerning the status of the circuit or call.

Traditionally, different techniques have been used for signaling over a telephone subscriber's local circuit or loop than have been used between offices, and over long distance toll circuits. Different techniques were used because of the different signaling requirements existing for each type of path.

Subscriber Loop Signaling

Signaling falls into three basic categories: supervisory signals, information signals, and control signals. *Supervisory* signals tell the telephone office that a connection is desired, or that it is no longer being used. *Information* signals tell the subscriber or the operator the status of the call, or the condition of the circuit. *Control* signals provide directions for establishing the desired connection.

One of the earliest methods of signaling employed a magneto or a-c generator located at the subscriber's telephone. The desired supervisory signal was transmitted by cranking the magneto. The resulting 15- to 20-cycle a-c voltage attracted the operator's attention by actuating a "drop" on the switchboard. The "drop" was an indicating device consisting of a panel light with a hinged metal plate in front of it. The metal plate was hinged at the bottom and latched at the top in such a fashion that a ringing signal would release the latch and allow the plate to swing down, uncovering the light. When the call was completed, another crank of the magneto notified the operator that she could break the connection. Magneto supervision is still employed in some locations, particularly where the subscriber is remote from the central office. In such a case, each telephone instrument may have its own battery for supplying voice current power.

A much more commonly used type of supervision is known as *common battery supervision*, which has nothing to do with the quality of the batteries used. When the customer lifts his instrument from its hook, a d-c path is completed through a loop consisting of the customer's instrument, a battery in the telephone office, and a line relay.

Figure 1. First automatic telephone employed both magneto and push-button signaling for establishing desired connection.

World Radio History

TYPE OF SUPER-VISION	CALLING END		CALLED END		BATTERY SOURCE
	SEIZURE (CONNECT) SIGNAL	DISCONNECT SIGNAL	ON-HOOK	OFF-HOOK	
HIGH-LOW	CLOSED LOOP	OPEN LOOP	HIGH RESISTANCE	LOW RESISTANCE	CALLING END
LOW-HIGH	CLOSED LOOP	OPEN LOOP	LOW RESISTANCE	HIGH RESISTANCE	CALLING END
REVERSE BATTERY	CLOSED LOOP	OPEN LOOP	NORMAL LINE POLARITY	REVERSE LINE POLARITY	CALLED END
REVERSE HIGH-LOW	CLOSED LOOP FOLLOWED BY BATTERY REVERSAL	BATTERY RE-TURNS TO NORMAL POLAR-ITY, FOLLOWED BY OPEN LOOP	HIGH RESISTANCE	LOW RESISTANCE	CALLING END
WET-DRY	CLOSED LOOP	OPEN LOOP	BATTERY AND GROUND CONNECTED TO LINE	BATTERY AND GROUND REMOVED	CALLED END

Table 1. Typical Loop Signaling Systems.

When the customer closes the loop, the relay is energized, thus connecting the customer's line (or loop) to the operator's board, or to automatic dial equipment. When the calling subscriber replaces his instrument on its hook, the circuit is broken, and the relay releases the circuit. This method of supervision gets its name from the battery which is common to both the office and the subscriber. Table 1 summarizes various loop signaling methods.

In order to know whether the called party's instrument is on-hook or off-hook, some means of supervision is required by the central office. The most widely used method is known as *reverse battery supervision.* When the called party lifts his instrument from the hook, a relay in the office trunking equipment responds to the closed loop, actuating a supervisory relay that reverses trunk polarity. This battery reversal is detected by a polarity-sensitive relay at the calling end of the circuit. The caller's trunking equipment may reverse the polarity of the caller's loop. Figure 2 illustrates a typical reverse-battery method of supervision.

Signaling Between Offices

Interoffice signaling usually involves different considerations than in the subscriber loop signaling described above. In many cases, the use of carrier or radio to provide the interoffice circuit precludes loop signaling methods. In

World Radio History

TYPE OF SIGNALING	DIRECTION OF SIGNALING				
	TO LOCAL OFFICE	TO CUSTOMER	INTEROFFICE	TO CALLED PARTY	FROM CALLED PARTY
INFORMATION SIGNALS	PARTY IDENTIFICATION	AUDIBLE RINGING	DEPENDS ON METHOD OF TRANSMISSION		RECORDER WARNING TONE
	COIN DEPOSIT	LINE BUSY			COIN DENOMINATION
	COIN DENOMINATION	PATHS BUSY			COIN DEPOSIT
	RECORDER WARNING TONE	NO-SUCH NUMBER			
	NUMBER CHECKING TONE	REVERTING TONE			
		MACHINE ANNOUNCEMENT			
		HOWLER TONE			
		TEST TONE, DIAL			
		TRANSMISSION TEST TONE			
SUPERVISORY SIGNALS	OFF HOOK	SEIZURE (CONNECTION) OR HOLD		RINGING	ON HOOK
	ON HOOK	DISCONNECT		DISCONNECT	OFF HOOK
	FLASHING	RECALL			FLASHING
CONTROL SIGNALS	DIGITS (VERBAL OR DIAL)	DIAL TONE OR OPERATOR REQUEST		COLLECT OR RETURN COIN	
	COLLECT OR RETURN COIN				

Table 2. Typical signals appearing in local loops.

World Radio History

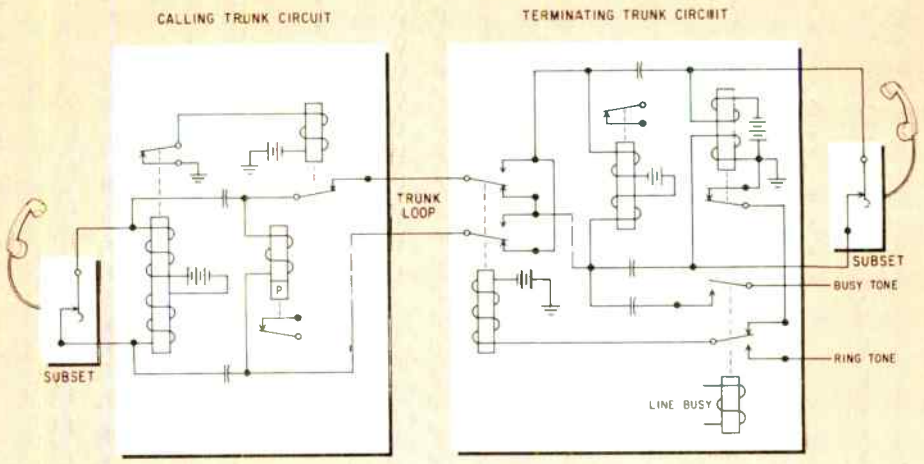


Figure 2. Typical reverse battery loop supervision circuit. Only supervisory relays are shown.

such cases, some form of *E & M* signaling may be used.

E & M signaling is characterized by the use of separate paths for the signaling and the voice signals. *E & M* signaling acquired its name from arbitrary letter designations appearing on early circuit drawings for systems using this type of signaling. The *M* lead transMits ground or battery to the distant end of the circuit, while incoming signals are recEived as either a grounded or open condition on the *E* lead. Thus, the *M* lead reflects local conditions, while the *E* lead reflects the conditions existing at the far-end of the circuit. Various simplex, duplex, composite, and other circuit arrangements have been devised to permit *E & M* signaling between offices on a d-c basis.

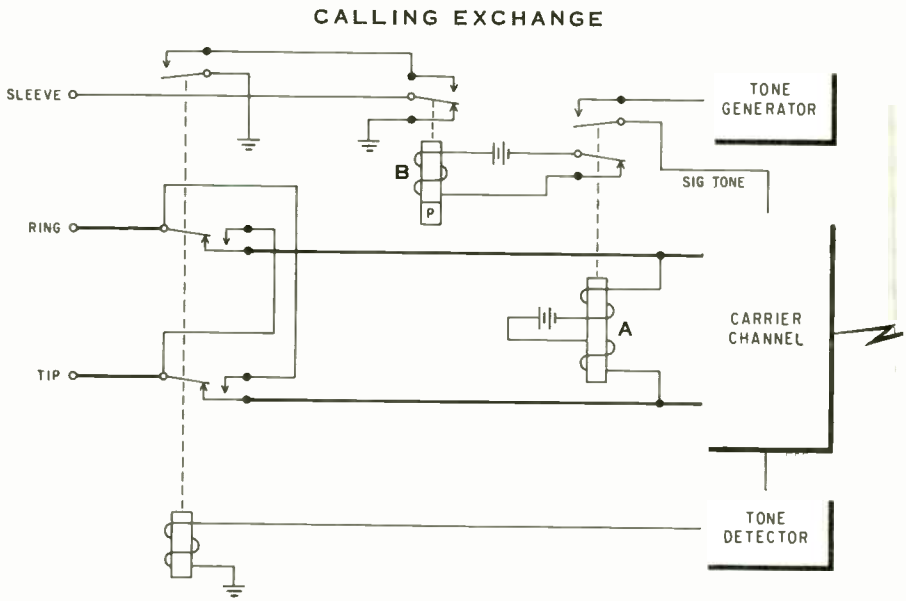
Carrier Signaling

Although many signaling functions depend on d-c or metallic signaling paths, obviously these paths cannot be carried over radio or carrier links. Carrier systems must, therefore, translate these d-c signals into a form that can be transmitted, then restore them to their original form for use by the office trunking equipment.

Figure 3 illustrates one approach to transmitting loop-dial, reverse-battery signaling over a carrier system. In this typical carrier system, a single signaling tone is used. Carrier signaling tones always lie within the passband of the carrier channel equipment, but may be either within the channel voice band or just outside it.

In the system shown, a signaling tone is transmitted over the carrier system in both directions. When a party at one end of the system picks up his telephone to make a call, his d-c loop is closed. This applies battery to relay *A*, causing it to close. One set of contacts removes the signaling tone from

World Radio History



the outgoing path, thus preparing the circuit for dialing and for the voice-frequency message that will follow the dialing. Another set of contacts on relay *A* energizes relay *B*, thus grounding the sleeve lead of the local telephone circuit. At the far end, incoming signaling tone is constantly monitored and used to keep the far-end *D* relay energized. When the signaling tone is interrupted by the sending telephone being removed from its hook, far-end relay *D* is de-energized. This connects a terminating resistor and a series diode across the tip and ring of the far-end loop. In addition, relay *C* and a series diode are also connected across the tip and ring of the far-end loop. The two diodes are connected so that current may flow through either the terminating resistor or the *C* relay, but not both. Loop polarity determines which will conduct. At first, current flows through the terminating resistor, used to provide a partial termination to the carrier terminal hybrid circuit.

When the calling subscriber pulses his transmit A relay by dialing, the signaling tone transmitted over the carrier system is also pulsed. This causes the receive *D* relay to repeat these pulses through the terminating resistor and the local trunking equipment.

When the called party answers, far-end trunking equipment in the office reverses battery polarity across the line. This reversed polarity cuts off the conduction through the terminating resistor because of the polarity of its series diode, and permits current to flow

World Radio History

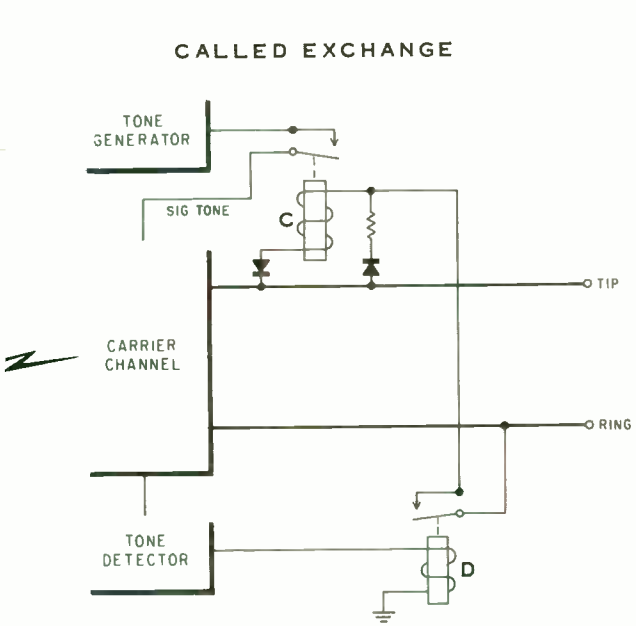


Figure 3. Typical arrangement for single-tone signaling over carrier channel between exchanges.

through the *C* relay. The *C* relay is energized, thus stopping the transmission of signaling tone back to the calling end. At the calling end, interruption of signaling tone permits the *B* relay to de-energize. This reverses polarity of the transmit line and provides the required line supervision to the transmitting office equipment.

Conclusion

In a system like the one just described, signaling units associated with the carrier equipment substitute for the metallic paths over which d-c signaling is normally carried. In other carrier systems, similar approaches are often used. Although the signaling principles may be similar, the specific circuit arrangement used at any of the nation's 3,600 independent telephone companies may differ from that used by a neighbor in some detail such as polarity. Carrier equipment linking two systems employing different signaling methods must be designed to accommodate and translate one signaling method to the other. This requires careful planning and design of carrier signaling equipment. The carrier equipment designer's problem is magnified by the need for adapting standard carrier equipment to private communications systems which may use signaling methods not normally encountered in public telephone service. For this reason, carrier equipment designed for subscriber service often includes provisions for ready substitution of signaling components.

World Radio History

World Radio History

the Lenkurt® Demodulator

VOL. 12 NO. 3 MARCH, 1963

Signaling in Carrier Channels

Considering the simplicity of the functions performed by signaling in a communications network, it is surprising that they are accomplished in so many ways. Although these jobs are simple, they are most vital to the basic operation of the network.

In carrier systems, the methods by which supervisory signals are achieved have an important effect on the quality of transmission and the cost of the equipment. This article reviews some of these fundamental considerations of carrier signaling.

Signaling provides the "nervous system" of the communications network — reporting needs and bringing response. It coordinates the various parts so that they can operate together. Signaling provides the means for managing and supervising the communications system; it establishes connections, helps select the route (when there is a choice), announces the incoming call, reports the fact if a line is busy. Without signaling, the system could not operate.

Signaling can be done in many ways. The body sets up separate networks of nerves, completely independent of those which report the senses and direct the muscles. Some communications systems do the same, using separate channels to convey information used in controlling the operation of others. More often, however, it is more economical and much more flexible if each channel carries its own signaling. In local (physical or metallic) telephone circuits, this can be achieved by direct currents which share the line with the signal voltages. In multi-channel carrier transmission, however, different techniques are required. Voice channels customarily occupy a bandwidth of 4 kc. Some of this bandwidth is used for isolation from adjacent channels. The rest, usually about 3700 cycles, must carry both speech and signaling. In some cases, this channel bandwidth is used not only for voice, but also for one or more telegraph or teletypewriter circuits — so-called "speech-plus." In such cases, channel filters must be designed to pre-

World Radio History

vent mutual interference between the speech, telegraph, and supervisory signals which share the channel.

Despite its importance, supervisory signaling has very little information content. Most signaling operations occur in establishing a connection, and at the end of the call. Generally, no signaling functions are required during the conversation itself (except that of maintaining the connection). This fact makes possible one of the most important signaling methods.

In most carrier systems, one of three basic methods may be used for signaling: "in-band", "out-of-band", or "separate-channel" signaling. In general, separate-channel signaling is only used on very high density "backbone" routes or under special circumstances where signaling cannot be conveniently handled with the communications channels themselves. Such an application is found in certain submarine cables where "TASI" (Time Assignment Speech Interpolation) is used to squeeze extra channel capacity from the idle time present in most voice communications. (For a more detailed description of TASI, see DEMODULATOR, August, 1961.) Although separate channel signaling has the advantage of leaving the entire voice channel free for communications, without the possibility of mutual interference between the speech and signaling, it is rather uneconomical since it requires that certain channels be set aside to handle only signaling functions. In addition, repair and maintenance is more complicated when signaling and speech are sent over separate channels. Reliability may be less since both channels are subject to failure independently of each other.

The two most widely used signaling methods are the so-called *in-band* and *out-of-band* methods. With out-of-band signaling, channel filters are designed with an upper cutoff frequency well below the top edge of the channel. This leaves a portion of the spectrum free to transmit signaling tones. Generally, a single tone is used and this is keyed to convey signaling information.

Some equipment takes advantage of the existence of a separate signaling channel above the voice frequency portion to perform other functions. In the Lenkurt 45-class equipment, for example, *two* tones can be used. Signaling information is transmitted by alternations between the two tones. Since one or the other of the two tones is always present, it becomes possible to use the signaling tone as a reference pilot for regulating the individual channel level.

By completely separating signaling from the speech portion of the channel, it is possible to maintain relative freedom from mutual interference between the speech and the signaling tones; signaling tones can be transmitted during the conversation, thus permitting extra functions such as regulation, which might be desirable during the period the speech channel is in use.

In addition to being more flexible, out-of-band signaling can be much easier and more economical to accomplish, particularly if some sacrifice in

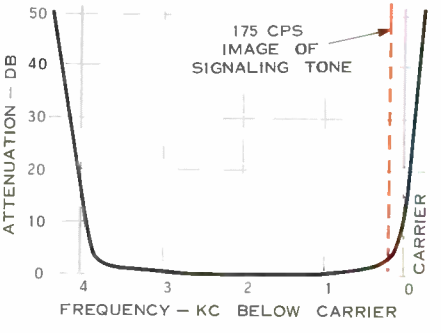


Figure 1. Typical carrier channel band-pass characteristic. Note that 175-cycle image tone, shown in solid red, is attenuated about 45 db.

World Radio History

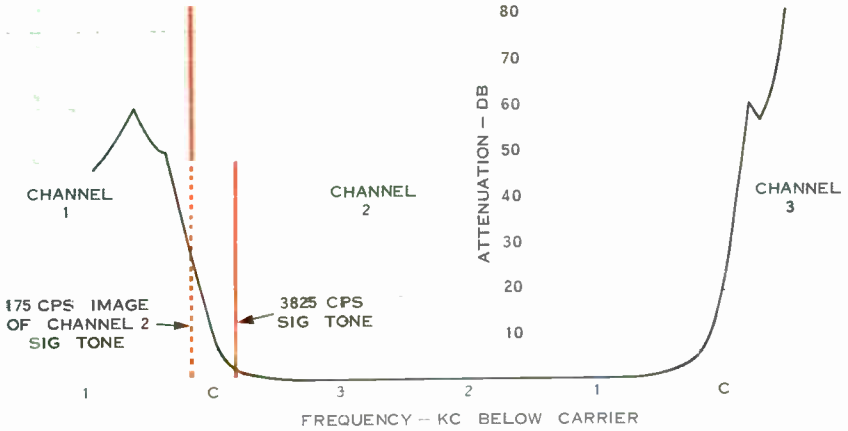


Figure 2. Relative response and attenuation characteristics of three adjacent SSB channels spaced 4 kc apart. When 3825-cps signaling is demodulated, images fall in adjacent channel at 175 cps. When transmitting, upper sideband images of 175-cps speech components could cause false signaling if inadequately filtered.

channel bandwidth is allowed. In telephone circuits, there is very little speech energy present at the upper end of the channel. Accordingly, filtering requirements may be somewhat relaxed (since telephone instrument weighting also provides a degree of "filtering"). This makes it possible to provide good quality transmission for relatively little equipment cost, since the greatest cost of carrier systems is in the channel filters. In general, efforts to increase bandwidth by approaching the channel edges more closely increases the cost of the carrier equipment.

Optimum Design

A careful compromise between speech quality and equipment cost is required. If the out-of-band signaling tone is too low in frequency, the restriction on bandwidth may impair speech quality. If the signaling tone is raised in frequency so that it lies close to the top edge of the 4-kc band, channel filters must be made more complex.

A case in point is the use of a signaling frequency of 3825 cycles per second, standard in many countries. Where channel filters are not sufficiently effective, a 3825-cps signaling tone appears as a 175-cps tone in the adjacent channel, but at a fairly low level. This results from the fact that the 3825-cycle tone falls at the same frequency as the "image" or unused sideband of a 175-cps tone in the adjacent channel. Although filters essentially eliminate this sideband, even the best channel filter characteristics are far from ideal. Figure 1 shows typical attenuation characteristics. Note that although the 175-cps component from the next channel can appear following demodulation, it is attenuated about 45 db. Although attenuated to a low level, the tone may still be audible and disturbing. In such cases, it is necessary to provide additional filtering following the demodulator in order to eliminate the tone.

Conversely, if a high-pass 175-cps filter is not used prior to modulation, speech energy at this frequency may

World Radio History

appear as 3825-cycle energy in the adjacent channel, thus causing false signaling. In order to attenuate the 175-cps component adequately with conventional filters, the low frequency channel cutoff is approximately 300 cps. In accordance with conventional practice, the high frequency cutoff remains at approximately 3400 cps. Although this frequency range provides acceptable performance over most communications circuits existing today, longer circuits would be excessively degraded. For instance, current CCITT* Standards call for a 2500 kilometer hypothetical path with 3 audio drops or links. At present, new standards are being prepared which call for a 25,000 kilometer path with 12 links. The cumulative effect of the repeated filtering that would be required to prevent adjacent channel interference would result in a channel characteristic like that shown in Figure 3. This type of frequency characteristic reduces intelligibility. Accordingly, in order to maintain high standards of speech quality, it would be necessary to employ channel filters having much sharper attenuation characteristics, thus raising the cost of equipment significantly.

One way of overcoming this problem is to lower the frequency of the out-of-band signaling tone so that it is farther from the edge of the 4-kc band. However, this lowers the highest speech frequency that can be transmitted. Thus, the voice frequency cutoff might be reduced from 3400 to about 3300 cps. This would allow the signaling tone to be reduced from 3825 cps to say, 3700 cycles. Although the top frequency would be lower, this permits the lower cutoff frequency also to be reduced, perhaps from 300 cps to about 200 cycles. This provides a dual advantage: the image frequency of the signaling tone goes from 175 cps to 300 cycles, well within the effective filter rejection capability. Because of the increased attenuation of the signaling image frequency, supplementary filtering following the demodulator is not required. Actually, quality is improved. Subjective tests by the British Post Office and Bell Laboratories have shown that a frequency band from 200 cps to 3050 cps provides better quality than the band ranging from 300 to 3150 cps, where channel bandwidth must be restricted, as in submarine cables. Studies have shown that voice intelligibility is more dependent on the low frequency end of the voice band than on the high frequency end.

One disadvantage of out-of-band signaling is that it requires some sort of d-c repeater at the end of each link. That is, the signal pulses are detected

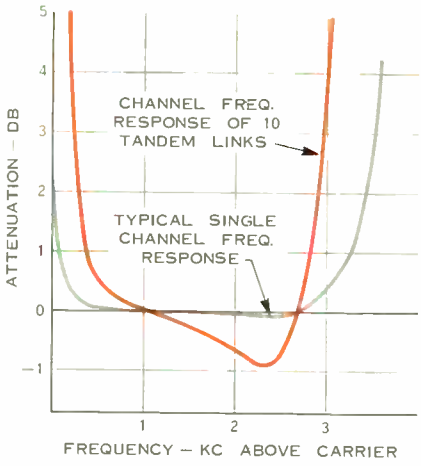


Figure 3. Typical end-to-end channel frequency response is shown by black curve. Variations are retained by each tandem link and tend to become exaggerated as they are repeated. Red curve shows typical degraded frequency response of ten better-than-average channels connected in tandem.

*Consulting Committee for International Telephony and Telegraphy.

World Radio History

and made to operate a relay. The relay, in turn, keys the signaling equipment in the succeeding link. Thus, signaling terminals are required at both ends of each link. This has the disadvantage of increasing the cost, complexity, and possible distortion of the signals.

In-Band Signaling

There is a growing trend toward greater use of in-band signaling. This appears to be a natural evolutionary step away from the use of separate channels for signaling. In the earliest days of carrier transmission, all signaling equipment was completely separated from the equipment used for voice transmission. Out-of-band signaling, an intermediate step in this evolution, brought the speech and signaling together in the same communications channel, while keeping a "barrier" between them.

With in-band signaling, the two are even more intimately merged. Signaling tones are transmitted at a frequency within the speech band, usually either 1600, 2400, or 2600 cps. The principal objection to in-band signaling is that the signaling tones lie right in the speech band. This leads to the possibility that speech energy at the signaling frequency may be able to "talk-down" the signaling, that is, cause false signals with voice energy. Conversely, signaling tones are audible and thus cannot be used during conversation.

The biggest advantage of in-band signaling is the extreme flexibility that it provides. The speech and supervisory signals share the same transmission facility, but at different times. The system is arranged so that supervisory signals are on the line only before and after a call. Since the signaling becomes a part of the transmission, it is not necessary to use d-c repeaters when going from one link to another. At branching points, a similar flexibility is obtained. The lack of d-c repeaters eliminates the delay and pulse distortion which characterize out-of-band signals sent through several links.

In-band signaling provides unusual flexibility and economy in large offices.

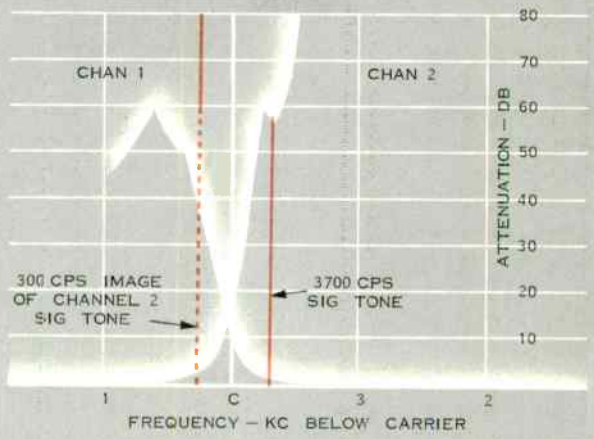


Figure 4. Greatly improved isolation between adjacent channel speech and signaling results when out-of-band signaling frequency is lowered to 3700 cps, thus shifting image to 300 cycles. Signaling and speech signals are much more effectively attenuated by channel filters.

World Radio History

Since the signals are carried over the speech circuit, it is unnecessary to cable the so-called E & M (receive and transmit) signaling leads through the office. The signaling equipment can be associated directly with the switching equipment, thus allowing a trunk circuit to be obtained from any available transmission medium, rather than being restricted to certain carrier systems.

Preventing Talkdown

In order to prevent spurious signaling by voice energy, a "guard" circuit is commonly used to distinguish between speech and signaling tones. Typically, the guard circuit consists of a network which detects the presence of other frequencies. When other frequencies are present, the guard circuit "assumes" that the signaling tone frequency is caused by speech and therefore prevents signal circuit response.

Further protection can be obtained by proper choice of frequency for the in-band signaling tone. In general, it is desirable to use the highest frequency that can be transmitted easily through

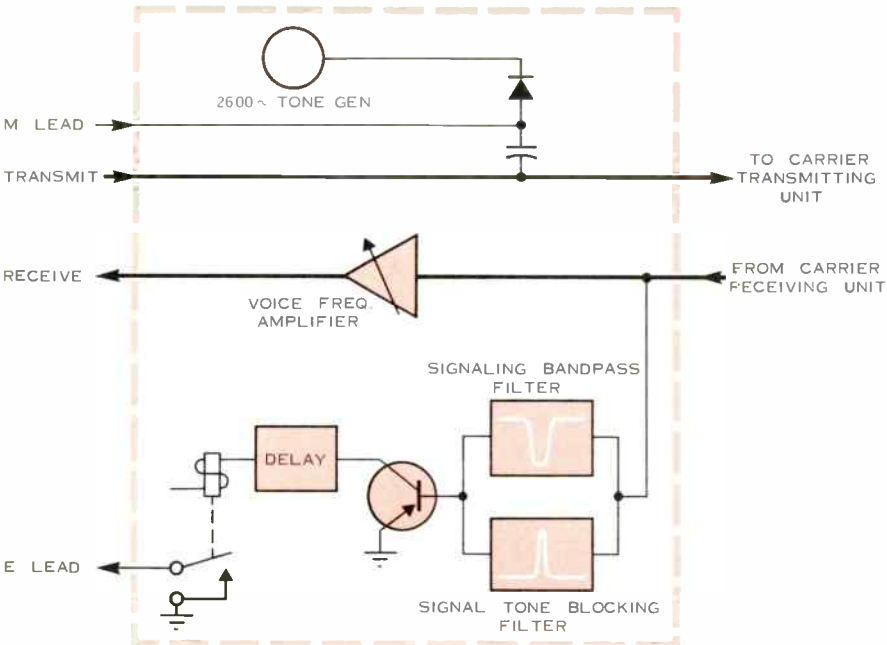


Figure 5. Simplified block diagram of a typical in-band signaling guard circuit designed to prevent "talkdown" or signaling imitation by speech. M lead in transmitting circuit biases diode to cut-off during off-hook condition, keys 2600-cps tone during dialing. At receiver, presence of 2600-cycle tone alone permits transistor to conduct, energizing the signaling relay, and lifting E lead from ground. Presence of other frequencies indicates speech. Energy from the signaling tone elimination filter causes transistor to cut-off, de-energizing the relay and grounding the E lead. Time delay in relay circuit reduces the chance of noise simulating speech and causing improper disconnect.

World Radio History

the worst transmission channel that might be used. Speech energy declines rapidly at the higher frequencies, thus reducing the likelihood of talkdown. Certain older carrier systems have a cut-off frequency near 2800 cycles. For this reason, one of the most commonly used in-band signaling frequencies is 2600 cycles per second. Speech energy at 2600 cps is relatively low.

As a further precaution, a brief time delay on the order of 30 milliseconds reduces the likelihood of speech or noise energy causing spurious signals. Normally, most noise frequencies are very transient. By introducing a delay, the circuit is made relatively insensitive to noise energy at the signaling frequency.

Time Division Signaling

In some new carrier or multiplex systems, a different method of transmitting supervisory signals may be employed. For instance, in Lenkurt's low-cost 81A Exchange Trunk Carrier System, all 24 voice channels share a common signaling channel, with time division providing separation between channels. Each of the signaling leads is connected to a sampling gate. Each channel is sampled in sequence, and the presence or absence of a signaling tone is transmitted to the receiver. In this particular system, signaling is transmitted by shifting the frequency of the level-regulating pilot (which would be transmitted anyway).

At the receiver, the incoming pulses are sorted and distributed to the appropriate channels. Although this arrangement lacks the flexibility of in-band signaling, it does provide unusually reliable and economical signaling without encroaching on the bandwidth available for each channel. By associating most of the signaling functions with the common equipment, cost of the system is substantially reduced without reducing quality or reliability, an important consideration in short-haul systems.

Similarly, time division multiplex systems now appearing commercially transmit PCM code pulses to represent speech information in each channel, then transmit an additional impulse for signaling information, in much the same manner as in the Lenkurt 81A equipment. The method used in the 81A system is the time-division equivalent of separate-channel signaling, while the time division multiplex system is the equivalent of out-of-band signaling. Perhaps this more sophisticated use of combinations of modulation methods for transmitting supervisory signals represents the next major step in the evolution of efficient signaling. •

BIBLIOGRAPHY

1. T. H. Flowers and D. A. Weir, "Influence of Signal Imitation on Reception of Voice-Frequency Signals," *Electrical Communications* (Technical Journal of I.T.&T.), Vol. 26, pp. 319-337; December, 1949.
2. H. Fletcher, "Speech and Hearing in Communication," Chapters 1 to 6, D. Van Nostrand Company, 1953.
3. R. S. Tucker, "Sixteen-Channel Banks for Submarine Cables," *Bell Laboratories Record;* July, 1960.
4. J. C. Christensen, R. W. Ruth, and S. A. Welk, "The Lenkurt In-Band Signaling System," Paper No. 62-1119, *AIEE Summer General Meeting;* June 17-22, 1962.

World Radio History

World Radio History

the Lenkurt Demodulator

VOL. 8 NO. 5 MAY, 1959

AMPLIFIERS

For Carrier Applications

The entire communications industry and many of the industries it serves depend on the existence of amplifiers. The nature and design of amplifiers used in carrier applications is important, for amplifiers have a profound effect on the cost and performance of a carrier system. This article considers some of the characteristics of amplifiers and how they are used in a modern carrier system.

Without the amplifier, vaudeville would still be restricted to theaters and showboats instead of illuminating millions of living rooms each Sunday evening. Long distance telephony would be impractical or impossible; indeed, all long distance communications would be reduced to the dots and dashes of telegraphy. The amplifier has similarly transformed many other industries, largely by improving their ability to communicate.

In communications, the amplifier's value stems from its ability to restore the strength of signals weakened by transmission loss. The basic characteristic of an amplifier is that it enables a very small electrical signal to control a much larger quantity of electrical energy. Thus, an amplifier can transform a weak signal to a large signal of the same form.

An ideal amplifier would produce an output signal exactly duplicating the original signal in all respects except magnitude. Practical amplifiers fall short of this ideal by introducing some form of distortion. Generally speaking, the greater the amplification required of an amplifier, the more likely it is to distort the signal. This may not be too important in amplifiers used alone, such as hearing aids, public address systems, or intercoms. But in applications where many amplifiers are used in tandem, such as long distance communications circuits for telephone and television, distortion becomes intolerable. In transmitting messages over long distances, amplifiers (repeaters) are spaced along the path at intervals as short as 2½ miles to restore lost signal strength. In crossing the country, a signal may pass through more than a hundred repeaters

World Radio History

and be amplified in power billions of times.

If the first repeaters introduce distortion, even in very small amounts, this distortion receives the same amplification as the signal. Since each additional repeater adds its own distortion to that already imposed on the signal, the message might well be unintelligible by the time it reached its destination if strong measures were not taken to limit distortion.

There are numerous ways of obtaining amplification. The familiar electron tube and transistor are used in the great majority of all amplifiers. Until the last ten years, the electron tube was almost alone in performing the world's amplifying chores. Today, the transistor is enjoying a rocket-like climb as state-of-the-art technical improvements permit the transistor to replace the electron tube in more and more applications. Increased life and reliability, as well as reduced power and space requirements, provide a powerful motive for increased use of transistors in carrier equipment.

Other less well known types of amplifiers include the magnetic amplifier, parametric amplifiers, and the maser. Magnetic amplifiers have been around for years, generally used for control and regulation. Battleships in World War I used magnetic amplifiers to control the position of gun turrets, as did bombers in World War II. In communications, magnetic amplifiers are used principally in regulating the output voltage of power supplies.

The newest amplifiers are the parametric or variable-parameter amplifiers. The parametric amplifier obtains its operating power, not from the conventional d-c source, but from a continuous-wave frequency, usually at some multiple of the signal frequency. This "pump" frequency continuously varies a reactance in the amplifier circuit in such a way that pump-frequency power is converted to signal-frequency power, thus producing signal amplification. This type of amplifier provides a great reduction of noise in UHF and microwave radio service.

The maser is a special form of parametric amplifier that takes advantage of

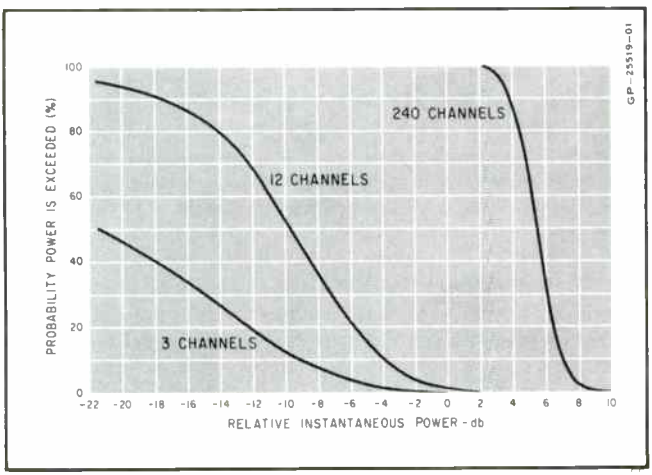


Figure 1. Probability of signal power being exceeded for different numbers of channels. Note that 3-channel system has 22-db range, while 240 channel system has less than 8-db range.

World Radio History

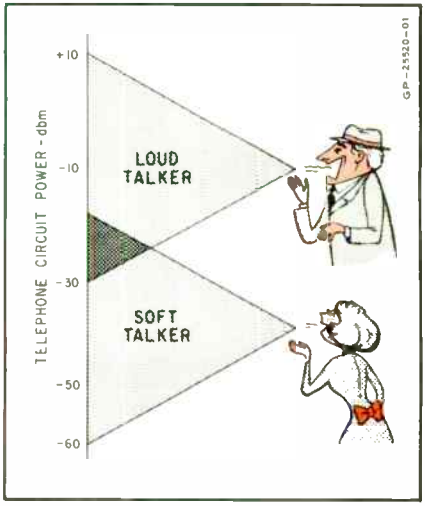
the special properties of certain substances at very low temperatures. Under the right conditions of temperature and magnetic field, these substances will absorb energy from a very high pump frequency and release it to the signal frequency, thus achieving amplification In this particular type of amplifier, thermal noise is reduced to levels which, until recently, were believed impossible to achieve. Despite the inconvenience of using liquid nitrogen or liquid helium to keep the maser properly cooled, the spectacular reduction in noise makes this type of amplifier appear very promising for applications in radio astronomy, radar, and "scatter" communications.

Despite the enhanced performance and reduction in size and power of some of these "solid-state" amplifiers, electron tubes still maintain a lead over other types of amplifiers because of their higher state of development.

Amplifier Design Considerations

Distortion is the principal enemy of the amplifier designer. An important cause of distortion in amplifiers is inadequate power capability. This is particularly important in carrier amplifiers because of the tremendous range of signal level that may be encountered. Since most carrier systems are operated under low impedance conditions to maintain compatibility with telephone systems or related radio equipment, considerable power may be required to develop the required signal voltage.

In a single carrier channel, there may be a tremendous power level range imposed on the amplifier. During listening periods, there is almost no signal



Channel amplifiers must accommodate 70-db power range without losing soft speech or distorting high levels.

power on the line. When the telephone user begins to talk, power jumps dramatically, varying over a range of about 30 db or 1,000 to 1, for an individual talker. The range for all talkers is 10,000,000 to 1, or 70 db!

When several channels are combined into a group, power variations tend to average out, so that in groups of 64 or more channels, the difference between average power and probable peak power is greatly reduced, and may be predicted on the basis of statistical probability. Careful attention to this fact in designing carrier amplifiers permits considerable savings in complexity, space, and power required for carrier systems. Figure 1 shows how power level range varies for various numbers of channels.

The question might well be asked, "Why not design for maximum possible loading based on all channels being used simultaneously by loud talkers?"

World Radio History

Such a system would require amplifiers to have much greater capacity than actually required. Equipment would require more input power, additional heat would be generated (a serious consideration in large, high-density terminals), and the system would cost more, for no significant improvement in performance. Consequently, modern carrier systems are designed to accommodate average load at the time of greatest traffic. To prevent a few loud talkers from upsetting the balance between probable maximum instantaneous load and the power capacity of the system, individual channel amplifiers are usually provided with peak limiters to prevent excessive levels from entering the system.

Amplitude Distortion

Amplitude distortion is the greatest single cause of poor amplifier performance. Amplitude distortion occurs when the amplifier is operated in a non-linear portion of its dynamic characteristic. Signal voltage variations in the output are not proportional to input signal voltage changes. In such a case, the amplifier tends to act like a modulator, producing harmonics and multiples of the various input frequencies. These intermodulation products cause interchannel crosstalk and noise which interfere with the desired signal.

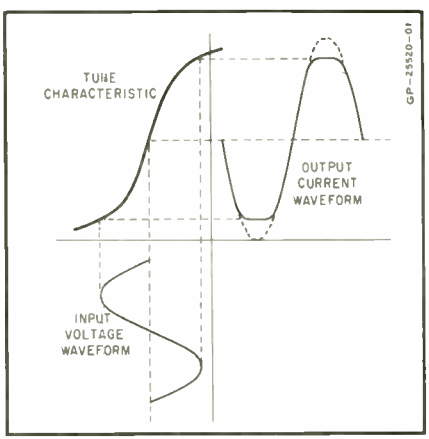


Figure 2. Distortion caused by excessive input voltage. Tube is driven into non-linear operation.

Negative Feedback

A powerful technique for correcting amplifier deficiencies was discovered by telephone researchers in 1927. This technique, called negative feedback or inverse feedback, reduces distortion, improves frequency response, and essentially frees the amplifier from the effects of power line variations and changes occurring within the amplifier itself.

Negative feedback gives an amplifier a great "reserve" which is used automatically to correct distortion and amplifier irregularities of all sorts. As an example of negative feedback, assume that an amplifier is required to have 60 db gain. If the amplifier is designed with 100 db gain, a portion of the output voltage may be applied to the input in such a way as to partially cancel the input signal. When the cancellation obtained in this way reduces the amplifier gain from 100 db to 60 db, the amplifier is said to have 40 db negative feedback.

The signal applied to the input of an amplifier represents the difference between the original input signal and the negative feedback signal. This difference is rather small if large amounts of feedback are employed. Now, if the amplifier has a tendency to amplify one frequency more than others within its passband, this frequency appears more strongly in the feedback signal. When

World Radio History

applied to the input signal, that particular frequency is reduced more than are the other frequencies, thus giving it a relative strength reduction just about equal to the excess amplification at that frequency. As a consequence, frequency response will tend to be uniform. Similarly, if amplification is reduced for any reason, the feedback signal is reduced, permitting the input signal to enter the amplifier at a higher level.

For a given output voltage, distortion is reduced by a factor equal to the reduction in amplification by feedback. Thus, the above amplifier with 40 db feedback will provide an output in which intermodulation or distortion products will be 40 db below those in the output of an amplifier producing the same output level, but without negative feedback.

An amplifier employing negative voltage feedback tends to maintain uniform output voltage regardless of the load placed across the output. As more and more voltage feedback is employed, the output impedance of the amplifier approaches zero. If, instead of feeding back output *voltage*, negative feedback is derived from output *current*, the amplifier tends to maintain a constant current regardless of load impedance. This is equivalent to increasing the output impedance of the amplifier.

Since both current and voltage feedback are beneficial to amplifier performance, they afford the carrier amplifier designer a convenient tool for adjusting the amplifier output impedance, while at the same time improving the performance of the amplifier. By carefully adjusting the ratio of current feedback to voltage feedback, output impedance can be raised or lowered to suit the needs of the circuit. In carrier systems this is particularly useful where the amplifier operates into one or more filters.

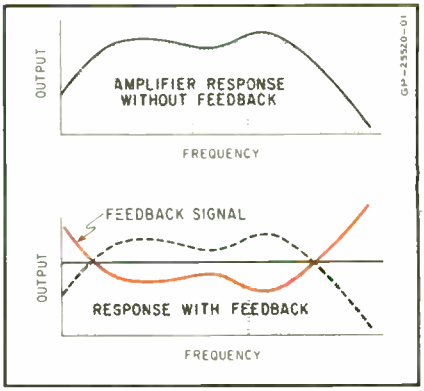


Figure 3. How negative feedback improves amplifier frequency response.

Amplifiers and Filters

Carrier systems employ many filters to divide the frequency spectrum into channels and groups of channels. The more efficient the filters, the more tightly packed channels and groups can be before they interfere with one another. Since high-performance filters are extremely impedance-sensitive, optimum filter performance can be obtained only when proper circuit impedances are carefully maintained. Thus, judicious use of feedback in amplifier design contributes yet another benefit to overall performance of a carrier system.

By carefully designing filters and amplifiers to work together, designers of modern carrier systems achieve performance that not long ago was considered to be impractical for economical design. Not only do these modern designs improve performance, but they permit reductions in complexity, power requirements, and space.

World Radio History

World Radio History

the Lenkurt® Demodulator

VOL. 12, NO. 12 DECEMBER, 1963

FILTER FLANKING

What and Why?

Frequency-selective filters, the very heart of modern carrier systems, have certain characteristics which may seem surprising. Why should some filters, which are designed to provide as much isolation as possible between their respective channels, require the presence of each other for proper functioning? The answer lies in certain unavoidable side-effects of the way in which filters operate.

These characteristics have considerable practical importance. In some cases, the removal of, say, the channel 5 filter will degrade the performance of channels 4 and 6, and a channel group which is only partially equipped may function quite poorly. This article discusses this "filter flanking" effect, how it occurs, and some means for minimizing adverse effects.

Where bandwidth requirements are not stringent, carrier channels are often separated by relatively large guard bands and filters can be simple. In such cases there is little interaction between adjacent channel filters. But modern carrier systems are usually required to transmit the maximum number of channels in the narrowest possible bandwidth, with the result that guard bands are narrow and filters must have very sharp cut-off characteristics. In this situation adjacent filters affect each other even though their passbands do not overlap. The smaller the guard band between adjacent filters, the more important this interaction or "flanking" effect is in achieving the required attenuation. In effect, adjacent filters serve as additional elements of each other. As a consequence, high-performance channel banks may not achieve their proper performance if one or more channels are removed. These characteristics are a natural result of the way in which a filter must operate to provide the necessary attenuation.

A conventional electrical filter consists of a network of reactive elements such as capacitors, inductors and possibly transformers which, ideally, are loss-free. They are able to store energy

World Radio History

(by virtue of their inductance and capacitance), but cannot dissipate it. Thus, all power delivered to the input of the network must eventually appear at the output. The reactance of capacitors and inductors varies with frequency as shown in Figure 1. This permits networks to be built which have high input reactance at some frequencies, but low input reactance at others. At frequencies where input impedance is essentially resistive, power is accepted and transferred to the output by the network, but is blocked at frequencies where input impedance is mainly reactive.

Figure 2 shows a reactance network operating between a generator, E, with internal resistance R_G and a resistive load, R_L. The reactance network is assumed to operate as an ideal filter — without internal losses. Within its passband, the filter presents to the generator an impedance which is essentially resistive and approximately equal in value to R_G. Thus, maximum power transfer can occur, and the filter absorbs virtually all the available power of the generator, transferring it to the load, R_L.

The out-of-band impedance of the filter, as seen by the generator, is almost a pure reactance. At frequencies outside the passband, the filter absorbs negligible power from the generator, and hence delivers virtually no power to the load. Therefore the presence of the load resistance, R_L, has essentially no influence on the input impedance. However, to say that no out-of-band power is transferred from the generator to the filter is not the same as saying that the filter has no input current at these frequencies; the input current flows, but since it is 90° out of phase with the input voltage, no power is absorbed (because the power factor is zero.)

Thus, the method of attenuation of a filter is not at all like that of a resistive pad. The pad accepts the signal and dissipates some (or possibly most) of the energy internally, whereas the filter refuses to accept the signal at all.

Filters in Parallel

Parallel filter operation can be illustrated by considering the simple case of two filters — one lowpass and one highpass, so designed that the passband of each coincides with the stopband of the

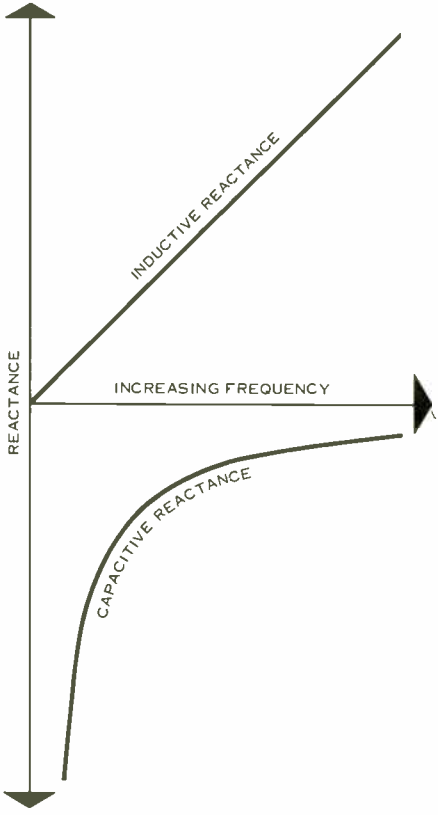


Figure 1. Higher frequencies cause inductive reactance to increase while capacitive reactance decreases. The two reactances are 180° out of phase.

World Radio History

other. When the inputs of two such filters are connected in parallel and driven by a common impedance-matched generator, it is possible, in principle, for each filter to absorb the maximum power available from the generator at frequencies lying within its own passband. It can do this because, at these frequencies, the other filter cannot accept any power. It is this ability to select power at certain frequencies that makes it possible for filters to be operated efficiently in parallel without "loading" the generator.

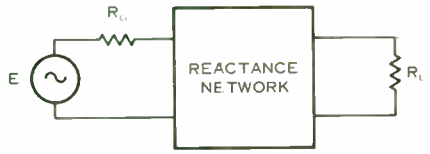


Figure 2. A filter is a network of reactive elements. For maximum power transfer, filter impedance within the passband should match the generator impedance (R_G).

The two filters, operating together, act as a frequency-selective power divider. Therefore, when the two filters are designed to operate in parallel, the input impedance at the common terminals should equal the generator impedance (R_G) over both passbands because maximum power transfer can occur only when the filter impedance matches the generator impedance. Accordingly, within its own passband each of the paralleled filters will also match the generator.

Since the net impedance of all filters as well as the individual impedance of each filter equals the generator impedance, it might seem possible to remove one or more filters without disturbing the impedance match between the signal source and the frequency-selective filters. Unfortunately, this is not so. Filter characteristics are affected by the overlapping electrical effects of adjacent filters, thus requiring careful design to balance one against the other.

Filter Admittance

In discussing the characteristics of filters to be connected in parallel, it is simpler to think in terms of *admittance* (Y), the reciprocal of impedance. Admittances connected in parallel add directly, while impedances do not. Like impedance, which is the sum of the two quantities, resistance and reactance, admittance is the sum of *conductance* (G), and *susceptance* (B). These two components of admittance are at right angles to each other, as indicated in mathematical expressions by the term j.

If the input admittance of the low-pass filter is assumed to be

$$Y_L = G_L - jB_L,$$

and that of the high-pass filter is assumed to be

$$Y_H = G_H + jB_H,$$

then the input admittance of the parallel combination is

$$Y = (G_L + G_H) + j(B_L + B_H).$$

Since, in effect, only the conductance portion of the admittance absorbs power, an *ideal* filter pair would have no net susceptance across the passbands, and the combined conductance of the pair would be equal to the reciprocal of the generator resistance, $1/R_G$. In other words, an ideal pair of filters would meet the mathematical conditions

$$G_L + G_H = 1/R_G$$

and

$$B_L + B_H = 0.$$

World Radio History

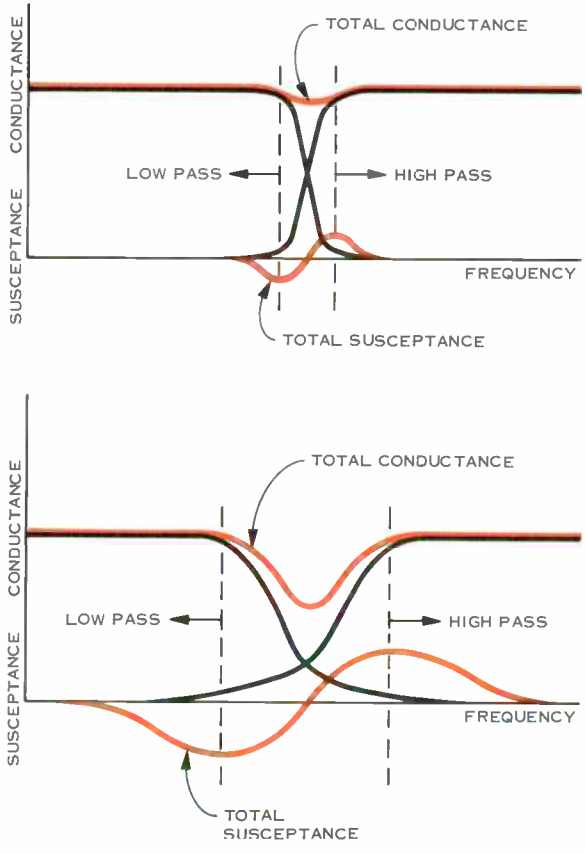


Figure 3. As the separation between filter passbands increases, the total susceptance deviates farther from zero and interferes more with passband performance.

For these conditions to be readily met, it is necessary for both filters to be designed for minimum susceptance. This does not mean that no susceptance can be present—but rather the susceptances of the two filters must have equal magnitudes and opposite signs at all frequencies within their passbands. Thus, the susceptances must always "cancel" each other. If they do not cancel, the input admittance will appear either capacitive or inductive and will cause the filters to reject part of the desired signal.

If the filters have minimum susceptance, it is only necessary to design them so that

$$G_L + G_H = 1/R_G$$

over the passbands. Then the condition

$$B_L + B_H = 0$$

will be closely approached when the gap between the passbands is small. As this gap becomes progressively larger, the total susceptance deviates farther from zero. (Figure 3 compares the effects of varying the gap between the passbands.) Eventually the error becomes intolerable and it is necessary to connect a network of reactive elements, producing essentially pure susceptance, across the input terminals to cancel the net susceptance of the filter combination. This *susceptance-annulling network* usually consists of one (or pos-

World Radio History

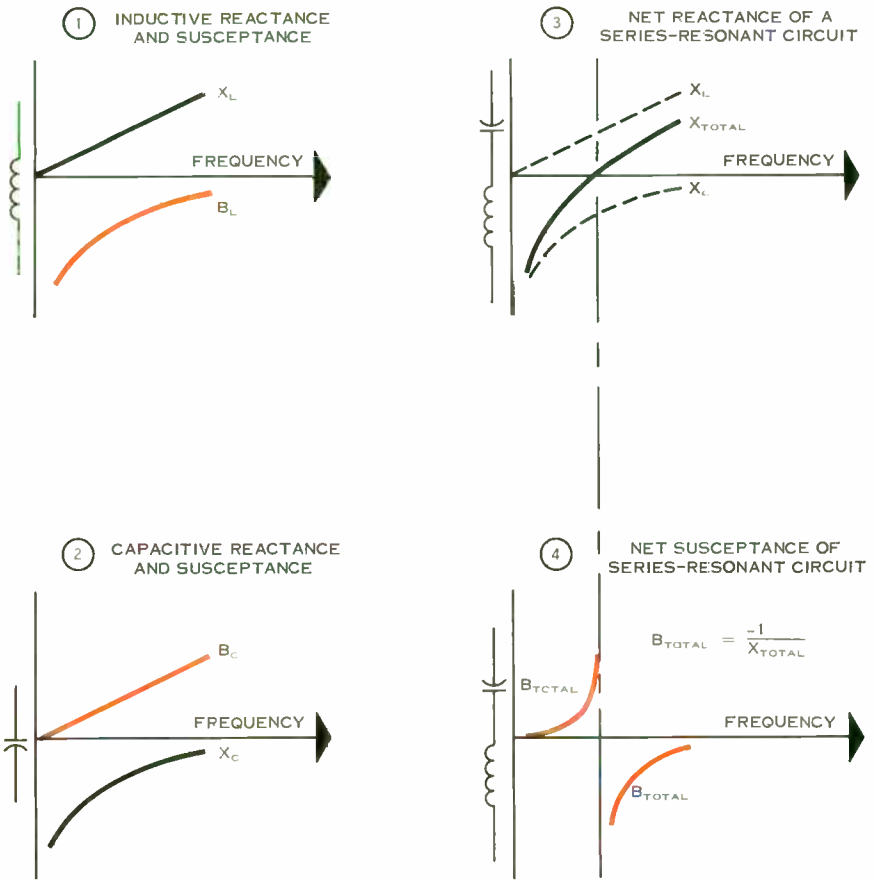


Figure 4. Relationship between reactance and susceptance of circuit elements as frequency varies.

sibly two) tuned circuits. Figure 5 shows the effect of the annulling network in cancelling the susceptance of the two filters.

Another network, essentially a third filter, can be used as a *conductance-correcting network* to maintain a constant value of conductance. Its use is not so widespread, however, because most of the conductance deviation occurs in the area of transition between lowpass and highpass, and the conductance here is not normally of interest — it is usually sufficient to maintain a constant value within the passbands, without concern for the transition region.

Thus, even in the simple case of two filters flanking each other, one depends on the other and often both depend on the annulling network for satisfactory operation. Filters designed for flanking do not usually provide satisfactory performance if they are not flanked; and a

World Radio History

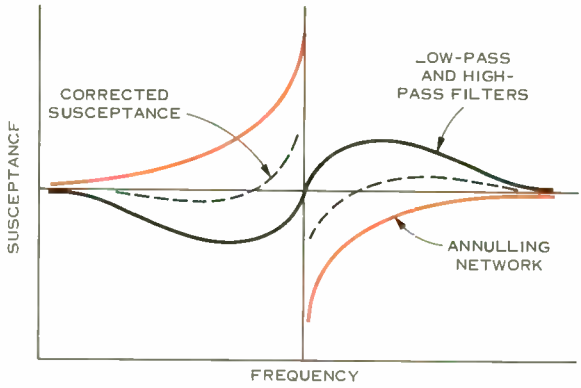


Figure 5. Annulling network provides susceptance opposite to that of the filters, partially cancelling excess susceptance.

filter designed for individual operation cannot be arbitrarily flanked without altering its characteristics.

Practical Filter Groups

This discussion has so far considered only the simplest case: a single low-pass and a single high-pass filter operating in parallel. Modern carrier systems, however, may use many bandpass filters in parallel, as indicated in Figure 6. Each base group shown consists of 12 filters, each with a nominal 4-kc passband, covering the frequency range of 60-108 kc. The actual voice-frequency input to each channel is 0.3-3.5 kc; thus, 4-kc channel spacing allows an 800-cps guard band between channels.

In the next modulation step, five 12-channel groups are combined to form a 60-channel supergroup occupying the 312-552-kc band. The bandpass filters used with these five groups are spaced 48 kc apart. This "stacking" of frequencies can be carried on almost indefinitely, but the principle remains the same. Within each level (group, supergroup, etc.) filter flanking occurs.

Each filter affects *all* the others to some extent, but the effect is most pronounced on the adjacent filters. For example, the arbitrary removal of the channel 3 filter would not have a serious effect on channels 1 and 5, and would affect channel 6 even less—but the characteristics of channels 2 and 4 would be severely altered. As shown in Figure 7, this would normally reduce the sharpness of the cutoff on the side toward the missing filter.

Since the effect of one filter on the adjacent one depends on the frequency separation between them, widely spaced filters can usually be flanked without fear of interaction. In practical carrier systems, this characteristic may make it desirable to split a group into two subgroups, composed of even-numbered and odd-numbered channels. In this arrangement, channels 1, 3, 5, 7, 9, and 11 operate directly in parallel and are connected through a hybrid to the parallel combination of channels 2, 4, 6, 8, 10, and 12. This technique is in common use because it minimizes the problems inherent in designing filters for flanked operation with close spacing. The filters can often be designed to operate singly, without considering their effects on each other.

Annulling Networks

The question naturally arises as to the flanking of the "end" filters, channels 1 and 12. These channels exhibit

World Radio History

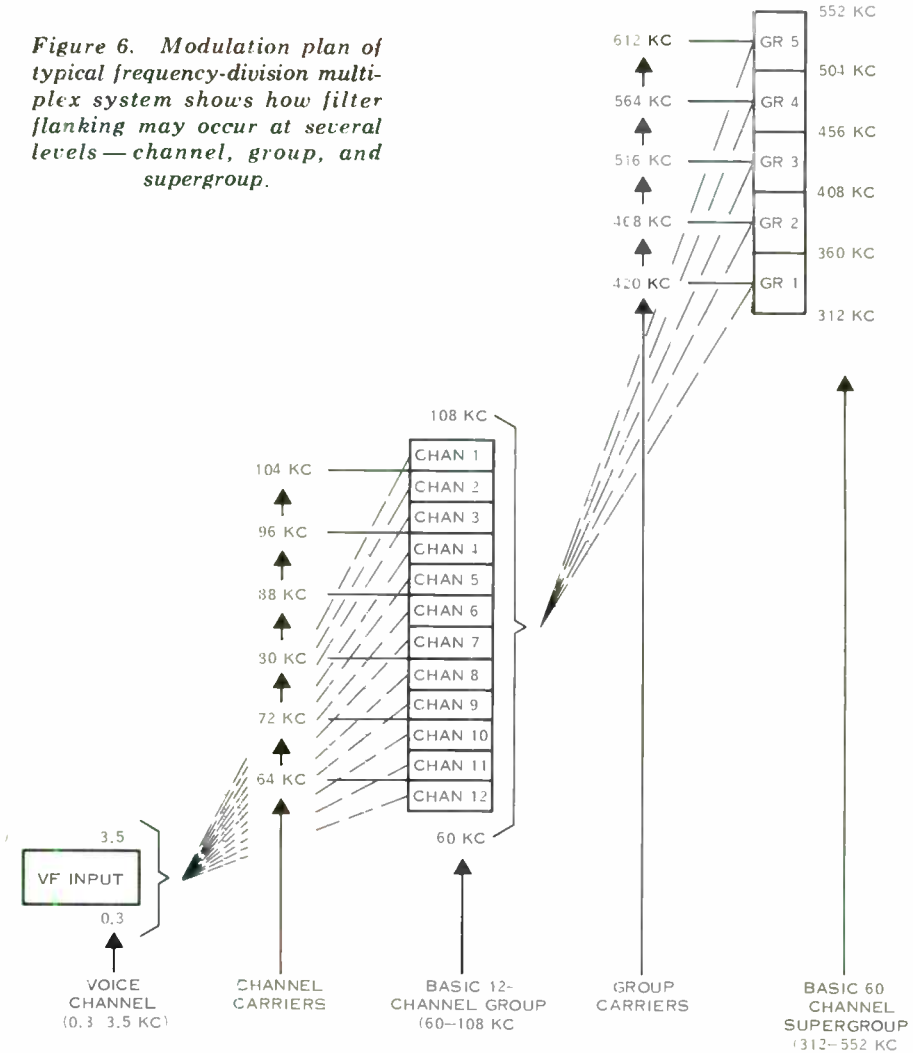


Figure 6. Modulation plan of typical frequency-division multiplex system shows how filter flanking may occur at several levels — channel, group, and supergroup.

characteristics much like those of the unflanked channels 2 and 4 shown in Figure 7. Without adjacent filters, the sharpness of the cutoff on the unflanked sides suffers. To correct this, so-called annulling networks are built to simulate the filters that would be used for channels 0 and 13, if these channels existed. Essentially, these annulling networks provide equal and opposite susceptance to that of the unflanked filter. Sufficient susceptance cancellation can be achieved by using a series tuned circuit to simulate the missing filters. The physical location of these networks is of little consequence so long as they are *electrically* across the input terminals. For example, "filter 0" may be built into the physical container for filter 11—and still provide flanking for filter 1.

The same type of flanking and the same interrelationships occur at other

World Radio History

levels in the modulation plan, but the principles involved are the same at the group and supergroup level as they are at the channel level.

Operational Considerations

When flanking problems arise, it is almost always because a carrier system is operated below design capacity. The installation may be only partially equipped, or a filter may be temporarily removed from a channel bank. Unless the missing filters are simulated in some way, the result may be degraded performance — at least in the channels immediately adjacent to the gaps left by the missing filters.

This is not to say that multiplex systems must not be operated partially equipped. The manufacturer may specify the minimum number of channels of a specific system which should be installed without using some type of compensating network to take the place of the missing filters. For example, the manufacturer of a 12-channel system might recommend that a partially equipped installation include no fewer than 6 channels (unless a compensating network is used). Such a network is not usually complicated or expensive — a typical one would consist merely of an inductor and a capacitor — but it is necessary in many cases for satisfactory system performance when only a few channels of a multi-channel system are installed.

Of course it may not be practical to design a network to simulate a single filter which is temporarily removed from the middle of a filter bank. In such a case, adjacent channel performance may be maintained by plugging in a non-operational channel to obtain the flanking effect of its filter.

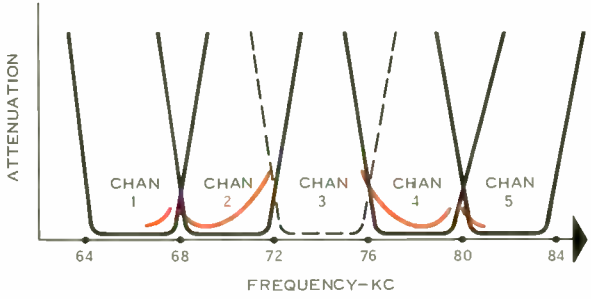


Figure 7. Removal of channel 3 filter degrades response of all other channels. Most noticeably affected are the near "corners" of the adjacent channels.

It appears that filter flanking and its effects will continue to concern both designers and operators for some time to come. In fact, flanking effects become more important when more complex filters are developed. Various other arrangements such as the use of active elements (amplification) or resistive networks to improve isolation are becoming more common. So-called active filters and phase equalizers achieve much better performance than the conventional passive networks, and this trend may profoundly affect future equipment designs. However, an awareness of the interaction between flanked filters goes far toward forestalling the problems which may arise in operating systems using this technique. •

World Radio History

For those not familiar with the "complex" numbers used to describe impedance and admittance, here is a brief explanation of the "real" and "imaginary" components of

Impedance and Admittance

Impedance is made up of two components, resistance and reactance. Although both are measured in terms of ohms, they are not the same quantity. Resistance is contributed by elements which *dissipate* electrical energy, passing it off as heat. Reactance, on the other hand, is contributed by elements which *store* energy (inductors and capacitors) without dissipation. It is the phase relationship between resistance and reactance of a circuit element which controls the phase angle between current and voltage — and hence controls the power transferred by a given current and voltage. The phase angle between the current through a resistor and the voltage across a resistor is zero. Thus, current and voltage are in phase, the *power factor* (the cosine of the angle between them) is unity, and the resistor dissipates power according to the relation

$$P = I^2R .$$

By contrast, current through an ideal inductor or capacitor is 90° out of phase with the applied voltage — current lags by 90° in the inductor and leads by 90° in the capacitor (assuming a sinusoidal steady-state condition). Thus, the power factor is zero ($\cos 90° = 0$) and no power is absorbed by the inductor or capacitor.

Since the phase or "directional" relationship between resistance and reactance controls power transfer, it becomes apparent that magnitude alone cannot completely specify these two quantities. They have both magnitude and direction. The direction is normally specified in terms of the phase angle between them. And since both resistance and reactance are required to define impedance, it too is a directional quantity. Impedance is the *vector sum* of resistance and reactance (shown graphically below). By convention, the resistance R is considered to have an angle of 0°. Hence, it is the reference and is called the *real* part of impedance. Reactance, with the symbol X, has an angle of $\pm 90°$ ($+90°$ for inductive reactance and $-90°$ for capacitive reactance) and is the imaginary part of impedance — although the concept of reactance is not fictitious. These are called *complex* quantities because of their composite nature.

Impedance, Z, is expressed mathematically as

$$Z = R + jX ,$$

where the "j" indicates the 90° phase difference between R and X. Impedance can also be expressed as

$$Z = \sqrt{R^2 + X^2}\angle\theta ,$$

where $\sqrt{R^2+X^2}$ is the magnitude and $\angle\theta$ indicates the direction.

The component parts of impedances *in series* add directly. For example, if

World Radio History

$$Z_1 = R_1 + jX_1$$

and

$$Z_2 = R_2 + jX_2,$$

then

$$Z \text{ total} = R_1 + R_2 + j\,(X_1 + X_2).$$

However, the same is not true of impedances *in parallel;* the *reciprocals* of the paralleled impedances must be added:

$$\frac{1}{Z \text{ total}} = \frac{1}{Z_1} + \frac{1}{Z_2}.$$

Resistance and reactance add at right angles to form impedance. Their relative magnitude determines the angle of the impedance.

The expression for total impedance then becomes less convenient:

$$Z \text{ total} = \frac{Z_1 Z_2}{Z_1 + Z_2}$$

$$= \frac{(R_1 + jX_1)\ (R_2 + jX_2)}{R_1 + R_2 + j(X_1 + X_2)}.$$

Because of the awkwardness of such expressions, it is often easier to speak in terms of *admittance,* (Y), the reciprocal of impedance. Admittances in parallel add directly just as do impedances in series. Like impedance, admittance is a complex quantity; that is, it has a "real" component and an "imaginary" component with a 90° angle between them. The real part is called *conductance,* (G), and the imaginary part is called *susceptance,* (B). Thus,

$$\text{Admittance } (Y) = \frac{1}{\text{impedance}} = \frac{1}{Z}$$

$$= \frac{1}{R + jX}$$

$$= \text{conductance} + j(\text{susceptance})$$

$$= G + jB.$$

In general, a large resistance implies a small conductance, and a positive reactance implies a negative susceptance. However, conductance is not the reciprocal of resistance, and susceptance is not the reciprocal of reactance because the definition of each is based on the complex quantities impedance and admittance.

For admittances in parallel, if

$$Y_1 = G_1 + jB_1$$

and

$$Y_2 = G_2 + jB_2,$$

then

$$Y \text{ total} = G_1 + G_2 + j(B_1 + B_2).$$

The admittance across the passband of an *ideal* filter would necessarily consist entirely of conductance. Any susceptance present would introduce a phase shift between current and voltage, thus lowering the power factor and reducing the power transferred. Since the filter is composed of less-than-ideal reactive elements, susceptance is inevitable. This can be countered, however, by cancelling each positive susceptance with a negative susceptance to produce a net result approximating zero. •

World Radio History

the Lenkurt Demodulator

VOL. 8 NO. 4 APRIL, 1959

The Use of FILTERS in Carrier Telephone Systems

In frequency-division multiplex, the frequency spectrum is divided into discrete frequency bands for the transmission of information. The frequency bands are normally separated by electrical wave filters. Although the design of electrical wave filters involves a combination of specialized knowledge and art, filter applications may be readily understood through a knowledge of the characteristics of basic filter types.

In this article, the characteristics of the basic classes of filters are described, and the use of filters in a number of applications is explained.

Filters are essentially devices which have particular resonance characteristics. Although any device which exhibits a resonance to electrical waves may be used, the most commonly employed filter components are capacitors and inductors. In the simplest case, a simple series circuit consisting of a capacitor and inductor is a filter.

Because of the exacting requirements of filters used in carrier telephone applications, most filters contain a number of meshes (or two-terminal networks) which consist of an inductor and one or two capacitors. Depending upon the characteristics desired, the meshes are arranged in various series and shunt combinations in relation to the metallic path of the transmission line.

Design parameters for a filter are obtained from such factors as: (1) the impedance, both input and output, into which the filter must work; (2) the frequency range to be passed; (3) the attenuation of frequencies which may be present on either side of the desired frequency range; (4) the loss that may be tolerated in the pass band; and, (5) whether similar filters will be operated in parallel (flanking). Each of these factors should also be considered in the application of filters.

Filters may be classified in a number of ways. A commonly used method is to classify the filter type by means of the relationship between the pass-band and the cut-off frequency. With this method, there are four basic types: (1) low-pass filters; (2) band-pass filters; (3) band-elimination filters; and, (4) high-pass filters. In Figure 1, typical circuit configurations and attenuation-

World Radio History

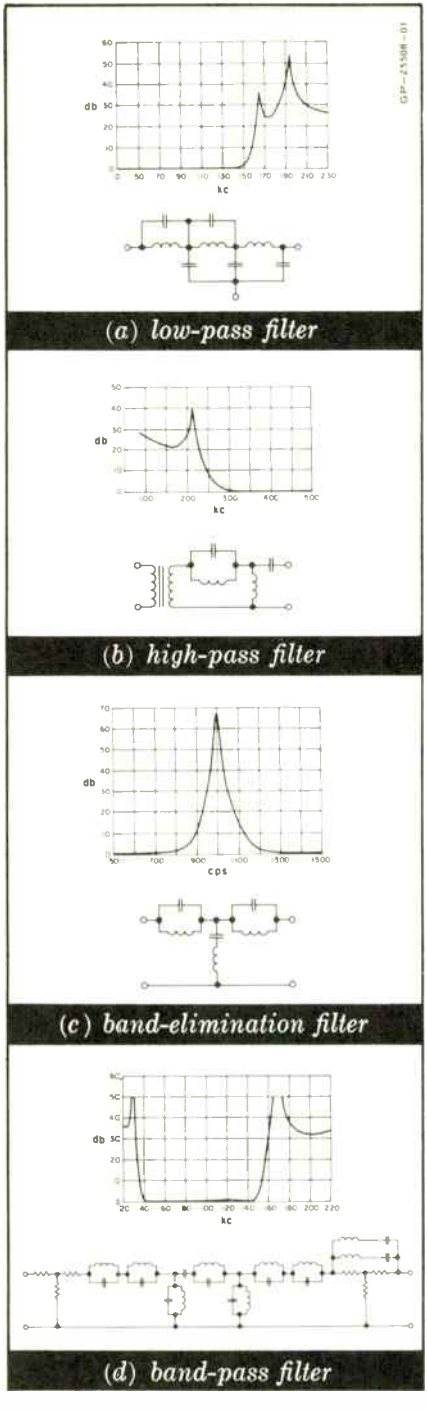


Fig. 1. Illustrating the four basic filter types.

frequency characteristics are shown for each filter type. From these general characteristics, filters may be selected for specific applications.

Low-Pass Filters

An ideal low-pass filter is one that has been designed to pass frequencies from 0 up to the cut-off frequency, and to effectively suppress all frequencies above the cut-off frequency. In a practical filter, the out-of-band attenuation will vary with frequency.

A typical low-pass filter application is shown in Figure 2. Here the filter is used at the input of the transmitting branch of a carrier channel. The function of this low-pass filter is to restrict the range of frequencies that may be passed into the carrier channel. In this application, the cut-off frequency is in the order of 3.2 kc. However, frequencies above cut-off will enter the channel, but will be considerably reduced in amplitude.

Where out-of-band signaling is employed, the maximum possible attenuation is required at the signaling frequencies in order to avoid false signaling. In the application shown in Figure 2, out-of-band signaling frequencies of 3400 cps and 3550 cps are used. Speech signals at these frequencies entering the circuit are adequately suppressed as shown by the attenuation peaks (often called infinite points) as shown in the attenuation-frequency characteristics.

Band-Pass Filters

Restricting the input frequency range is only one step in the process of minimizing the bandwidth required for the transmission of speech over carrier-derived channels. Additional filtering is also required at the output of the channel modulator.

World Radio History

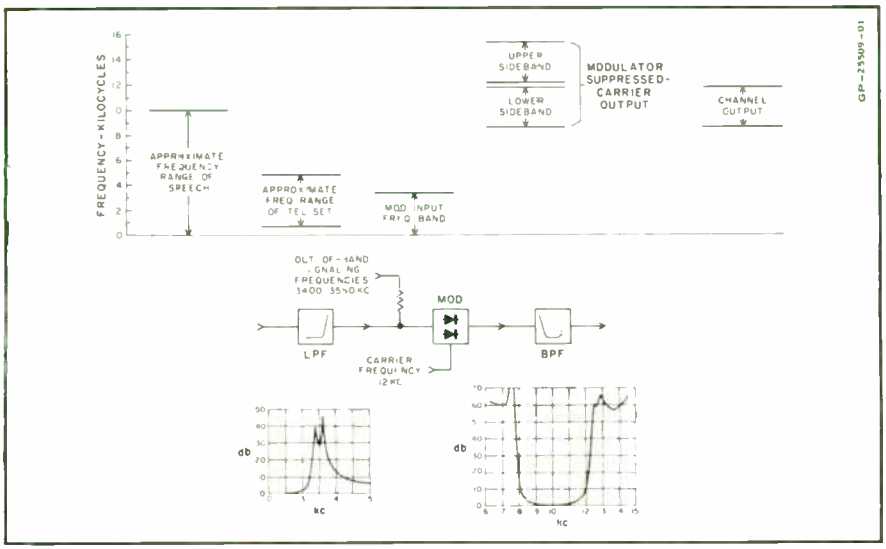


Fig. 2. Use of a low-pass and a band-pass filter in the transmitting branch of a carrier channel. In the upper portions of the diagram, the frequencies which may occur in the various parts of the circuit are shown. Also shown are typical characteristics of the filters used.

Modulation of the voice and signaling frequencies can be accomplished in any one of a number of ways. Where suppressed-carrier, amplitude modulation is employed, the modulator output contains two principal sidebands. One sideband is above and the other one is below the carrier frequency. As shown in Figure 2, the frequency range occupied by each sideband depends upon the frequency range of the original modulating frequencies. Each sideband contains all of the intelligence that was present in the original modulating wave. Either one or both sidebands may be transmitted depending upon the type of carrier system.

When a number of carrier channels are included in a system, the channels normally occupy adjacent frequency bands. Interference between channels is kept to a minimum by limiting the channel frequency band. This is readily accomplished by using a band-pass filter. In Figure 2, the band-pass filter selects only the lower sideband. Since the unwanted sideband would cause crosstalk in an adjacent channel band, the relative attenuation of the filter must be sufficient to keep crosstalk to within acceptable limits.

Band-Elimination Filters

Occasionally, it is necessary to prevent either a range of frequencies or a single frequency from passing beyond a certain point in a circuit. For example, the frequency range of a voice channel may be reduced to permit transmission of a carrier telegraph channel over the same circuit. A more common application is the elimination of a single-frequency pilot tone—such as is used for regulation or frequency control. This application is illustrated diagrammatically in Figure 3.

In this case, the pilot signal is used both for regulation and frequency con-

World Radio History

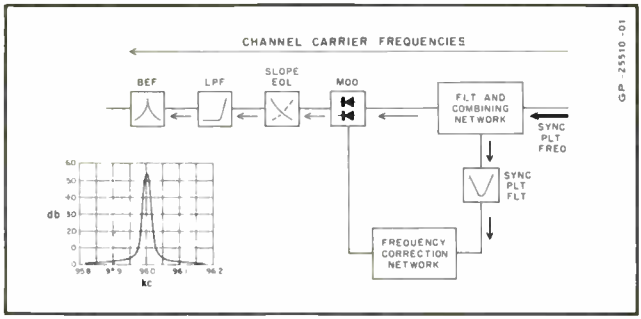


Fig. 3. It is sometimes necessary to severely attenuate a pilot signal to avoid interference. For this purpose, a band-elimination filter is inserted in the common line after the pilot pick-off point.

trol, and is transmitted at a relatively high level when compared to the message level. At pilot pick-off points, a portion of the signal is accepted by the regulator and control circuits on a bridging basis. The remainder of the pilot signal passes through along with the carrier channel frequencies. Level reduction of this pilot frequency to an acceptable minimum is achieved by inserting a band-elimination filter directly into the receiving circuit.

Where low-level pilots are used, it is not always necessary to eliminate the pilot. However, when systems which use pilot signals are interconnected on a carrier frequency basis elimination of pilot frequencies is often necessary. In this case, band-elimination filters are used with the interconnecting equipment.

High-Pass Filters

The ideal high-pass filter passes with a minimum of attenuation all of the frequencies above the filter cut-off frequency, and effectively suppresses all frequencies below cut-off. High-pass filters may be used separately, but are quite often used in conjunction with a low-pass filter. Such filter combinations are used where it is necessary to separate two frequency bands being transmitted over a common line.

The requirements for these filter combinations are different for each application. For this reason, rather than to call such filters by a name which is a combination of the filter types used, the filters are often denoted on the basis of their application. Examples are: (1) directional filter; (2) line filter; and (3) junction filter.

Directional Filters

For two-wire carrier systems, separation of the transmitting and receiving bands is necessary at terminals and repeaters. The filter combination normally used for this purpose at the two- to four-wire junction is called a directional filter. In addition to separating the two frequency bands, the filter provides impedance matching at the junction.

Directional filters are normally designed as a part of the carrier equipment, and therefore are only required to pass the frequency bands used in the carrier system. For this reason, directional filters may consist of two band-pass filters or a low- and high-pass filter combination. Figure 4 illustrates the use of directional filters at a repeater location.

Line Filters

The most effective means of utilizing a wire pair is to sub-divide the available

World Radio History

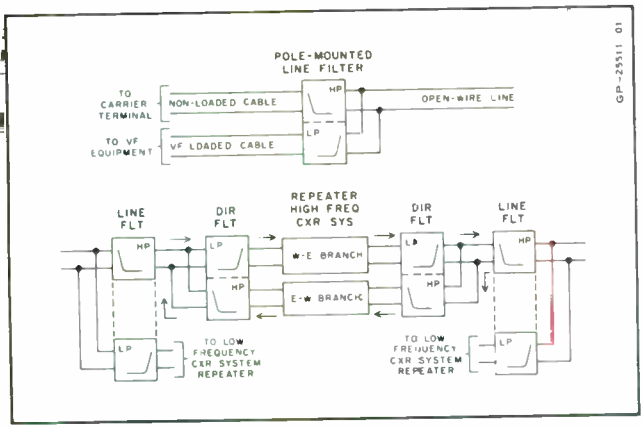


Fig. 4. A line filter is used to separate the frequency bands used by two carrier systems or by a carrier system and a voice circuit. Directional filters separate the frequency bands of a carrier system used for the two directions of transmission.

frequency band into carrier channels. Quite often this can only be achieved by using two or more carrier systems. At terminals and repeaters, and sometimes at intermediate points, it is necessary to separate the frequency bands of the different carrier systems. The devices most commonly used for this purpose are called line filters and junction filters.

Line filters normally consist of low-pass and high-pass filter sections. The low-pass section will effectively transmit all frequencies below cut-off, and the high-pass section will pass frequencies above cut-off which are in the normally used carrier frequency spectrum. Two types of line filters are commonly used: (1) main-station line filters; and, (2) auxiliary-station line filters.

Main-station line filters are used at terminals and main repeater stations where all carrier systems on the line are repeatered. These filters provide the desired out-of-band metallic suppression, but have relatively low longitudinal suppression characteristics. Therefore, these filters may be used for phantom operation, and are intended for use when low-frequency equipment is connected to the equipment side of the low-pass filter section. Two applications of main-station line filters are shown in Figure 4.

Auxiliary station line filters are designed for use at intermediate repeater stations where only the high-frequency carrier system is repeatered. For this reason, the low-pass sections are designed to be connected directly together. To prevent excessive longitudinal coupling around the repeater, these filters are designed to provide longitudinal suppression which is in the same order of magnitude as the suppression in the metallic path. This high order of longitudinal suppression virtually negates operation of auxiliary-station line filters on phantom circuits.

Junction Filters

Junction filters are similar to main-station line filters, but are normally designed for applications at non-gain points—such as at the junction of open-wire lines and intermediate cables. Since a number of these filters may be used between terminals, the suppression is kept low to prevent appreciable degradation of the frequencies near the limits of the pass band.

World Radio History

World Radio History

the Lenkurt Demodulator

VOL. 10 NO. 4 APRIL, 1961

Military Versus Commercial CARRIER SYSTEM DESIGN

For several years, Lenkurt has been developing two of the most advanced carrier systems available today. Despite the fact that both were developed nearly simultaneously by two teams within the same company, the systems are remarkable for their differences rather than their similarities. Each was designed to satisfy rigorous — but different — needs. This article contrasts the differing equipment requirements of military and commercial communications and shows how in good design, "form follows function."

Science and technology press forward constantly, again and again opening the way for achievements previously not practical, and revealing better ways of doing old jobs. Eventually, even perfectly adequate equipment is gradually replaced by designs which take advantage of newer techniques. In commercial communications, however, new features or techniques are rarely the basis for retiring old equipment for new. Instead, old equipment is usually retired as it passes the point where it can continue to serve reliably and economically. Often, well-proved designs may even be preferred because of the refinement and dependability of the equipment.

This is not generally so in military systems. Yesterday's fighter plane, although still a superb aircraft, may be virtually useless against another which has only a slight performance advantage. In military communications, a similar philosophy applies. Although communications systems are not pitted against each other in combat, commu-

World Radio History

Figure 1. Lenkurt 45BX carrier equipment in telephone toll center. Protruding equipment on right is Lenkurt 33A carrier, a classic older design still in great demand because of its reliability and performance. Relatively controlled environment in such centers reduces an important problem forced on military equipment.

nications dependability and performance could provide the decisive edge which makes the difference between victory and defeat.

Contrast in Economics

The cost of civilian communications equipment may be spread over a long, predictable life span, thus permitting a careful estimate of the performance required. Commercial equipment can be designed to perform a certain function to a certain standard of quality — no more, no less. This careful blending of economics and engineering results in systems of very high efficiency, and pro-

World Radio History

vides communications of the lowest cost consistent with reliability and high quality.

The extremely high stakes involved in military action prohibit this philosophy of prudent economy. No matter what the cost of a military communications system, it proves to be a tremendous bargain if it helps prevent a war or win a battle. Equipment which falls short of doing its required job under the strenuous conditions of crisis, might just as well not be built in the first place.

The Military Problem

Unlike civilian communications networks, which evolve in a deliberate, carefully-organized pattern to match the gradually changing needs of our civilization, military needs are most likely to erupt full-grown in an hour, and under the worst possible circumstances.

Military equipment must be designed to be "at home" in whatever situation it finds itself. Temperature extremes, sandstorms, blizzards, unskilled handling — these and more, are typical of some of the adversities which may hamper proper communications.

Because the circumstances under which a military system may be used are largely unpredictable, it must be assumed that the traffic load will be severe. More facsimile, digital data transmission, and other pulse-type communications are used than in civilian systems. These types of signal impose a much heavier load than ordinary voice messages because the randomly-occurring speech sounds average out statistically, when many channels are involved. For this reason, a military system must

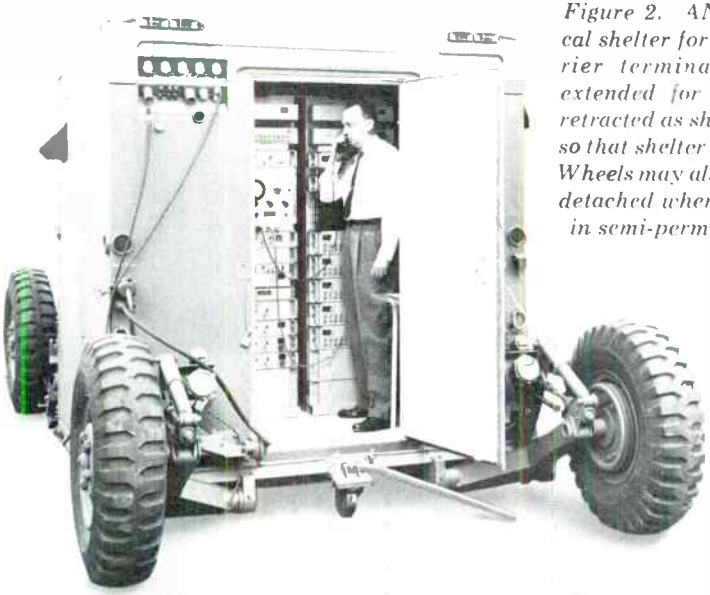


Figure 2. AN/FCC-17 tactical shelter for 60-channel carrier terminal. Wheels are extended for travel, may be retracted as shown during use, so that shelter rests on ground. Wheels may also be completely detached when shelter is used in semi-permanent location.

World Radio History

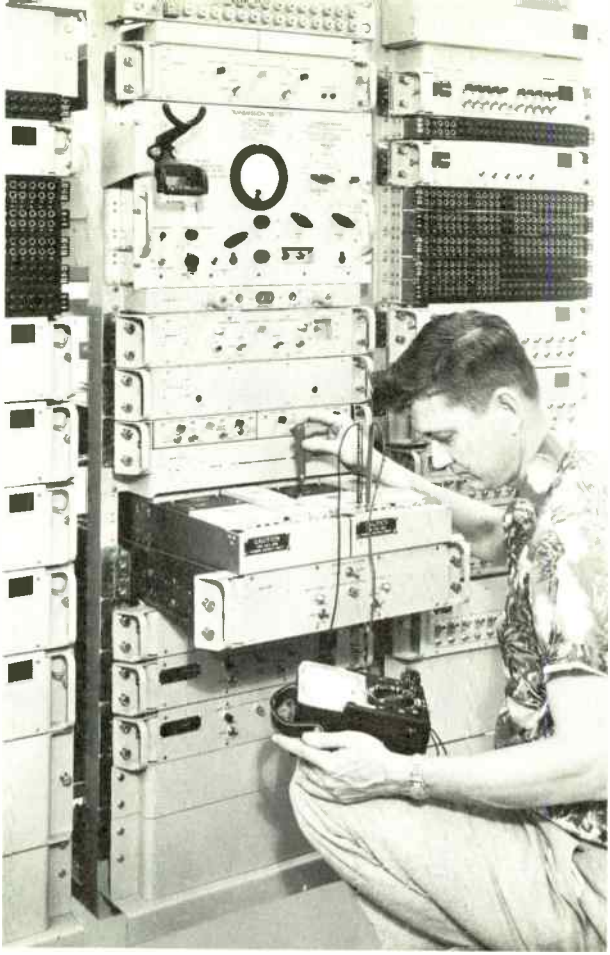


Figure 3. Roll out shelves in military system allow easy access and maintenance to all circuits. Military system requires greater rack depth than commercial system, but permits shorter racks. Deep racks permit roll-out shelves and provide additional mechanical strength.

be able to accomodate a considerably heavier load than its civilian counterpart, for a given number of channels.

As a final difficulty, the *circumstances* of military use make particularly severe demands on the physical structure of the equipment. It must be able to withstand the severe mechanical shock and vibration associated with transportation of all types, as well as shock that might result from hostile action. In addition to being ruggedly built, the equipment must be designed so that maintenance is particularly easy, since there is never assurance that well-trained personnel will always be available to keep the system running properly.

The Problems of Commercial Design

There is quite a difference between good design in a military system and in a system suitable for commercial communications. The military system must achieve certain performance characteristics regardless of cost. A commercial system, on the other hand, must meet very definite performance requirements, but within rigorous cost limitations. This can be done, not by cheapening

World Radio History

the design, but by making it much more specialized in its function.

In a military system, quality of performance means ability to overcome adverse conditions; in a commercial system, high quality invites and stimulates more use, thus permitting the system to justify itself economically. A very careful balance is involved. Improved performance certainly invites more patronage, but increases the cost of the system at a very rapid rate. The quality of the system must not be so high that no one can afford to use it. Nor should the quality of communication be so poor that other means prove more satisfactory.

Recent Examples

Two new carrier systems provide an unusual opportunity to show how these problems were met in commercial and military engineering practice. The military system, designated "*Multiplexer Set AN/FCC-17*," will be the standard carrier system used by the United States

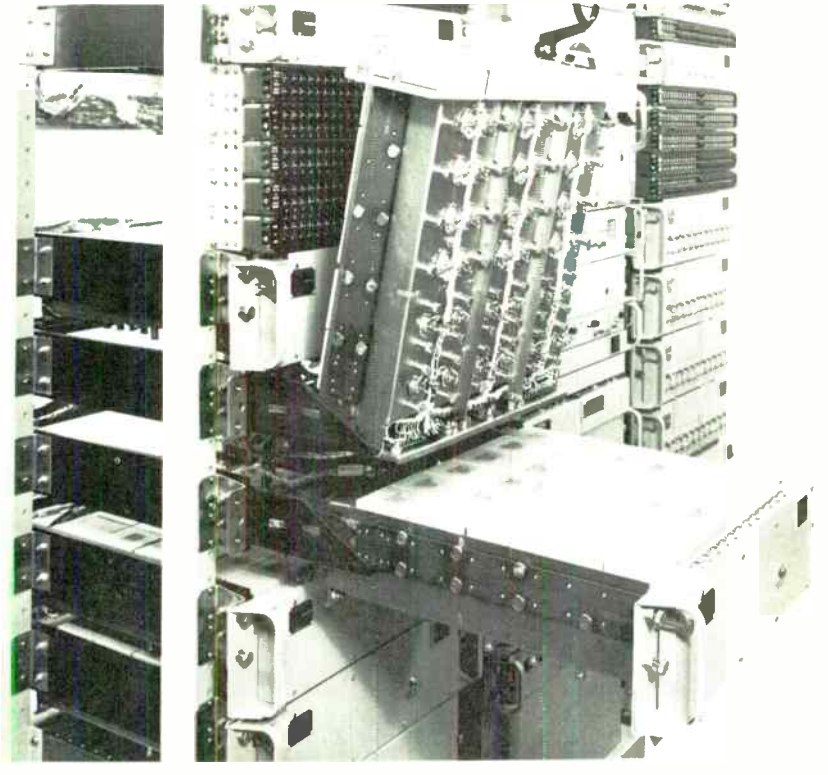


Figure 4. Military system employs deep, rugged racks to support equipment. Roller slides permit inspection and maintenance of all circuits without interrupting service. Extended drawers each contain channel filters and modulators for 12 transmitting channels. Each drawer may be locked fully extended and may be tilted upward for access to bottom interconnections.

World Radio History

Air Force in filling its carrier communications needs all over the globe. The commercial system, known as *Type 46A*, supplements and extends the usefulness of Lenkurt's 45-class carrier systems, used by more organizations around the world than any other carrier equipment.

Both systems are transistorized, both provide up to 600 channels. Both are designed to provide circuits of very high quality over great distances. Beyond this, the similarities end.

Versatility

Both systems are extremely versatile, each in a different way.

The AN/FCC-17 was designed for *universal* use, permitting transmission over radio and, with auxiliary equipment, wire or cable. There is no restriction on the type of information which it can accept for transmission; speech, digital data, facsimile, and related types of signals may be transmitted over any or all channels without requiring special treatment.

The 46A has been designed to permit coordination with virtually all other types of carrier system used for long-haul service. A variety of standard options provide pilot frequencies and operating levels recommended by C.C.I.T.T. or used in such systems as the

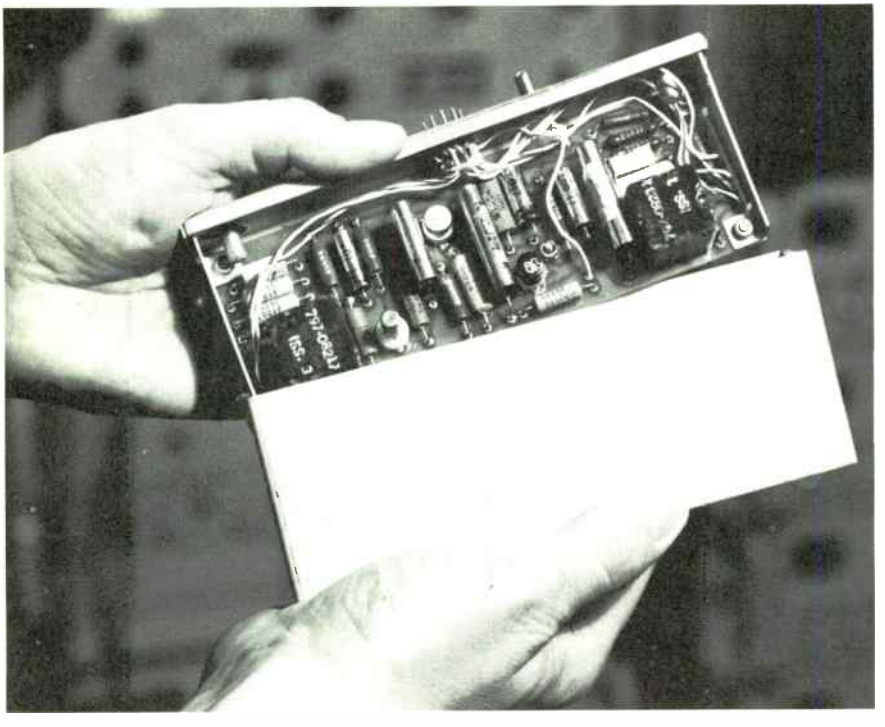


Figure 5. Conventional etched circuit techniques are used within sealed plug-in units of AN/FCC-17 system. Units are sealed to protect them from environment and handling. Internal repairs are performed at maintenance depot.

World Radio History

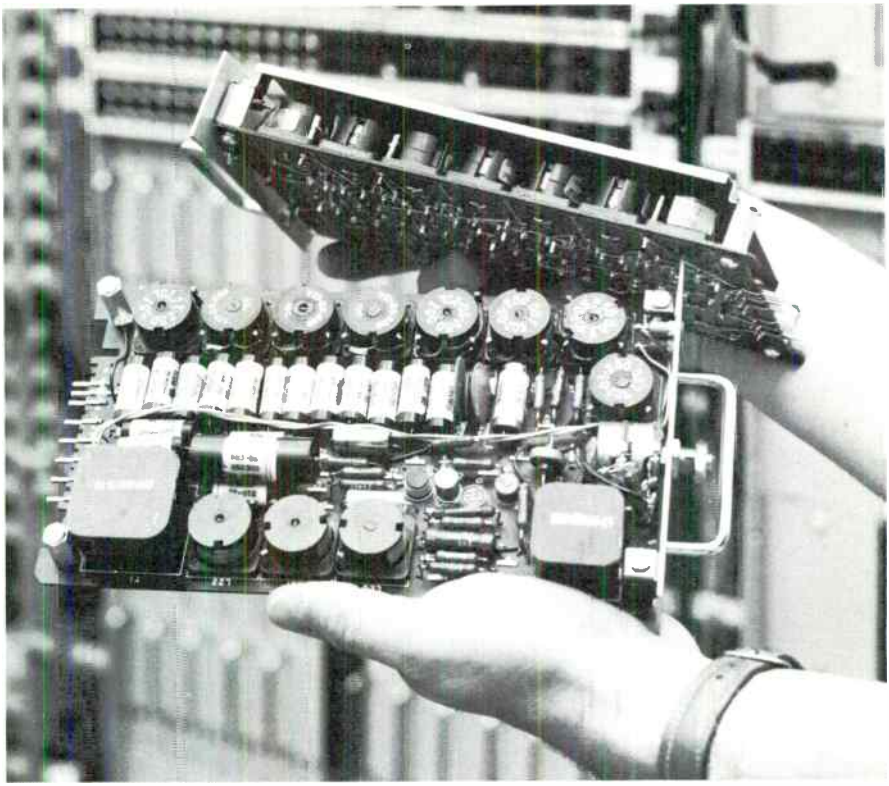


Figure 6. Typical 46A plug-in units. Shield has been removed from lower unit. Round objects shown at edge of both units are ferrite inductors used in filters. Note stitched wiring visible on upper unit. This type of construction permits automatic fabrication, yet provides even better repairability and reliability than expert hand wiring.

Western Electric "L" carrier system. To enhance the versatility of the equipment, 46A channel banks are designed for use with the group and common equipment of such systems.

Mechanical Construction

Radical new construction methods are used by each system to meet its special objectives. One of the prime requirements of the Air Force system is mobility and the ability to function reliably despite physical abuse. All the equipment is suitable for use either in fixed installations or in mobile communication centers. One tactical version is illustrated in Figure 2. This is a 60-channel terminal mounted within a special mobile shelter to permit travel over rough terrain.

Heavy four-post racks support the roll-out shelves in which all equipment is mounted. Each shelf may be extended and tilted up for inspection, and adjust-

World Radio History

ment without interfering with the operation of the system. Virtually all electrical components are contained within sealed units which plug into the sliding shelves. This technique permits defective units to be replaced immediately, and protects the components from dirt, moisture, and improper handling. In addition to speeding field maintenance, plug-in units permit equipment repairs to be handled at efficient central depots.

The 46A system does not require the extreme mechanical ruggedness of the AN/FCC-17, but is designed for conventional mounting on racks of the sort found in all telephone and communication centers. Instead of the sealed circuits found in the military system, all components are mounted on insulating boards which plug into equipment shelves. Figure 7 shows a close view of a typical 46A shelf with several of the plug-in units removed. Typical plug-in units are visible to the right and left in the picture.

As in the military system, plug-in modules permit very rapid maintenance in case of equipment failure. Unlike the AN/FCC-17, all circuit components are exposed. This is permissible because the 46A does not have to cope with the unusually wide range of environmental conditions for which the AN/FCC-17 is designed.

In the military system, printed or etched wiring is used to achieve uniformity and maximum space economy, thus enabling all active circuits to be enclosed in sealed containers to protect them from the environment.

The 46A equipment uses Lenkurt's unique stitched wiring process (de-

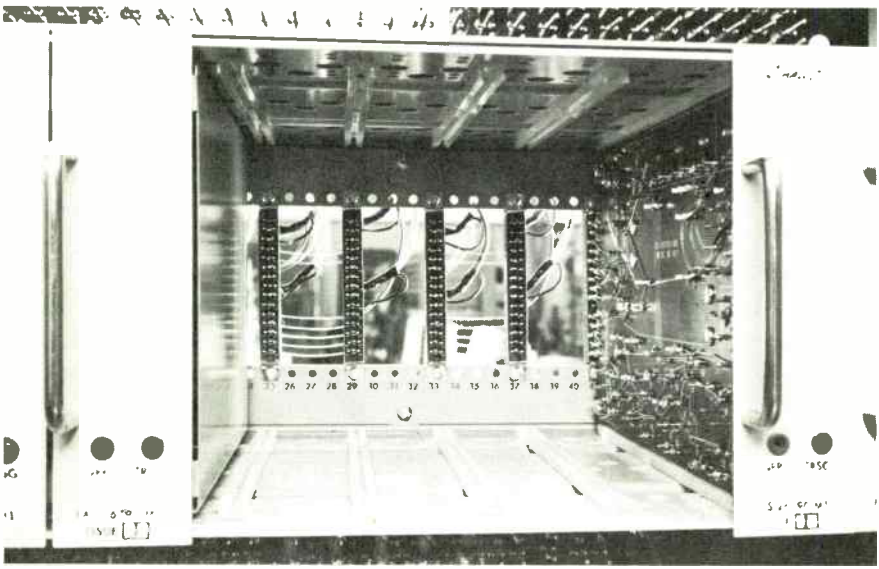


Figure 7. 46A plug-in units are held by nylon slides, and plug into receptacles at rear of shelf. In this view, four plug-in units have been removed. Note stitched wiring on unit at right.

World Radio History

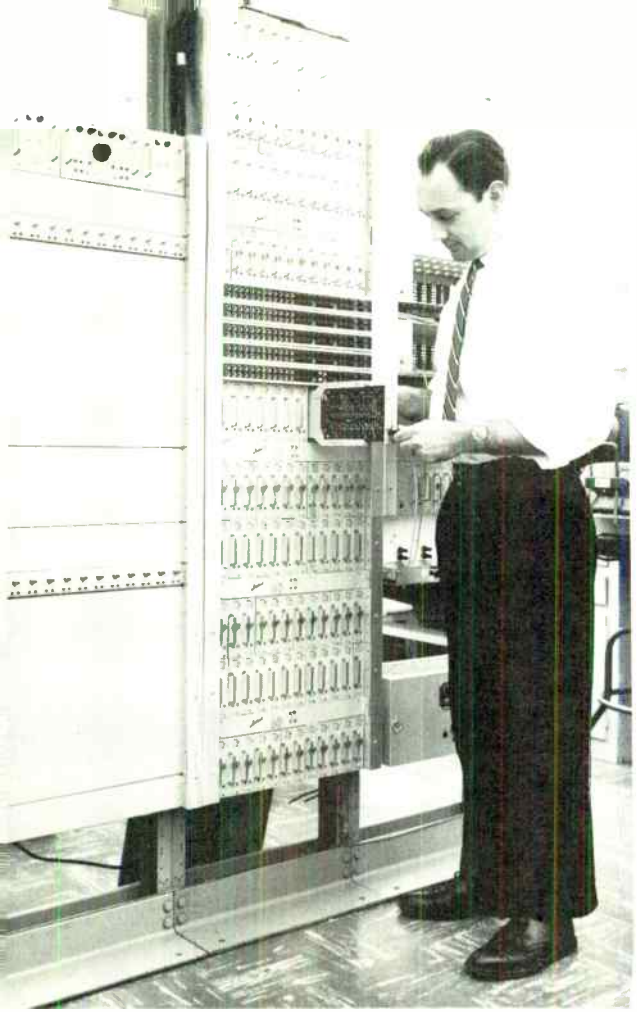


Figure 8. A typical 60-channel 46A terminal with additional common equipment. Left rack includes carrier supply for 288 channels (240 plus 20%). Group equipment for 60 channels is contained in two of the shelves in the rack on right. Engineer is shown using shelf extender which permits access to both sides of plug-in unit while it is connected into system, permitting easy testing and maintenance.

scribed in DEMODULATOR, June, 1960). This construction method is similar to etched wiring in that insulating boards are used to support the components and interconnections. The stitched wiring process combines the uniformity and ease of manufacture of printed circuits with the easier maintenance and repairability of hand wiring. Automatically-inserted staples serve as connection points for components and wire interconnections, thus permitting components to be replaced by less skilled personnel. The penalty for this is the slightly greater space required by the projecting staples (Note the unit to the right in Figure 7).

Load Capacity

The load-handling capacity of a carrier system is one of its most critical design features, since this has an important bearing on the cost, size, and noise performance of the equipment. It is difficult to predict exactly how much load will be imposed by a single voice channel because of the wide range in individual speech characteristics. As the

World Radio History

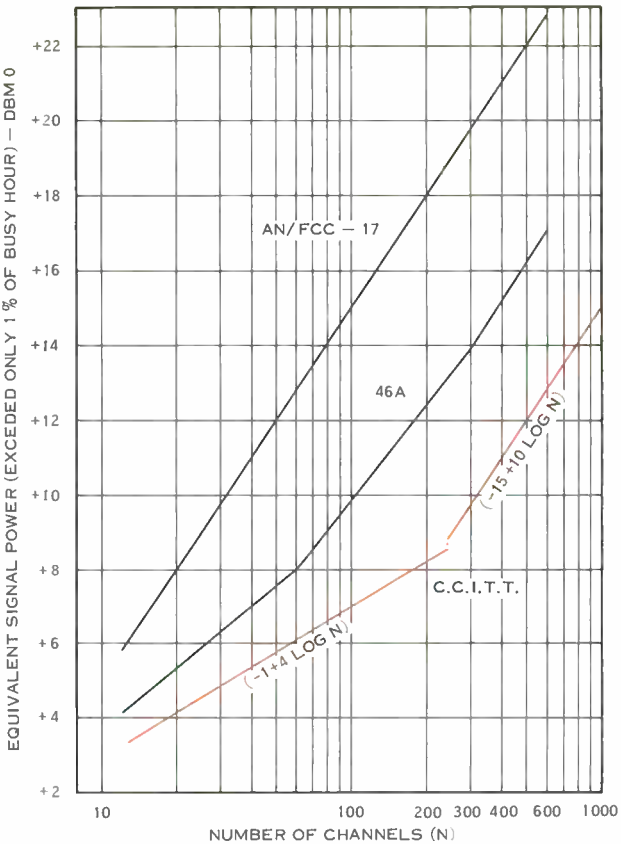


Figure 9. Comparison of load-handling characteristics of AN/FCC-17 system, 46A system, and the load characteristic recommended by CITT. The 46A system will handle more data transmission than provided for in CITT recommendations. Military system will handle 100% data load with no loss of performance.

number of channels increases, however, individual characteristics average out so that the total load becomes easier to estimate.

Careful measurements of the load presented by various numbers of voice channels have led to the C.C.I.T.T. load estimates shown in Figure 9 (red curves). The values shown represent the approximate load presented by speech plus a normal amount of signal power due to signaling tones, carrier leak, and pilot tones.

Because of the steadily increasing amount of data transmission, the load handling ability of the 46A system has been made substantially greater than the "normal" capacity indicated by the C.C.I.T.T. recommendations. The 46A load capability shown in Figure 9 was determined by assuming that ¾ of the channels would use out-of-band signaling tones (an option), $\frac{1}{16}$ of the channels would propably carry telegraph or similar data traffic, and $\frac{1}{12}$ of the channels would be used for SAGE data. Under these conditions of loading, the 46A system meets all noise and distortion standards for long-haul, toll-quality carrier.

World Radio History

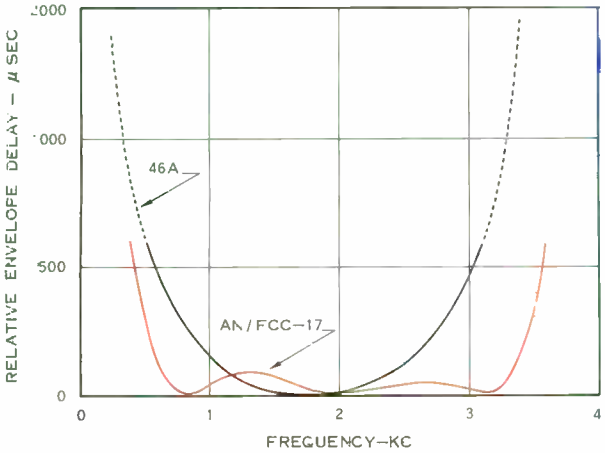


Figure 10. Comparison of relative envelope delay of 46A and AN/FCC-17 channels. Delay equalization is provided for all channels in the military system. 46A-channel filters are designed so that no channel varies more than 25 microseconds from characteristic shown.

It cannot be assumed that the military carrier system will be permitted to enjoy such an equitable and well-rounded load distribution. Being a military system, it must be designed to accommodate the worst possible load that might be applied to it. Accordingly, the AN/FCC-17 was designed to carry 100% data or facsimile.

Delay Distortion

Envelope delay has a much greater apparent effect on the quality of data and facsimile transmissions than on voice circuits. Since both systems place special emphasis on the ability to accommodate pulse transmission, special pains were taken to restrict envelope delay as much as possible.

In any carrier system, almost all the envelope delay experienced is contributed by the channel filters rather than the filters associated with groups of channels. In the 46A system, channel filters were designed to have *uniform* delay characteristics, regardless of their frequency. A "nominal" delay characteristic was determined and all channels must not deviate from this characteristic by more than 25 microseconds. This uniformity permits a standard delay equalizer which can be used with any channel, regardless of frequency.

Since the AN/FCC-17 was designed specifically to handle data, special care was taken to minimize envelope delay over a large portion of the channel passband. Figure 10 compares typical channel envelope delay characteristics of the two systems.

Conclusions

Although both systems are notable for their particular electrical and mechanical features, more outstanding is the way in which each design has been carefully tailored to meet the special needs of its particular application. Although the military system includes many "extras" to obtain its extraordinary performance, it is remarkably compact. Similarly, the 46A improves upon performance standards available in previous equipment, and manages to improve size, cost and reliability at the same time. ●

World Radio History

World Radio History

SECTION II
MICROWAVE RADIO

World Radio History

World Radio History

the Lenkurt Demodulator

VOL. 10 NO. 9 SEPTEMBER, 1961

THE NEW MICROWAVE

Part 1

In the dozen years or so since the first commercial microwave systems went into service, microwave radio has undergone astonishing growth and change. To meet new demands, channel capacity has steadily increased, performance has been improved, and power consumption and physical size reduced.

Many desirable new features, however, tend to oppose each other. For instance, although the use of transistors is very appealing because of their structural characteristics and low power requirements, transistors usually introduce more noise and distortion than electron tubes. Greater bandwidth imposes a severe problem of linearity in all circuits, but particularly in modulated klystrons and in the receiver discriminator.

This article is the first of two which discuss some of the problems encountered in designing a new, high-capacity transistorized microwave system, and how they were solved.

Just a very few years ago, it seemed unlikely that there would soon be any substantial need for microwave systems able to carry more than one or two hundred voice channels (except, of course, in the great transcontinental "backbone" routes). Today, however, there is a surprising demand for equipment capable of handling 600 channels or more. Even if some of the "need" may be premature, or stimulated by the glowing promises of newcomers to the field, there still remains a small but growing body of microwave users whose needs are very real and for whom the higher capacity means lower costs and greater efficiency.

Many different types of problems appear in designing a system to accommodate 600 voice channels or a wide-band video signal. Part of the difficulty is economic, part is technical. "Brute force" methods which have been used for very long toll circuits over high density routes are generally too expensive for use in equipment intended for a

World Radio History

broader range of applications. One way out of this problem is to devise simpler, more sophisticated techniques which yield the same performance as the older methods. Actually, this approach is virtually mandatory since efforts to attain greater channel capacity by merely scaling up older designs, or to "transistorize" the equipment by substituting equivalent transistor circuits for their original electron tube counterparts, will prove disappointing.

Transistor Problems

The basic nature of transistors introduces problems of a sort not usually encountered in electron tube equipment. One of the outstanding problems is the remarkable sensitivity of transistors to all sorts of outside influences. Very small changes in operating bias may cause rather large changes in the gain, stability, and frequency response of a transistor circuit, particularly in wideband circuits. Even the life expectancy of transistors seems to be greatly influenced by the way in which they are used and the temperature at which they operate.

Electron tubes are essentially "one-way" devices, maintaining almost perfect isolation between the output circuit and the input. Changes in load have little or no effect on the input. In transistors, however, small changes in load impedance may cause rather large variations in the total performance of the circuit. Numerous techniques have been developed to compensate for these transistor characteristics: temperature-sensitive resistors in the power source of the transistor circuit alter the operating bias to compensate for the temperature characteristics of the transistor. Similarly, voltage-sensitive diodes and varistors are employed to reduce the effect of power supply variations. Despite the availability of these techniques, the problem becomes exceedingly difficult as transistors are used in high frequency

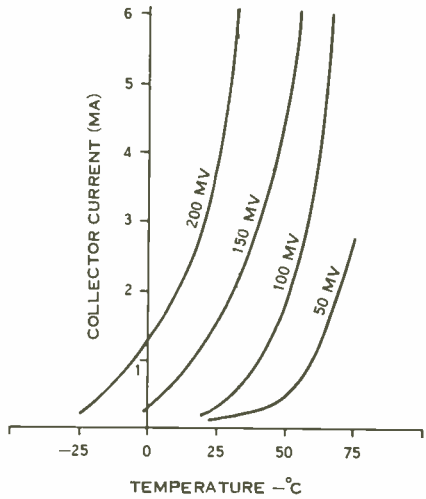


Figure 1. Output current versus temperature in typical common-emitter transistor circuit for several values of collector bias.

circuits with much greater bandwidths. Figure 1 illustrates how transistor characteristics may vary with temperature.

Transistorized IF Amplifier Design

Problems of this type become most acute in the design of a wideband IF (intermediate frequency) amplifier for use in a microwave receiver. It is in this section of the microwave system that the signal is weakest, where the greatest amplification occurs, and where phase and amplitude distortion are most likely. In short, this is the portion of the system where skimpy or inadequate engineering can be most harmful.

In conventional design practice, each IF amplifier stage is tuned to achieve maximum gain over the desired band of frequencies. The overall bandpass and phase shift characteristics of the receiver are determined by the chain of tuned amplifiers, each of which is important to receiver performance.

At the standard 70-mc IF frequency, component values in the tuned circuits

World Radio History

are quite small, and even the inter-electrode capacitances of electron tube amplifiers are an important part of the circuit tuning. Although these values don't change appreciably with the age of the tube, they do vary from tube to tube, and the phase and amplitude characteristics of the IF amplifier can be considerably altered merely by replacing an aging tube.

In transistor amplifiers, the problem may be worse. Here, replacement is less of a problem than maintaining the stability of the circuit in the face of changing conditions. Transistors introduce a considerable capacitive or inductive reactance, depending on the operating frequency and the way in which the transistor is used. To a much greater extent than in electron tubes, this reactance varies according to the bias applied to the transistor.

Normally, this wouldn't have much importance except that in a microwave system, wide variations in signal level exist because of fading, and these are normally compensated for in the IF amplifier. This is achieved by applying an AGC (automatic gain control) bias voltage to various stages of the amplifier.

When this technique is used in transistorized amplifiers, the desired gain control may be achieved by varying transistor bias, but the resulting changes in transistor reactance affect the tuning characteristics of the stage. The overall effect is to degrade the receiver band-pass and phase shift characteristics and introduce a subtle form of intermodulation distortion.

There are several ways of reducing this type of difficulty. One is to provide far more gain than is actually required for the signal, then use most of it for negative feedback and AGC. The negative feedback helps stabilize the individual stages, and the AGC controls the operating point of the amplifiers. This conventional approach encounters serious difficulties in transistorized equipment as the bandwidth is increased. The tuned transistor circuits tend to become more and more unstable. The common-emitter transistor connection (that will most likely be used in order to achieve the necessary gain) must be neutralized, and the tuning and line-up of the equipment becomes increasingly critical. In very broad-band circuits, feedback may not be very effective due to the difficulty of maintaining the required 180° phase shift over the entire bandwidth. Because of the critical tuning required, actual field perform-

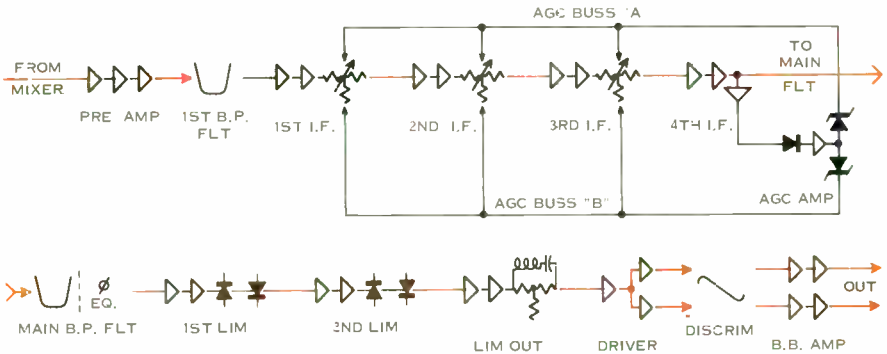


Figure 2. Simplified block diagram of Type 76 receiver IF section. Separation of tuned circuits and transistor amplifiers provides more stable operation, better delay equalization.

World Radio History

ance is likely to become degraded.

An entirely different approach was taken in Lenkurt's new *Type 76* microwave equipment. Rather than fight the complicated interaction of transistor behavior and tuned circuits, the two were entirely divorced from each other. As shown in Figure 2, all tuned circuits were removed from the amplifier stages and isolated in two bandpass filters. This provides several advantages: the bandpass characteristics of the receiver become independent of the amplifiers, the signal level, or the way in which the amplifiers are operated. Replacement of amplifier components no longer has any effect on the tuning characteristics of the receiver.

In addition, this arrangement makes it possible to include an envelope delay equalizer which exactly matches the characteristics of the IF bandpass filters, thus permitting much tighter control of envelope delay, a very important consideration in the transmission of color television and high-speed pulse or data signals.

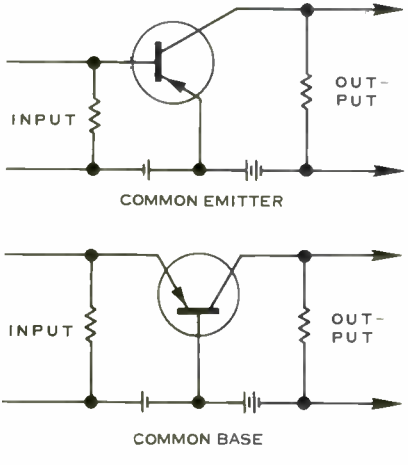


Figure 3. Comparison of common-emitter and common-base transistor arrangements. CE provides high current gain, but may be unstable in wide-band circuit. CB is very stable but requires transformer output.

With the tuning elements eliminated from the amplifier circuits, it became possible to take a fresh approach to the design of the transistorized IF amplifier circuits. The object was to improve circuit stability and frequency response, and to reduce the effects of temperature, time, and other external influences. Since most of the trouble in very wide-band transistor amplifiers comes from the variation in reactance within the transistors themselves, circuits were devised which overcame this effect, thus making the IF amplifier essentially a resistive, constant-impedance network. In this approach, three closely-related techniques were used: a stable transistor configuration, an improved interstage coupling technique, and transistor reactance cancellation.

Transistor Characteristics

Throughout the IF amplifier section, a common-base transistor circuit was used, instead of the more typical common-emitter arrangement. The common-emitter configuration is the most widely used transistor circuit because it provides the greatest power gain of the three possible arrangements, and is easy to use because of biasing and impedance-matching considerations. By contrast, the common-base arrangement is little used because it has a current gain less than unity and therefore always requires some form of transformer for coupling it to following stages, thus increasing manufacturing cost.

Unfortunately, the popular common-emitter circuit is the least stable of the three from the viewpoint of temperature and bias variations. In addition, age shows its greatest effect in the common-emitter version. Figure 3 shows these basic arrangements.

The reason for this contrast in stability lies in the basic nature of transistors. The current gain or α of a transistor connected in the common base configuration is only slightly affected by age or

World Radio History

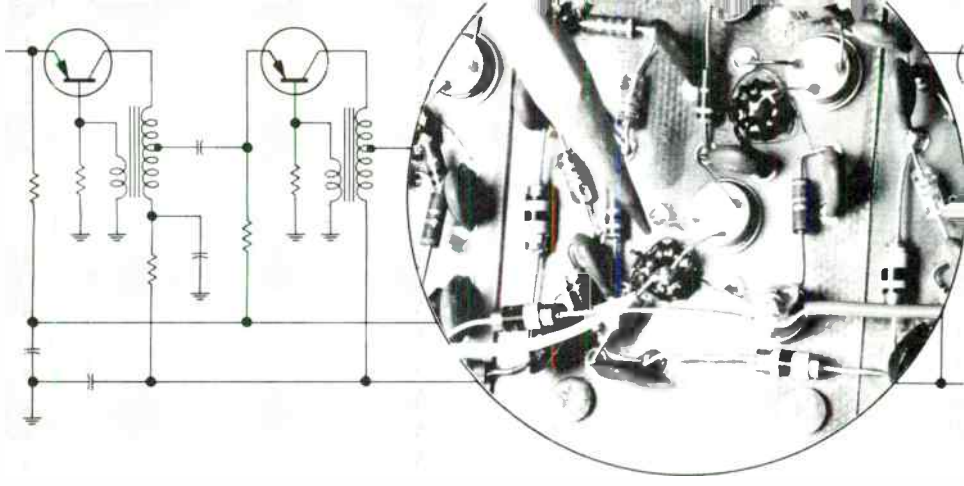


Figure 4. Simplified schematic and photograph of typical 76 IF amplifier stage. Toroidal interstage transformer (at tip of pencil) provides flat frequency response to 200 mc, allows reactance feedback, thus stabilizing the transistor and extending its frequency response.

such external factors as temperature and bias. In the common-emitter arrangement, however, gain is proportional to β which is controlled by α according to the relationship

$$\beta = \frac{1}{1-\alpha} .$$

A quick bit of arithmetic shows that very small changes in α result in very large changes in the value of β. For instance, should the α of a transistor change from 0.995 to 0.994, the gain of the common-base configuration will only vary by 0.1%. In the common-emitter arrangement, however, the gain would drop from 200 to 166, a change of 17%. At lower frequencies, the problem of compensating for this change in gain is much less difficult and the greater economy of the common-emitter circuit usually recommends it over the common-base arrangement.

Transformer Coupling

In order to take advantage of the superior stability of the common-base connection, it was necessary to provide an interstage coupling that would match the output impedance of one stage to the input impedance of the next and, at the same time, convert the voltage gain of the amplifier to useful current gain.

This requirement can be satisfied by a transformer, but conventional transformers cannot operate efficiently at 80 mc and still provide a good impedance match to the transistors.

In the 76 receiver, a novel interstage coupling transformer is used which satisfies all these requirements with remarkable efficiency. Not only does this device provide efficient coupling and excellent impedance-matching characteristics, it opens the way to sharply-improved transistor performance. Figure 4 shows a simplified schematic and a photograph of typical IF amplifier stages in the 76 receiver. Interstage coupling is achieved by autotransformer action in the main winding of the toroidal transformer. This produces a stepdown in voltage but an increase in signal current, thus correcting the low current gain of the common-base connection. The special transformer design used in the Lenkurt interstage coupling provides uniform frequency response

World Radio History

out to at least 200 mc, so that IF frequency response is limited by transistor characteristics rather than by the coupling network. The black curve of figure 5 shows typical frequency response of the resulting IF amplifier.

Transistor Reactance Cancellation

Although transformer coupling is unusual at these frequencies, this particular kind of transformer circuit provides still other benefits not yet attainable by more conventional techniques.

Transistors, like most components, unavoidably introduce reactance, and this limits the ultimate frequency response of the transistor. Because of the special nature of the Lenkurt interstage transformer it is possible to apply a special form of feedback to the transistor base which tends to neutralize the reactance appearing in the transistor itself. Unlike conventional negative feedback which corrects *amplitude* variations, this feedback arrangement tends to offset the changes in circuit *reactance* caused by the changing signal voltage across the transistor junction. The net result is an amplifier circuit that is essentially resistive and remarkably free of phase distortion.

AGC Technique

One of the important characteristics of transistors is that variations in load resistance cause a corresponding change in input impedance. Similarly, variations in input impedance affect the transistor output impedance, and these changes may have an undesirable effect on the over-all performance of wide-band circuits. Even though these undesirable effects are largely overcome in the Lenkurt 76 IF strip, due to the interstage coupling and type of feedback used, an improved automatic gain control technique was developed to maintain the inherent stability of the IF amplifiers under all conditions.

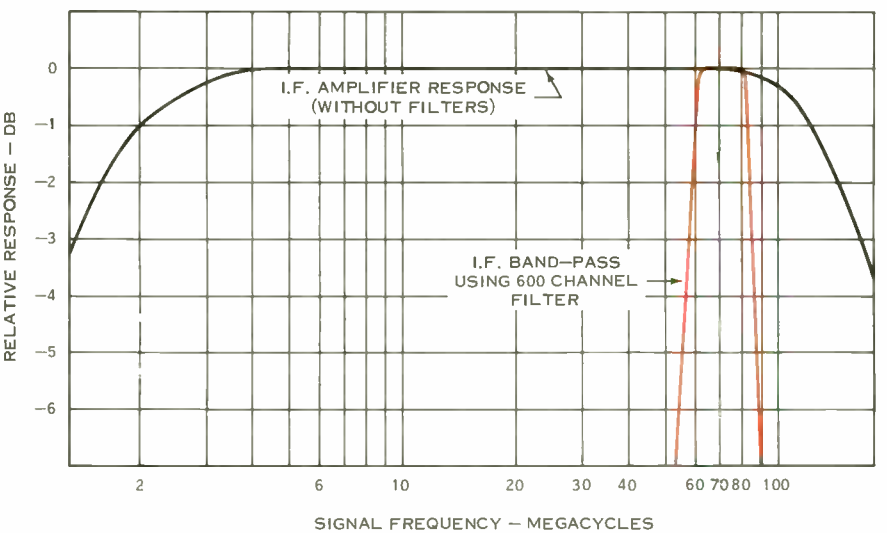


Figure 5. Frequency response of Lenkurt 76 IF amplifier with (red curve) and without (black curve) lumped bandpass filters. Other filters and equalizers may be substituted in applications having different bandwidth requirements.

World Radio History

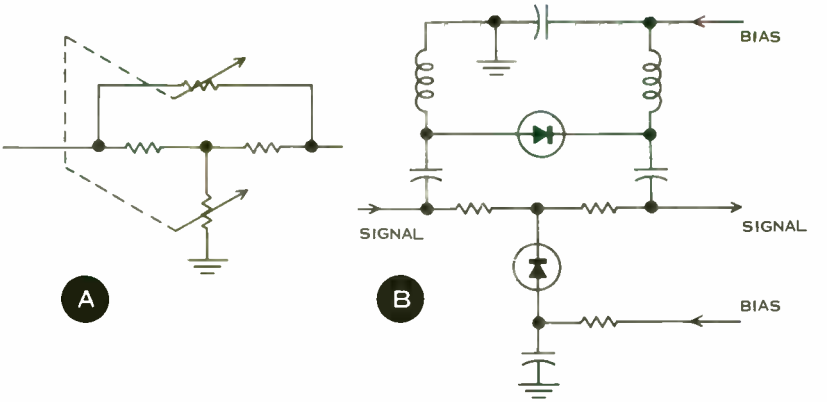


Figure 6. Typical bridged-T attenuator shown in A maintains constant impedance in both directions as one variable resistance increases, and other decreases. Lenkurt circuit uses variable resistance of semiconductor diode at different current values to achieve same effect in AGC circuit. Additional components shown isolate dc bias and IF signal from each other.

Instead of altering the gain of the controlled stages by changing transistor bias, a form of "variolosser" or variable-loss circuit was introduced to control the output level of the controlled stages. In this arrangement, an electronically-controlled variable attenuator is inserted in the signal path following each pair of transistor amplifiers. This permits the transistors to operate at a constant fixed point on their characteristic, with the benefit of improved performance and possible extended life.

In the Lenkurt circuit, the variolosser takes the form of a bridged-T attenuator with variable elements, as indicated in Figure 6. If both the series and shunt elements are varied inversely, input and output impedance of the network will remain constant over a wide (20 db) range of attenuation values. Two bias voltages are required, one to control the series resistance, and another opposite-going voltage to control the shunt resistance. By deriving these voltages from the same source, the two variable elements "track" very well and maintain the desired constant impedance over the operating range of the circuit.

Conclusions

The design approaches described above for a particular piece of equipment illustrate the trend toward increasing sophistication and ingenuity that is required to achieve superior performance. As the number of channels transmitted over a microwave system increases, it becomes more important to preserve or improve performance quality because of the increased value of the resulting communication. The upsurge in high-speed data and wideband video transmission increases the stringency of the required transmission standards, and makes it particularly important that state-of-the-art limitations of transistors and similar components not be allowed to determine performance capabilities. •

Next month, the second article in this two-part series will discuss ways of overcoming intermodulation distortion in the critical modulation and demodulation circuits of a microwave system—the transmitting klystron and receiver discriminator.

World Radio History

World Radio History

the Lenkurt Demodulator

VOL. 10 NO. 10 OCTOBER, 1961

THE NEW MICROWAVE

Part 2

The third-generation microwave equipment now appearing reflects the growing demand for very high channel capacity and increased reliability. One way that reliability is sought is by substituting transistors for electron tubes. Transistors, however, may introduce new problems which require skillful engineering to solve satisfactorily — as described in last month's article.

This article discusses additional design problems which stem from increased bandwidth and transistor characteristics — the problems of nonlinearity in the transmitting klystron and the receiver discriminator.

Although heavy-duty microwave equipment of extremely high quality has been in service for over ten years carrying television and long-distance telephone circuits from coast to coast, this equipment is too costly for most of the newly-emerging industrial and other medium-length applications such as educational television. The early equipment was forced to obtain the necessary performance by complex circuits and elaborate modulation methods.

Simpler ways of doing much the same job are now available, but the intense desire to achieve greater reliability through the use of transistors provides additional complications. Transistors aren't *worse* than electron tubes — they are just *different*. Transistors and transistor techniques are advancing rapidly, and only now have reached the point where they can be applied to high-quality microwave equipment with assurance.

World Radio History

As bandwidth is increased, extremely careful design is required to preserve linearity, not just in transistor circuits, but throughout the system. Noise, the bane of communication, finds more opportunities to creep into the system when the bandwidth is increased. Ironically, noise thrives and grows whenever the techniques used for overcoming it are misused. For instance, thermal or background noise may be reduced by increasing the modulation index (and frequency deviation) of an FM radio system. If this is carried beyond a certain limit set by the equipment, nonlinear or intermodulation distortion provides far more noise than is eliminated.

When additional channels must be accommodated, the required extra bandwidth "uses up" some of the frequency deviation capability of the equipment and limits its ability to fight the ever-present background noise.

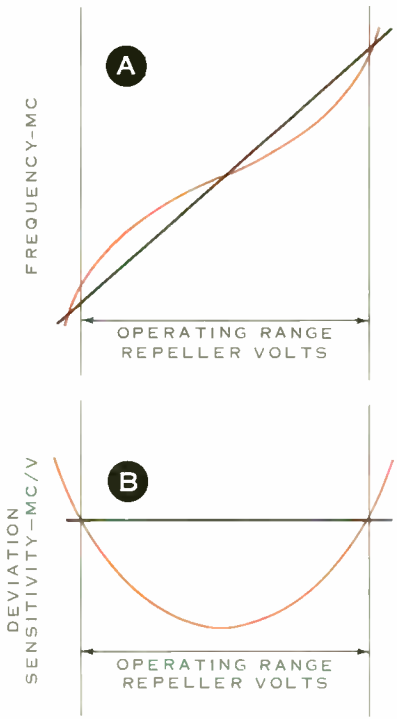


Figure 1. Non-linear modulation characteristic of reflex klystron operating into a matched load is shown by red curve in A. Resulting non-uniform modulation sensitivity is shown in B. Black curves represent ideal linear performance.

Klystron Linearity

In ordinary amplifiers, intermodulation distortion can be overcome by negative feedback — which causes nonlinearities to cancel themselves out. Such a technique, although possible, is not economically practical where the microwave signal is obtained from a modulated klystron — as is the case in virtually all commercial microwave systems operating at frequencies of 6000 mc or higher.

Unfortunately, these klystrons are inherently non-linear in their modulation characteristics, and this non-linearity increases as the frequency deviation becomes greater. The reason for this is that klystron output frequency is controlled both by the modulating voltage applied to the klystron repeller, and by the impedance of the klystron load. The tuning of the klystron to a given output frequency is essentially an impedance-matching process. Now, as the modulating signal alters the output frequency of the klystron, an impedance mismatch is created which changes with the instantaneous frequency deviation, and this mismatch has a de-tuning effect which either opposes or aids the effect of the modulating signal. The result is non-linear distortion which becomes greater as the deviation is increased. Figure 1*A* diagrams a typical klystron modulation characteristic in terms of repeller voltage versus frequency. Figure 1*B* shows how *deflection sensitivity* —

World Radio History

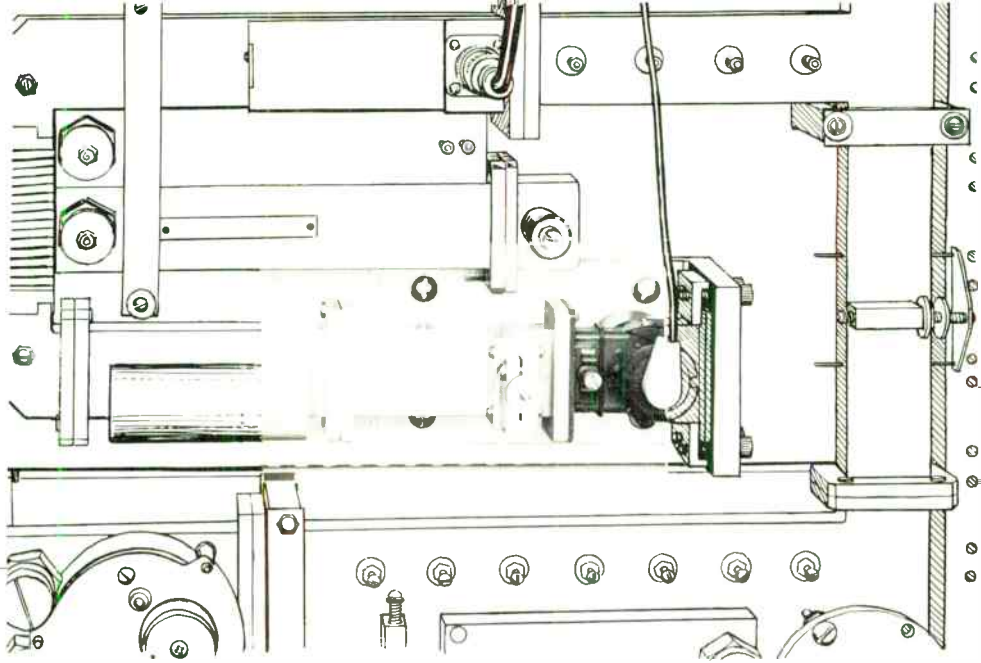


Figure. 2. Klystron linearizer (shown in photo insert) is attached directly to the transmit klystron. Unit is completely passive, may even improve power output from individual klystrons. Attaching device to left of linearizer is a ferrite isolator to eliminate effects of possible impedance mismatch with load.

the amount of frequency *change* for a given change in repeller voltage — varies with frequency.

This non-linear klystron frequency response is the principle cause of intermodulation distortion in a microwave transmitter, and thus a very important source of system noise at times of peak load. The greater the frequency deviation, the more distortion that occurs. Obviously, the distortion problem becomes much worse when additional channels must be accommodated by the klystron, as in 600-channel systems. If the overall frequency deviation is not increased to match the increased channel load, each channel's share of the frequency deviation is proportionately lessened, thus reducing the FM noise advantage. Conversely, if per-channel deviation remains the same as in smaller systems, the increased total deviation may cause intolerable intermodulation distortion.

The problem was overcome nicely in the high-capacity Bell TD-2 microwave system by the use of a so-called klystron linearizer. This is a device that causes a portion of the outgoing wave to be reflected back to the klystron so as to cancel the impedance-mismatch caused by frequency deviation. The exact phase and amplitude of the reflected wave are extremely important, and in the Bell linearizer, the phase is controlled by the location of a plunger or short circuit in a long (three to eight foot) section of waveguide or coaxial cable; amplitude of the reflected wave is controlled by a variable attenuator. Although a substantial portion of the klystron output is lost in the attenuator, this is of

World Radio History

little importance in the TD-2 system, since the signal undergoes further conversions and amplification before it is transmitted.

A related device is used in the Lenkurt *Type 76* 600-channel microwave transmitter. Figure 2 shows the linearizer attached to the transmit klystron. Although much smaller and of a different design than the TD-2 linearizer, it performs exactly the same function. In the Lenkurt design, a portion of the transmitted energy is reflected from a wide-band filter section in the linearizer waveguide, the phase of the reflected signal being controlled by an adjustable stub in the cavity. The amplitude of the reflected signal is determined by an adjustable waveguide transformer, located near the klystron end of the linearizer. By adjusting both the transformation ratio of the transformer, and the phase of the reflected signal, it is possible to obtain any desired degree of linearity for the transmit klystron. Figure 3 compares the deflection sensitivity of a klystron operating with and without the linearizer. The presentation shown in the photographs is comparable to Figure 1*B*, and is directly related to the linearity of the klystron; the greater the curvature, the greater the intermodulation noise added to the signal.

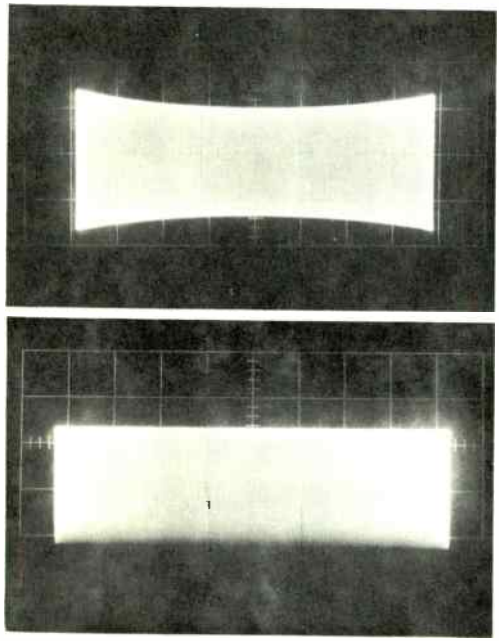


Figure 3. Actual unretouched photographs of effect of linearizer. Top photo shows variation in klystron sensitivity across a 20-mc bandwidth. Same klystron with linearizer shows no measurable variation across identical deviation range. Difference in width of the two displays is caused by oscilloscope adjustment.

Transistorized Discriminators

At the receiver, a similar problem exists. After passing through a series of IF amplifiers and limiters, the signal must be restored to its original form and range of frequencies. This is usually accomplished by a phase discriminator, the most popular kind of which is named after its inventors, Foster and Seeley.

The discriminator has the ability of detecting the *frequency rate of change* of the FM signal, and converting this to a signal voltage identical to the original modulating signal. In the discriminator, as in the klystron, any departure from perfect linearity in this conversion process results in intermodulation distortion much more serious than that generated elsewhere in the receiver.

One way of achieving discriminator action is to apply the IF signal to two resonant circuits, each tuned to a frequency near one end of the band of interest. For instance, one could be tuned to 58 mc, the other to 82 mc, as shown by the gray curves in Figure 4. If the circuit is well-designed and properly adjusted, the two characteristics will combine to yield the characteristic

World Radio History

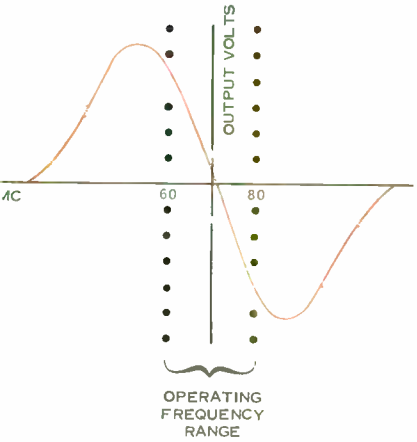


Figure 4. Typical discriminator characteristic (red) is sum of gray curves, which represent response of tuned circuits or phase shift of two sections of discriminator. Linearity becomes increasingly difficult to maintain as bandwidth increases.

shown in red. Note that curvature in one of the resonance curves just offsets that from the other so that the over-all response is a straight line from 60 to 80 mc. Obviously, if the circuit is not precisely balanced, or if one resonant circuit differs slightly from its companion, the resulting characteristic will exhibit irregularities which cause distortion. The phase discriminator accomplishes exactly the same function in a somewhat different manner, but is still subject to the same need for maintaining exact balance between the two symmetrical halves of the circuit.

As bandwidth becomes greater, it becomes increasingly difficult to match the two halves of the circuit exactly and maintain linearity. The most direct way is to lower the circuit Q so that there is less curvature to the resonance characteristic, and thus reducing the variation from the desired straight-line discriminator characteristic. However, this reduces the output level obtained from the discriminator.

Low output from the discriminator produces two distinct problems; semiconductor diodes, which replace electron tubes in the so-called "solid-state" equipment, have a characteristic similar to that diagrammed in Figure 5. Note that forward conduction does not begin until there is a potential of several tenths of a volt across the diode, and that in this region, diode conduction is extremely non-linear. At low signal levels, therefore, there is a likelihood of increased distortion. In order to avoid this type of distortion, it is very desirable to drive the discriminator at as high a signal level as possible.

A second difficulty arising from low signal output from the discriminator relates to the introduction of excessive noise in the demodulated baseband signal. Although it is more convenient in

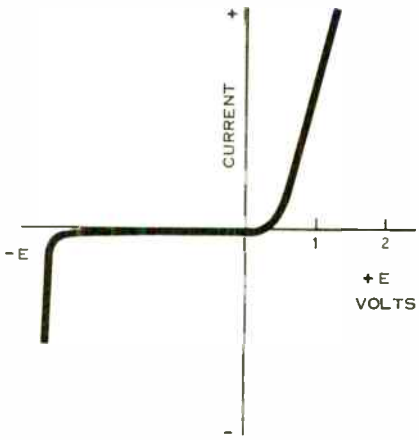


Figure 5. Typical semiconductor diode voltage-current characteristic. Nonlinearity at very low signal voltage can cause serious distortion. High-level driving signal minimizes this effect.

World Radio History

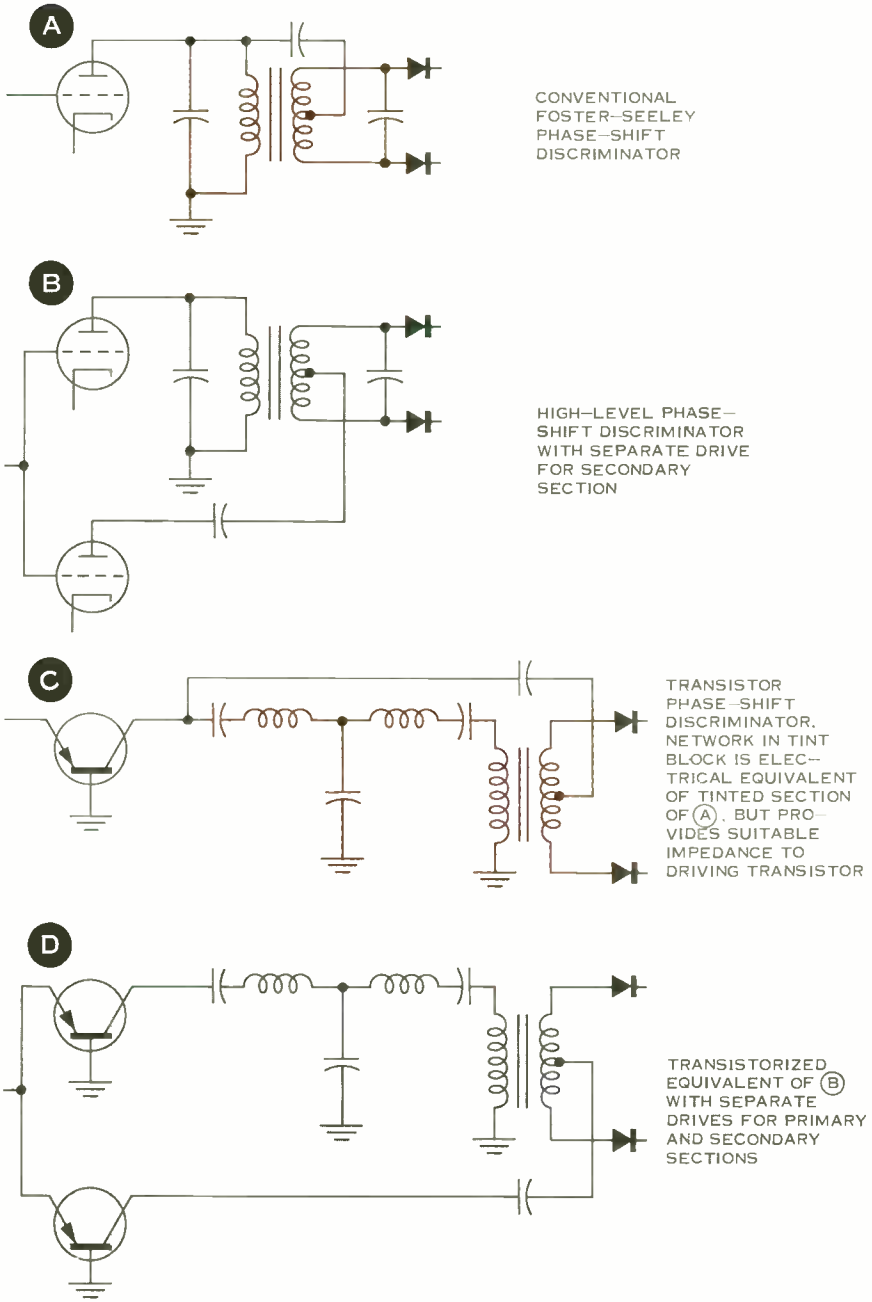


Figure 6. Evolution of a high-level, transistorized discriminator from a conventional Foster-Seeley phase discriminator using electron tubes.

World Radio History

conventional practice to obtain amplification at the baseband frequency, rather than the IF frequency, this can drop the signal level dangerously near the noise level of the diodes.

Several other factors complicate the design of a discriminator for transistorized equipment. At 80 mc, conventional transformers introduce too much loss and non-linearity. If air-core inductances are used, efficiency drops sharply, making it difficult to obtain an adequate output level from the discriminator. In the Lenkurt 76 discriminator, this problem was overcome by using the same high-frequency toroidal transformer technique as used in the IF amplifier sections for interstage coupling and reactance feedback (see DEMODULATOR, *September,* 1961) in constructing the discriminator hybrid circuit. These transformers provide efficient coupling and highly linear frequency response to frequencies above 200 mc, and thus prove to be highly satisfactory in the wide-band discriminator.

Another problem in designing a transistorized discriminator is that of achieving a suitable impedance match between the load, the hybrid, and the driving amplifier. In the 76 receiver it was considered very desirable to use the same common-base configuration for the driving amplifier as was used in the IF amplifiers, in order to maintain utmost stability and freedom from impedance variations which would tend to unbalance the discriminator and introduce distortion.

A special discriminator circuit was developed to match the output impedance of the transistor to that of the transformer hybrid. This circuit and a conventional discriminator arrangement are contrasted in Figure 6, which shows the "evolution" of the high-level discriminator used in the 76 receiver. The circuits shown within the pink-tinted areas are essentially equal in their electrical function, but are adapted to the particular circuits with which they are used. The T-network shown in Parts *C* and *D* of the illustration has the property of providing a suitable match between the transistor and the hybrid, and also achieving the desired phase shift required for discriminator action.

The net result of this technique is a discriminator of unusual efficiency, and which provides an unusually high output for its bandwidth. Separate driving amplifiers for the two branches of the hybrid permit an input level of +15 dbm, and thus allowing a very high output. An indication of the improvement is provided by the 0.15 volt-per-megacycle output obtained from the 76 discriminator, as contrasted with the 0.03 volt-per-megacycle more typically obtained from other discriminators operating across the same bandwidth.

Conclusions

Despite the widespread appeal of transistors, they are no panacea, and their use can result in degraded performance under some conditions, if suitable engineering skill is not employed in overcoming their present limitations. Radio bandwidth requirements are increasing, and this trend can be expected to grow, thus making it more difficult to maintain high standards of performance in communications equipment. Increased channel capacity and increasing circuit lengths for light and medium radio systems make it imperative that performance standards be *improved* rather than relaxed in the face of technical difficulties. Performance and reliability will improve continuously under the pressure of diligent research so long as the users of microwave equipment continue to appreciate the long-term values so obtained. •

World Radio History

World Radio History

the Lenkurt Demodulator

VOL. 6 NO. 12 DECEMBER 1957

6,000 MC RADIO SYSTEMS

Some Equipment and Operating Considerations

The 6,000 mc common carrier and industrial bands are valuable because they can accommodate very large numbers of voice channels. The characteristics of 6,000 mc radio waves, however, call for circuitry which differs in several important respects from conventional radio. Further, propagation is affected more by the atmosphere at this frequency than at lower frequencies.

This article describes two major differences between circuits for 6,000 mc and conventional radio. It also discusses the effect of the atmosphere on propagation at 6,000 mc.

Conventional radio circuits are used for communication at frequencies up to about 2,000 mc. Above this region they will not operate satisfactorily. Components such as standard radio tubes and two-wire or coaxial transmission lines either fail completely or have too much attenuation. In recent years, however, new components have been devised that make microwave communication in the region above 2,000 mc practical and economical.

One of these components, the klystron vacuum tube, is now widely used at frequencies up to 25,000 mc or greater. In some respects it performs more efficiently at microwave frequencies than an ordinary radio tube does at broadcast frequencies.

Another component, the waveguide, makes use of the very short wavelengths of microwaves to provide an effectively shielded transmission line having relatively low loss. The waveguide is not only capable of operating as a transmission line but also can be designed to operate as a capacitance, inductance, filter, or hybrid. When used with auxiliarly devices such as magnets and ferrites, it can be made to operate as an isolator, circulator, modulator, discriminator, or attenuator.

World Radio History

Besides having different terminal and repeater components, microwave radio systems are affected more by atmospheric conditions than broadcast or other low-frequency radio. Special consideration must be given to 6,000 mc equipment location to overcome the deeper and more frequent fades which occur as frequency increases in the microwave region. Additional importance is also given to the sensitivity and noise figure of receivers because of the deep fading.

Vacuum Tubes

In conventional vacuum tubes operating below microwave frequencies, the time required for an electron to travel from the cathode to plate is very small compared to the time required for a signal on the control grid to go through one cycle. However, if a microwave signal is applied to the grid, its cycle may be short compared to an electron's *transit time,* and will become shorter as the frequency increases. The flow of electrons in transit toward the grid will not be able to follow the signal variations exactly.

The plate-current wave will be very distorted and will be out of phase with the grid voltage. In addition, power will be dissipated at the control grid and, to a lesser extent, at the cathode. The result is distortion of the output signal and loss of gain.

To overcome this problem a klystron vacuum tube uses a stream of electrons to excite a resonant cavity within the tube. The tube output is taken from this cavity at its resonant frequency. Klystrons do not depend on a plate current to create an output voltage.

Klystrons used in microwave transmitters generate the carrier frequency which is modulated and fed to the sending antenna. In receivers they generate the radio frequency which is mixed with the incoming signal to produce an intermediate frequency for amplification and detection. Klystrons have now been developed to the point where they can be made about as reliable as the best standard industrial tubes. Fig. 1 is a schematic representation of the internal and external circuits of a klystron oscillator.

Waveguides

One of the characteristics of electromagnetic waves is that they can be confined in, and propagated along, hollow metal tubes. Such tubes may be circular or rectangular in cross-section, but a rectangular tube is generally the most practical because of its wide frequency range and ability to maintain wave polarization. The power loss in a waveguide is about one-third the loss in a comparable air-insulated coaxial line, and very small compared to the loss in a flexible coaxial cable with solid insulation (rubber, plastic, etc.). Waveguides can be made rigid or flexible and have an infinite life as long as they remain free of corrosion or dents.

One factor limits the use of waveguide. This is the physical dimension required for propagating given frequencies. The lowest frequency a waveguide can transmit is determined by its width. The wavelength at this lowest frequency, or *cut-off frequency,* is twice the width of the waveguide. This means, for example, that a waveguide designed to transmit a 30 mc signal must be at least 17 feet wide. Needless to say, this is impractical.

World Radio History

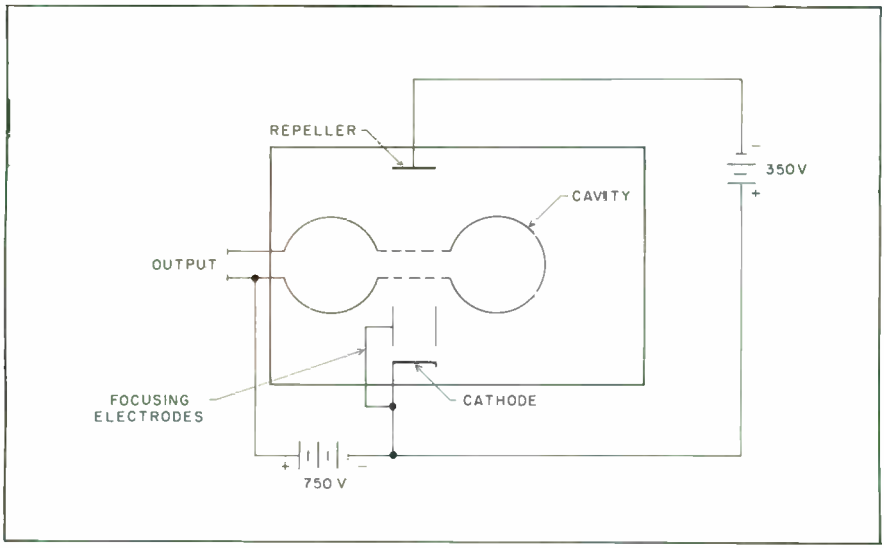


Fig. 1. Diagram of a reflex klystron. Electrons flow from the cathode across the cavity gap toward the repeller. The repeller returns the electrons in bunches to the cavity gap. These electron bunches induce voltages across the cavity at its resonant frequency and maintain a condition of oscillation.

At 2,000 mc, however, half a wavelength is about three inches, and at 6,000 mc only one inch. Waveguides can be built economically to handle these frequency ranges, and they lend themselves to convenient equipment arrangements. Fig. 2 shows the compact waveguide assembly used in Lenkurt Type 74A *Microtel* equipment.

Propagation

Radio signals travel from a transmitter to a receiver by three principal means: *ground waves, sky waves,* and *space waves.* Ground waves cannot be used at microwave frequencies because they are completely attenuated within a few feet of the transmitter. Sky waves are reflected, refracted, or scattered back to earth by ionized layers of the upper atmosphere (the ionosphere), but only to a small extent above 100 mc. At microwave frequencies they are usable only in very high-power expensive systems which must bridge great distances in a single hop. Space waves, which travel through the atmosphere immediately above the earth, are the most practical propagation means for microwaves.

To be usable, the space waves must arrive at the receiver with a certain minimum signal strength. Below this minimum, known as the *threshold level,* the signal is drowned out by receiver noise. When the received signal drops below this threshold in a common carrier radio system, telephone circuits connected to it through a dial exchange will disconnect. To restore service they

World Radio History

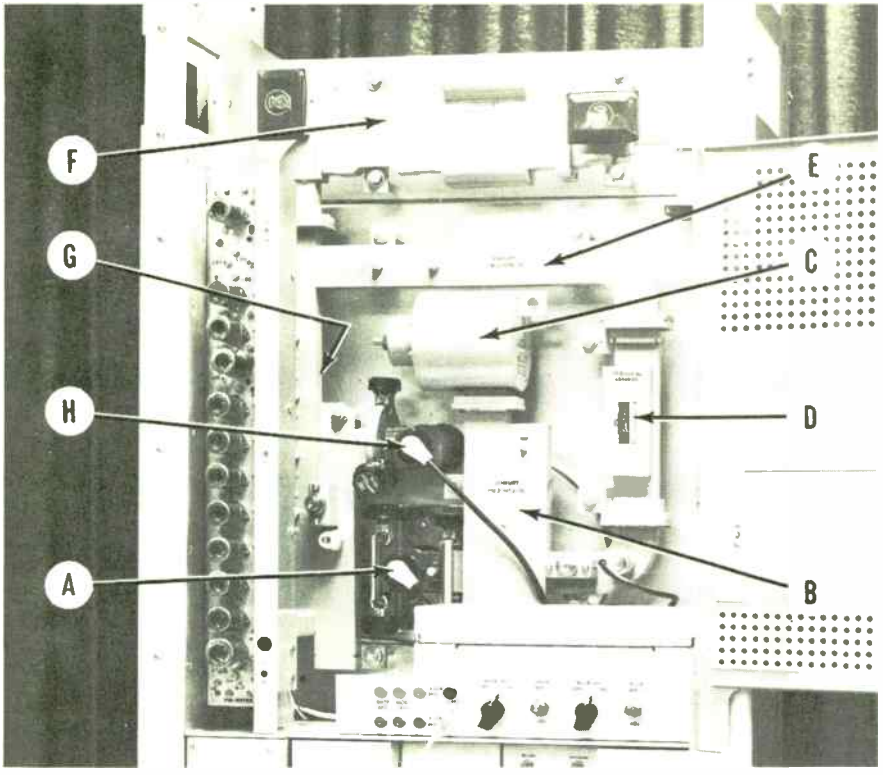


Fig. 2. Type 74A transmitter-receiver with waveguide assembly exposed. The principal microwave components shown are: (A) transmitting klystron; (B) waveguide discriminator; (C) reference cavity for controlling transmitter frequency; (D) isolator; (E) waveguide run to circulator panel; (F) circulator panel; (G) waveguide run to r-f mixer; and (H) local oscillator klystron.

must be redialed. Further, for toll quality communication, the signal must remain several decibels above the threshold level to maintain the desired signal-to-noise ratio.

The maximum distance allowable between transmitter and receiver for toll quality service is determined by transmitter power output, receiver threshold, and the sum of the losses between them. There is a relatively small loss from the transmitter or receiver to its antenna, an appreciable antenna gain, and a varying path loss. Fig. 3 shows gains and losses of a typical radio section.

The total path loss between antennas is made up of two parts: (1) path attenuation and (2) fading. These losses are determined by path length, path clearance above the earth, atmospheric conditions, and frequency.

Path Attenuation

If a space wave is radiated from a point (isotropic) antenna it spreads out equally to all directions in the shape of an ever-expanding sphere. As the surface of the sphere moves farther

World Radio History

and farther from the point antenna, the radiated energy is spread over a larger area and the amount of energy per square-foot of wave front decreases. Mathematically speaking, the energy concentration at a point on a wave front is inversely proportional to the square of the distance from the antenna.

The power that can be extracted from a wave front by a similar point antenna is inversely proportional to the square of the frequency. Thus, the power received by a point antenna is inversely proportional to both the square of the distance from the source and the square of the frequency. The ratio of this power to the total power radiated is called *path attenuation*.

When the receiving antenna is something other than a point (a parabola-shaped dish, for example), the amount of power extracted from the wave front is greatly increased. The ratio of the amount of power received by a practical antenna to the amount extracted by a theoretical point antenna is called *antenna gain referred to an isotropic radiator*.

The gain of a parabolic antenna increases with antenna area. It also increases with operating frequency. So, for a given radio path with fixed-size antennas, the path attenuation increases with frequency. But so does the antenna gain. One tends to offset the other. Table 1 compares path attenuation and antenna gain for two radio sections operating at different frequencies over the same path length with antennas of the same size.

Fading

Fading of received signal strength is caused primarily by variations in atmospheric conditions. These variations

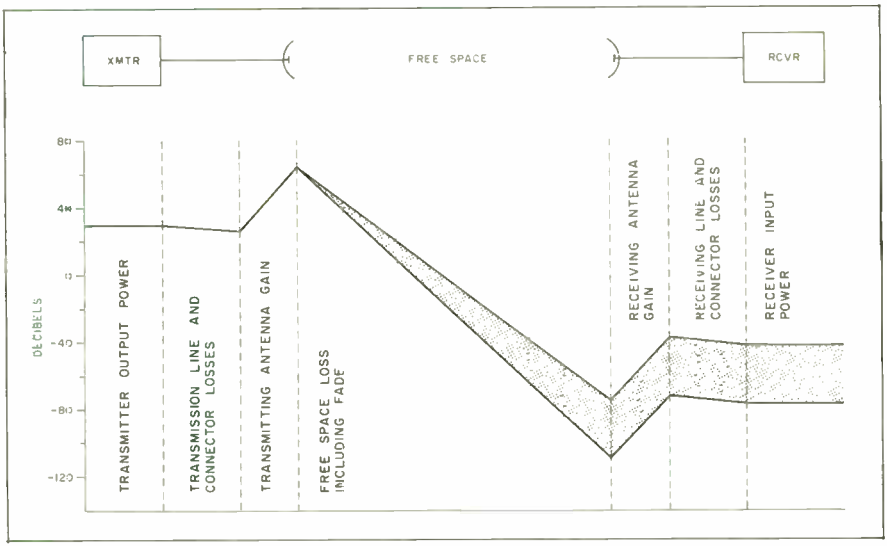


Fig. 3. Gains and losses in a typical radio section. The shaded area indicates the range of losses expected from fading during 99.9% of the time.

World Radio History

TABLE I

COMPARISON OF RADIO SECTION CHARACTERISTICS
AT 1,000 and 6,000 MC

	1,000 MC	6,000 MC	REMARKS
Path Length	25 mi	25 mi	
Antennas	6' Parabolic	6' Parabolic	
Free-Space Path Loss	124.5 db	140.0 db	Computed from: $L = 10 \log f^2 d^2$
Antenna Gain (2 antennas)	46.0 db	77.0 db	Computed from: $G = 10 \log f^2 + 10 \log D^2 - 52.6$ for one antenna
Normal Transmitter Power	+37 dbm	+30 dbm	
Normal Misc. Losses (trans. lines, combining filters, circulators, etc.)	8 db	5.2 db	
Net Received Signal Power	-49.5 dbm	-49.5 dbm	

cause radio waves to bend away from their normal lines of propagation. The fades resulting under a given set of conditions occur in greater number and greater severity as signal frequency increases. The reasons for more frequent fading at higher frequencies can be explained by examining the mechanics of wave propagation.

A signal beamed toward a receiving antenna consists of a series of wave fronts whose centers are on the line of sight from trasmitting antenna to receiving antenna. The surface of each of these wave fronts consists of an infinite number of isotropic radiators sending signals in all directions away from the wave front. Thus, at any instant there are an infinite number of paths from a given wave front to its receiving antenna.

For example, in Fig. 4 if any two paths differ by one-half wavelength or any odd multiple of a half wavelength, the energies received over the paths will cancel. If they are the same length or differ by any whole number of wavelengths, they will reinforce each other. The paths AR and A'R from the wave front to the receiver R are one-half wavelength longer than the line of sight path OR. All secondary waves emanating from within the area defined by AOA' as diameter will reinforce the direct wave, OR, at the receiver because they are less than one-half wavelength out of phase with OR. This area is known as the first Fresnel zone and provides one-quarter of the received field energy.

The path lengths BR and B'R are one wavelength longer than the direct path OR. All secondary waves emanating from the shaded area between the first Fresnel zone and the circle whose diameter is BOB' will act at the receiver

World Radio History

to partly cancel the waves from the first Fresnel zone because they are between one-half and one wavelength out of phase with OR. This shaded area is the second Fresnel zone. All odd-numbered Fresnel zones will reinforce the direct wave and all even-numbered Fresnel zones will cancel odd-zone energy.

The third Fresnel zone is defined by diameter COC′, and the fourth by DOD′. There are an unlimited number of Fresnel zones, with each succeeding one contributing less energy than the one before. The area of the Fresnel zones is determined by their distance from the transmitter and receiver, and the operating frequency. The higher the frequency, the smaller is the difference in path lengths which is equal to a half wavelength; hence, the smaller the first Fresnel zone and the others surrounding it. However, each Fresnel zone still contributes the same proportion of energy.

Fresnel zone sizes are important because they determine the effect of wave bending (refraction) on path clearance above the earth and on reflections from smooth earth surfaces. Smaller Fresnel zones cause obstacles in the radio path to obstruct a greater percentage of radiated energy. They also cause more severe and more frequent cancellations of energy between reflected and directly transmitted waves when the latter are bent or *refracted* by the earth's atmosphere.

Refraction

Refraction occurs when a wave changes velocity in passing from one medium into another. This occurs in air when two layers have different densities. As a radio wave travels upward at an angle through the atmosphere it normally encounters air of decreasing density. Since the top of the wave reaches the lighter air first, it increases

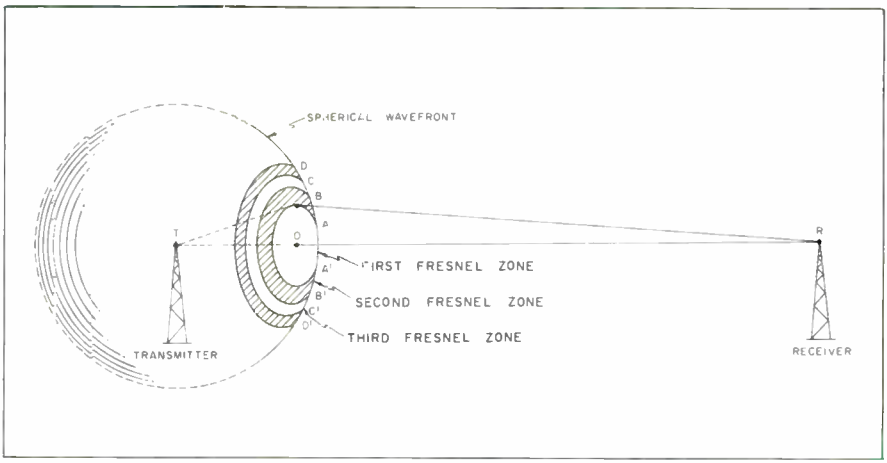


Fig. 4. Fresnel zones of a transmitted signal's wavefront at distance OR from a receiver. The difference in path length to receiver from edge of one zone and edge of adjacent zone is one-half wavelength. Odd-numbered zones reinforce the signal and even-numbered zones cancel it.

World Radio History

its speed first, and the wave bends back toward the earth.

In this way radio waves normally tend to follow the earth or, in effect, the earth appears to flatten and have a larger radius. Correspondingly, if inversion occurs and the air becomes heavier as altitude increases, the wave bends away from the earth.

When atmospheric conditions are such that the air density increases with height, the earth between transmitter and receiver appears to bulge up into the wave fronts. If this effective bulge reaches the line of sight between the transmitter and receiver, microwaves of any frequency will be attenuated about 20 db. However, if the refraction increases so that the earth bulge rises above the line of sight, attenuation will be even greater but no longer equal for all frequencies. Loss in this "shadow zone" increases rapidly with frequency.

For example, Figure 5 shows an effective earth bulge of 67 feet above the line-of-sight caused by atmospheric refraction. The first Fresnel zone radius for a 6,000 mc signal is 81 feet. The 67-foot obstruction results in a −0.83 clearance of the first Fresnel zone at 6,000 mc and causes a loss of 50 db over normal propagation conditions.

In addition to changing the clearance above obstacles, refraction causes fading by changing the relative path length of the direct and reflected waves. As the direct wave is bent above or below the line of sight, its path-length increases so that part of the time the two waves reinforce each other and part of the time they tend to cancel. This is shown in Fig. 6.

If the terrain between the transmitter and receiver is a good reflector, the

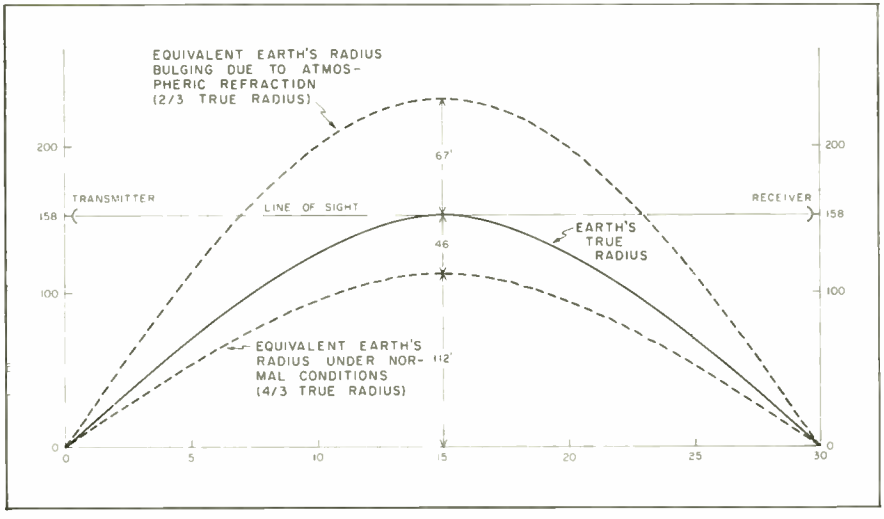


Fig. 5. In this example normal atmospheric refraction permits a signal clearance of 46 ft. above the earth. This is represented by an equivalent flattening of the earth's radius to 4/3 true radius. When atmospheric density increases with height the earth's surface appears to bulge upward. Here a bulge of 67 ft. above line-of-sight gives an equivalent earth's radius of 2/3 true radius.

World Radio History

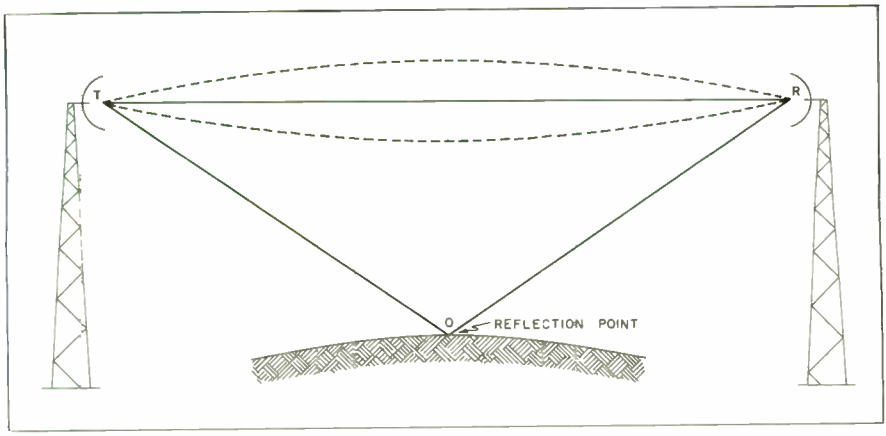


Fig. 6. Refraction of direct wave which results in partial cancellation by reflected wave. TR is the direct path and TOR is the reflected path from transmitter to receiver. Dotted lines show bending of direct wave due to refraction. This changes path length TR and causes alternate canceling and reinforcing by reflected wave TOR.

cancellations may be nearly complete and very deep fades will result. High-frequency waves will cancel each other more frequently than low-frequency waves because smaller changes in refraction are required to cause a difference in paths of one-half wavelength. Thus a 6,000 mc radio system will have more fades than a lower frequency system but they will be of shorter duration.

Fade Margin

To insure that a transmitted signal reaches a receiver with at least a minimum strength for a certain percentage of the time (for example, 99.9%), enough extra signal strength must be available during normal propagation to compensate for most fades. This is called *fade margin*, and is usually determined by actual field experience because there are no reliable formulas for predicting atmospheric conditions.

A fade-margin figure dictated by experience for a typical path at 6,000 mc is 30-40 db. The exact figure used depends on field studies of the particular location. When the signal path is over a good reflecting surface such as water, additional fade margin must normally be allowed.

Conclusions

The frequency bands in the 6,000 mc range available for common carrier and industrial use are valuable because of their high channel capacity. The use of these bands, however, involves special microwave circuitry and increased losses to the space wave. The necessary circuit elements—such as klystrons and waveguides—are readily available and present little difficulty in application. In fact, klystrons and waveguide circuits are more rugged and simpler than low-frequency tubes and coaxial cable. The additional path attenuation and fading at these frequencies can be overcome by shorter sections and antennas of larger gain.

World Radio History

World Radio History

the Lenkurt® Demodulator

VOL. 13, NO. 8 AUGUST, 1964

HETERODYNE REPEATERS For Microwave

Among the most vital elements of a microwave system are the repeaters which amplify and redirect the signal. As more stringent performance demands are made on the system, repeater quality becomes even more important. However, such qualities as low noise and distortion cannot be easily achieved without sacrificing such other factors as flexibility. This article considers the advantages and disadvantages of the so-called "heterodyne" repeater, and compares it to the widely used demodulating, or baseband, repeater.

The nature of microwave radio transmission requires that all but the shortest systems use intermediate repeating stations to provide gain and direction for the signal. Since each repeater consists essentially of a receiver and a transmitter through which the signal must pass, the repeaters are just as important to system performance as is the terminal equipment.

Because a perfect repeater has never been built, the signal is degraded somewhat each time it is retransmitted. Each repeater distorts the signal to some degree while retransmitting the distortions introduced by the preceding ones. This cumulative effect is so important that the length of a system is usually limited by the number of repeaters through which the signal passes. If the repeaters used in a particular system were replaced with ones providing better performance, the system could be extended while maintaining the same end-to-end performance. Conversely, a shorter system could tolerate repeaters with higher distortion.

But other factors also affect the choice of repeaters. Chief among these is flexibility. Some communication systems require that groups of channels, or even a single channel, be dropped or inserted at the various repeater points. This means that the baseband must be available to permit the various channels to be separated.

World Radio History

Other systems, however, may require little or no drop and insert capability. The signal is simply applied at one end, amplified several times en route, and taken off at the other end.

The choice of repeaters for a microwave system, therefore, is somewhat more complex than a simple cost-versus-performance comparison. Before making such a choice it is necessary to define the qualities required of a repeater for use in a specific application.

Performance Criteria

One of the major factors which must be considered in defining the performance of a microwave system is noise. Noise may come from a number of different sources, and it may appear in various ways. As far as the repeater is concerned, however, noise is generally of two types — thermally generated random (or "white") noise, and intermodulation products. The repeater itself generates both types of noise, adding its contribution to that already present. Thermal noise is picked up by the antenna and generated in the electronic circuitry, while intermodulation noise is produced by every non-linearity through which the signal passes.

Thermal noise is independent of system loading, but intermodulation noise increases as the loading increases. Thus, a repeater adequate for 120 channels might produce unacceptably high noise when loaded with 600 channels.

Another factor which is becoming increasingly important in defining the performance of a microwave system is its ability to carry a video signal. While a number of criteria apply here, among the more sensitive are differential gain and differential phase (see "Performance Testing of Television Channels," *The Lenkurt Demodulator*, October and November, 1963). *Differential gain* is the variation in the *gain* of the transmission system as the luminance or

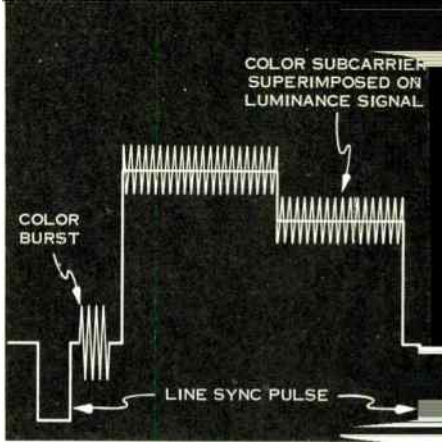


Figure 1. In the American and Canadian (NTSC) color system, "color burst" is transmitted as phase reference for color subcarrier superimposed on luminance signal. Any change in phase or amplitude of subcarrier causes color change in picture. This is called "differential phase" or "differential gain" when caused by change in luminance signal.

brightness signal varies between the values for "black" and "white." Any variation in *phase* of the color subcarrier as a result of changing luminance level is called *differential phase*. Ideally, variations in the *luminance* signal voltage should produce no changes in either the amplitude or the phase of the color subcarrier.

Both of these parameters are directly concerned with color information. In the American and Canadian system, a color subcarrier at a frequency of about 3.58 Mc is superimposed on the luminance signal. Different colors or *hues* are indicated by shifting the phase of the color subcarrier. The *saturation* or richness of the color is transmitted by varying the amplitude of the color subcarrier. In an ideal system, which would have no differential gain or differential phase, changing brightness values in the picture would have no effect on the phase or amplitude of the subcarrier.

World Radio History

However, when differential phase is present, a change in the brightness of the scene could change skin color from pink to purple, while differential gain could change the color saturation from, say, pink to red.

While differential gain and differential phase significantly affect only color signals, such requirements are usually established for any system engineered to carry video signals, even though initially it will only carry black and white.

In addition to specific performance requirements which must be considered when choosing microwave repeaters, other factors concerned with system engineering, such as drop, insert, bridging, and branching requirements must also be evaluated. Furthermore, the contribution of the terminal to signal distortion is much more important in a short system, while repeater performance assumes increasing importance in longer systems. Thus, no one repeater type is likely to provide the best performance in all respects, and a compromise is often necessary.

Types of Repeaters

Every microwave repeater performs two essential functions. Obviously it must provide gain. The power output of a typical repeater is from 55 to 105 db higher than the received power. The other function is not quite so obvious, but it is just as important. Each repeater must also perform a frequency change, transmitting at a slightly different frequency from that at which it receives, to provide enough isolation to minimize interference between the hops. This frequency shift is usually 252 Mc in the common-carrier band, and may be somewhat less in other bands.

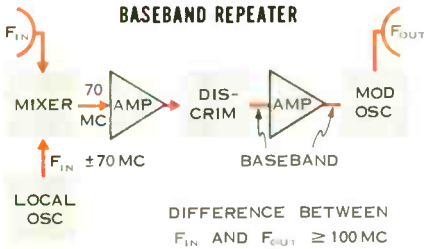
The necessary amplification can be accomplished at any convenient frequency. One widely used method is to connect two ordinary terminals "back-to-back." In this arrangement the signal is translated to an intermediate frequency, amplified, then demodulated and amplified again at the baseband frequency. Finally it is remodulated for transmission in the microwave frequency range. This type of repeater is often called a demodulating, a baseband, or a back-to-back repeater.

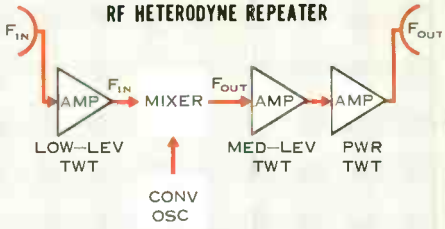
However, amplification can also be provided at the intermediate-frequency, or IF, stage without going through the demodulation and remodulation processes. This is what occurs in an IF heterodyne repeater. As in the baseband repeater, the signal is first "heterodyned" to the IF stage. Here it is amplified before passing through the up-converter to be translated to the microwave frequency. Since the desired power output of a heterodyne repeater is usually beyond the capability of present solid-state devices, the output stage is normally a traveling-wave tube operating at microwave frequency.

Still another type of repeater is the RF heterodyne. In this repeater the amplification is provided directly at the microwave frequencies. Typical block diagrams of the three types of repeaters are shown in Figure 2. There are inherent advantages and disadvantages to each of the three types, even if all are well designed and manufactured.

The RF heterodyne repeater is seldom used for several reasons. Perhaps the most important of these is the present cost of providing gain at microwave frequencies. Typically, gain is provided in three stages, using either traveling-wave tubes throughout, or a parametric amplifier followed by two traveling-wave tubes. Both traveling-wave tubes and parametric amplifiers are quite expensive when compared to the more conventional transistor amplifiers. Other problems, such as designing filters with the required selectivity at microwave frequencies, providing adequate limiting and automatic gain control, and cor-

World Radio History





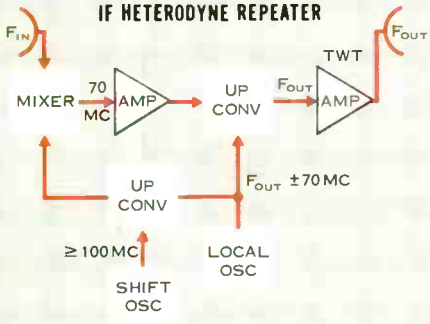


Figure 2. Microwave repeaters are classified by whether they provide amplification at baseband frequency, intermediate frequency, or radio frequency. Heterodyne repeaters eliminate the distortion produced by modulation and demodulation, but baseband repeaters have more drop and insert flexibility.

recting delay distortion, also arise. These problems are not unsolvable, but they are difficult — and hence expensive to overcome. In any case, a one-step frequency conversion must be made to separate the received signal from the transmitted signal.

Thus, the choice is usually between the IF heterodyne (often called simply the "heterodyne") and the baseband repeater. Both are widely used and the choice is dictated by the requirements of the specific system. Probably the most significant advantage of the heterodyne repeater is its improved noise-performance. Each time a signal is modulated or demodulated it picks up a certain amount of intermodulation noise. Since the heterodyne repeater "heterodynes" the signal down to the 70-Mc intermediate frequency and then heterodynes it back up to the microwave frequency without demodulating the FM signal, much of this intermodulation noise is avoided. In a typical case this means about a 3 to 4 db distortion noise improvement in favor of the heterodyne repeater.

Because this same modulation and demodulation process also introduces a large part of the differential gain, the heterodyne repeater has an inherent advantage for video transmission.

The heterodyne repeater also offers better baseband level stability than does the baseband repeater because level variations occur almost entirely in the modulation and demodulation processes. In a system of baseband repeaters these level variations tend to be cumulative, making it more difficult to meet end-to-end objectives such as those required for drop level stability in Direct Distance Dialing.

Another significant advantage of the heterodyne repeater is its increased power output. A typical baseband repeater has a power output of about 1

World Radio History

watt, whereas the output power of the heterodyne repeater is normally 5 watts or more. This five-fold increase in power output means an additional 7 db to improve the signal-to-noise ratio, permitting longer hops or more channels.

But there are also advantages to the baseband repeater. The biggest of these is its flexibility. Since the full baseband is available at every repeater site, it is comparatively simple to drop or insert any desired number of message channels. Channels to be dropped are simply separated from the rest of the baseband by appropriate filters. Other channels can then be inserted into the "slot" left by the dropped channels.

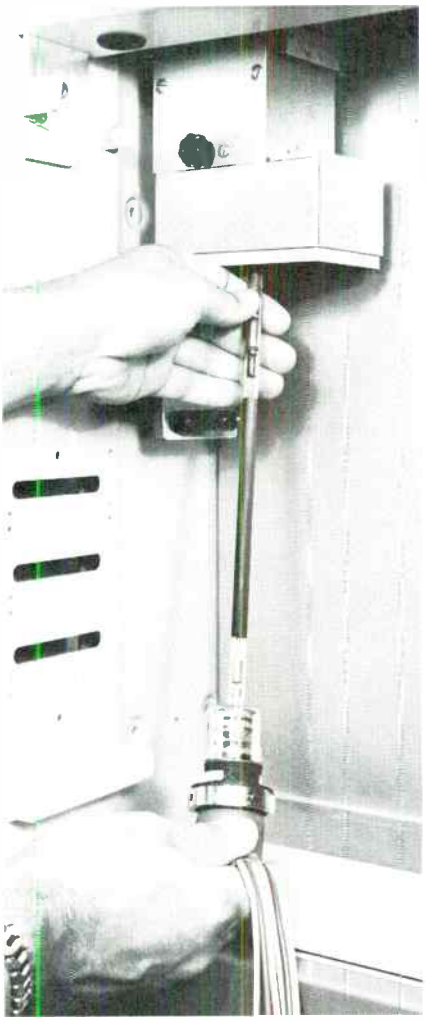


Figure 3. Traveling-wave tube (TWT), shown here being withdrawn from its mount, acts as output amplifier for heterodyne repeater.

The flexibility of the heterodyne repeater, on the other hand, is considerably restricted by the fact that the baseband is usually available only at the end terminals. Since demodulation does not normally occur at intermediate points, channels cannot be dropped as is done in the baseband repeater. Access to the baseband can be provided at repeater sites by bridging in the IF system and then demodulating the portion of the signal which is bridged off. But even so the full FM spectrum appears at each station, with no portion blocked. This means that if channels are to be inserted, idle frequencies must be available in the baseband spectrum. In other words, a number of channels cannot be dropped at some point and a similar number inserted in their place.

The baseband repeater also has the advantage in price — at least when only initial cost is considered. Because of its traveling-wave tube and the required power supply associated with the tube, the heterodyne repeater is usually somewhat more expensive than the baseband repeater.

Thus, the comparison between the baseband and the heterodyne repeaters becomes partly one of cost versus performance, but the choice is usually dictated by the particular requirements of the specific system under consideration.

Applications

In some cases, the choice of repeater type is clear-cut. For a video system 1000 miles long, heterodyne repeaters would be used almost without question. Con-

World Radio History

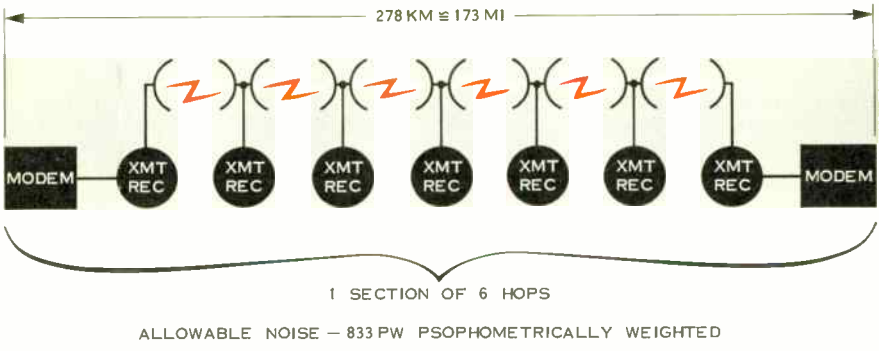


Figure 4. CCIR hypothetical reference circuit consists of nine sections of six hops each, with the sections interconnected at baseband frequency.

versely, for a 120-channel message system 100 miles long, with channels to be dropped and inserted at the repeaters, baseband repeaters would be used — again, almost without question. It is in the middle area between these extremes, where the guidelines are less firm, that questions often arise.

One logical place to start in making a comparison of different systems is the hypothetical reference circuit established by the CCIR (International Radio Consultative Committee). This reference circuit provides a common meeting ground for systems engineers and equipment designers. This 2500-kilometer (1550-mile) circuit is divided into nine equal sections of six hops each. The sections are interconnected at baseband frequency, resulting in nine modulation-demodulation processes in 54 hops. The CCIR-recommended limitation for mean noise power in any hour is 7500 picowatts, psophometrically weighted — equal to 32.7 dba, F1A weighted. (This is the radio contribution only, and does not include noise contributed by the multiplex equipment.)

Breaking this down, each section is 173 miles long, consisting of six 28.8-mile hops, as shown in Figure 4. The allowable noise for the section is then 833 pw, or 23.2 dba. The six feeder sets introduce perhaps 150 pw; the group delay (equalized on a per-section basis) accounts for about 100 pw; and the modulator/demodulator contributes approximately 37 pw. Subtracting these leaves an allowable contribution for the repeaters of 546 pw — 91 pw (13.6 dba) per repeater.

While this hypothetical system does not match any "real-life" system, it does provide a basis for comparing equipment performance against CCIR recommendations and it sets up performance objectives for equipment designers. For example, Lenkurt's heterodyne system is designed to exceed all CCIR recommendations when carrying 960 message channels or a monochrome or color video signal.

While a typical baseband repeater system might be hard put to meet this CCIR objective, it should be remembered that the per-hop objective used here is for a long-haul system. A much higher noise figure might be tolerable for a shorter system. By way of comparison, a 960-channel system using baseband repeaters might have as much as 27 dba per section, considerably

World Radio History

above CCIR's 23.2 dba. But this only means that this particular system could not meet the CCIR noise objective for *nine tandem sections* — 54 hops in all. For shorter systems it could provide perfectly acceptable service; or the same system might well meet CCIR recommendations when carrying only 600 channels.

The other major performance advantage of the heterodyne repeater, decreased differential gain and phase, is considerably more difficult to state quantitatively. Differential gain is inherently lower in the heterodyne repeater because there is no contribution from the modulation and demodulation processes. However, a portion of the differential phase and gain is contributed by other parts of the system, usually through phase and amplitude distortion. Thus, it can be reduced by equalization at the IF stage in either type of repeater. Since the heterodyne repeater is built with more precise delay equalization to combat the systematically accumulated delay distortion, it offers less differential phase—and more performance "margin." For example, suppose a six-hop system has a differential gain requirement of 1 db and a differential phase requirement of 1 degree. Initially, this could probably be met by either baseband or heterodyne repeaters. The difference would come in maintaining the system within these tolerances. On a system of several hops, the lower maintenance costs to keep the heterodyne system within the specifications might well offset the higher initial cost.

Conclusion

The heterodyne repeater is not a new development. It has been used in various applications almost since the advent of microwave communication systems. Until recently its use was confined primarily to long-haul "back-bone" routes, but two factors are combining to change this. One is the vastly increased need for communications of all types—voice, data, telegraph, facsimile, etc.—which is resulting in heavier-density systems. The other major factor is the tremendous increase in video transmission. This includes not only commercial broadcast television, but such other services as educational television and industrial applications.

The heterodyne repeater is not a replacement for the baseband repeater. The baseband repeater will long be an important part of many microwave systems — particularly where drop and insert capabilities at intermediate points are important. The two types of repeaters are, of course, complementary, not competitive. Often both types are used in the same system to take full advantage of the best features of each. For example, a long-haul system may be composed primarily of heterodyne repeaters, with baseband repeaters used where channels are to be dropped or inserted.

Thus, while the heterodyne repeater does not replace the baseband repeater, it does offer the systems engineer an important tool in planning heavy-density, long-haul, or video systems. Not every system needs its capabilities, but for those that do it is invaluable. •

BIBLIOGRAPHY

1. *Transmission Systems for Communications,* Bell Telephone Laboratories; New York, 1959.
2. J. P. Kinzer and J. F. Laidig, "Engineering Aspects of the TH Microwave Radio Relay System," *The Bell System Technical Journal:* November, 1961.
3. *Documents of the Xth Plenary Assembly, Volume IV,* International Radio Consultative Committee; Geneva, 1963.
4. "Performance Testing of Television Channels," *The Lenkurt Demodulator;* October and November, 1963.

World Radio History

World Radio History

the Lenkurt Demodulator

VOL. 3 NO. 6 JUNE, 1954

Some Factors Affecting

THE PROPAGATION OF MICROWAVES

Over Point-to-Point Radio Systems

Reliable communications can be obtained over point-to-point radio systems just as easily as they can be obtained over conventional wire and cable lines. Just as a wire-line system is made reliable by engineering the system to compensate for predictable variations in line losses, a radio system can be made reliable by engineering the system to compensate for predictable variations in propagation losses.

In this article, the important factors which affect propagation of radio waves are discussed and some of the methods of utilizing them or compensating for them are described.

In the outside telephone plant, wire and cable have long been the standard transmission facilities for toll and exchange routes. Until after World War II, no extensive use of radio was made in the telephone industry. It was normally used only where land lines or submarine cables were impractical. This situation has now changed. Radio equipment designed specifically for telephone toll plant usage is presently available. Operation and maintenance of this equipment is fully compatible with normal telephone practices. Because of its many advantages, radio is finding wide application for expansion and replacement of existing outside plant wire lines and cables.

While much public attention has been given to national microwave networks for the transmission of hundreds of telephone channels and several television channels, the recent development of radio and channelizing equipment for light to medium traffic routes has made the use of point-to-point radio economically practical for expansion and extension of toll and exchange facilities.

Many telephone companies are now finding that microwave has a definite place in their outside plant. Numerous installations already made have demonstrated that microwave systems can be engineered to be equally or more reliable than conventional wire lines or cables.

To engineer a wire line system

World Radio History

requires a knowledge of the transmission characteristics of wire lines at the frequencies used. In the same manner, to engineer a microwave system requires a knowledge of the propagation characteristics of radio waves at microwave frequencies.

Fading, a phenomenon encountered in radio links, is comparable to increased attenuation under severe weather conditions, a basic factor in wire line engineering. Fading is caused by the effect of air and terrain on radio wave propagation.

Radio waves at microwave frequencies and light waves have many of the same characteristics. Since the behavior of light waves is well known through the science of optics, and since radio waves and light waves have many of the same properties, certain optical principles are useful in describing radio wave propagation. The most important of these are reflection, refraction and diffraction.

Optical Properties

Because they behave like light, radio waves can be reflected from smooth conducting surfaces and focused by reflectors or lenses. When radio waves pass from one medium to another (such as from dry air to moist air) they are bent or refracted in the same manner as light waves are bent by a lens or prism. Radio waves tend to bend around large obstacles in their path by a process known as diffraction. They also are scattered by small particles such as rain and snow.

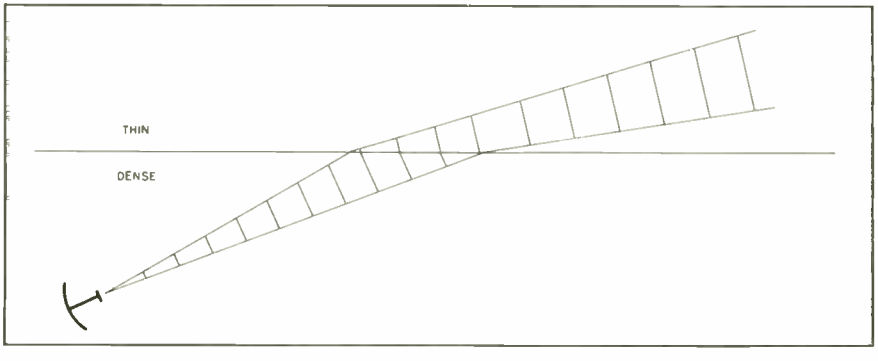
Each of these properties can cause variations in the received signal strength. They must be considered and allowances made for their effects when engineering a radio system.

Reflection

Very short radio waves are usually focused by dish-shaped metal reflectors. Such reflectors concentrate all the energy into a relatively narrow beam that can be directed like a light beam of a searchlight. This concentration of radio energy allows transmission over longer paths with much less power than would otherwise be required with non-directional antennas. (See Demodulator, Vol. 1, May, 1952.)

While the ability to reflect radio waves is very useful for focusing them into a beam, reflection is also a primary source of received signal variation. Reflections occur when radio waves strike a smooth surface such as water or smooth earth. If both reflected and direct waves reach the receiving antenna, it is possible for the two waves to cancel each other and reduce the received signal strength.

FIGURE 1. Refraction at a boundary between air at different densities. The speed of radio waves is slower in the denser medium.



World Radio History

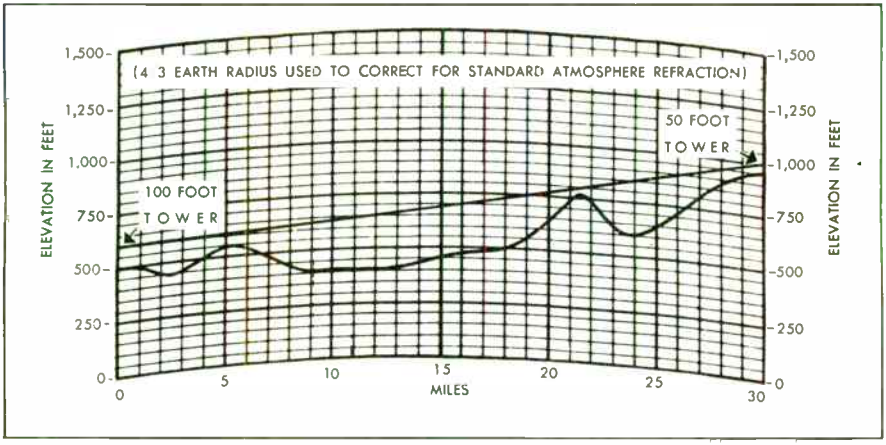


FIGURE 2. Profile charts are often prepared with 4/3 true Earth's radius to allow for normal atmospheric refraction. Charts prepared with true Earth's radius are also widely used. True Earth's radius provides a more conservative method of system engineering.

Depending on the length of the reflected path compared to the direct path, the reflected wave may arrive at the receiving antenna either in phase, out of phase, or partially out of phase with the direct wave. Under conditions where the reflecting surface is very smooth and the reflected wave and direct wave are exactly out of phase at the receiver, the reflected wave may temporarily almost completely cancel the direct wave and cause a very deep fade in received signal strength. Cancellation is worst when the reflecting surface is a calm body of water, smooth moist earth or the thin layer of hot air that lays just above the surface of desert sand in the daytime.

In general, reflected waves are undesirable. Changes in the refractive qualities of the air cause the point of reflection to shift and the reflected and direct waves pass in and out of phase with each other causing wide variations in received signal strength.

Rough terrain, such as a rocky or wooded area, is generally a very poor reflector of radio waves. Such terrain either absorbs much of the radio energy or scatters it so that little reflected energy reaches the receiving antenna. For this reason, radio paths with reflection points in rough terrain have very little interference from reflected waves.

Refraction

Refraction occurs because radio waves travel with different speeds in different media. In free space (a vacuum) the speed is maximum. In any other medium, however, radio waves travel slower. As shown in Figure 1, when radio waves pass from dense air to thin air, their direction is changed. As the upper part of a wave front enters the thinner air it starts traveling faster than the lower portion which is still in the dense air. The result is that the path of the waves is bent or refracted.

When considering refraction of radio waves through the Earth's atmosphere, it is usually assumed that under normal conditions the atmosphere is densest at the Earth's surface and becomes thinner at higher altitudes. This variation in air density above the Earth causes radio waves near the surface of the Earth to travel more slowly

World Radio History

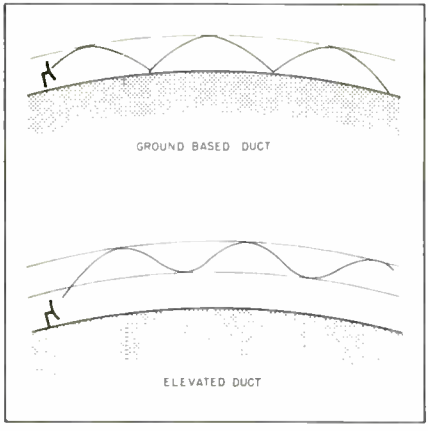


FIGURE 3. Ducts formed by stratification of the Earth's atmosphere. Ducts tend to trap radio waves and guide them around the Earth's Surface.

than those considerably above the surface. The result of these different velocities is a bending of the direction of wave travel which causes the waves to tend to follow the Earth's surface.

Because curved paths through the atmosphere are difficult to represent graphically, it is customary to draw profile charts of the Earth's surface with the Earth's radius represented as 4/3 actual size. The use of this fictitious radius approximately compensates (under average conditions) for the bending of the waves by the Earth's atmosphere and permits the illustration of radio paths on a profile chart as straight lines. An example of a profile chart with 4/3 Earth's radius is shown in Figure 2.

Simple refraction causes no great difficulty in the engineering of radio routes. Occasionally, however, refraction effects can seriously disturb the transmission of signals over line-of-sight paths. Under unusual conditions, the atmosphere becomes stratified with definite boundaries between layers of different densities. This causes the path of the radio waves to be bent first down and then up so that the waves become trapped in a layer of dense air. As a result the waves are guided along the air layer in much the same manner as microwaves travel in a wave guide. Layers of dense air that trap radio waves are commonly referred to as ducts.

The existence of a duct may either increase or decrease the received signal strength depending on whether the duct guides the waves toward or away from the receiving antenna. Ducts are more frequent near or over large bodies of water and in climates subject to frequent temperature inversions (stratification of air). While ducts may cause fading, their most important effect is that they sometimes guide radio waves well beyond the optical line-of-sight so that they are detected by distant repeaters of the same system. This type of interference can be avoided by changing transmitted frequencies at repeaters and locating repeater stations along a zig-zag path. Examples of two common types of ducts are shown in Figure 3.

Diffraction and Fresnel Zones

Ordinarily, radio paths are selected so that there is a direct line-of-sight between the transmitting and receiving antennas. However, a direct path between transmitting and receiving antennas is not necessarily a sufficient condition for good radio transmission. If a radio wave passes near an obstacle such as a hilltop or a large building, part of the wave front will be obstructed, and the amount of energy received will differ from that received if no obstacle were there. The cause of this difference is known as diffraction.

A simplified physical explanation of diffraction is shown in Figures 4 and 5. In Figure 4 a succession of unobstructed radio wave fronts are shown progressing from

World Radio History

the transmitter to the receiver. The whole surface of each individual wave front contributes energy to the receiving antenna. However, energy from some portions of the wave front tends to cancel energy from other portions because of differences in the total distances traveled. The shaded areas in Figure 4 show the paths of energy that cancel some of the energy transmitted by the paths shown unshaded. The cancellation is such that half of the energy reaching the receiver is cancelled out. Most of the energy that is received is contributed by the large unshaded central area of that portion of the wave front that is closest to the receiver. If an obstacle is now raised in front of the wave so that all of the wave front below the line-of-sight is obstructed, (this is shown in Figure 5) half of the broad central area is obstructed and a greater loss of energy occurs. Under this condition the radiated power reaching the receiver is reduced to one fourth normal or by 6 db. If the obstruction is lowered (or the receiving antenna raised) so that all of the central zone is exposed, the power received by the receiving antenna is even greater than it would be if the obstacle were not there. This is shown in Figure 6.

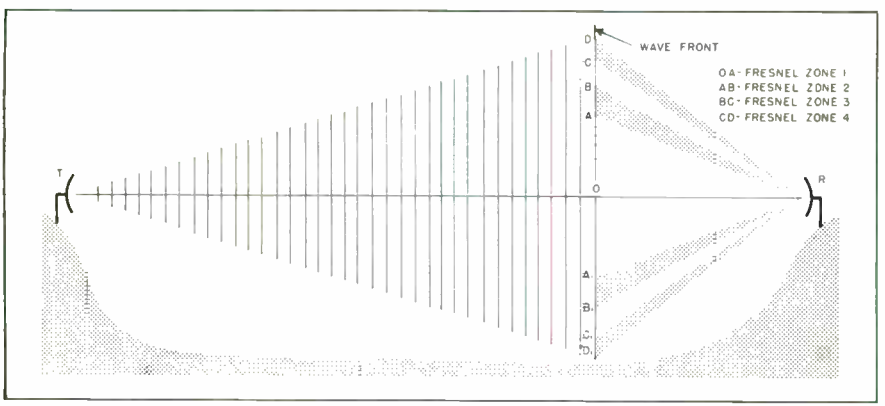
The various zones of the wave front that contribute either in-phase or out-of-phase energy are called Fresnel zones after their discoverer Augustin Jean Fresnel (1788-1827). The large central zone is called the first Fresnel zone and zones farther removed from the line of sight are called the second, third, fourth zones, etc. If the obstruction is such that the first zone is above the obstruction, the radio path is said to have first Fresnel zone clearance. In general, first Fresnel zone clearance is considered to be very desirable, although clearance of only one half the first zone is adequate. Of course, clearance greater than first Fresnel zone is also adequate.

Paths without adequate clearance are not desirable because refraction by the atmosphere may change. If a path just clears an obstacle under normal conditions, a change in refraction may cause the path to be obstructed. Careful construction of a profile chart and visual examination of the proposed route should show whether adequate path clearance is available.

Absorption and Scattering

Because no transmitting media other than free space is perfect, some radio energy is absorbed

FIGURE 4. Energy contribution from a wave front to a receiver. The dark areas are out of phase with the light areas.



World Radio History

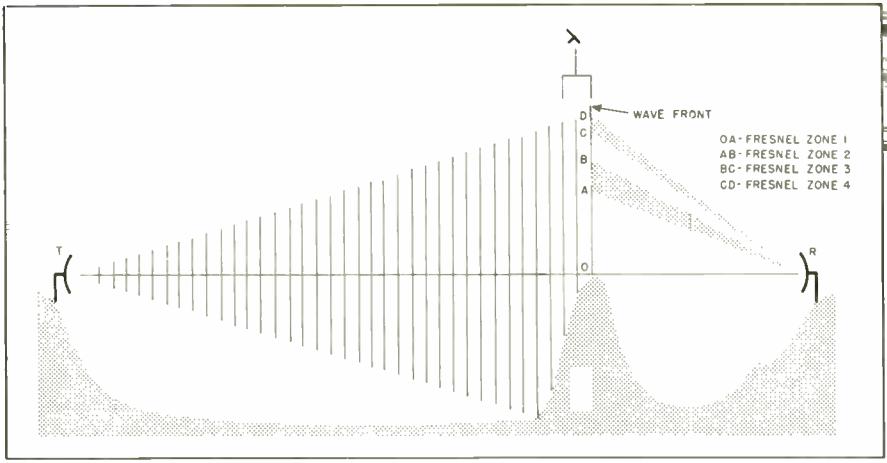


FIGURE 5. When an obstacle is raised in front of a receiver to the line of sight, half of the wave front is obstructed and the received signal strength is reduced by 6 db.

from a wave when traveling through the Earth's atmosphere. In clear weather, absorption is very slight for radio waves of less than 10,000 megacycles frequency. Rain, snow, and fog, however, can absorb or scatter large amounts of radio energy, especially at the higher microwave frequencies. Below 1000 megacycles, however, reduction of received signal strength by scattering and absorption by fog or precipitation is not a serious problem over paths of the usual length.

Radio Route Considerations

Each of the factors that affect radio propagation must be taken into consideration when planning a radio route. In many cases, visual examination of the route topography and a knowledge of weather conditions along the route are sufficient to determine the feasibility of proposed transmitter, repeater, and receiver locations. Where there is a doubt, profile charts can be used to determine more precisely the transmission conditions to be expected. In exceptional cases it may be desirable to make propagation tests.

Where visual examination indicates that the radio path is completely or partially over smooth terrain or water, the point of reflection from the terrain of the transmitted wave should be determined from a profile chart. The reflection point is located where the angle the transmitted wave makes with the Earth's surface is equal to the angle of the reflected wave.

If the reflection point is found to be on a smooth surface such as a flat field or water, relocation or a change in height of the transmitting or receiving antenna may be desirable. Often one of the antennas can be located so that it is masked from smooth ground or water reflection but still in line-of-sight to the other antenna. An example of using nearby terrain to mask unwanted reflections is shown in Figure 6.

How Bad is Fading?

Fading can only be determined absolutely over a particular path on the basis of experience just as the ice or frost attenuation of an open wire line must be learned absolutely by experience. It is the combined effects of all the various factors described above that can cause variation in received radio

World Radio History

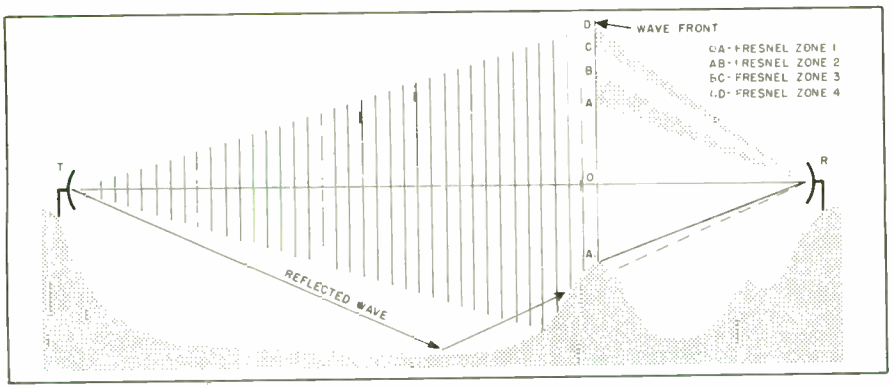


FIGURE 6. When the obstruction is lowered (or the receiving antenna raised) until all of the first Fresnel zone is exposed, the received signal strength is greater than if the obstruction did not exist.

frequency signal strength. Observations of fading over practical systems operating at frequencies below 1000 megacycles have shown that fade margins above normal space losses of 0.5 to 1.5 db per mile (depending on the terrain) will provide satisfactory transmission for 99.9 percent of the time. The 0.1 percent of time that transmission quality is below standards of such systems amounts to only 9 hours per year; which compares favorably with the performance obtained from many wire line and cable systems. In most cases a well planned system can be expected to be completely out of service due to excessive fades for less than 1 hour per year.

Conclusions

The factors affecting radio propagation over line-of-sight paths, while differing greatly in details, have much the same effect on transmission quality and reliability as do weather and temperature changes on open-wire lines and cables. The problems created by these factors are as amenable to solution as are the transmission problems of conventional wire and cable.

Each of the factors affecting the propagation of radio waves contributes somewhat to variations in received signal strength. However, by understanding the way in which each factor affects propagation, by engineering the radio system to minimize their effects, and finally by allowing a sufficient margin for unavoidable fading, very high quality circuits can be obtained with maximum reliability and reasonable cost.

World Radio History

World Radio History

the Lenkurt® Demodulator

VOL. 2 NO. 1 JANUARY, 1953

How to Prepare and Use

PROFILE CHARTS OF RADIO LINK ROUTES

With the help of an accurate profile chart of a radio link route, competent engineers often can predict performance closely enough to make actual propagation tests unnecessary.

In this article a few aspects of preparing and using profile charts are discussed to give the reader a practical acquaintance with some of the problems encountered when installing a radio link.

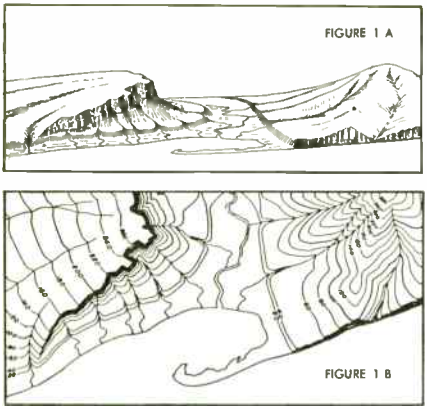
Radio waves in the higher frequency ranges used for point-to-point radio links (commonly called microwaves) exhibit many of the properties of light. They travel in relatively straight lines, and they are bent (refracted) by the atmosphere, reflected by solid objects or surfaces, and diffracted by physical objects in or near the transmission path. To predict the effect of these properties upon the propagation of energy between two antennas, the nature of the terrain between antennas must be considered.

The first step in estimating the propagation characteristics between two antenna sites is to assemble elevation data about the intervening terrain. Using this data a profile chart is prepared to show the elevation of all hills, ridges, tall buildings, or other obstacles that might interfere with line-of-sight transmission of radio waves. A satisfactory transmission path can then be intelligently selected by analyzing the information on the profile chart along with any other pertinent data.

Sources of Data

Several different sources can provide the data required for preparing a profile chart. In many cases it can be obtained from topographic maps with contour lines showing elevation of land at convenient intervals. Information on maps of this type, prepared by the United States Department of the Interior, is available from the Director, United States Geological Survey, Washington, D. C. A section of a typical map of this type is shown in Figure 1B. Figure 1A is a sketch of the area which this map represents.

World Radio History



COURTESY U. S. DEPT. OF THE INTERIOR

FIGURE 1. *Contour maps provide an excellent data source for preparing profile charts. Figure 1B is a contour map of the area sketched in Figure 1A.*

For locations where these topographic maps are unavailable, local county surveyors can often provide the required data.

If no previously prepared maps can be obtained, a special survey may be required although, in many cases, sufficient data can be obtained through one or more "common sense" procedures. In one such method an altimeter is used to determine the relative heights of land along a proposed transmission path. In some cases a spotlight (or sun reflecting mirror) can be used to determine if a line-of-sight path exists and, if conditions permit, the light source can be moved vertically to determine path clearance.

If data obtained from topographic maps shows that only marginal clearance exists, the elevations of high points should be checked by survey or altimeter to insure their accuracy.

Consulting engineering services are ordinarily available to make either ground or aerial surveys of proposed radio link routes.

Preparing the Profile Chart

After tentative antenna sites have been selected, and the relative elevation of land between these sites has been determined, a profile chart can be prepared. In some cases a complete profile such as those shown in the examples will be necessary; in other cases only certain hills or ridges need be indicated to be sure adequate path clearance exists.

An important factor influences the shape of a profile chart. Although the surface of the earth is curved, microwaves tend to travel in straight lines. However, they are bent (refracted) a small amount by the atmosphere. The amount of bending (refraction) varies with atmospheric conditions. The effect of refraction is such that if a profile chart is prepared on the basis of four-thirds (4/3) of true earth radius, a straight line between antenna sites will indicate clearance between the actual transmission path and the earth. However this factor of 4/3, which would increase the permissible distance between antennas, is not always accurate. Under some atmospheric conditions the refraction caused by the atmosphere will diminish and the actual transmission path will approach true line-of-sight conditions.

Because reliability and continuity of service are very important for a multichannel radio link, many radio engineers prefer to be conservative and base propagation predictions on the basis of path clearance shown on a profile chart prepared with true earth radius.

The effect of using 4/3 and true earth radius for the same path is shown in the two sketches of Figure 2. Figure 2B shows that when the amount of atmospheric bending is normal, a clear path is indicated when planning is done with a profile chart drawn with 4/3 true earth radius. Under abnormal conditions, however, if a true line-of-sight path exists, the transmission path between antennas is

World Radio History

interfered with by the ridge shown on the chart in Figure 2A, drawn with true earth radius.

Since the choice of earth radius varies with topography and climate and is influenced by the amount of fading allowable, the advice of a competent radio engineer should be obtained when a radio link is being installed over a path where some question about clearance or other factors exists.

Printed forms are available for plotting profile charts. A form used by Lenkurt's engineers has been used for the examples in this article. This form, which uses true earth radius, is based on the relationship between the height of an observer and the distance to the horizon where, if "h" is in feet and "d" is in miles, $h = 2/3d^2$. If 4/3 earth radius is used, this relationship becomes $h = 1/2d^2$. The scale of either chart can be changed if desired by doubling the horizontal interval and quadrupling the vertical interval (or by dividing the horizontal interval by 2 and the vertical interval by 4).

Using the Profile Chart

An accurately drawn profile chart will show whether or not

FIGURE 2. *Profile charts prepared with true earth radius provide a more conservative estimate of propagation conditions than those prepared with 4/3 true earth radius. The clear path indicated in Figure 2B (4/3 true earth radius) is interfered with when true earth radius is used (Figure 2A).*

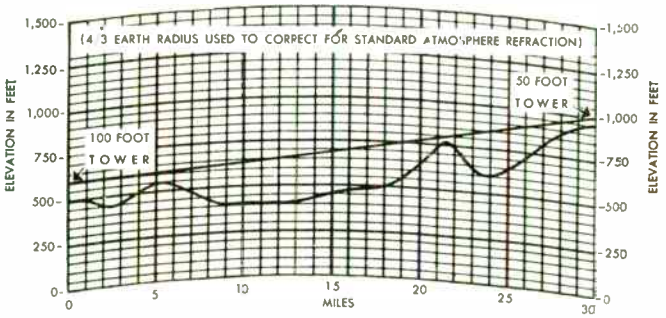


FIGURE 2 B

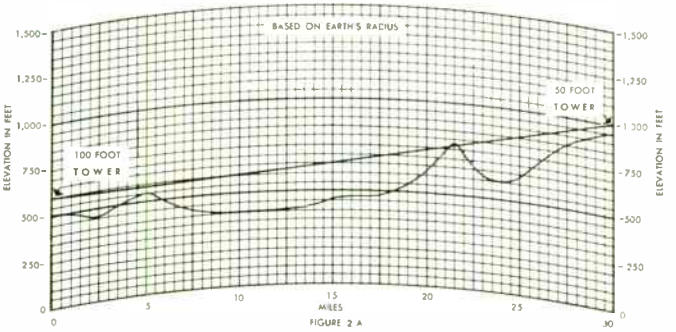


FIGURE 2 A

World Radio History

adequate path clearance exists for the transmission path between antennas. The chart can also be used to determine the "reflection point" as shown in Figure 3.

The path clearance desired varies with frequency and with distance from the transmitting antenna. Lenkurt's engineers usually consider that about 75 feet minimum clearance is acceptable for a system operating at 900 megacycles.

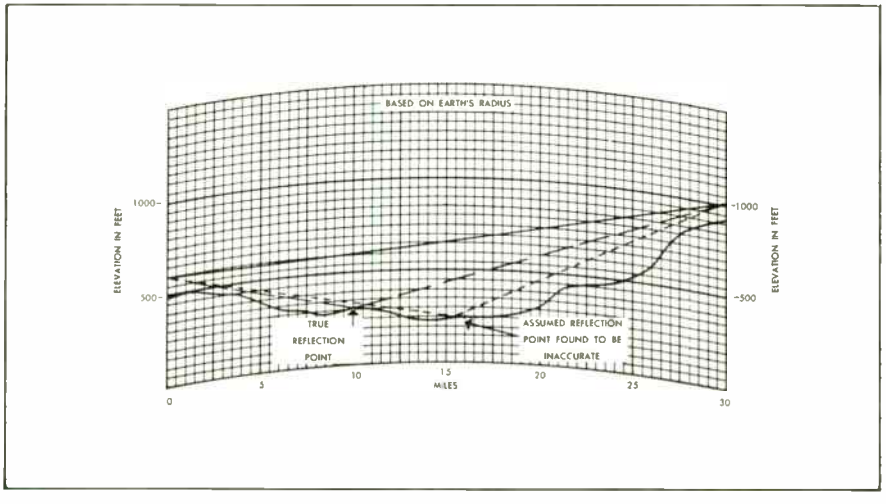
The effect that a ray reflected by the earth will have depends to a great extent upon the character of the surface at the reflection point. A strong reflection will be caused by a smooth body of water or by smooth earth while a weaker reflection will come from wooded terrain. In general, a strong reflected wave is undesirable because it can cause fading and distortion.

The reflection point can be found from a profile chart by using a "cut and try" method illustrated in Figure 3. By inspection, an assumed reflection point is selected and straight lines are drawn between this point and the two antennas. The assumed reflection point is the true point if the two lines rise the same number of feet in going an equal number of miles to the right and left of it. In Figure 3 the dotted line indicates an incorrect reflection point because the lines do not rise equal amounts in equal distances to the right and left of the assumed point. The dashed lines, however, indicate the true reflection point.

Conclusions

An experienced radio engineer often can predict the effects of available path clearance and reflection with sufficient accuracy to determine whether or not proposed antenna sites can be used. In many cases, the losses caused by unfavorable topography can be overcome by using higher gain antennas, higher transmitter power or lower loss cables between radio equipment and antennas. Sometimes different antenna sites or higher antenna mountings might be required. In any case an accurate profile chart is an invaluable tool for the experienced engineer planning radio links.

FIGURE 3. *Reflection point is easily determined from a profile chart by using a "cut and try" method.*



World Radio History

the Lenkurt Demodulator

VOL. II, NO. 4 APRIL, 1962

Rapid Microwave Switching

Increasing congestion of microwave frequency allocations, plus greater need for protection against transmission failures as system capacities increase, is placing new emphasis on so-called "hot-standby" techniques for assuring continuous microwave transmission. This article reviews some methods of microwave switching, essential to hot-standby protection.

The conventional way of protecting against the failure of a microwave communications system is *frequency diversity* transmission, a method which requires two fairly well separated frequency allocations. This method protects against fading, since the most common type of fades rarely occur simultaneously on separate frequencies. Equipment failure is guarded against by the duplication of radio equipment. Unfortunately, two separate frequencies are required, and these are becoming increasingly scarce in some areas due to the burgeoning growth of microwave communications. For this reason, frequency diversity transmission may not be permissable.

When only a single transmission frequency is available, complete protection is more difficult. *Space diversity* provides some protection against fades, but is much more costly than frequency diversity. Two well-separated antennas, and a receiver for each, are required to receive the same transmission. The separation of the two transmission paths prevents most fades from occurring on the two simultaneously. The extra antennas, reflectors, greater tower height, and additional land may make the cost of space diversity prohibitive.

Nevertheless, some means of protecting the system against equipment failure is required. This is often achieved by some form of "standby" transmission equipment. In some older systems, the standby equipment sat idle without power applied until there was an equipment failure. When a failure or loss of

World Radio History

power was sensed, the standby equipment was turned on and soon took over the communications load. In more recent times, however, even brief interruptions of service during the period required for the standby equipment to warm up, have become intolerable, with the result that so-called "hot-standby" transmission is gaining much wider usage.

In hot-standby systems both the operational and the standby equipment are energized ("hot") at all times. If some part of the system fails, or if power drops below some arbitrary value, transmission is immediately turned over to the standby equipment. Service is interrupted for the time required to switch from one transmitter to the other. Early hot-standby switching arrangements required one or more seconds for the switching operation. In many operations today, the large volume of communications carried over microwave make even this brief interruption extremely undesirable, if not intolerable. As a result, greater emphasis is being placed on reducing the time required for switching.

Problems of Switching

The problem of microwave switching is fairly old. In even the earliest radar systems, the sensitive receiver had to be effectively disconnected from the antenna while the transmitter sent its brief, powerful pulse, then re-connected quickly to listen for echoes. The problem was generally solved by the use of gas discharge tubes, thyratrons, and the like. Although speed of switching was fairly important, the degree of isolation required was not great, since it was only necessary to protect the receiver from damage caused by the high-powered transmitter pulses.

In a communications system, however, it is necessary to have a high degree of attenuation of the undesired signal. This would not be required if both the standby and operating transmitter signals were precisely in phase with each other. If there are slight discrepancies in the phase or frequency of the two transmitters, destructive interference of the two signals results, causing a large increase in system noise.

Although klystron oscillators can be phase-locked very easily, practical considerations of equipment design make this approach undesirable. Accordingly, all presently-available hot-standby methods allow only one signal to be present at a time. This is accomplished by the use of some sort of switch which, in one state, can connect the transmitter to the antenna with very little loss, or, in the other state, block the RF signal and provide a high degree of isolation. The actual amount of isolation required depends on the nature of the microwave system. Basically, the switch should reduce the standby RF carrier enough that noise caused by mutual interference is less than the "idle noise" of the system. This, of course, varies with the bandwidth of the system, its loading, crosstalk performance, and similar factors. In general, isolation or attenuation of the standby carrier should be about 80 db for most microwave equipment used in this type of service.

Switching Requirements

In data transmission, the loss of a single symbol can have expensive consequences. The data transmission speed determines how fast the transmitter switchover must be made in order to avoid errors.

In general, switching time should be less than one-half the duration of a single data pulse—as much less as possible—thus reducing the likelihood of that pulse being lost at the instant of switchover. In the case of 100-speed (100 words-per-minute) teletypewriter

World Radio History

Figure 1. Carrier equipment like shown above enables hundreds of channels to be transmitted over a single radio beam. Increasing traffic density makes it imperative that some sort of standby protection be provided for microwave equipment.

signals, about 80 bits or pulses are transmitted each second. Accordingly, the switchover should take place in less than 1/160 second—that is, about 6 milliseconds. Similarly, 60-speed teletypewriter signals must be switched in less than 10 milliseconds to avoid loss of characters. For higher speed data transmission, faster switching speeds are required.

Microwaves are considerably more difficult to switch than lower frequencies because they are almost always transmitted through waveguide instead of wire. The problem is not only physical but also electrical. Microwave signals are propagated through the waveguide as electrical and magnetic fields, instead of a simple flow of current, as in a wire. Slight imperfections or discontinuities in the waveguide cause the signal to be reflected instead of just being interrupted. When the signal is reflected back towards its source, it affects the

World Radio History

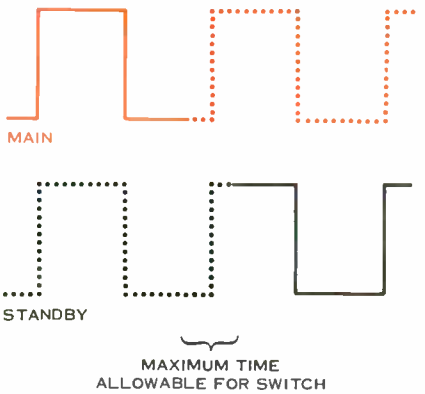


Figure 2. Maximum switching time of half the pulse length is required to permit switchover with little chance of causing data errors. Shorter switching time is desirable.

voltage and power distribution in the waveguide, thus providing wrong indications at measurement points.

In telecommunications, it is very desirable to maintain the standby equipment in a condition which duplicates actual transmission. This allows the standby equipment to be monitored continuously, thus revealing any need for maintenance and providing positive indications of the state of the system.

Methods of Switching

The switching technique used has an important bearing on the behavior of the standby microwave equipment. Thus, a microwave switch that reflects the radio energy back to the klystron sets up standing waves. These interfere with klystron operation, and tend to give a misleading power indication. This can be avoided by using an isolator, a device for transmitting energy in one direction but absorbing that returning from the opposite direction. However, an isolator may not be necessary if a switch is used which absorbs the energy rather than reflects it.

The earliest method of switching microwave transmitters was by the use of a so-called waveguide switch, such as diagrammed in Figure 3. In this device, the antenna is directly connected to the operating transmitter-receiver by means of a movable waveguide section. The output power of the standby transmitter is connected to an absorptive load. After a fault or equipment failure, a motor-driven mechanism operates a movable waveguide section that removes the absorptive load from the standby unit and connects it to the other unit (now disconnected from the antenna). At the same time, the antenna is physically connected to the output of the standby unit. With the exception of switching time, this method of switching can have very good performance characteristics. If the switching machinery is well built, so that mechanical tolerances are exceptionally close, a very high degree of isolation is obtained, and there is very little "forward loss" in the transmitting condition. The disadvantages of this method are that it is mechanically complex, and switching time is extremely slow—sometimes requiring several seconds to complete the switch.

In an effort to improve switching time, simpler versions of the mechanical waveguide switch have been produced which are able to switch from one transmitter to the other in from 25 to 100 milliseconds. However, losses are somewhat higher. At best, these rotary switches provide from 40 to 60 db reverse isolation. Large amounts of power are required for rapid switching—on the order of about 50 watts typically.

Waveguide Plungers

A very popular method of waveguide switching, and one that is still widely

World Radio History

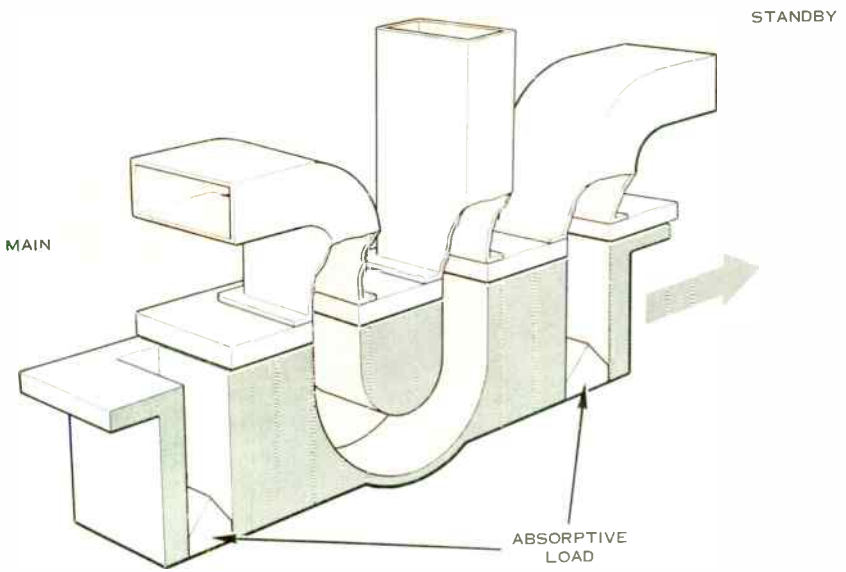


Figure 3. Early systems used mechanical waveguide switch like shown here. When switchover is required, electric motor drive slides shaded portion sideways; waveguide loop connects antenna to the operating equipment. Absorptive terminations are provided for the disconnected transmitter.

used, is the solenoid-driven waveguide plunger. In this method, a thin rod is inserted through the waveguide to stop the signal. Although the rod actually blocks only a very small area of the waveguide, it almost totally reflects the signal, thus providing an attenuation of about 40 db. By using two plungers appropriately spaced, additional attenuation is possible.

The plungers are driven by solenoid magnets so that the plunger is completely withdrawn from the waveguide of the operating transmitter and inserted through the waveguide from the standby transmitter. If the operating transmitter fails or loses power, one plunger is withdrawn and the other inserted in about 30 or 40 milliseconds. This is faster than the waveguide switch because less mass must be moved in the switching operation.

The mechanical plunger switch is still used by some manufacturers because it is reasonably reliable and provides effective switching action. It is stable and has good power handling capacity. Its bandwidth is relatively limited — only about 1% — because it is essentially a form of waveguide filter. However, this is still quite adequate for most communications applications.

The disadvantage of the plunger switch is that it is mechanical, and therefore inherently limited in its potential speed. Typical plunger switches

World Radio History

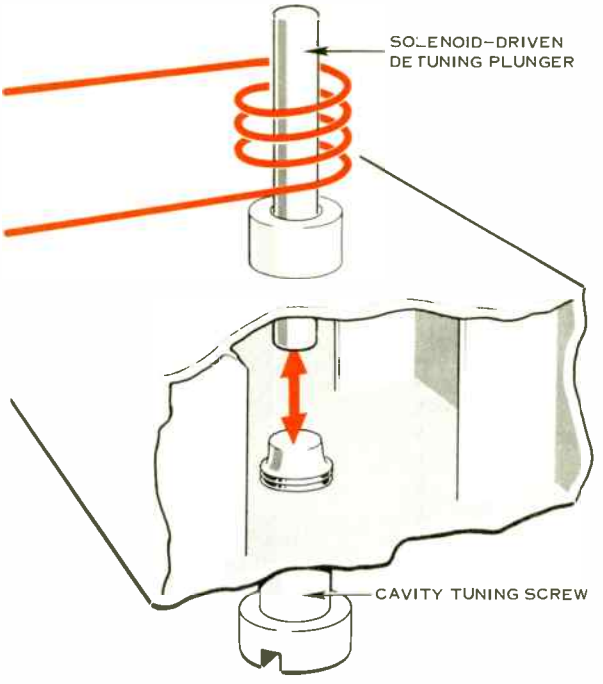


Figure 4. One form of waveguide plunger switch, still used in some equipment. When plunger is withdrawn, energy passes with little loss. With plunger inserted in cavity, microwave energy is mostly reflected.

operate at 1/5th the speed required to avoid loss of teletypewriter characters. If only voice circuits are to be carried over the microwave system, the mechanical plunger is entirely adequate, and this is also true for slow telemetering and control signals.

Electronic Switching

In order to avoid the limitations of mechanical switches, engineers have turned to fast-acting electronic devices. Special versions of some commonly-used ferrite devices have proved to be extremely effective as microwave switches. For instance, ferrite isolators make excellent high-speed switches. Microwave circulators have been successfully used as two-way switches.

The "Faraday" rotation type of ferrite isolator is widely used to absorb microwave reflections in waveguide, and thus reduce distortion. When microwaves travel through a ferrite slab lying in a magnetic field, the plane of polarization of the microwave signal is rotated. Signals coming from the other direction are also rotated. By designing the device carefully, each signal can be made to rotate exactly 45°; the signal traveling in one direction passes through with hardly any attenuation, while the opposite-going signal is rotated until it is at an angle of 90° with the outgoing waveguide, but is aligned with a vane of "lossy" or resistive material which absorbs its energy. Since the signal traveling in the desired direction is polarized at right angles to the lossy vane, little energy is absorbed.

In the switching isolator, *two* 45° sections are used, and the permanent magnet of the fixed isolator is replaced with electromagnets. By reversing the magnetic fields, it becomes possible to rotate the plane of polarization of the

World Radio History

microwave signal parallel to the output waveguide and perpendicular to its vane of absorbing material—thus letting the signal pass through—or perpendicular to the waveguide and parallel to the vane, thus blocking passage of the microwave energy.

The switching isolator has no moving parts and thus can switch much faster than devices which depend on the mechanical motion of a mass. The only basic limitation to the speed of a switching isolator is the time required to overcome and reverse the magnetic fields of the electromagnets. In the Lenkurt 74B hot-standby equipment, this problem is easily overcome by storing a large amount of "reserve" electrical energy in a series inductance. In the 74B arrangement, the effect of the standby RF carrier is reduced more than 80 db, insertion loss is 0.5 db to 1 db, and switching time is about one millisecond—five time as fast as necessary for teletypewriter service. The switching isolator provides an attractive solution to the problem of rapid switching because it is also able to serve as a conventional isolator to prevent waveguide reflections from reaching the transmit klystron.

Other ferrite devices have been used for switching microwaves. In one arrangement a rod of ferrite is suspended in the waveguide, but without an applied magnetic field. Normally, it has very little effect on the microwave signal passing through. When it is subjected to a magnetic field, however, the

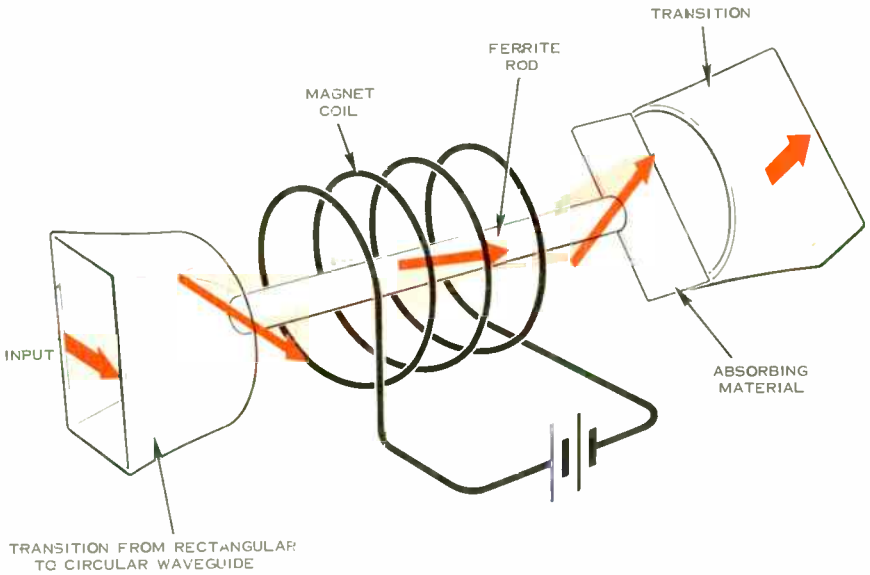


Figure 5. Faraday rotation switch uses property of ferrite in a magnetic field to rotate microwave plane of polarization. In one plane, signal is passed through with minimum attenuation; in other plane, signal is absorbed in vane of resistive material. In practice, two 45° sections are used to achieve 90° rotation.

World Radio History

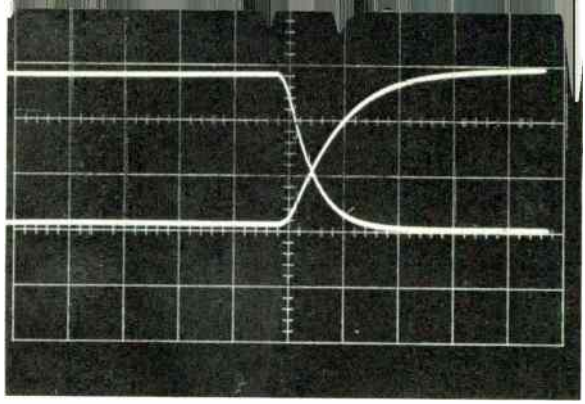


Figure 6. Unretouched photograph of Lenkurt 74B "Hot Standby" switching action, using Faraday rotation switch. Time scale is one millisecond per division. Ninety percent of power is transferred in first millisecond.

ferrite magnetic permeability becomes saturated, and the microwave signal is strongly reflected. This technique can provide 60 db isolation, with about 0.5 db forward loss in the "on" condition.

Diode Switching

The rapid growth of computer technology has led to considerable work on the switching of microwave energy with semiconductors. One form of semiconductor microwave switch is diagrammed in Figure 7. In this device, a microwave diode is placed across the narrow dimension of a section of waveguide. When the diode is biased so that current flows through it, microwave energy is attenuated only about one db. When the diode bias is reversed, microwave energy is strongly reflected, thus providing up to 25 db attenuation. Essentially this occurs because the diode introduces a capacitive reactance when non-conducting—the equivalent of a short circuit across the waveguide.

Another version, shown in Figure 8, employs a probe into the waveguide. When the probe is just ¼ wavelength long, it is "series resonant" and reflects the microwave signal. If the probe is physically contacted by another probe at right angles to it, it is necessary to readjust the length of the vertical probe to achieve series resonance and block the signal. If the side probe is retracted slightly so that electrical contact is broken, the structure is detuned, and the microwave energy passes freely. Now, if a biased diode is used to make and break electrical contact, the device becomes an electrically-operated microwave switch which provides about 20 db isolation and 0.2 db insertion loss.

In most diode switches of the type described, not more than about 25 db isolation can be obtained by a single switch section. By spacing two or more diodes in tandem, greater isolation becomes possible. Such tandem switches are sensitive to the spacing between diodes and to applied voltage since the voltage determines the effective capacitance of each diode. In general, the most effective switching is obtained when diodes are spaced approximately a quarter wavelength apart in the waveguide, thus effectively creating an anti-resonant cavity.

The use of diodes in this fashion has suggested still other ways of switching. In one configuration, a diode is used to shunt each cavity of a multi-cavity tuned filter like those commonly used in the output of a microwave transmitter to suppress extreme modulation sidebands.

Such diode switches can provide extremely fast microwave switching—one microsecond or less. Isolation is typically 20 to 25 db per section, thus permitting very effective suppression of the carrier for hot-standby applications.

World Radio History

Two major difficulties must be overcome, however. The most effective location for the shunting diode is in the center of the cavity where the electric component of the microwave transmission is strongest. In this location, however, the microwave diode is subjected to strong fields which overload the diode and shorten its life considerably. If the diode is moved away from the center of the resonant cavity, it is subjected to less overload, but produces poorer switching action. Another disadvantage of this particular method is the relatively high insertion loss. In addition to the losses inherent in the filter cavity, each diode provides between 0.5 and 1 db loss per section, depending on its placement within the cavity. Thus, a compromise must be found between adequate isolation and excessive loss of transmitted power.

Surprisingly enough, the *type* of diode used may have an important bearing on the switch performance. In most switching arrangements, germanium and silicon diodes behave quite differently. Usually, expensive microwave diodes are required, and these are quite delicate and sensitive to overload. The nature of the semiconductor and the amount of bias used to obtain switching action can affect insertion loss, isolation, and power handling capabilities. Typically, this type of microwave switch has relatively low power handling capability. Because of this and the compromise between isolation and insertion loss, these switches are used more in laboratory applications and certain types of radar modulators than in practical communications applications.

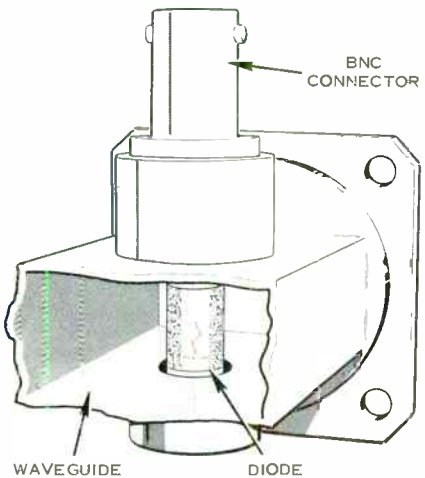


Figure 7. Semiconductor diode may be substituted for plunger to pass or block microwave energy. This switch configuration requires special, high-performance diodes, and is limited in power-switching capability.

A related device, developed by Lenkurt for its new 76B microwave system, overcomes most of these objections. The 76B hot-standby switch provides 75 db isolation at the worst frequency (but effectively infinite isolation at the best frequency), while introducing 0.5 db loss. Switching time is somewhat less than 1 microsecond. This does not, however, include the time required to detect the need to switch and energize the control circuit. The 76B switch does not require microwave diodes, but uses rugged switching diodes of the type used in computers. In tests, the new switch was able to control 8 to 10 watts of microwave power without harm to the diodes.

Possible Trend

In the United States, hot-standby arrangements have been required only in private and industrial microwave systems. Despite the inherent privacy of a microwave link (due to the narrow beams employed) it is becoming more and more difficult to find frequency allocations that do not interfere with other systems.

World Radio History

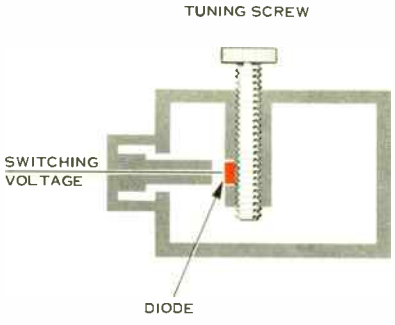


Figure 8. Alternate version of switch uses semiconductor diode to detune vertical probe in waveguide. Electrical "contact" between horizontal probe and tuned vertical probe is made by biasing diode "on" or "off."

Frequency diversity is, of course, the preferred method of achieving system reliability because of the inherent protection it provides against both equipment failures and propagation fades. However, as microwave continues its present growth, an increasing shortage of frequency allocations may make hot-standby necessary even in services where it is not required today. Under these circumstances, protection against fading will depend upon the use of larger fade margins—obtained by the use of greater transmitting power, larger antennas, and shorter path lengths. Faster hot-standby switching will be increasingly important to minimize loss of information—a vital consideration as the capacity of the systems increase. •

BIBLIOGRAPHY

1. R. B. Carver, E. G. Spencer, and M. A. Harper, "Microwave Semiconductor Switching Techniques," *IRE Transactions on Microwave Theory and Techniques;* October, 1958.
2. C. M. Johnson and J. C. Wiltse, "A Broad-band Ferrite Reflective Switch," *IRE Transactions on Microwave Theory and Techniques,* (correspondence); July, 1960.
3. R. V. Garver, J. A. Rosado, and E. F. Turner, "Theory of the Germanium Diode Microwave Switch," *IRE Transactions on Microwave Theory and Techniques;* January, 1960.
4. Robert V. Garver, "Theory of TEM Diode Switching," *IRE Transactions on Microwave Theory and Techniques:* May, 1961.
5. F. Reggia, "A New Broad-band Absorption Modulator for Rapid Switching of Microwave Power," International Solid State Circuit Conference, University of Pennsylvania, Philadelphia, Pa.; February, 1961.
6. A. Calvin, "Rectangular Waveguide Switches," *IRE Transactions on Microwave Theory and Techniques:* July, 1961.

World Radio History

the Lenkurt Demodulator

VOL. 10 NO. 3 MARCH, 1961

Microwave Diversity . . .
how it improves reliability

Microwave radio, promising infant of 1951, has become the dependable, durable worker of 1961. No longer limited to such special-purpose uses as transcontinental television and very long telephone circuits, microwave is already demonstrating a remarkable versatility in filling the diverse, ever-growing communications needs of the world.

Microwave's extremely high information capacity provides a communications bargain, but places extra importance on the need for reliability. With today's improved equipment, the greatest cause of signal failure is no longer the equipment, but the characteristic fading encountered in the transmission path itself. This article discusses microwave fading and how it may be overcome.

Today's microwave equipment has become so dependable that the most important single cause of transmission failure is the propagation characteristics of the radio waves themselves. Because of their very short wavelength, microwaves have many of the properties of light. Microwave signals can be concentrated by suitable reflectors into tight beams which conserve power and increase privacy of transmission. Like light, microwaves are refracted by the atmosphere through which they travel, and are subject to obstruction or reflection by intervening objects. These particular phenomena often result in *microwave fading*, an important cause of transmission failure.

Although fading occurs in many different ways, only two types have much importance in microwave transmission. These are *multipath fading* and *inverse beam bending* or *ducting*. Of the two, multipath fading is much more common and troublesome.

Multipath Fading

Microwaves travel at the ultimate speed of light only in a vacuum. Through air, radio waves are slowed

World Radio History

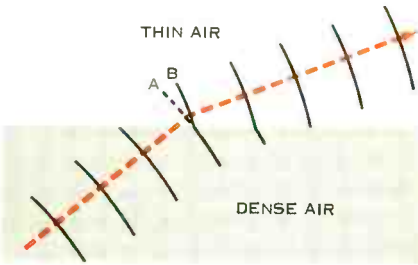


Figure 1. Wave direction is shifted by change in propagation velocity. Wavefront at B would have only reached A had it continued in slower medium.

down very slightly. The denser the air and the greater its moisture content, the slower the velocity of the radio or light rays. When the wavefront passes diagonally from a dense layer to a thinner layer of air, the direction of travel of the wave is altered. The first part of the wavefront to enter the thinner layer begins traveling slightly faster than the portion of the wave still in the denser medium. As the remainder of the wavefront passes into the thinner air, it increases velocity, but cannot overtake those parts of the wavefront which entered the thin air first. As a result, the direction of travel of the wavefront is deflected slightly, and the amount of this deflection or refraction is proportional to the difference in the velocity of the wave through the two masses of air.

If the change in air density is gradual, the refraction or bending of the radio beam may be continuous, so that the beam is gently curved away from the thinner toward the denser atmosphere. Since the atmosphere is normally more dense near the earth, and usually thins out with increased altitude, radio and light rays do not follow a true straight-line path, but are usually deflected downward and tend to follow the curvature of the earth. For this reason, so-called line-of-sight radio paths often extend beyond the visual horizon.

Frequently, atmospheric irregularities cause a portion of the radio beam to be bent away from the most direct path, so that one part continues on the direct path to the receiving antenna, and other parts are deflected upward or downward. As a result, two or more separate components of the original transmission may arrive at the receiving antenna, each having traveled a path slightly different from the others. This is called multipath transmission.

When the component traveling the longer path reaches the receiving antenna, it arrives slightly later than the direct beam because of the difference in path length. Consequently, the different components of the signal will be somewhat out of phase with each other, because of the difference in the length of path each has traveled.

If two equal signal components travel paths having a difference of $\frac{1}{2}$ wavelength, they will arrive at the receiving antenna 180° out of phase and will cancel each other. This, of course, destroys the signal. If several components reach

Figure 2. In normal atmosphere, beam is gently deflected toward earth by gradual change in density of atmosphere above the earth's surface.

World Radio History

Figure 3. Irregularities in atmosphere cause radio beam to break up into several components traveling different paths. Those which manage to reach receiving antenna will have traveled paths of different lengths, causing phase cancellation.

the antenna *in* phase, they add and increase the received signal. It should be noted that a 6000 mc signal has a wave length of only about two inches. Thus, a one-inch difference in the path length of two signal components makes the difference between a very strong signal and one that cancels itself out.

Where partial signal cancellation occurs due to atmospheric multipath transmission, the message content or modulation of the signal is essentially unaffected, so long as enough signal power remains to be detected by the receiver. Differences in path length due to atmospheric refraction and reflection usually vary from a fraction of an inch to not much more than six or seven feet. Although this much difference provides many opportunities for phase cancellation of the *RF* signal, it is only a very tiny fraction of a wavelength of the *modulating* signal. For instance, the highest modulating frequency in a 600-channel carrier system is about 2.4 mc, and has a wavelength of about 410 feet. If path-length variations of ten feet are experienced, this represents a phase shift of only about eight degrees — not enough to produce significant distortion. Since microwaves are transmitted as very narrow, concentrated beams, it is unlikely that path differences greater than a few feet will occur.

Physical Causes of Multipath

Multipath transmission can result from any of several different causes. One path may be the direct optical path, while one or more other paths may result from reflection off the surface of the earth or water lying in the transmission path. Reflection and refraction from layers of air having different densities are most common.

Where there is constant wind and continuous mixing of air, multiple transmission paths rarely occur. Where the air is very still, it tends to collect in layers, each layer having a different temperature and usually a different moisture content. Sometimes the division between layers is quite sharp, and a microwave beam may be reflected from this interface as though it were a mirror.

Frequently, the air is particularly still during the night and just before dawn. Under these conditions, there is a great difference in the transmission characteristics of air close to the surface and air some distance above the surface. Heavy, moisture-laden air collects close to the

World Radio History

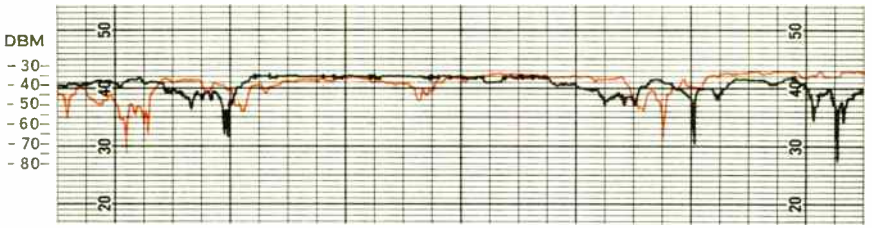


Figure 4. Actual recording of frequency-diversity transmission. Signals followed identical paths, were spaced 180 mc (3%) apart. Note that deep fades never occurred simultaneously. Small simultaneous fades are accommodated by fade margin.

ground. Humidity rises and may reach 100%, thus producing a ground fog. At various heights above the earth, the air will have different temperature and moisture characteristics, and each region will have a different index of refraction. Under these circumstances, a microwave signal may be split into many paths.

Ducts

Occasionally, a temperature or humidity inversion will occur so that instead of thinning with increased altitude, the atmosphere may actually become denser at some point above the earth. Instead of being deflected toward the earth, microwaves will experience less downward deflection, and may actually be bent upward. This is known as *inverse beam bending*. Although relatively rare, it may cause very long, deep fades. These fades do not result from multipath interference, but from the deflection of the signal above the antenna. If additional path clearance is provided between the signal path and possible obstructions, some rays may reach the receiving antenna despite the inverse beam bending. Fades due to inverse beam bending are also said to be caused by *earth bulge*, since the effect is analogous to the earth bulging upward so as to obstruct the transmission path.

Above the inversion, the normal thinning of the air with increased altitude is restored. Microwaves deflected upward by the temperature inversion are again deflected downward when the air begins to thin with increased altitude. As they reach the temperature inversion, they are again refracted or reflected upward, so that they are trapped within a

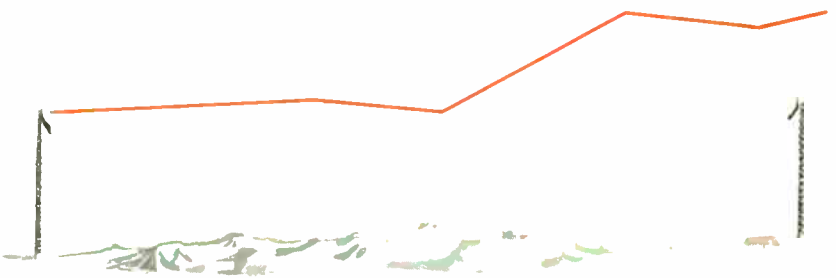


Figure 5. Temperature or humidity inversion may cause duct which traps signal.

World Radio History

narrow layer or "duct" of air. This is known as *trapping* or *ducting* of the signal. If both the transmitting and receiving antennas lie within the duct, the received signal strength may be considerably increased. Normally, however, such ducting or trapping usually results in a deep fade which may block communications for hours. This type of fading is much less common than atmospheric multipath fading.

Overcoming Microwave Fading

Since multipath fading is an almost perfectly random phenomenon, it occurs on a chance basis. Lord Rayleigh, in his important work on sound, showed that random phase cancellations occurred in a predictable manner, and followed the relationship shown in Figure 6. This curve has been shown to describe the likelihood of microwave fades of various degrees of severity. As shown in the curve, fades of 35 db or more may be expected .02% of the time. In the course of a year, this amounts to about an hour and 45 minutes during which transmission is interrupted.

One way of overcoming this loss is to design the system with enough performance reserve to offset all but the very deepest fades — those of 40 db or more. This approach is useful, but can become quite costly. Larger antennas, reflectors, and towers are required to obtain enough system gain to offset the deep fades. In addition, the distance between microwave repeaters may have to be decreased substantially.

Increased transmitter power will also provide additional reserve against deep fades. If a 20 db fade margin is considered ample, a 100-watt output signal is required to assure continuous performance of the quality obtained by a 1-watt output signal in the absence of fades.

One objection to this approach concerns interference. Although there is far more available bandwidth at microwave frequencies than at lower frequencies, the supply is not unlimited. One of the very desirable features of microwave transmission is that low powers and narrow beams permit the use of the same frequency spectrum by many different communications systems, so long as they are separated physically. With microwave, this separation need not be great.

A far more practical approach to overcoming microwave fading is to use some form of *diversity transmission*. In general communications experience, three types of diversity have been found useful in overcoming fading: *polarization* diversity, *space* diversity, and *frequency* diversity. The first of these, polarization diversity, while effective in systems where propagation is largely by sky wave, has been found to provide no advantage in point-to-point, line-of-sight microwave systems.

Space Diversity

Space diversity takes advantage of the fact that simultaneous fading is not likely over two well-separated paths. In a typical microwave space diversity system, the signal from a single transmitter is received at two antennas having a large vertical separation. The two independently-received signals are connected to a diversity combiner which selects the signal having the greater freedom from noise.

This method of diversity has the advantage of conserving frequency spectrum, since only one transmitting frequency is used. Its greatest value is in overcoming multipath fading in which one of the paths is caused by a specular or direct reflection such as from water, a building, or the earth itself.

A major objection to space diversity is its cost. Since vertical separation is required, additional antennas and waveguide are necessary. If a single tower is

World Radio History

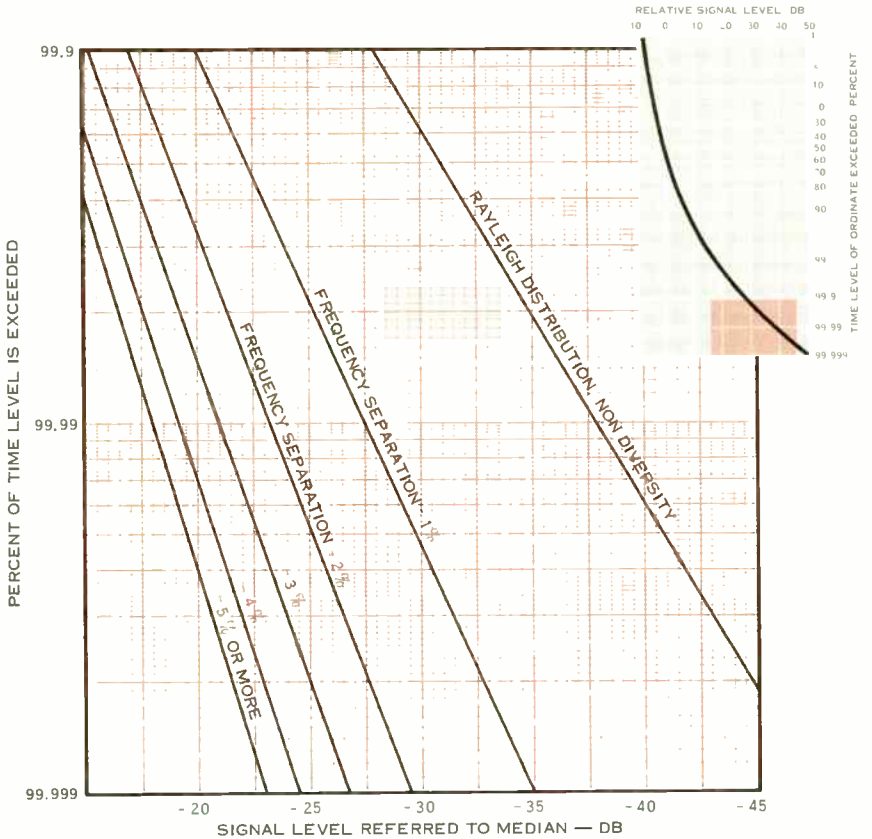


Figure 6. Rayleigh distribution curve (inset) shows probability of fades of various depths. Expanded section shows diversity improvement for various frequency spacings.

used, it must be much stronger than one designed for a single antenna, and will need to be much higher to achieve enough vertical separation between antennas.

Another objection is that space diversity does not provide nearly as much freedom from fading as does frequency diversity, for instance. Although separate paths are involved, they may or may not experience simultaneous fading. The greater the vertical separation (and the higher the tower), the less likely that the two signals will fade together. Although only a single transmitter is needed, two receivers are required.

Frequency Diversity

For all practical purposes, the refraction or deflection of a microwave signal is independent of its frequency; signals of even quite widely separated frequencies are refracted about the same amount under the same transmission conditions. Thus, several microwave signals of different frequencies will experience identical refraction and splitting into separate components. If a single pair of antennas is used for transmitting these frequencies, all components of both signals will follow identical paths.

One might expect that all signals traveling exactly the same paths would

World Radio History

experience identical phase cancellations and would therefore tend to fade simultaneously. This is not so, however. Microwave signals of different frequencies fade independently of each other if they travel identical paths. The greater the frequency separation, the less chance the two frequencies will fade simultaneously. If the frequencies of the two signals are brought closer together, they will tend to fade more and more nearly simultaneously.

The reason for this frequency-selective fading is that signals having different frequencies also have different wavelengths. A signal of *x* frequency may have components that arrive exactly out of phase at a given moment. Since signals of *y* frequency (of longer wavelength) travel over the identical paths as those of *x* frequency, they *cannot* be exactly out of phase with each other at the same moment as the components of *x* frequency. When one frequency fades, the other will usually be at near-normal strength.

As in the case of space diversity, a diversity combiner is required to select the better signal and reject the faded transmission. (Diversity combiners were discussed in the DEMODULATOR, *March*, 1959 and *September*, 1959.)

Although two complete transmitters and receivers are required at each diversity terminal, this is much less costly than a space diversity system which would be required to provide comparable protection from equipment failure and fading. Since both transmitters operate continuously, no switching equipment is required to substitute one transmitter for another in case of equipment failure. The requirements for towers, reflectors, antennas, and waveguide are the same as for non-diversity systems.

Conclusions

Most microwave communications systems operate full-time carrying important data and other information vital to business and industry. The many needs which efficient communications satisfy make it imperative that interruption be held to an absolute minimum.

Of the many ways for increasing transmission reliability, frequency diversity has been found to provide the best protection at the lowest overall cost. Not only does it provide almost sure protection against atmospheric fading, it is the only system which permits ready maintenance of the microwave equipment without withdrawing the system from service. Single-frequency hot-standby systems, are generally unsatisfactory for continuous-duty service in systems requiring very high reliability because they provide no protection from fading. In addition, their protection against equipment failure is reduced because complete periodic maintenance is impractical without rerouting the communications and taking the system out of service. This same objection, while not true of the receiving end of a space diversity system, is true of the transmitting portion. •

BIBLIOGRAPHY

1. A. B. Crawford and W. C. Jakes, "Selective Fading of Microwaves," *Bell System Technical Journal;* January, 1952.
2. O. E. De Lange, "Propagation Studies at Microwave Frequencies by Means of Very Short Pulses," *The Bell System Technical Journal;* January, 1952.
3. R. L. Kaylor, "A Statistical Study of Selective Fading of Super-High Frequency Radio Signals," *The Bell System Technical Journal;* September, 1953.
4. *Microwave Path Engineering Considerations — 6000-8000 Mc,* Lenkurt Electric Co., Inc., San Carlos, California; April, 1960.

World Radio History

World Radio History

the Lenkurt Demodulator

VOL. 8 NO. 9 SEPTEMBER, 1959

How to increase microwave reliability

Microwave systems are being used more than ever before to satisfy a seemingly insatiable demand for more and better communications of all types. Larger numbers of channels are transmitted over individual microwave paths than was common even quite recently. The use of data transmission is growing so fast that within a decade or so it may provide the greatest volume of all communications traffic. Both of these trends place a very high premium on the reliability of microwave communications. This article discusses some of the more important factors which influence the reliability of a microwave communications system.

Reliability has two meanings in the microwave communications business. First there is the conventional meaning concerning the ability of the equipment to stay in service for extended periods with minimum attention. The other aspect of reliability is transmission reliability. It is a measure of system engineering, propagation characteristics, and the ability of the equipment to yield high quality performance through adverse transmission conditions. It is also influenced by the equipment's freedom from distortion or tendency to degrade a signal, especially when numerous repeaters are employed in the transmission path.

Transmission reliability is particularly important where numerical data are transmitted. Pipelines use data for monitoring and controlling remote unattended pumping stations and even refineries. Railroads are increasing their use of microwave for transmitting routing, scheduling, and other information. Great quantities of accounting

World Radio History

and other management data are transmitted by many companies, and this is increasing as more companies use business machines that handle and transfer data directly, rather than going through human channels. For such use, transmission must be as nearly perfect as possible. Unlike speech, which is little affected by "static" or other momentary degradation of transmission, data usually has little redundancy. Where a word might be easily recognized despite interference, a noise pulse could alter a data character to a different value. Such errors could remove the wrong car from a train, raise a man's pay to a ridiculous figure, or increase an order beyond the wildest dreams of the sales manager!

Path Engineering

One of the big factors affecting microwave transmission reliability is fading caused by changing propagation characteristics over the transmission path. Some types of fading may reduce signal strength only a few db. Others may drop the signal level 40 db or more. Fade characteristics vary from place to place and from time to time.

Since propagation variations cannot be controlled except, perhaps, by choice of operating frequency and by care in selecting the path, the best way of overcoming such fades is to design the system with enough operating margin to assure a strong signal during all but the deepest fades. It is necessary to choose receiver sensitivity, transmitter power, path length, and antennas so that the desired signal-to-noise ratio is achieved even under severe fading conditions.

Microwave Antennas

At microwave frequencies, transmission is achieved with very low powers by concentrating most of the radio energy into a narrow beam. The larger the antenna (and reflector), the narrower the beam and the stronger the received signal. Large antennas increase system cost, but provide increased reliability. Large reflectors require much heavier towers, not so much because of their weight as because of increased wind loading. Since large, high-gain antennas reduce beam width, they also increase the requirement for accurate beam positioning. This may cause fades due to deflection of the radio beam by wind action on antennas or reflectors. This is usually resolved by using heavier towers and more extensive guying, often a very expensive procedure.

Transmitter Power

The development of klystron oscillators capable of more power than previously available affords the system designer a rather painless way of increasing transmission reliability. Other aspects of transmission being equal, doubling the transmitter power adds 3 db to the fade margin. Raising output power from 100 milliwatts to 500 milliwatts provides a 7 db signal-to-noise advantage. This means that the 500-mw transmission will tolerate a 7 db deeper fade than the 100-mw signal, for equal performance. This extra operating margin may be used to lengthen the transmission path or reduce the size of antennas and reflectors if the margin is not required to maintain the desired reliability.

Noise Threshold

Another way of keeping a tight rein on interference is to restrict receiver bandwidth. All FM receivers strongly suppress noise as signal strength increases. During a deep fade, noise usually increases in direct proportion to the reduction in signal strength, db for db, until the receiver detection

World Radio History

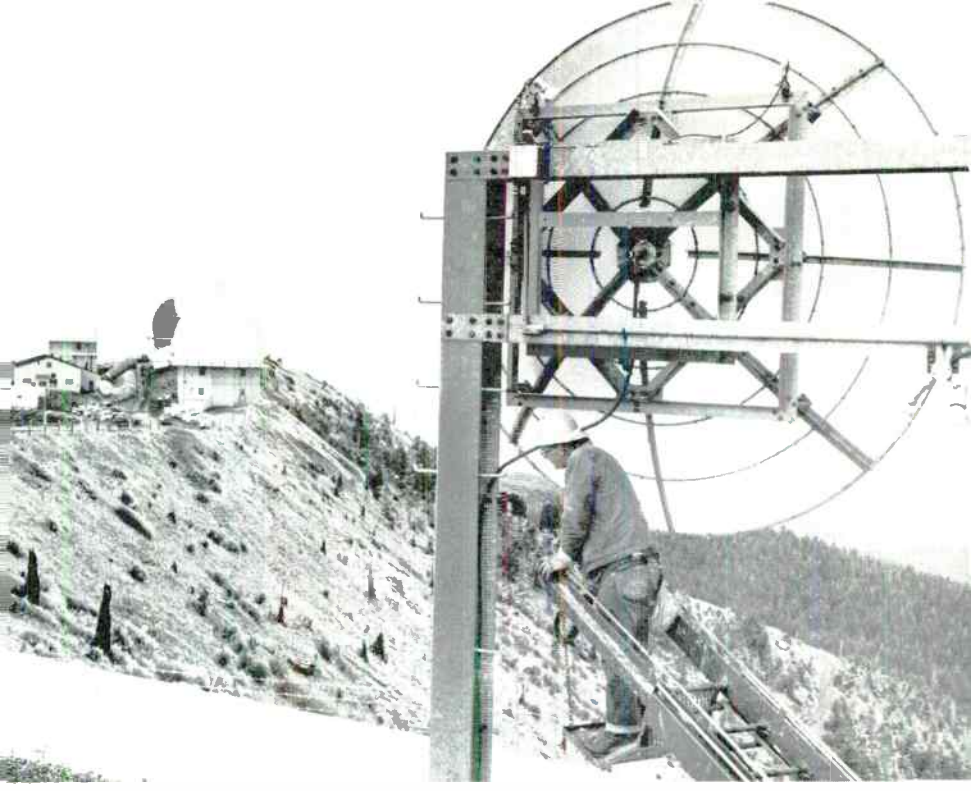


Figure 1. Vital SAGE data is transmitted from defense warning radar to computer center by microwave relay. Microwave equipment must be reliable and relatively immune to interference from the radar itself.

threshold is reached. At this point, noise jumps dramatically, drowning out whatever residual signal may be present.

One of the controlling factors of receiver noise threshold is the bandwidth of the receiver, largely controlled by the i-f amplifier. Although receiver quieting on stronger signals soon equalizes the difference between a broadband and a narrower band receiver, the narrow band receiver is less susceptible to noise—that is, its noise threshold is lowered in direct proportion to the reduction in bandwidth. Assuming equal i-f amplifier performance, a receiver with half the bandwidth of another will enjoy a 3 db noise threshold advantage over the wider bandwidth receiver. These are the most important 3 decibels in the performance of the receiver, for these "bottom" dbs—not the high level ones—determine the limits of good receiver sensitivity and reception reliability.

Diversity Combiners

Fades usually vary with the operating frequency and the physical path characteristics. Even fairly closely spaced frequencies or physically spaced paths will not fade simultaneously. While one fades, the other usually maintains a good signal, as shown in Figure 2. Al-

World Radio History

though space or frequency diversity normally can yield about the same protection from fades, space diversity costs more because of the additional antennas, reflectors, and tower expense. Diversity advantage is diagrammed in Figure 3.

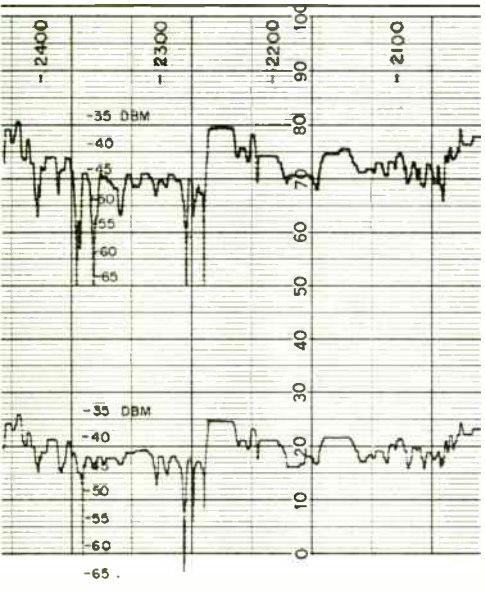


Figure 2. Actual recording of frequency diversity reception at 6,000 megacycles. The two signals are spaced approximately 118 megacycles apart. Note that most fades do not occur simultaneously on the two frequencies.

In a diversity system, some sort of combiner is required to select the stronger of the two signals and reject the noise from the faded receiver. Several forms of combiner are in common use.

The *linear combiner* is usually a passive device that adds the two receiver output signals together. This form of combiner has the disadvantage of being unable to reject the tremendous increase in noise from a receiver during a very deep fade, and the signal-to-noise ratio will not be more than 6 db better than the worst channel, except when the two signals are equal. When this occurs, the linear combiner provides a 3 db improvement in signal-to-noise ratio over either individual signal. On the good side, intermodulation or distortion products from the two paths can be arranged to nearly cancel out, thus improving noise performance.

The *switching combiner* senses noise, pilot tone, or both, and selects the better signal of the two. A relay in the signal path usually performs the actual switching. This method suffers the disadvantage of introducing transients and transmission errors at the moment of actual switching, even when fast-acting relays are used.

The *ratio-squared combiner* adds the two signals in proportion to their freedom from noise. Thus, the output from a noisy receiver is largely suppressed, while the other, with less noise, supplies a larger portion of the final signal. Theoretically, this method can yield optimum performance as a combiner. The disadvantage of this approach is complexity, cost, and the necessary use of electron tubes or similar active elements in the signal path. Electron tubes are always a source of potential failure, for they are the least reliable of all electronic circuit elements. Practical combiners based on this principle do not always achieve their potential advantage because of design compromises made to reduce complexity and cost.

A very practical and reliable arrangement is a combination of the linear combiner and the switching combiner. In such an arrangement, the linear combiner's reduction of intermodulation is usually available, and under fair transmission conditions, the combination of both signals provides a 3 db increase in the signal-to-noise figure. When either

World Radio History

receiver fades below normal good performance, it is muted and the unfaded receiver provides the signal, thus eliminating the major defect of the linear combiner — the lack of noise suppression. Relay failure does not interrupt reception, and relay operation may be arranged to avoid the errors that are characteristic of the switching combiner. And the passive linear combiner portion is inherently more reliable than any combiner that uses electron tubes in the signal path.

Internal Noise

In systems where many repeaters are used, noise from within the equipment becomes an important factor in transmission reliability. Such noise adds up to reduce the signal-to-noise ratio and reduce the margin against fades.

An important source of noise in microwave systems is the imperfect matching of characteristic impedances of coupled transmission elements. In an ideal system, there is perfect transfer of energy from the antenna to the waveguide, from the waveguide to the receiver mixer, and from the mixer to the intermediate frequency amplifier. This would be possible only with perfect matching of characteristic impedances at each junction. Unfortunately there are several conflicting conditions in a microwave receiver that make this difficult.

The first stage of an intermediate frequency amplifier is usually designed for the highest gain possible, consistent with bandwidth and noise considerations of the circuit used. Since the gain of the stage is partially a function of the input impedance, impedance is usually high. The antenna and waveguide, however, exhibit a much lower impedance. Normally, a coupling be-

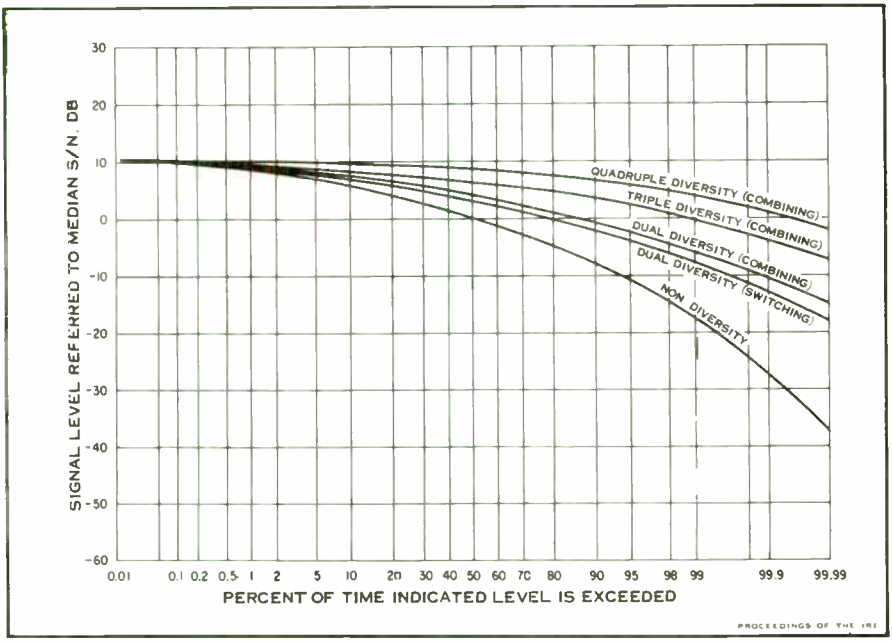


Figure 3. Theoretical advantage of various diversity systems over non-diversity system.

World Radio History

tween the low impedance source and the high impedance load would result in part of the energy being reflected back up the waveguide to the antenna. This reflection may be expected because perfect power transfer is never realized. In the antenna and waveguide portions of the circuit, a perfect impedance match can be achieved at only one discrete frequency. Since microwave and carrier transmissions are essentially broadband, some impedance mismatch may occur. This gives rise to a particular form of distortion peculiar to microwave frequencies.

Radio energy picked up by the antenna is coupled to the waveguide. Because of a slight impedance mismatch between the antenna and the waveguide, a very small part of the energy received is reflected and lost from the antenna. Most of the energy, however, is transmitted down the waveguide to the receiver. At the receiver, most of the energy is coupled to the mixer, where it is absorbed. However, once again, impedance mismatch causes part of the energy to be reflected, this time back up the waveguide to the antenna. Again, a small portion is reflected because of the impedance mismatch between the antenna and the waveguide. As a result, incoming signal and twice-reflected signal travel down the waveguide together.

The phase relationship between the original signal and the reflected signal will vary continuously with the component frequencies, causing phase distortion, which increases the background noise. Since the phase distortion is a function both of the signal frequencies and the travel time in the waveguide, this phase distortion is much greater for long sections of waveguide. The chances of impedance mismatch are far greater in the mixer than in the waveguide and antenna, because of the conflicting impedance requirements of the various circuits that join in the mixer.

An excellent way of overcoming the problem of energy transfer in the mixer, and at the same time greatly reducing strong adjacent channel interference and locally-originated noise, is the balanced mixer.

A balanced mixer for microwave use is diagrammed in Figure 4. Incoming signals are received in the waveguide chamber on the right. Local oscillator energy, differing in frequency from the incoming signal by the desired i-f frequency, enters the left chamber. An aperture between the two chambers is so arranged as to split input power from either branch equally between the two chambers. Energy from both chambers is picked up and mixed by two balanced diode probes connected together so that there is cancellation of input signal and local oscillator signal. The i-f frequency produced by the mixing of the two frequencies does not cancel out, and is applied to the i-f amplifier. Input energy not coupled to the i-f amplifier is reflected only down the local oscillator branch, where it is absorbed in an attenuator used to adjust local oscillator injection voltage. Local oscillator energy in the input branch of the mixer is rejected by a preselector filter which prevents it from being radiated.

This type of mixer has the advantage of balancing out all energy except that derived from both branches. This reduces noise sidebands from the local oscillator and strong adjacent channel interference that might get through the RF preselector filter. It is particularly beneficial at locations where there are strong sources of pulsed energy such as near airports or other locations where radar is used.

Somewhat similar performance may be obtained by using an unbalanced

World Radio History

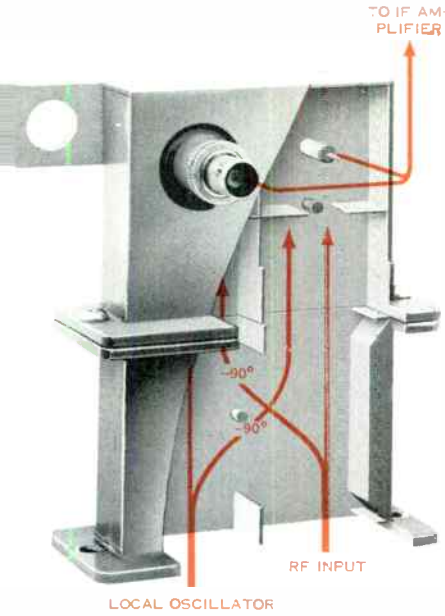


Figure 4. A balanced mixer for microwave signals. Such a mixer suppresses noise or interference appearing in one branch only, but provides low-noise mixing of desired input signal and local oscillator frequency.

mixer, but including an isolator and a filter in the local oscillator branch. The filter reduces noise from the local oscillator, and the isolator tends to absorb reflected energy from the mixer chamber. Such an arrangement is still vulnerable to adjacent channel interference, even with additional filtering in the input branch, for any two signals spaced by the amount of the i-f frequency, will mix and enter the i-f amplifier as interference. Pulses, such as those from radar, provide a rich source of frequencies suitable for such mixing, and often appear at sufficiently high peak powers to cause interference, despite heavy input filtering.

Equipment Reliability

All major suppliers of microwave equipment take special pains to select components and materials that will provide the utmost in reliable operation under a variety of operating conditions. However, the most carefully built and tested component is not entirely free of the possibility of failure, as missile designers are learning the hard way. The most reliable system will be one that provides the fewest opportunities for failure. The more active a component, the more likely it is to fail. Electron tubes are most likely to fail. Transistors don't operate reliably at temperature extremes. Although we can't eliminate electron tubes or transistors yet, we can concentrate on other unreliable items such as blowers, heaters, and thermostats. Parts of this type may not be part of the actual electronic circuitry, but their failure is no less damaging than the most critical electronic component.

Under the pressure of business and competition, it may be tempting to reduce initial cost of a system by cutting corners. The cost of achieving reliability may seem high, but with increased use of data and high-density multiplex or carrier systems, this apparent high cost may be a real bargain, particularly when compared to the cost of unreliable transmission. Such "extravagances" as diversity transmission and redundant common equipment (such as power supplies) may pay for themselves many times in systems where communications reliability is important.

World Radio History

World Radio History

the Lenkurt Demodulator

VOL. 14, NO. 1 JANUARY, 1965

Microwave for TV TRANSMISSION

A few years ago the transmission of television signals over microwave radio was limited almost entirely to the facilities used by the national broadcast networks. Since these facilities primarily connected major population centers, they were engineered and maintained by a relatively small number of experienced personnel. Now the situation has changed radically. Many small telephone companies are being called on to carry television for the broadcasting industry, educational television systems, and others. New applications are continually appearing for industrial television, and community antenna television is growing very rapidly.

The result is that many people suddenly find themselves in some part of the television business, bombarded with strange terminology, and sometimes even forced to make decisions about things they may not fully understand. Even those with sound technical backgrounds in communications often find themselves wondering about some of the "whys" of television transmission. This article is written in an attempt to answer some of the questions concerning the transmission of television over microwave radio.

A microwave system has been likened to a pipeline or a railroad. Each provides a hauling service, transporting a commodity from one point to another. But even though the basic idea is similar, the physical arrangements of these transportation systems are vastly different. Each is tailored to the characteristics of the commodity to be transported.

Even more specifically, different microwave systems have different characteristics, depending on the types of

World Radio History

traffic they may be called upon to handle, just as a railroad uses one car for lumber and another for passengers. So it is that microwave systems which carry television signals differ from those which carry multiple voice channels. Often the same basic system can be used for either type of service, but the arrangement is different.

Video Signal Characteristics

A logical place to start a study of a transportation system is with an analysis of the traffic. In this case, this means the characteristics of a television signal. There are two parts to the signal, video and audio; and they are often transmitted independently. The video signal is more complex and it sets the transmission standards.

The television picture is produced by a varying voltage which represents the various shades of gray, from white at one extreme to black at the other. A higher voltage produces a whiter image, and a lower voltage a darker image. This voltage is divided into "sections" 63.5 microseconds long. Each section contains the information for one scanning line on the picture. The lines are separated by synchronizing pulses which cause the electron beam to retrace and start a new line. The last line at the bottom of the picture completes a "field," which represents half the picture. The trace then starts at the top again, interlacing the new lines between the lines of the previous field. This second field completes the "frame," forming an entire picture. In the NTSC (National Television Standards Committee) system, there are 30 such frames per second. Each frame consists of 525 scanning lines.

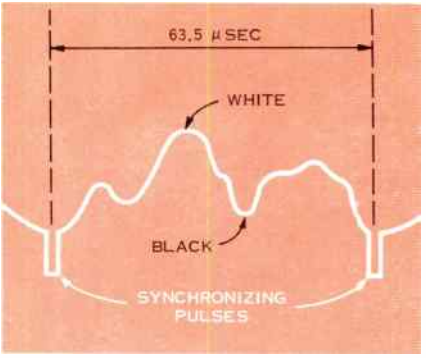


Figure 1. Voltage waveform for one scan line of a television signal. Higher voltage produces a whiter image.

A color signal contains the same basic intensity information as a black and white signal, in addition to the color information. The color information is carried by a color subcarrier — a continuous signal at 3.58 Mc. Different colors are indicated by shifting the phase of this subcarrier, while the saturation or richness of the color is indicated by varying the amplitude of the subcarrier. This is one of the most critical parts of the television signal. Distortion of the color subcarrier may result in unusual color combinations. For example, a normally attractive, blonde, rosy-cheeked actress appears grotesque with bright green hair and a fiery-red face. Viewers tend to blame the receiver manufacturer when this occurs. But often at least part of the signal distortion is in the transmission. (For a discussion of the tests which establish acceptable performance, see "Performance Testing of Television Channels," THE LENKURT DEMODULATOR, October and November 1963.)

World Radio History

Transmission Techniques

A basic microwave transmitter for television transmission (shown in Figure 2) is essentially the same one that would be used for message traffic. In either case, the "heart" of the transmitter is the klystron. The klystron is usually the only vacuum tube used in modern solid-state equipment. Its function is to take the amplitude-varying baseband input and translate it to a frequency-varying radio signal at a much higher frequency. The klystron is tuned to produce a nominal center frequency

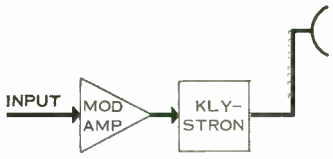


Figure 2. The most basic microwave transmitter consists of a klystron and a modulation amplifier.

when no modulating signal is applied. When a modulating voltage is applied, it causes the klystron output to swing back and forth across the center frequency. The higher the voltage, the greater the frequency swing.

A modulation amplifier is normally used ahead of the klystron to increase the amplitude of the input signal so that it can cause enough frequency deviation in the klystron. For example, in a typical system, a 0.36 volt peak-to-peak input to the modulation amplifier produces a 4-Mc deviation in the output of the klystron. This means that if the nominal center frequency is 6000 Mc, the deviation across the frequency band is from 5996 Mc to 6004 Mc.

The Video Receiver

The basic microwave receiver (shown in Figure 3) is only slightly more complex than the basic transmitter. From the receiving antenna the signal goes to a bandpass filter which selects the desired frequency band, rejecting all other frequencies. The signal then goes to a mixer where it is literally "mixed" with another frequency from the local oscillator. One of the results of this mixing or "heterodyning" process is a lower sideband at some intermediate frequency. This intermediate frequency is the difference between the frequencies of the incoming radio signal and the local oscillator signal. For example, if the incoming signal is at 6000 Mc and is mixed with a local oscillator frequency of 5930 Mc, the resulting intermediate frequency will be 70 Mc. This 70-Mc intermediate frequency ("IF") goes to a special low-noise preamplifier. (The other frequencies produced by the mixing process are filtered out.) The IF preamplifier is one of the most critical parts of the system, as far as noise is concerned, because at this point the signal is very weak and any noise added by the amplifier is of the same order of magnitude as the desired signal. Thereafter, noise and signal are amplified together. Thus, special care is taken to avoid introducing noise at the IF preamplifier.

World Radio History

The signal at the IF stage is still frequency modulated, even though it has been translated down from the 6000-Mc radio frequency. The function of the discriminator is to demodulate the signal and return it to its original baseband form. An ideal frequency-modulated wave has a constant amplitude. However, various types of interference and transmission irregularities may cause the signal amplitude to vary. The limiter removes any such unevenness and produces a constant-amplitude FM signal. The recovery of the baseband signal then occurs in the discriminator.

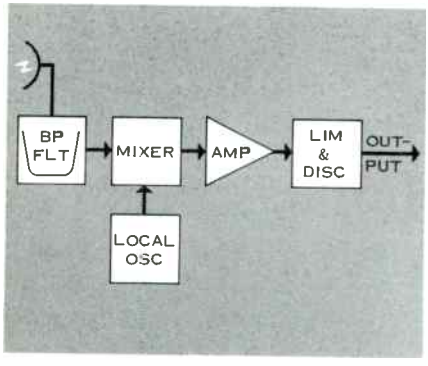


Figure 3. Basic microwave receiver first "heterodynes" the signal down to the 70-Mc intermediate frequency. The discriminator then demodulates the signal, recovering the original baseband frequencies.

Equipment Refinements

The basic equipment described thus far is essentially the same for both TV transmission and message traffic. While this basic system would transmit and receive signals, the quality of transmission would probably not be acceptable by most standards. Therefore, modern transmitters and receivers include a number of auxiliary items whose purpose is to refine the transmission quality. Figures 4 and 5 show simplified block diagrams of a transmitter and a receiver equipped with some of these auxiliary items.

In a message transmission system a technique called "emphasis" is used to overcome some of the effects of noise. This involves a "pre-emphasis" network in the transmitter and a corresponding "de-emphasis" network in the receiver. The effect of emphasis is to improve the signal-to-noise ratio at the more critical high frequencies.

A similar technique is used in television transmission (where it is often called "pre-distortion"). The technique is much the same for both message and video channels, but the reasons for using it are quite different. It provides a high-frequency noise advantage when used for message traffic. However, low-frequency noise interferes with a television signal much more than high-frequency noise. Therefore, emphasis offers little *noise* advantage to video transmission. Instead, emphasis is used to improve the quality of color transmission. (Standard practice is to engineer a video link for color transmission, even though initial plans may only call for it to handle black and white signals.) Intermodulation between the color subcarrier and the low-frequency components of the video signal is produced by the transmission variations in the system. The unwanted products of this intermodu-

World Radio History

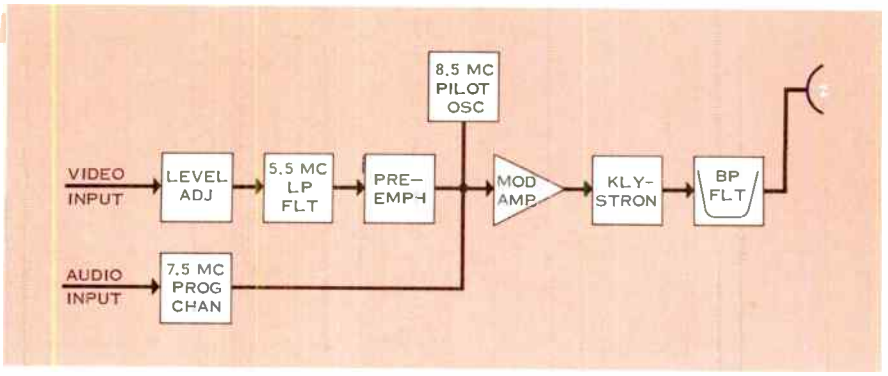


Figure 4. Performance of the basic transmitter shown in Figure 2 is improved by adding auxiliary items. The program channel carries the audio portion of a television signal or can be used for other services.

lation are of such a frequency that they interfere with the color information. Emphasis reduces the amplitude of the intermodulation products by reducing the low-frequency components of the signal. The result is better color performance.

Another refinement to the basic microwave system is the addition of a pilot-frequency oscillator to the transmitter and a pilot-frequency detector at the receiver. With the addition of these two items a single-frequency pilot tone is transmitted at all times, even when no signal is applied to the system. This permits a continuous check of both equipment and propagation path continuity, even under "no-signal" conditions.

Further refinement often consists of level adjustment facilities. The standard interface video signal has an amplitude of 1 volt peak-to-peak. Most transmitters are designed to accommodate this level. A signal at any other level must be adjusted. A lower-amplitude input signal goes through an amplifier, while a signal of higher level goes through an attenuation pad.

Level adjustment in the receiver normally takes the form of additional amplifiers. An intermediate-frequency amplifier often follows the IF preamplifier to provide a fairly high-level signal for the limiter and discriminator. A baseband amplifier placed after the de-emphasis network provides the proper output level for the baseband signal.

An important part of most microwave systems for video transmission is the "clamper" at the output of the receiver. Ideally the baseband frequency response of such systems would extend down to dc on the low-frequency side. However, it is not practical to extend the response below a few cycles per second. Therefore, no dc reference potential is included in the video signal. The *relative* amplitudes of the different

World Radio History

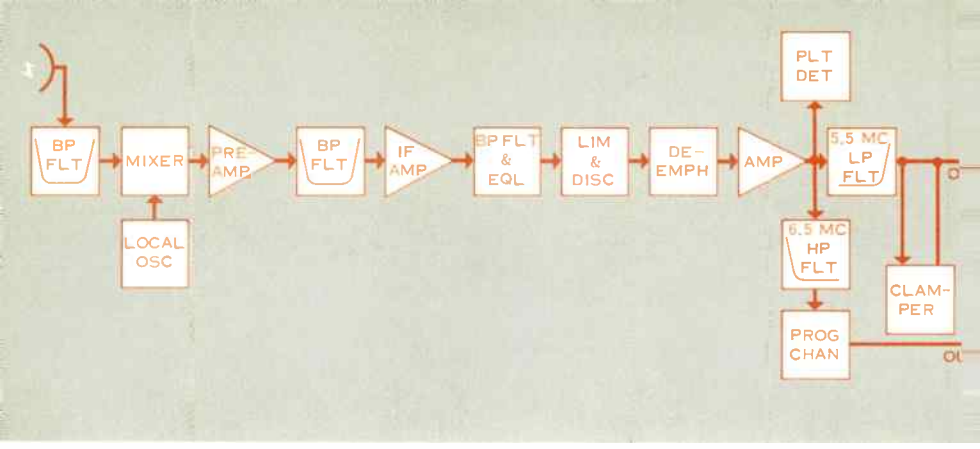


Figure 5. Receiver performance is also improved by adding auxiliary items. Particularly important in television transmission is the clamper, which provides a dc reference to maintain constant picture brightness.

frequency components in the video signal are accurate, but the whole signal may move up or down in relation to a constant dc reference. The result is a brightening or darkening of the whole TV picture as seen on the viewer's screen. The clamper is a device which provides the necessary dc reference potential. Thus, when a clamper is used, the brightness of the viewer's picture remains constant regardless of any variations in the transmission.

Bandpass filters inserted in the IF section before and after the IF amplifier limit the bandwidth of the IF signal, removing extraneous signal components outside the desired frequency band.

An equalizer in the IF stage also improves performance. Since different frequencies are delayed by different amounts as the signal passes through the system, the time relationship between the frequency components is altered, with the result that the demodulated video signal is distorted. The equalizer's function is to adjust the time relationship between the various frequency components in the IF band. The effect is better reproduction of the transmitted signal.

The Audio Signal

The audio signal may be carried on the same microwave system as the video signal by means of a program channel. The program channel is a separate FM transmitter which accepts an audio signal with a bandwidth of about 15 kc (enough bandwidth to give high-fidelity reproduction of both music and speech). In a typical arrangement such as that shown in Figures 4 and 5, the program channel carrier is centered on 7.5 Mc. This puts it well above the frequency range of the video signal, and it is simply bridged onto the transmission path near the input to the transmitter. In other

World Radio History

words, the signal from the program channel is frequency-division multiplexed above the video signal. A common arrangement uses a lowpass filter (5.5 Mc in the illustration) ahead of the program channel insertion point. This lowpass filter "cleans" a slot for both the program channel and the pilot frequency, removing any unwanted interference from the frequencies above the video baseband.

At the receiver both the pilot frequency and the program channel are "picked off" just ahead of a similar lowpass filter. This lowpass filter removes the pilot and program frequencies from the video output. A highpass filter keeps the video signal out of the program channel.

For an "off-the-air" pickup such as that used for community antenna television systems, the program channel would not normally be used to carry the audio portion. In a standard broadcast signal the audio is transmitted 4.5 Mc above the video carrier. When such a signal is picked up by the receiving antenna, both the audio and the video portions can be transmitted through the normal transmission path without separating the audio signal. In such a case, a program channel may be used for other purposes, such as transmitting the signal from an FM broadcast station. In fact, two or three program channels with carriers at different frequencies may be used to transmit FM broadcast signals.

Conclusion

A typical microwave system for television transmission is capable of handling 960 voice channels in place of the single video channel. In both cases, the baseband is about four megacycles wide, and the power-handling requirements are approximately the same. Both arrangements use the same basic system. The difference is in the baseband treatment and the auxiliary items.

Such things as the emphasis networks and the various filters, used in both types of systems, are similar versions of essentially the same item. On the other hand, the clamper used for a video system has no application in a message system because no dc reference is required for message traffic. However, one auxiliary item developed for television transmission can be very useful in a message system. This is the program channel, which can also be used for wide-band data service.

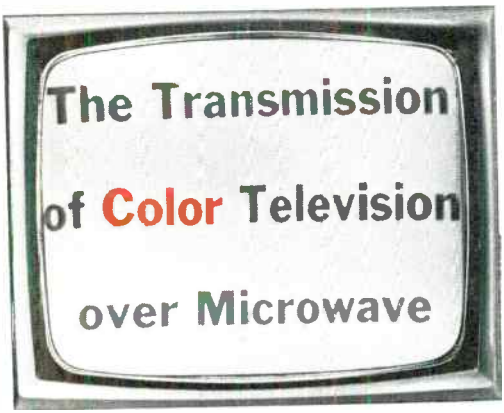
Individual systems differ, of course. The attempt here has been to explain how a basic system operates, then to describe a few of the auxiliary items that tailor the system to carry a specific type of traffic. Microwave systems designed for television transmission are basically the same as those designed for message transmission. Their general purpose is even the same — to provide satisfactory service to the user, without distracting noises in one case and without brilliantly colored "snow" in the other case.

World Radio History

World Radio History

the Lenkurt Demodulator

VOL. 11, NO. 2 FEBRUARY. 1962



A new and important period of growth has begun for television. After fifteen years primarily as a medium of mass entertainment, television is achieving new status as a powerful technique for extending and improving education. Closed-circuit television networks, covering large areas, have been built in several states, and many others are in the planning stages. For such systems, microwave radio is the most likely method of transmitting the signals from their point of origin to outlying areas, except where the distance is very short. This article discusses monochrome and color television signals, and the characteristics required of microwave equipment used in their transmission.

Educational television is emerging from several years of experimentation as one of the best solutions to the severe shortage of good teachers in the United States. Educators believe that television will greatly improve the quality of teaching in almost all types of schools. In areas that are sparsely populated or where school budgets are low, television permits a broader curriculum and lessons of a better quality than otherwise available. Where teachers and finances are more abundant, television relieves teachers of the chore of presenting routine, generalized subjects and allows them more time to spend with small, specialized classes where their teaching skills may be used more effectively.

The present trend of commercial television is toward more and more use of

World Radio History

Figure 1. Principal value of educational TV is in improving quality rather than quantity of instruction. Television permits teaching-specialists a larger audience, and allows classroom teachers more time for personal contact with students, thus increasing effectiveness of both the specialists and local teachers.

color and this is believed to add even more to the effectiveness of television as an educational tool. Many national educational programs are now broadcast regularly in color.

Television has proved to be such an effective teaching medium that rather elaborate methods are sometimes employed to reach as many students as possible. In Indiana, for instance, two large four-engined aircraft loaded with television transmitters and video tape reproducers climb daily to about 23,000 feet where they broadcast on several channels to an estimated five million students in three states. Despite the high cost of the aircraft and their equipment (one plane is in reserve), the number of students served is very large, thus making the cost per student very low.

This approach, while effective in the particular region where it is used, is not economically suitable for most areas—for instance, where population is sparse, or concentrated in a few large cities. In addition, this transmission method is particularly vulnerable to such occurences as bad flying weather, mechanical and electrical difficulties in the aircraft, and the like.

For these and other reasons, most television networks transmit their signals over coaxial cable or point-to-point microwave radio. Because of the high cost of coaxial cable, however, it is generally used only in special situations or where the transmission distance is quite short—three miles or less, typically. To date, most television networks, both commercial and educational, use microwave transmission.

The Television Signal

There are important differences between a television signal and most other types of signal commonly transmitted over microwave. Voice or other audio signals consist of only a single variable—a voltage or current which varies with time to represent the original sound vibrations. Reproduction of a picture or scene, however, requires that at least three independent variables be transmitted—information about the relative brightness of all points in the picture, and also their horizontal and vertical position.

These three variables are encoded into a single time-varying signal by systematically scanning the picture area very rapidly. Information about the

World Radio History

scanning method is transmitted with the brightness information so that the picture can be "reassembled" at the receiver in the same manner in which it was "taken apart" at the transmitter. This scanning information takes the form of synchronizing pulses which tell the receiver when to begin each "frame" or picture image, and when to begin each of the line scans which make it up.

In the standard North American television system described in this article, the image is scanned 60 times a second. These image *fields* are paired and interlaced with each other to provide 30 *frames* or complete pictures a second. The 60-cycle field rate is rapid enough to be undetectable to the human eye, thanks to the eye's "persistence of vision" or poor high-frequency response.

As shown in Figure 2, picture brightness is transmitted by varying the amplitude of the video signal during each line scan. At the end of each line, it is "blanked" during its return by a pulse having an amplitude "blacker than black." During this blanking interval, the horizontal synchronizing pulse is transmitted which starts the next line scan on its way. Note that black or darkness results in *increased* modulation of the television signal, and white causes less modulation. The seemingly inverted presentation used in Figure 2 is based on the IRE (Institute of Radio

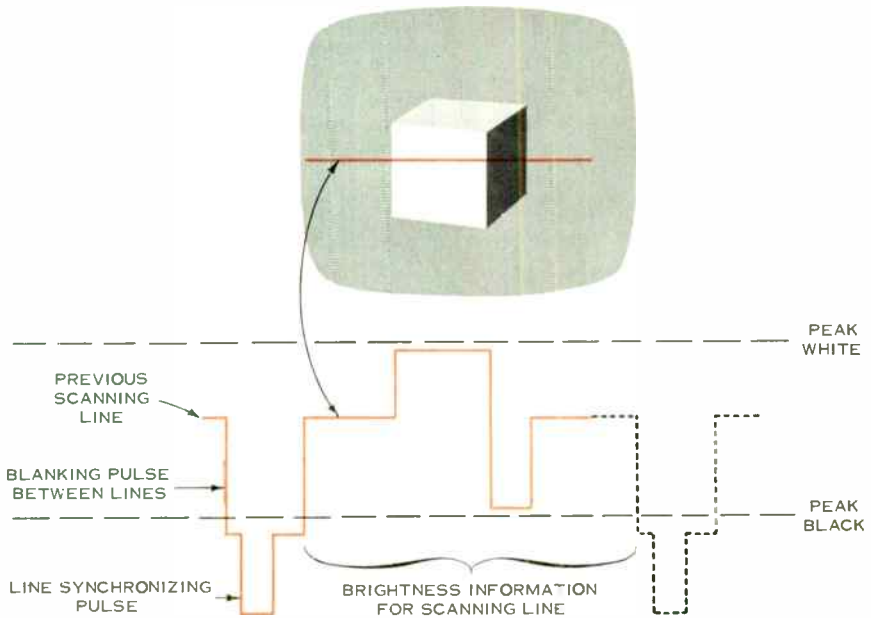


Figure 2. Basic television signal consists of a blanking pulse for darkening the tube between scanning lines, a synchronizing pulse for starting scan in step with the camera, and brightness or video information. The partial scan represented by the red line in the picture causes a waveform like that shown directly below it. The end of each frame is identified by a special combination of pulses, not shown.

World Radio History

Engineers) scale of amplitude relationships adopted as standard in the United States.

Image Detail

The fineness of detail that can be transmitted is called the resolving power or *resolution* of the system. It may seem surprising that resolution of the television system may be different in the horizontal direction than in the vertical. This is because of the way in which the complete television image is formed. Each complete frame consists of 525 horizontal scanning lines, about 40 of which are blanked out between frames and fields. The number of "active" or visible lines remaining determines how many discrete objects or picture elements can be distinguished in the vertical direction. The greater the number of scanning lines, the better the vertical resolution. Since the smallest object that can be resolved must be somewhat larger than the distance between lines (so that it can't fall between successive lines and be missed) the number of vertical picture elements that can be resolved is about 342.

Horizontal resolving power depends on the speed with which the scanning spot sweeps across the picture, and the ability of the electronic circuits to respond faithfully to very close transitions between dark and light. This ability is a direct function of the bandwidth of the system. The finer and closer the detail, the more rapidly the signal must vary, and the greater the bandwidth required. The sweep rate is determined by the need to scan all 525 lines in 1/30 second and the desirability of obtaining a horizontal resolution approximately equal to vertical resolution. These requirements are met in a television system having a bandwidth of about 4.3 megacycles.

This bandwidth is necessary only

Figure 3. When bandwidth is reduced, horizontal definition is impaired without affecting vertical resolution. Compare the horizontal and vertical converging lines in this example in which bandwidth is limited to about one mc.

where 30 frames a second are required. If it is acceptable to transmit more slowly—say, five frames a second—the bandwidth can be reduced in proportion—to 700 kc in this example. This concept of "slow-scan" television is used in some industrial and commercial applications where the image does not change rapidly. It has been used by banks for examining signatures from a distance and by electric power companies for monitoring remote meters. Conversely, some technical processes may require *greater* bandwidth because of the need for more resolving power than is possible with the standard television system.

Color Television

Early, unsophisticated approaches to color television required three separate pictures to be transmitted, one for each of the primary additive colors. This required three times as much bandwidth as a monochrome signal, and was obviously undesirable, both from the standpoint of cost, and excessive use

World Radio History

of bandwidth. In addition, such a system did not provide a signal suitable for viewing on conventional monochrome receivers.

In the color television method that finally evolved, a single signal is transmitted which strongly resembles the standard monochrome signal, and which can be received in a normal fashion on a conventional black-and-white receiver. Color information is added to the basic signal in the form of a special *chrominance* signal which is combined with the regular video transmission in such a way that it does not noticeably affect the picture as it appears on a monochrome receiver.

Actually, the chrominance signal adds two new independent variables to the basic television signal. One is the *hue* or color of the subject, and the other is its richness of color or *saturation*. A red object of highly saturated color would appear intensely red or crimson. If saturation is decreased without changing the hue, the subject would appear as some shade of pink. When the saturation of a color is low, it is the same as diluting the pure color with white.

The chrominance signal takes the form of a special sub-carrier, the sidebands of which modulate the regular video or luminance signal which supplies brightness information about the image. The color sub-carrier has a frequency of 3.579 mc, but is shifted in phase to indicate the exact hue of the scanned object at any instant. The amplitude of the color sub-carrier determines the saturation or richness of the resulting color.

The phase and amplitude values of the chrominance signal are extremely important, since they carry all color information. Even slight distortions affect the quality of color reproduction. In order to establish a suitable phase and frequency reference, a short "burst" of the unmodulated color sub-carrier is transmitted during the blanking pulse for each line. Although the color burst consists of only eight or nine cycles of the color sub-carrier, this is enough to correct the phase of an oscillator in the receiver, as required. In this way the color television receiver is effectively phase-locked to the transmitter.

In order to prevent the chrominance signal from affecting the black-and-white image on a monochrome receiver, its frequency (3.579 mc) is chosen to be an *odd* multiple of *half* the line scanning frequency. This has the effect

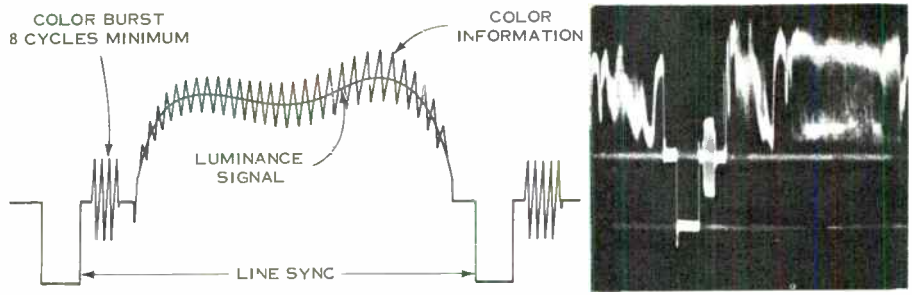


Figure 4. Basic color television signal is diagrammed at left, actual signal is pictured at right. So-called "color burst" is transmitted during blanking pulse to serve as a phase reference for the color information which modulates the basic video or luminance signal.

World Radio History

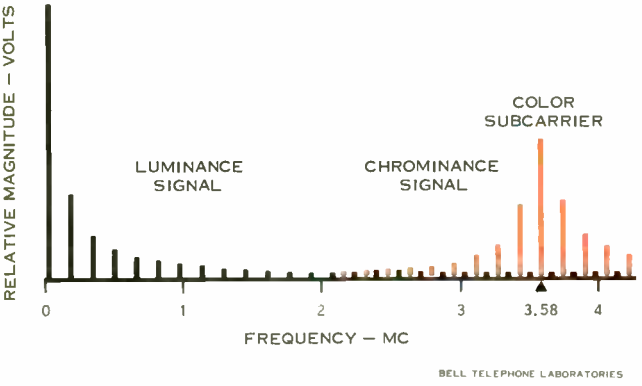


Figure 5. Energy distribution of the color television signal across the spectrum. Line scanning and color sub-carrier frequencies are so chosen that sidebands fall into portions of spectrum unused by each other, thus using bandwidth most efficiently.

of allowing the color modulation of the luminance signal to cancel itself out (as far as the eye can detect) in consecutive frames. For instance, in one frame, the color modulation on the luminance signal may tend to darken a certain spot. In the following frame, the same spot is lightened, thus neutralizing the darkening of that spot in the preceding frame. The *average* brightness remains the same.

This ingenious feature also conserves the bandwidth of the system. Since the television signal is interrupted by individual line scans and by 60-cycle fields, it consists of many harmonics of the line scan frequency, each of which is surrounded by 60-cycle sidebands. In monochrome television, the frequency spectrum between these sidebands is unused. By choosing a color sub-carrier frequency that is an odd multiple of half the line scanning frequency, all the chrominance signal components fall into these unused spaces between the luminance signal components. Accordingly, this system of color television requires no more bandwidth than the conventional monochrome system. Figure 5 shows the energy distribution of the color and monochrome signals across the video spectrum.

Microwave Design Considerations

The extreme complexity of the television signal—especially the color signal—imposes very stringent requirements on the transmission system. Distortion, poor frequency response, transmission irregularities all tend to degrade the quality of the final image. Since such distortions tend to be cumulative as the number of repeaters in the system increases, it is necessary to provide extremely good performance in each link of the system.

Ideally, the entire television system, including the microwave or other transmission equipment, should be free from amplitude distortion and non-linear phase shift from almost zero frequency to at least 4.5 mc. In practice this is very difficult to achieve, primarily because of the very great number of octaves which the signal must cover. By definition, an octave covers a range in which the highest frequency is just twice the lowest. Thus, the frequency range 10 to 80 cycles covers three octaves, and the television video spectrum covers nearly 18 octaves. Over such a broad range, the excessive use of negative feedback to maintain uniform frequency response may result in ampli-

World Radio History

fier instability and serious phase distortion. When transistors are used, the problem is complicated even more by the varying phase shift introduced by the transistors themselves, as a function of signal level.

The technical demands on microwave are different for television and message service. In most message applications, non-linear phase shift and envelope delay distortion are relatively unimportant—but the random noise and intermodulation distortion contributed by the radio are of great importance. Monochrome and color television are greatly disturbed by single-frequency interference, particularly at low frequencies. For this reason, special emphasis must be placed on eliminating hum from any source. By contrast, this has hardly any effect on message service microwave.

One of the most important differences is the extreme sensitivity of the television signal to non-linear phase shift. This is almost always associated with variations in the amplitude frequency response of the system, particularly abrupt changes or transitions. The resulting non-linear phase shift delays some frequencies more than others, with the result that the signal waveform is distorted. Although delay distortion of a speech or music signal is not readily detected by the ear, similar distortion of the television signal is very noticeable and affects the quality of reproduction greatly. Minor waveform aberrations show up as noticeable changes in the image.

Color television is particularly vulnerable to *differential phase* and *differential gain*. Simply speaking, differential gain means a change in the gain of the system as a result of signal level variations. In a color television broadcast, the presence of differential gain may cause some colors to appear unsaturated or washed out, while others might appear excessively brilliant. These extremes may be produced merely by changes in the brightness of various scenes.

Similarly, differential phase refers to phase shift which occurs as a function of the video signal level. This is important in color television because the colors which appear on the screen are determined by the exact phase relationship between the color burst and the color sub-carrier. Differential phase causes color distortion which is not constant, but which varies according to the brightness of the scene. Thus, a crimson object might change to orange or some other color if the overall illumination level of the scene were to change.

Low Frequency Effects

Ideally, the television system should have uniform frequency response down to zero frequency—that is, dc—itself. This is generally impractical because of the complications it introduces into the design of amplifiers used in the system. Instead, the direct-current component required to provide a brightness reference or base is reinserted at various points in the transmission system by a so-called clamper or dc restoration circuit. Without the clamper, the brightness of the scene tends to vary according to the overall range of brightness in the picture. Thus, a scene of average brightness would be darkened all over by the addition of a single small area of much greater brightness.

Where the low-frequency "roll-off" (of the video amplitude response) lies in the region 30 to 60 cycles or lower, the picture may show some vertical shading because of the system's inability to achieve the correct amplitude level immediately following the field synchronizing pulses. When frequency re-

World Radio History

Figure 6. Examples of streaking caused by poor response at the lower frequencies — positive streaking above, negative streaking below.

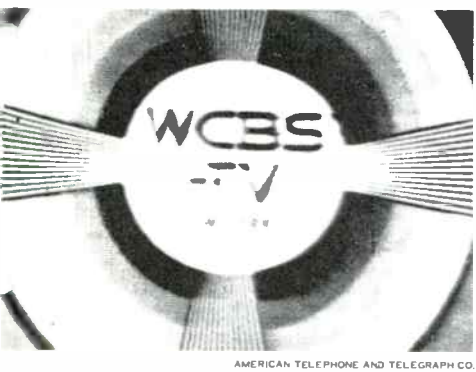


Figure 7. Smearing results from frequency response deficiencies in the frequency range between 150 kc and one mc.

sponse is irregular in the range below about 150 kc, an image defect called *streaking* occurs. If the gain drops one or more db in this range, the waveform may show a long-duration "overshoot" at tonal transitions from dark to light and light to dark. This shows up on the picture tube as "negative" streaking in which a dark streak trails a bright object (or light streak following a dark object). If the gain increases in the lower frequency region, the following streak is of the same color as the preceding image area, and is called "positive" streaking. These are illustrated in Figure 6.

If the frequency aberration occurs in the middle frequency range, that is, from about 150 kc to nearly a megacycle, the picture may show *smearing;* vertical images are blurred along the horizontal axis to give an overall smeared effect to the picture. Like streaking, this effect is caused by the inability of the waveform to follow rapid transition from one brightness level to another, as required.

High Frequency Errors

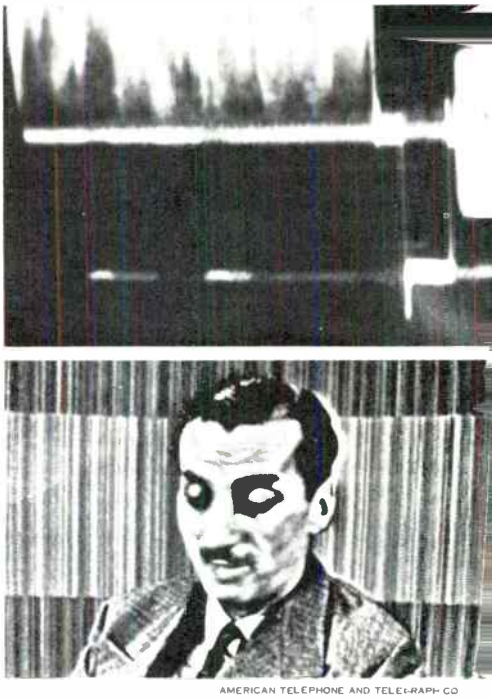
When the high-frequency gain of a television system is altered, the fine detail of the image is affected. If the high frequencies are depressed, there is some loss of horizontal detail and a general softening of vertical edges. If high-frequency gain is increased slightly, the picture assumes a much crisper quality due to the waveform's increased ability to follow very rapid changes and to make "squarer" pulses. If the high frequency gain is increased even more, the waveform may have a tendency to overshoot, thus producing a "spike" as it is often called. Figure 8 shows such a spike at the leading edge of a line synchronizing pulse. When this occurs in the picture, the overshoot may pro-

World Radio History

duce a very narrow contrasting outline at the right edge of contrasting objects. The higher the frequency of the increased response, the narrower the outline. This phenomenon is generally called a "following white" or "following black." If the overshoot occurs at lower frequencies, the outline is broadened, and smearing results.

A sharp "roll-off" of the high-frequency response of the system, or an irregularity of the frequency response within the pass-band of the system can result in "ringing" or damped oscillation immediately following an abrupt tonal transition. On the screen, this appears as a series of fine alternate light and dark lines just to the right of contrasting edges in the image. In long transmission systems, ringing may occur from the use of many tandem microwave links, even though the frequency cut-off of each is outside the video band. In effect, as the number of tandem microwave links increases, the ringing frequency becomes lower until it is noticeable in the picture. This can be minimized by avoiding a sharp frequency roll-off in the microwave frequency response, and keeping the over-all response free of irregularities which can be exaggerated by similar faults in succeeding links of the system.

Color television in particular is extremely sensitive to minor irregularities in the phase and frequency response in the upper part of the video spectrum. Even small frequency response deviations result in non-linear phase shift to some degree, and this has an important effect on the ability of the waveform to follow brightness variations faithfully. More important is the effect of such phase shift on the quality of color reproduction. Although the color subcarrier is at a frequency of approximately 3.6 mc, phase modulation of the carrier produces sidebands as low as



AMERICAN TELEPHONE AND TELEGRAPH CO

Figure 8. Overshoot results from excessive gain at fairly high frequencies. Compare the "spike" at the leading edge of the synchronizing pulse and the "following white" to right of man's head.

AMERICAN TELEPHONE AND TELEGRAPH CO

Figure 9. Ringing, shown as alternate bands to right of vertical lines, is caused by sharp discontinuities in frequency response characteristic, usually at high frequencies.

World Radio History

2 mc and as high as 4.2 mc. If some of these sidebands are delayed more than others by non-linear phase shift, the resulting colors produced at the receiver will be distorted in a noticeable and objectionable way. Figure 10 indicates some of the recommended tolerance limits for frequency response in an overall television system, including transmission facilities. Note that near the color sub-carrier frequency, requirements are as stringent as they are at the very low frequencies.

Both monochrome and color television have similar requirements at the low end of the video spectrum. Most of the transmission problems previously described affect both equally. Although it is extremely important to have excellent low-frequency response in the over-all system, the requirements imposed on the microwave system can be relieved greatly by the use of clamper circuits. Clampers, in effect, restore the lost direct-current component that establishes the brightness reference level, by measuring the voltage change or drift of the line-synchronizing pulse and supplying a compensating bias voltage to restore the reference base. This frees the television image from changes in brightness of the over-all picture due to the brightness content of the picture.

Effective clampers, however, are generally complicated and expensive. Overly-simple clampers may substitute one trouble for another. If the microwave system is able to transmit direct-current information, clampers can be dispensed with altogether. Although this is generally impracticable, the need for clampers can be minimized by using a transmission system that can transmit extremely low frequencies that closely approach direct current.

The most stringent requirements at the upper end of the video spectrum are dictated by the need for as little differential phase and gain as possible in the transmission of color. This is best done by preserving a very wide bandwidth, free from irregularities well beyond the actual frequency limits of the television signal itself. The farther the cut-off frequency of the system is from the signal, the less distortion there

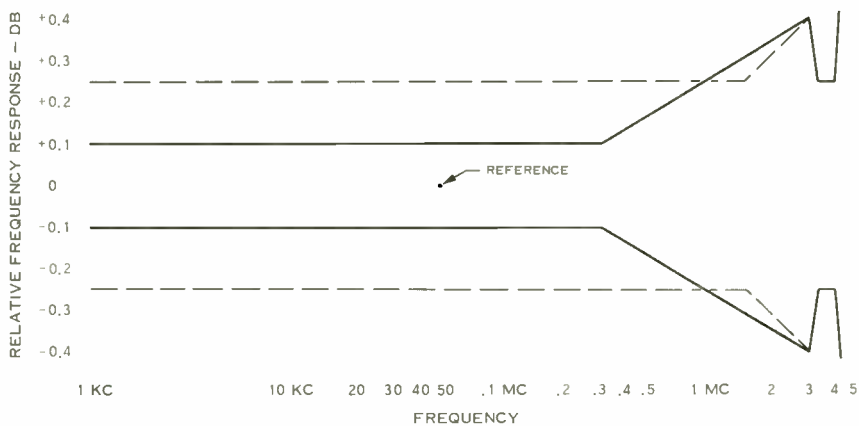


Figure 10. Typical frequency response limits specified for microwave systems used in transmitting color television. Variations within these limits are expected to be gradual, rather than abrupt.

World Radio History

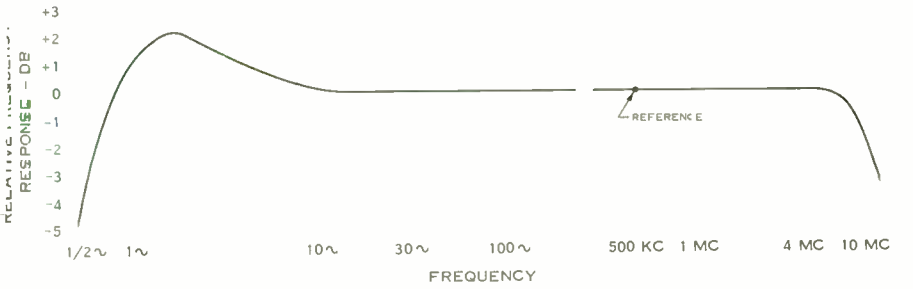


Figure 11. Typical frequency response characteristic for the Lenkurt 76A microwave system used for transmitting color television. Rising characteristic at very low frequencies compensates for non-linear phase shift effects, improves "square-wave tilt" and reduces need for clamping circuit.

will be in the final image. Figure 11 shows the frequency response characteristics of the Lenkurt 76A microwave system, especially developed to accommodate color television.

The Future of Color

Compatible color television is about seven years old in the United States and Canada, but is only now gaining wide public acceptance, primarily because of the high cost of color receivers. Now that costs are coming down and color broadcasting is increasing rapidly, it may be expected that even cheaper and better receivers will be developed. It seems reasonable to expect that color television will experience the same rapid growth that characterized the first television broadcasting fifteen years ago. Despite the extra complexity of the color receiver, large volume production will reduce prices considerably.

Although color enhances the attractiveness and entertainment value of commercial television, it should be particularly effective in educational television. The additional impact of color adds considerably to television's ability to convey information and impress the student viewer. Closed-circuit color television has already been used in medical and surgical teaching with excellent effectiveness.

Even if color television is not to be used in the immediate future, microwave transmission facilities should have the capability of transmitting color without distortion, since the lack of this capability will be one of the most important factors in retarding the future obsolescence of the system. •

BIBLIOGRAPHY

1. *Teaching by Closed Circuit Television*, Report of a joint conference, Committee on Television, American Council on Education, Washington, D. C., and State University of Iowa, Iowa City, Iowa; 1956.
2. *Television Signal Analysis*, American Telephone and Telegraph Company; New York, 1955.
3. *Transmission Systems for Communications*, Bell Telephone Laboratories; New York, 1959.
4. John W. Wentworth, *Color Television Engineering*; McGraw-Hill, New York, 1955.

World Radio History

World Radio History

the Lenkurt Demodulator

VOL. 12, NO. 10 OCTOBER, 1963

Performance Testing of Television Channels

Part One

Television, having completely conquered home entertainment, is now maturing into a powerful industrial and educational communications tool. Schools, businesses, and utilities are finding in closed circuit TV a valuable means of operating more effectively.

The technical side of relaying television signals from one point to another may provide a few surprises, even to organizations well experienced in other forms of communication. Although television is transmitted over facilities very similar to those used for multi-channel communications, performance requirements and testing procedures are quite unlike those for speech signals. This article is the first of two which review the basic methods of testing and evaluating television transmission channels.

Television, especially color television, imposes new performance requirements on communication circuits. Unlike most other kinds of transmission, television is extremely waveform-sensitive. The ability of a television circuit to transmit an image which faithfully represents the original subject is directly dependent on the ability of the system to reproduce a signal waveform precisely. By contrast, in speech or even data transmission, the accurate reproduction of the signal waveform is relatively unimportant so long as the magnitudes of the various spectral components of the signal are reproduced accurately.

World Radio History

In television, the reproduction of the brightness of individual picture elements is determined by the voltage of the television signal at a given moment. To faithfully reproduce a contrasting edge, say, from black to white or white to black, the television signal must be able to make an abrupt change or "step" voltage without distortion. The shape of the video waveform depends entirely on the content of the picture, possibly consisting of impulses, steps, or level plateaus instead of the sine wave combinations which characterize speech or music waveforms.

This inherent difference between video and audio waveforms naturally leads to differing transmission requirements and capabilities. An effective transmission system for speech must have good amplitude-versus-frequency linearity. That is, amplitude response must not vary appreciably across the frequency band occupied by the audio signal. However, phase irregularities are relatively unimportant in speech transmission.

In a television system, phase characteristics become extremely important. Non-linear phase shifts as a function of

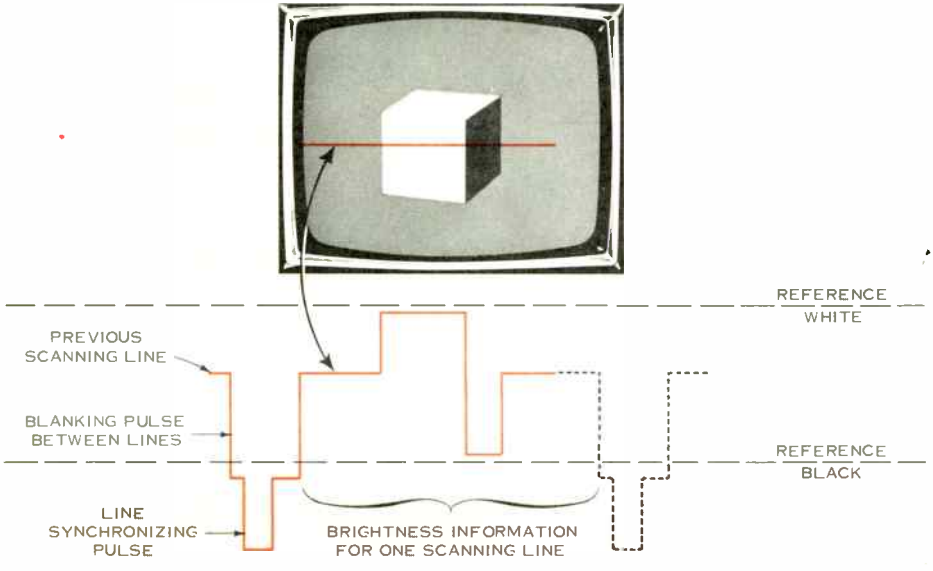


Figure 1. Basic television signal for a single scanning line. The partial scan shown in red in the picture raster produces a waveform like that shown directly below it. At the end of each scan, the blanking pulse darkens the tube while the scanning beam is returned. The scanning sweep is triggered by the synchronizing pulse which occurs during blanking. Ability of transmission system to pass such abrupt waveforms without distortion is vital to good picture reproduction. Clean edges between contrasting colors require excellent ability to transmit square waves and other such transients.

World Radio History

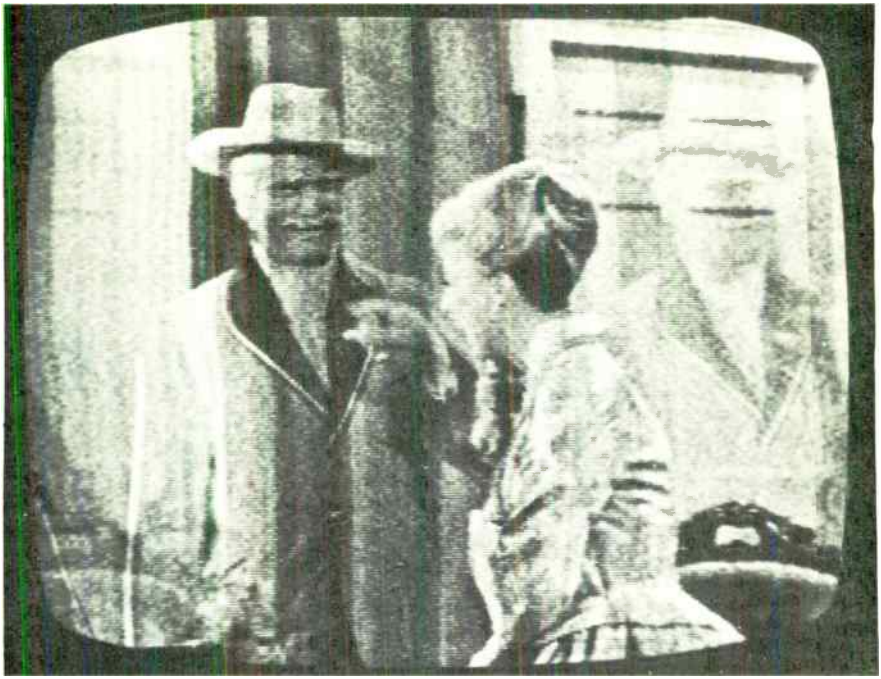


Figure 2. Poor frequency response or other irregularities lead to distortion of various sorts in the reproduced picture. Echoes usually result from non-linearities in baseband frequency characteristic, or from multipath transmission. Since transmission errors are rarely as bad as this exaggerated example, sensitive test methods are required to permit accurate evaluation of quality.

frequency serve to distort the all-important waveshape, and these distortions are directly visible on the television screen as some form of picture distortion. Amplitude-versus-frequency linearity is also extremely important because it determines the ability of the equipment to reproduce accurately the brightness values of the subject. Furthermore, it determines the ability of the waveform to make rapid changes from one level to another in response to fine detail in the picture. Thus, good frequency response affects the ability of the television signal to accurately achieve the desired voltage levels and to reach these values at the required time.

When linearity of phase shift or amplitude response is poor, various forms of distortion may be produced. For instance, where low frequency (such as 60 cps) phase and amplitude response is poor the signal is unable to adjust itself to the desired value for a considerable portion of the scanning sweep. This shows up on the picture monitor as "streaking"—errors in the brightness of the image. Response errors at the higher frequencies may result in "ringing," "smearing," or echoes. The mag-

World Radio History

nitude and distribution of frequency and phase errors have an important bearing on the way the picture is affected. The accurate location of picture information on the screen is a function of the ability of the system to respond within a certain time, but not *beyond* the proper time. Frequency and response time are reciprocal functions of each other; one bandwidth over which the amplitude response error extends, the smaller the separation between the desired detail and its echo. If the bandwidth occupied by the irregularity becomes smaller, the echo is more widely separated from the main image.

Most of these effects which are of great importance in the transmission of

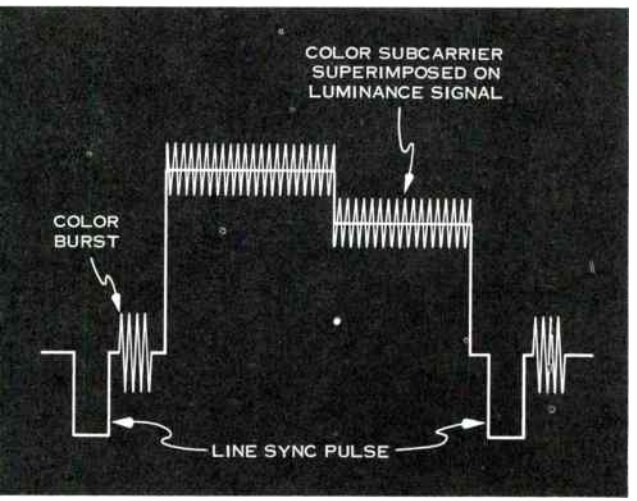


Figure 3. In NTSC color system, "color burst" is transmitted as phase reference for color subcarrier superimposed on luminance signal. Any change in phase or amplitude of subcarrier causes color change in picture. This is called "differential phase" or "differential gain" when caused by changes in luminance signal.

can be transformed into the other, and this transformation determines how the picture is affected by frequency response errors.

For instance, if there is a "narrow" irregularity in the television baseband frequency response—that is, an amplitude variation that is restricted to a rather narrow band of frequencies, the waveform will exhibit a "ringing" of low amplitude but of long duration. This shows up as an echo or "ghost" following important transitions of dark and light in the picture. The greater the

television have only negligible effect on the transmission of single or multiple-channel speech signals. For this reason, conventional methods of testing speech channels are able to reveal little about the suitability of a transmission channel for carrying television signals. Testing methods are required which are sensitive to those characteristics which directly affect picture quality.

Television Test Signals

A rough idea of transmission quality can be gained simply by observing the

World Radio History

transmitted picture, and often a skilled technician can diagnose difficulties directly from the raster. Since such evaluation is entirely subjective, problems arise in defining transmission quality. An impairment barely noticeable to one person may be highly objectionable to another. Under these circumstances what constitutes "acceptable" picture quality, and how should it be expressed?

In the final analysis, of course, a transmission system must provide a picture acceptable to the viewer. Thus, the amounts of tolerable distortion are based on viewing tests by critical viewers. Large amounts of distortion may be tolerated if the effect on subjective picture quality is negligible; but other, more potent types of distortion may be allowed much narrower tolerances. This, of course, suggests the need for standardized test methods which can be readily reproduced under a variety of conditions, but which do not require subjective judgements of quality. Many such tests have been developed by the television industry.

In general, the tests developed for black and white television are concerned with the measurement of three transmission parameters:

1. Amplitude-versus-frequency linearity,
2. Phase-versus-frequency characteristics, and
3. Transient response.

The first of these, amplitude linearity, implies the ability to reproduce signal voltage accurately regardless of the frequency (within the band of interest). The last, transient response, is the ability of the system to "follow" sudden, impulsive changes in the signal waveform. This ability is largely controlled by a combination of the two previous characteristics, amplitude and phase response. In general, good transient response requires excellent amplitude and phase characteristics, but is not necessarily assured, since minor perturbations of one may combine with the

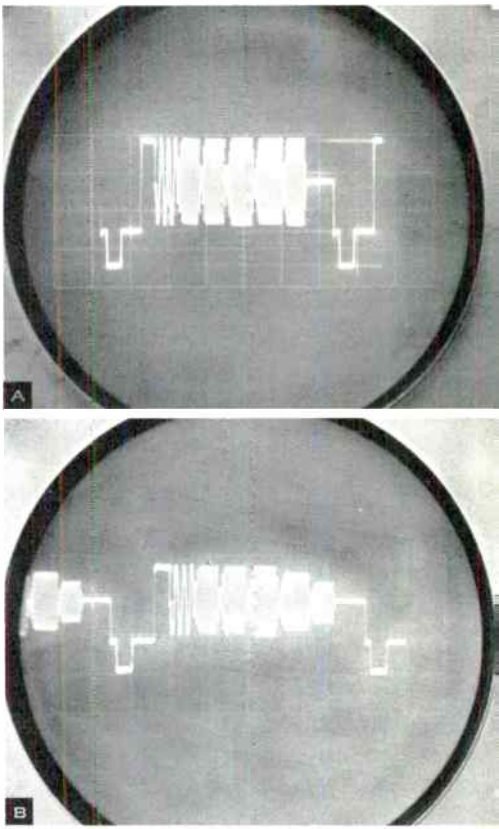


Figure 4. Multiburst signal consists of consecutive "bursts" of ascending frequencies, all transmitted at reference white level (A). Changes in frequency response show up as variations in relative amplitude of different frequencies (B). Note frequency roll-off at upper frequencies.

World Radio History

other in "the wrong way" to cause distortion of transients.

Color television transmission requires the consideration of these three factors, plus two more:

4. Differential gain and
5. Differential phase.

These last two parameters are perhaps the least understood, and yet they are the ones which place the most stringent requirements on the transmission system. *Differential gain* is the variation in the gain of the transmission system as the luminance or brightness signal varies between the values for "black" and "white." Any variation in phase of the color subcarrier as a result of changing luminance level is called *differential phase*. Ideally, variations in the *luminance* signal voltage should produce no changes in either the amplitude or phase of the color subcarrier. Thus, the presence of either one implies distortion—and thus it is redundant to speak of "differential gain distortion" or "differential phase distortion."

Both of these parameters are directly concerned with color information. In the American and Canadian NTSC system, a color subcarrier at a frequency of about 3.58 mc is superimposed on the luminance signal. Different colors or *hues* are indicated by shifting the phase of the color subcarrier. The *saturation* or richness of the color is transmitted by varying the amplitude of the color subcarrier. In an ideal system, which would have no differential gain or differential phase, changing brightness values in the picture would have no effect on the phase or amplitude of the subcarrier. However, when differential phase is present, a change in the brightness of the scene could change the color of a green object to yellow, while differential gain could change the color saturation from, say, a dark green to a pale value.

Even before color television, there was no simple, easy-to-use test signal

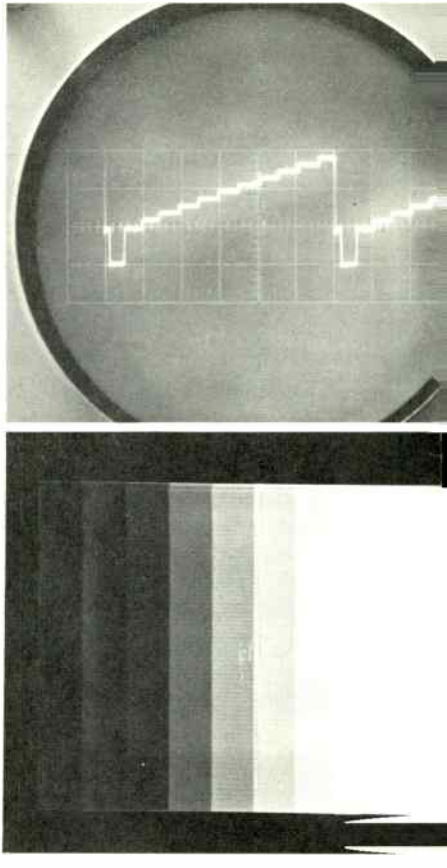


Figure 5. Typical stairstep test signal as shown on the "A-scope," and how it appears on picture monitor. Note that each step represents a shade of light. When applied to system, each step is equal in amplitude. Non-linearities in system are revealed by compression or expansion of individual steps or parts of waveform.

World Radio History

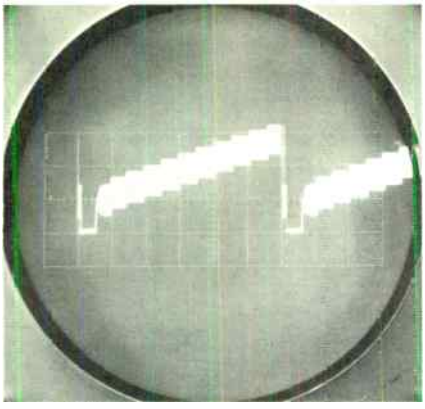


Figure 6. Stairstep test signal with superimposed 3.58-mc color subcarrier can be used to measure differential phase or gain. Differences in amplitude of modulation reveal differential gain. Differential phase measurements require synchronized phase detector.

that would give quantitative as well as qualitative evaluation of *all* transmission impairments. With the addition of differential gain and differential phase to the characteristics to be monitored, the problem of testing became even more difficult. Several test methods, however, have gained wide acceptance for measuring specific impairments, and some can be used for more than one of the parameters listed above.

Multiburst Signal

The *multiburst signal* is used to make a quick check of the amplitude-versus-frequency characteristics across the baseband. The signal consists of a series of "bursts" of equal-amplitude sine waves, each at a different frequency. In addition to the burst frequencies, the test signal includes a horizontal synchronizing pulse and a burst of peak white—the so-called "white flag"—to provide a white reference level. The complete signal is transmitted during one line interval. Typical burst frequencies are 0.5, 1.5, 2.0, 3.0, 3.6, and 4.2 mc. A transmitted multiburst signal appears in Figure 4A, and the received signal is shown in Figure 4B. A quick glance

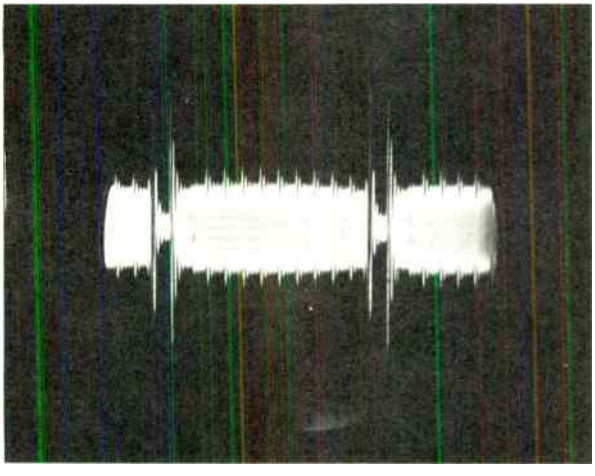


Figure 7. By passing stairstep with subcarrier through a high-pass filter, low-frequency steps are removed, leaving only subcarrier. Variations in amplitude reveal differential gain more clearly than in original unfiltered form (Figure 6).

World Radio History

at the oscilloscope ("A-Scope") waveform presentation of the received signal reveals a substantial decrease in gain with increasing frequency.

Obviously, the multiburst is not a complete check of the amplitude-frequency response. Dips and peaks occurring entirely between the burst frequencies may not show at all. This test is very useful, however, because it provides a spot check of the overall system response which can be evaluated visually in a few seconds.

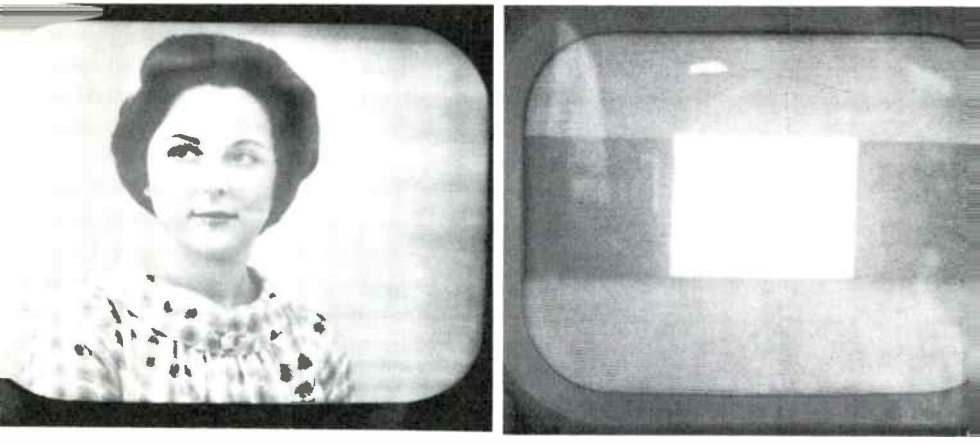


Figure 8. Streaking of picture is caused by poor low-frequency response. Note that relative duration of brightness values determines the amount of visible error in picture at left. At right is a typical window signal as it appears on the picture tube when the same amount of low frequency error is present.

Stairstep

The *stairstep signal* is so called because the A-Scope presentation resembles a staircase consisting of 10 steps extending from black level to white level. On the picture monitor this signal produces a series of vertical bands. Both presentations are shown in Figure 5. In the undistorted signal these steps are equally spaced. Thus, a visual check of the relative height of the steps after passage through the transmission system provides a quick and easy method for qualitatively evaluating system linearity.

A sine wave of 3.58 mc (the nominal color subcarrier frequency) impressed on the stairstep signal provides a method for measuring differential gain and differential phase. If this composite signal is passed through the transmission system and then through a high-pass filter, the low-frequency step components are eliminated and the 3.58-mc signal remains—distorted by any differential gain or differential phase which

World Radio History

may be present in the system. Any differential gain at the 3.58-mc subcarrier frequency shows up as amplitude variations in the horizontal presentation, as shown in Figure 7.

Differential phase can be measured by using a phase detector to compare the phase of the 3.58-mc signal impressed on a stairstep luminance signal with the phase of a reference signal of the same frequency. This reference signal may be provided in several ways. One method is to use a local free-running oscillator synchronized by the horizontal synchronizing pulse transmitted with the stairstep. The phase detector then measures the phase differences between this locally generated signal and the received signal at the various luminance levels. Any difference, measured in degrees, is the differential phase.

Window Signal

The *window signal* takes its name from its appearance on the picture monitor — a rectangular white area on a black background. The signal is generated as a line frequency (15.75 kc) square wave having a peak value equal to reference white. Since only half of each cycle is at reference white, the other half cycle being near reference black, only half the width of the screen is white. If the line frequency square wave is modulated by a 60-cycle square wave, the window signal will occupy only half the height of the picture raster. Most commercial window signal generators permit the resulting window to be adjustable in size and position.

The window signal is particularly useful in testing for low-frequency distortion. Phase distortion in the frequency range below about 200 kc produces "streaking"—one of the more objectionable forms of picture impairment. Streaking is the appearance of an incorrect luminance in the picture because of the waveform's inability to reach the correct value promptly. It usually spreads to the right from a point of sharp transition between light and dark or vice versa, as shown in Figure 8. In addition to being visible on the raster

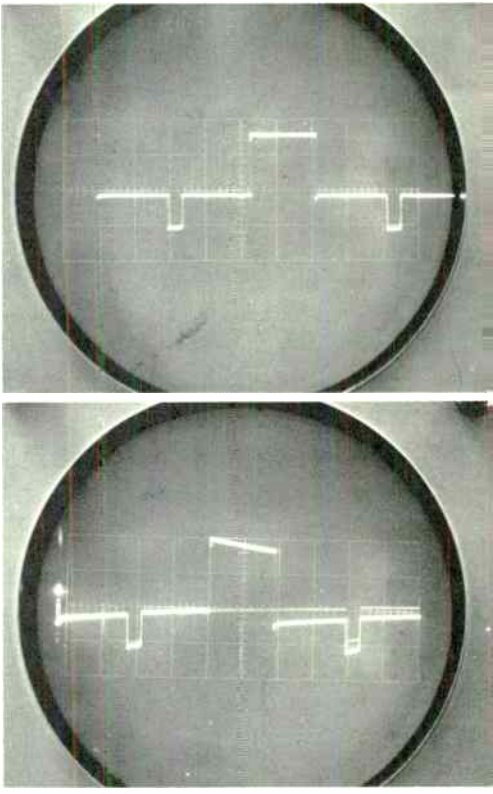


Figure 9. Oscilloscope presentation of square wave response in system with poor low-frequency response. Upper waveform is reasonably good (note tilt of sync pulse). Bottom picture illustrates very bad low-frequency response. Streaking would be severe.

World Radio History

presentation of the window, low frequency phase distortion can also be seen on the A-Scope, where it is indicated by a tilting of the square wave from the horizontal. As little as 2% tilt may be detectable in the picture, and 5% tilt indicates distortion which is readily noticeable. Figure 9B shows an extreme amount of "square-wave tilt."

The window signal is also used to test for "ringing"—a waveform "overshoot" or damped oscillation. Ringing is usually produced by sudden voltage transitions in a system with a sharp upper frequency cutoff, or by a transmission discontinuity below the cutoff frequency. The frequency of oscillation approximates the cutoff frequency or the frequency of the discontinuity. The "sharpness" of the discontinuity determines the duration of the ringing. As shown in Figure 10 each oscillation due to ringing shows as a light or dark band following the tonal transition which induced the ringing. By measuring the overshoot of square-wave transitions, it is possible to estimate the effect on picture quality and to determine the degree of correction required. Figure 11A shows the A-Scope presentation of the same degree of ringing shown in Figure 10, and Figure 11B shows the same effect on a sine-squared test pulse.

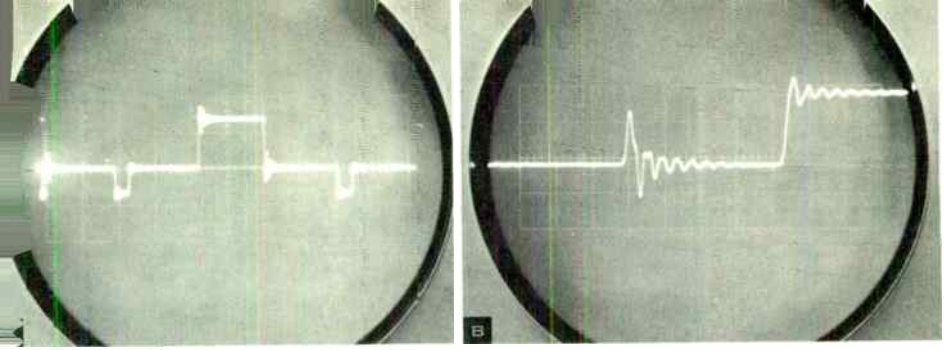
Figure 10. "Ringing" shows in picture as echoes displaced to the right of transitions between dark and light.

Transient Response Tests

Most of the tests described above are essentially "steady state" tests—that is, tests which employ sine wave signals or other signals which are inherently repetitive. Although these tests have significant value in evaluating the overall response of a television transmission system, they have definite limitations.

The most typical television signal is not necessarily repetitive at all, but may consist of a number of transients or instantaneous changes in amplitude. Such signals impose performance requirements on a transmission system which cannot be adequately simulated by sine wave substitutes. A television subject may consist of "optical transients" in which a small bright object may appear against a contrasting dark background. For perfect reproduction, the signal waveform should rise instantly to the value representing the

World Radio History



'gure 11. Same amount of ringing as in Figure 10 produces severe distortion window signal (A) and sine-squared test pulse (B). Note severe oscillations beginning and end of square wave. Sine-squared test is the more sensitive of e two, but is harder to evaluate qualitatively. Next month's article discusses this test in detail.

rightness of the small object, then just , quickly return to its former value as ıe scanning beam moves on.

In steady-state types of testing, the ɔw-frequency components of the wave- ɔrm that are necessarily present as a re- ılt of the long duration of the test 'gnal may obscure or modify the re- ɔonse of the system to transients. Since eady state signals are not necessarily ypıcal of those which the system may e called upon to transmit faithfully. hey are not fully adequate in evaluating erformance of the system. Furtner- nore, steady-state tests are less suitable or establishing *tolerances* on the distor- ion which may be allowed in a televi- ion system.

Because of these more or less inherent disadvantages of the "traditional" steady-state test methods, waveform or transient response tests have been developed, and these are becoming more widely accepted, particularly in England and other European countries. The most widely used transient response test is a combination of a "sine-squared" pulse and a modified square wave which together form the so-called *pulse and bar* test signal. This test method offers many advantages under some circumstances and has been recommended, along with some other tests, by the CCIR (International Consultative Committee for Radio). The sine-squared pulse and bar test will be discussed in the second article of this series, which will appear in next month's DEMODULATOR. •

BIBLIOGRAPHY

1. A. Ste-Marie, "Video Testing Techniques in Television Broadcasting," *Communications and Electronics*; March, 1958.
2. *Electrical Performance Standards for Television Relay Facilities*, Electronic Industries Association Standard RS-250; 11 West 42nd Street, New York 36, N.Y., October, 1961.
3. *Requirements for the Transmission of Monochrome Television Signals Over Long Distances*, C. M. T. T. (Annex 15/1) recommendation at the Xth Plenary Assembly of the C. C. I. R.; Geneva, 1963.
4. *Television Signal Analysis, Second Edition*. American Telephone and Telegraph Company; New York, 1963.

World Radio History

World Radio History

the Lenkurt Demodulator

VOL. 12, NO. 11 NOVEMBER, 1963

Performance Testing of Television Channels

Part Two

Last month's article in this two-part series discussed some of the more conventional test signals for evaluating video transmission systems. Many of these "steady-state" testing methods share certain disadvantages because they are based on sine waves or other repetitive signals. As transmission requirements have become more severe, particularly with the widespread acceptance of color television, more rigorous testing is in order. So-called "waveform" testing—a means of testing the system's ability to reproduce typical waveforms is generally required. This article discusses the most widely accepted waveform method, the sine-squared or "pulse and bar" test.

The chief disadvantage of sine-wave test signals is that they are not really representative of typical video signals, and thus may not fully reveal how a transmission system actually responds. Television transmission is waveform-dependent, which means that a true representation of the subject cannot be obtained from the magnitudes of the signal spectral components (energy distribution) alone, as in the case of speech or music. A typical television signal is more likely to consist of abrupt "steps" (corresponding to a sudden transition from one value of gray to another), or sharp impulses (light or dark spots against a contrasting background).

An appropriate test signal should include these typical elements in a way that will clearly reveal system performance, but not respond to system characteristics which do not affect the picture. In general, the test signal should:

1. Be representative of the commonly occurring parts of a television signal;
2. Have a spectrum confined to fre-

World Radio History

quencies of interest, so that distortion outside the band of interest is not indicated;

3. Have a simple "mathematical" shape, thus simplifying calculations;
4. Be easy to generate or reproduce accurately;
5. Have a simple shape that permits easy identification of distortion on an oscilloscope presentation;
6. Be sensitive to the kinds of distortion met in actual television picture transmission, thus allowing the detection of very small errors.

These requirements are well satisfied by a test method which employs a *sine-squared pulse*. This type of pulse simulates transient picture elements well. Unlike a pure sine wave, which has an infinitely narrow bandwidth, or a square wave, which ideally has an infinitely extended bandwidth, the sine-squared pulse has a bandwidth or spectral content which is quite restricted and easily controlled. This is important because unwanted frequency components outside the television band of interest would certainly be distorted and cause misleading test results.

The sine-squared pulse is so-named because its amplitude varies as the *square* of the sine of the phase of the signal: $A = \sin^2 \theta$, where θ is the phase angle. This is nearly as simple a waveform as the sine wave itself, which is defined as $A = \sin \theta$. Both waveforms are shown in Figure 1.

An idealized square wave contains a series of harmonics that extend indefinitely. Practical square waves, the type that are obtained in a test instrument, have very extensive harmonics, but they vary unpredictably with variations in circuits and test conditions. The sine-squared pulse, by contrast, has a limited, easy-to-reproduce spectrum controlled by pulse width. A "sharper" and narrower pulse contains higher frequencies than a broader pulse. Hence, the bandwidth of the transmission system to be tested determines the duration or width of the test pulse to be used.

Sine-Squared Spectrum

Normally, the nominal pulse width T is defined as its half-amplitude duration. When the sine-squared pulse has a duration of $\frac{1}{2f}$, energy content of the pulse is 6 db below peak value at

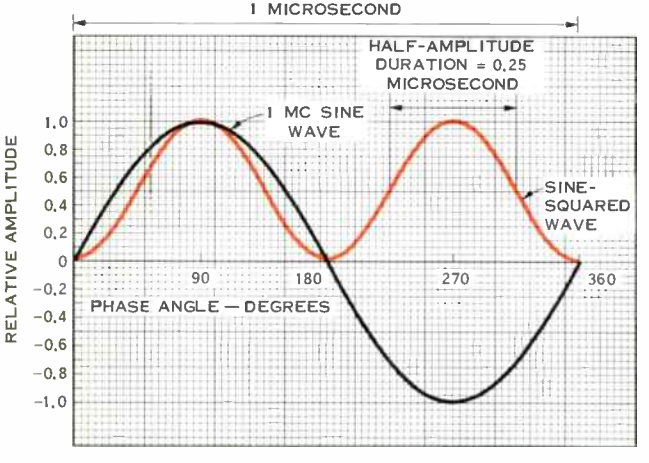


Figure 1. Sine-squared wave can be derived by numerically squaring a sine wave. The sine-squared wave shown contains no significant spectral components above 4 Mc.

World Radio History

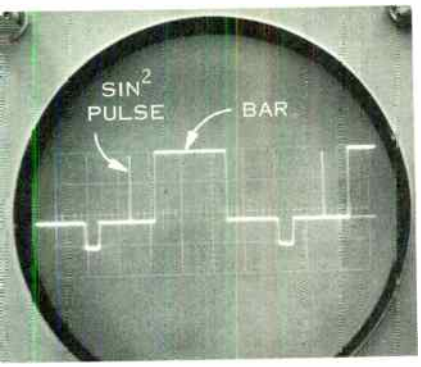


Figure 2. Pulse-and-bar test signal consists of sine-squared pulse and square wave with leading and trailing edges modified to sine-squared shape.

frequency f, zero at $2f$, and has no significant energy at higher frequencies.

Square waves provide the best test of low-frequency distortion, but are still more useful if the spectrum can be limited to the frequencies of interest. The square wave used in a window signal can be modified so that the leading and trailing edges have the same shape as the sides of the sine-squared pulse. If a sine-squared pulse is added to each line of a window signal which has been shaped in this way, a so-called *pulse-and-bar* test signal is obtained. This is illustrated in Figure 2.

The pulse and bar test permits sensitive performance evaluation across the entire frequency band. Because of the large amplitude of its low-frequency components, the bar gives the most sensitive indication of distortion at the lower frequencies—up to several hundred kc — just as does the traditional window signal. Unlike the window signal, however, the modified bar contains no significant out-of-band frequency components to produce spurious distortion indications.

The upper regions of the frequency band are tested by selecting a sine-squared pulse from either of two widths. The narrower pulse has a half-amplitude width of T, where T is the reciprocal of twice the upper frequency limit of the transmission system. For a 4-mc system, T is 0.125 microsecond. An analysis of its power spectrum shows the power to be 6 db below peak value at 4 mc and zero at 8 mc. Therefore, this pulse is particularly valuable for the upper frequencies, especially since considerable phase shift may occur near the upper cutoff frequency.

For frequencies between 0.5 mc and about 2 mc, a $2T$ pulse of 0.25 microsecond half-amplitude duration is often used. It contains no significant energy at frequencies above 4 mc. It is perhaps the most used of the three test signals because it is particularly suitable for use in routine adjustments where a detailed evaluation is not required.

Phase Sensitivity

Because the sine-squared pulse provides a dynamic test covering a broad frequency band, its usefulness is not limited to a particular type of distortion or to one frequency range. For example, the sine-squared pulse is a very good indicator of phase distortion. The sensitivity occurs because the pulse is highly symmetrical about its vertical axis; any transmission phase distortion skews the waveform and makes it unsymmetrical in a way that is immediately evident. If the high-frequency delay is greater than the low-frequency delay a ripple appears following the pulse, while greater low-frequency delay produces a ripple preceding the pulse.

In a low-pass transmission system, the *area* under the pulse remains constant because the area represents the dc

World Radio History

component of the pulse. If the amplitude of the pulse is decreased by restricting bandwidth with a slow-cutoff filter, the pulse becomes wider to maintain the same area, thus losing some fine detail. If the bandwidth is restricted by a *sharp*-cutoff filter, overshoot or ringing occurs and is indicated by a damped oscillation following the pulse. Overshoot adds negative area so the pulse height automatically increases to maintain the same area. Pulse height is significant because it represents the brightness of fine details in the picture. In general, because detail is retained, a relatively sharp frequency cutoff is more desirable unless the ringing becomes objectionable.

Although ringing is produced by a sharp-cutoff filter, it may also be caused by a frequency response discontinuity elsewhere in the band. Ringing caused by a dip in the response of the system is evidenced by damped oscillations following sudden transitions in the pulse and the bar. The frequency of this oscillation is the same as the frequency at which the response dip occurs, while the amplitude and damping of the oscillation are controlled by the width of the dip.

Echoes are perhaps the most easily recognized type of distortion— the "ghosts" they produce on the television screen are familiar to most viewers. An echo occurs when a signal reaches its destination via two paths of different electrical lengths. In broadcast television, this usually happens when the main signal arrives directly and a portion of the signal is received after reflection from a mountain, building, or other object. Echoes may also be caused by electrical discontinuities such as impedance mismatches in the transmission system. It has been shown that all forms of distortion can be represented by characteristic patterns of echoes in the received signal. Sine-squared test signals show these effects particularly clearly. By correlating the echo effects imposed on a test signal with subjective judgements of the picture degradation caused by the same distortion, it has been possible to create a quantitative rating system for all types of distortion.

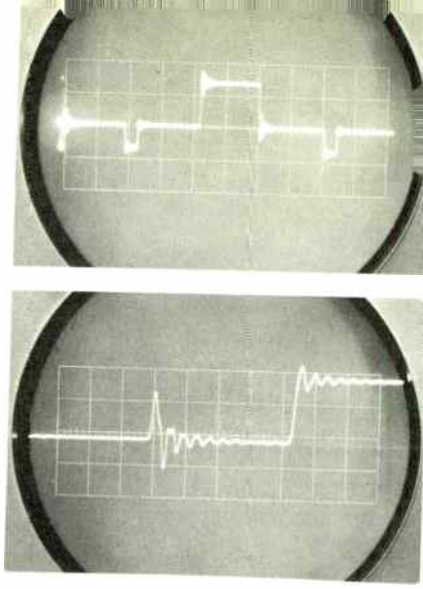


Figure 3. Identical ringing as indicated by square wave (top) and pulse-and-bar signal. Pulse and bar provides a much more sensitive indication of this distortion.

Sine-Squared Rating System

Test methods based on waveform distortion were pioneered primarily in Europe, particularly by the British Post Office Department and the BBC. Early European investigators were quick to recognize the shortcomings of steady-state testing. For example, a broad dip of 2 db in the amplitude-versus-frequency response may cause much more waveform (and hence picture) distortion than a "sharp" dip of 6 db;

World Radio History

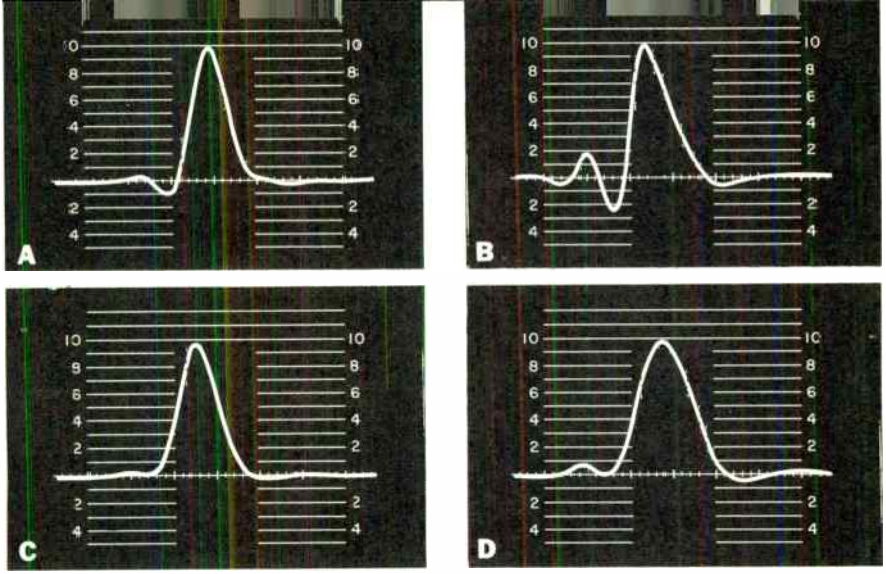


Figure 4. Response of local and long-distance television links to T *and* 2T *sine-squared pulses: (A) local link,* T *pulse; (B) long link,* T *pulse; (C) same local link,* 2T *pulse; (D) same long link,* 2T *pulse. Lack of symmetry about center of pulse indicates phase distortion. Greater effect on* T *pulse indicates distortion at higher frequencies.*

but steady-state tests may indicate that the "slight" 2-db dip is relatively unimportant.

Since the final test of transmission performance is subjective — the viewer's judgment of picture quality — the investigators used the reactions of a number of critical observers to establish performance limits. The tests revealed that the pulse-and-bar signal was in fact a sensitive indicator of the types of distortion which are most noticeable on the raster. But more important, the degree of distortion of the test signal gave a good indication of the amount of picture degradation. This allows the system performance to be specified in terms of a single factor — often called the *K* factor. The *K* factor is a quantitative measure of the distortion suffered by the pulse and bar.

Two test methods are generally used, one for *routine* testing and one for system *acceptance* testing. Both use the pulse-and-bar test signal and the *K* rating factor; the difference is in the method of interpretation. As the name implies, the routine-test method is used for everyday tests and adjustments where complete accuracy is not as important as speed and simplicity. By contrast, the acceptance-test method is a precision test used mainly to evaluate new systems or systems that have been modified

Routine Testing

For routine testing, an oscilloscope graticule marked as shown in Figure 6 is used to determine the permissible distortion, usually of the bar and the *2T* pulse. Essentially, the test amounts to a visual inspection to see whether the received test signal fits into the

World Radio History

limits engraved on the graticule. These graticules often show two limits. For example, a graticule now being marketed by a United States manufacturer indicates the limits for both $K = 2\%$ and $K = 4\%$. These limits are established by subjectively comparing distorted pictures with a picture impaired by a single undistorted echo delayed by more than $1\frac{1}{3}$ microseconds. (This arbitrary echo delay assures that the echo is not masked by the main signal. Echoes more widely separated tend to increase in annoyance value until a separation of about 10 microseconds is reached). Then, if the echo has an amplitude of 2% of the original pulse, the rating factor K is 2%. Any other type of distortion producing the same amount of picture impairment would also have a rating factor of 2%.

Other rating methods such as the "echo rating technique" use empirical bandwidth and frequency weighting curves to achieve a similar evaluation based on echo simulation. Thus, the use of the echo as the standard of comparison permits the placing of quantitative limits on all types of distortion.

Precision Testing

The acceptance-test method uses the same rating factor, but achieves more precise evaluation by mathematical analysis. A microscope is used to sample photographs of the transmitted T pulse and the received pulse waveforms at short intervals—equivalent to the waveform sampling in a time-division multiplex system. The series of samples forms a "time series" which can be used to describe the waveform mathematically. If the time series for the received pulse is divided by the time series obtained from the original transmitted pulse, the result is a "filtered" time series which is free from the imperfections of the test equipment. In a distortionless system, the mathematical expression for this filtered series would contain only one term. Therefore, any additional terms represent distortion. These distortion terms represent echoes of the undistorted pulse displaced in time. They appear both before and after the main pulse, but not necessarily in matched pairs because they do not represent attenuation or phase distortion separately.

As the name implies, the acceptance-test method provides a means for specifying the required performance of a transmission system and for ensuring that the system meets the specifications. By performing various fairly elaborate computations on the mathematical time

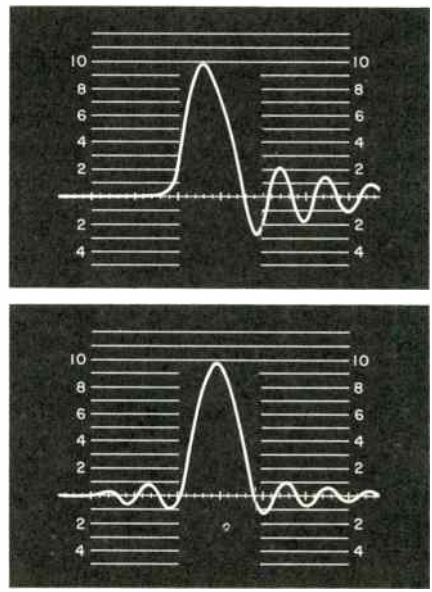


Figure 5. Ringing caused by low-pass filter follows pulse. Phase distortion is indicated by lack of symmetry. Partial phase equalization transfers some of the ringing to the other side of the pulse, improving symmetry.

World Radio History

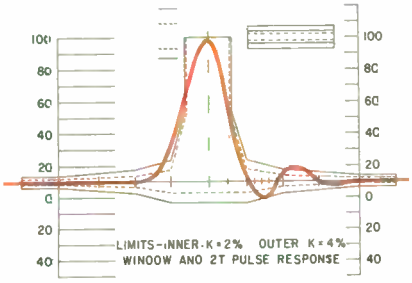


Figure 6. Graticule with limits for 2 percent and 4 percent rating factors. Response shown indicates that system ringing exceeds a K factor of 4 percent.

series for the received waveform it is possible to determine system performance in great detail, even to distinguishing the nature and magnitude of the distortion present. One of the special advantages of this technique is that it permits the effects of connecting transmission links in tandem to be calculated. If the time series of the individual links are multiplied together, the time series for the tandem connection is obtained and the overall rating factor can be calculated. In some cases the rating factors of the links add directly, but this is not true for random distortion (such as that due to component tolerance).

Conclusions

Although progress has been made in establishing quantitative limits on the distortion as revealed by various test signals, particularly with waveform testing, the fact remains that most transmission evaluation is done by visual inspection of a picture monitor and oscilloscope presentations of the waveform. This suggests the possibility of classifying *all* test methods in two general categories: those used for routine or maintenance testing, and those used for system specifications and for acceptance testing. Thus, most of the widely used steady-state tests such as multiburst, stairstep and the like would go into the maintenance-test category, along with the routine-test version of the sine-squared pulse and bar, without rendering present test equipment and methods obsolete. The steady-state tests are generally quite adequate for maintenance testing, although the pulse and bar method is gradually being accepted as a supplemental test.

The area where waveform testing is strongest is in the specification of performance requirements in terms of *numbers*. The acceptance-test permits the performance standards to be specified for a single link or for a complete transmission system; when the system is installed, the method provides assurance that it actually meets the specifications or reveals how it fails. Although this test method may not be accepted "overnight," it is showing signs of wider acceptance in the United States to match its growing use in Europe in recent years. •

BIBLIOGRAPHY

1. N. W. Lewis, "Waveform Responses of Television Links," *Proceedings of the Institution of Electrical Engineers* (British); July, 1954.
2. I. F. Macdiarmid, "Waveform Distortion in Television Links," *Post Office Electrical Engineers' Journal*; July-October, 1959.
3. R. Kennedy, "Sine-Squared Pulses in Television System Analysis," *RCA Review*; June, 1960.
4. H. Schmid, "A Graticule to Measure the Waveform Performance of TV Facilities," *IEEE Transactions on Broadcasting*; February, 1963.
5. K. H. Potts, "The Performance Evaluation of Television Line and Microwave Links Using Sine-Squared Test Techniques," *The Radio and Electronic Engineer* (formerly *Journal of the British IRE*); April, 1963.

World Radio History

World Radio History

the Lenkurt Demodulator

VOL. 11, NO. 7 JULY, 1962

Take The Mystery Out Of Microwave Literature!

Microwave communication is booming; new applications and better equipment are helping a phenomenal growth that promises to become even greater. Along with the many new users of microwave, new manufacturers are appearing on the scene. The result is a Babel of technical literature and promotional material which speak in many "languages." As many as four or five different terms for the same effect may be used in various publications. This article summarizes some basic considerations of microwave radio, and relates some of the words used to describe them.

What are the basic factors which determine microwave quality and distinguish one system from another? Is it possible to come up with a "figure of merit" which can be used to evaluate microwave equipment? Although the microwave art is too complex for any single scale or figure of merit, it is possible to "boil down" and combine some of the many diverse factors used in technical literature to define equipment performance.

Five basic factors can be used to evaluate a microwave system:

Performance Quality. In essence, this is freedom from the noise and distortion which obscures the signal or tends to create transmission errors.

Load Handling Ability. The information capacity of the system—number of

World Radio History

channels, data transmission rate, usable bandwidth, and the like. Capacity is dependent on the performance standards required, and the linearity of the system components.

System Length. Since noise and distortion are cumulative, this refers to the number of repeater sections or "hops" that can be used in tandem before performance becomes unacceptable. System length generally must be decreased with heavy loads (large number of channels).

Reliability. Only freedom from equipment failure is intended here, since protection against fading is largely a function of system or path engineering and the use of such techniques as space or frequency diversity.

Economy. This is the factor against which the other four are measured.

In general, these qualities tend to be incompatible with each other, so that one or more can be improved only at the expense of others. Thus, if load handling ability or noise performance is improved, economy will probably be reduced.

Importance of Noise

In microwave, as in other forms of electrical communication, noise is the principal enemy. Noise obscures the signal and causes transmission errors. Although noise is constantly introduced into the communications channel from the transmission medium and the equipment itself, this can be overcome by suitable design. Actually, the amount of noise present is not as important as the relative strengths of the signal and the noise; the greater the signal-to-noise ratio, the better and clearer the transmission. Accordingly, the level or amount of noise present in the receiver —regardless of its source—determines the signal threshold or minimum signal that can be received, as well as the signal level required for good transmission quality.

Noise Sources

Two basic types of noise exist within a microwave system: *idle noise* and *intermodulation noise.* Idle noise, which is always present despite the absence of modulation, consists of thermal noise generated within mixer diodes or low level amplifiers, shot noise from klystrons, or the noise often generated by semiconductor multiplier chains used in some receivers for the local oscillator.

Intermodulation noise is introduced into the system as a result of heavy signal load or increased operating level. The greater the traffic load, or the higher the operating level, the more intermodulation noise that is introduced. Usually intermodulation increases relatively slowly until a "break point" is reached, after which it increases very rapidly. However, it is desirable to operate the system at as high a level as possible (but short of the break point) in order to improve the signal-to-noise ratio.

In an FM microwave system, higher operating levels cause greater frequency deviation, which is very effective in overcoming some of the idle noise. However, only a limited amount of deviation is possible before non-linearities in the equipment increase intermodulation noise.

If the fixed amount of permissible deviation is shared by only a few channels, the signal-to-noise ratio in each channel will be quite good. However, as the number of channels is increased, inter-

World Radio History

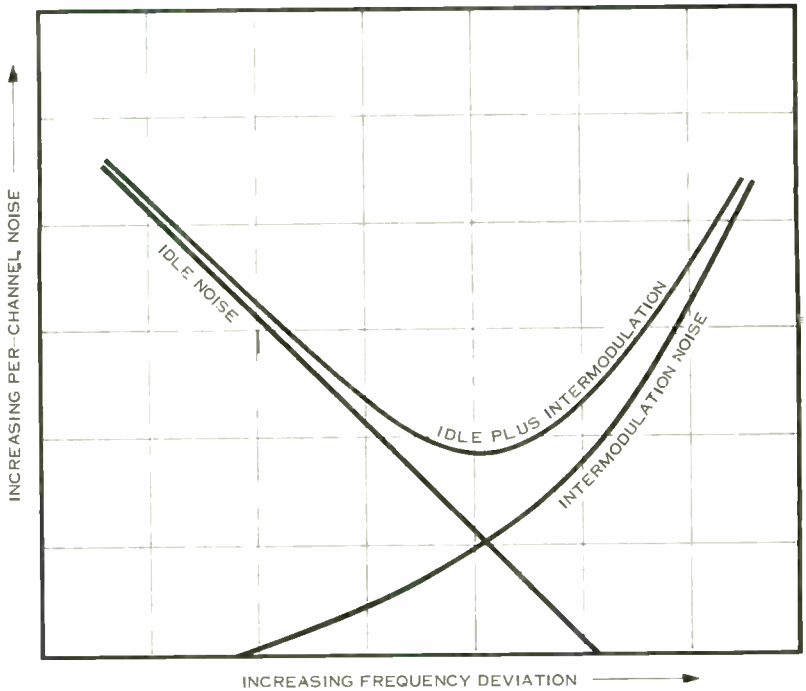


Figure 1. "Front end" idle noise is reduced in direct proportion to increase in frequency deviation due to signal level or system load. However, beyond "break point" of equipment, increasing intermodulation noise rapidly overcomes and reverses this advantage. The level at which intermodulation noise becomes dominant varies from equipment to equipment.

modulation noise limitations demand that the *per-channel* frequency deviation be reduced, with a consequent increase in idle noise. If system linearity can be improved, greater deviation can be used, thus restoring quality. Such improvements may be costly, thus placing an economic limit on the number of channels that can be handled for a given transmission quality.

How to Rate Equipment

Although the basic concept of overcoming noise to improve transmission quality is simple, many individual factors enter the problem in practical equipment. Many of these factors are sometimes used to define equipment performance, even though individually, they describe performance incompletely or even improperly. In some cases, this may originate with the design engineer, who must design "pieces and parts" of a system to standards which will result in the desired overall performance. Although these factors are meaningful to the designer, and may provide interesting comparisons, some individual fac-

World Radio History

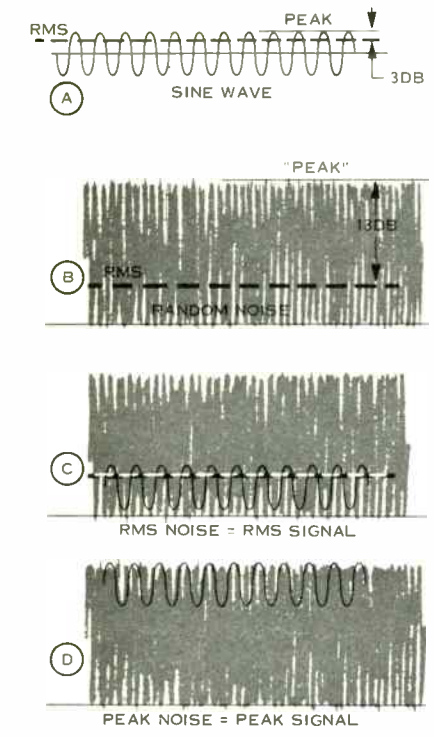


Figure 2. Effective or RMS power of sine wave is just 3 db less than peak power (A). Peaks of random noise are roughly 13 db above RMS power (B). Noise or "AM" threshold (C) is that level at which RMS signal power equals RMS noise power. FM improvement or "practical" threshold (D) occurs when signal peaks equal or exceed noise peaks, approximately 10 db above noise threshold.

tors may not provide adequate information about actual performance of the overall system. When diverse terms are used to define a single characteristic, the confusion is made even greater.

Some of the characteristics often used to describe the performance of microwave equipment are:

Receiver Noise Figure. This expresses the actual contribution of thermal noise in the "front end" of the receiver, compared to an ideal receiver. The figure is usually expressed in decibels; a receiver having a noise figure of 3 db is just twice as noisy as the equivalent ideal receiver. This is a relative expression, and by itself does not indicate receiver sensitivity. When the effective bandwidth is known, the absolute threshold of the receiver can be calculated. Thus, two receivers having identical noise figures, but different bandwidths will have different thresholds.

Noise Threshold. This is the RF input level at which signal power just equals the internally-generated front end noise power. It is determined by the bandwidth of the receiver and its noise figure according to the relationship

$$power\ (dbw) = noise\ figure\ (db) + 10 \log kTB,$$

where k is a constant (1.37×10^{-23}), T is effective antenna temperature in °Kelvin (the value 290° K is standard in current practice), and B is bandwidth in cycles per second. To convert dbw to the more conventional dbm, add +30 to the numerical value of dbw. Note: the effective bandwidth of the receiver is usually the bandwidth of the intermediate frequency amplifier (IF bandwidth), *not* receiver preselector bandwidth or the bandwidth of the transmitter waveguide filter.

AM Threshold. Synonymous with noise threshold. This term is occasionally used (even though the radio system employs FM), because there is no FM

World Radio History

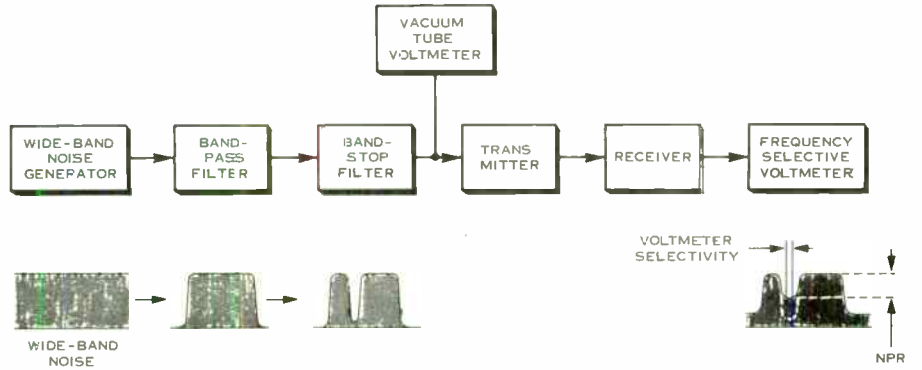


Figure 3. Typical arrangement for measuring intermodulation distortion in terms of noise power ratio. Random noise of the same effective power and bandwidth as baseband signal (but with narrow "slot" removed) is applied to transmitter. Added noise in slot is measured at output of receiver. Note that all noise contributions by both transmitter and receiver are included in a single figure; practice of quoting separate figures for transmitter and receiver is misleading.

"noise improvement" at and near threshold, and the relative *peak amplitudes* of the background noise and the signal are controlling.

Tangential Threshold. Also synonymous with noise threshold. The term stems from radar system usage and refers to a means of defining the minimum radar pulse amplitude that can be detected above the background noise.

FM Improvement Threshold. At the noise threshold level, the RMS or effective power of the signal is just equal to the RMS noise power. However, the *peak* value achieved by the noise impulses is approximately 13 db higher, as contrasted with the 3-db difference between peak and RMS values of the RF carrier. This is significant because of the unique characteristic of FM radio that signals with greater peak values dominate or "capture" the receiver, and literally suppress signals having somewhat lower peak values. Accordingly, at the *noise* threshold, noise is dominant and performance is much worse than it would be in an amplitude-modulated system at this same level.

When the signal level is increased about 10 db so that RF carrier peaks equal noise peaks, the so-called **FM Improvement Threshold** is reached. Above this level, the RF carrier peaks dominate, and noise is suppressed effectively. As the signal level increases, noise in the output is literally reduced — not just masked.

Typically, a baseband signal-to-noise ratio of about 30 db is obtained at the FM improvement threshold, and this improves about one db for each db increase of signal level above the improvement threshold. Normally a system is designed so that there is only a very small chance of the signal level dropping below the improvement threshold, despite propagation fading.

Practical Threshold. This is a widely

World Radio History

used synonym for FM improvement threshold.

System Gain. Since the signal-to-noise ratio or performance quality of individual channels depends, in part, on the strength of the received microwave signal, it is important to design the overall system so that even with fading, the received signal will not drop below a certain minimum value. Since transmitter power and the receiver threshold have fixed values, only the distance between antennas, and the size of the antennas and reflectors used, are left as variables which can affect receiver input level. The attenuation of the signal as it travels through space is proportional to distance, and can be indicated in decibels of loss. Similarly, the focusing and concentrating effect of the antennas can be shown as decibels of gain. Accordingly, these two variables can be selected to just match the difference in power between the transmitter output and the signal level required by the receiver. This range of power is known as *system gain* or *equipment gain*. For example, if the transmitter has an output of one watt (+30 dbm) and the receiver has a practical threshold of −81 dbm, system gain is 111 db. Assuming that a reserve of 40 db loss is required as protection against fading, the length of the transmission path and the sizes of the antennas should be selected to introduce no more than 71 db net loss.

White Noise Loading

As indicated above, such details as receiver noise figure and thresholds do not provide adequate information about the overall performance of a microwave system. Even when the received signal is strong and satisfactory, intermodulation may become controlling, due to system load, and thus degrade performance.

One of the most definitive tests of the overall performance of a microwave system is the "white noise loading" test. (For a detailed discussion of noise loading, see DEMODULATOR, *December,* 1960.) Essentially, it requires that a band of "white" or random noise of the same frequency range and power level that simulates a multichannel signal, be applied to the transmitter. A relatively narrow portion of the noise signal (10% of the baseband or less) is blocked by a "slot" filter before the noise signal is applied to the transmitter. At the output of the receiver, the noise which has "spilled" into the slot because of intermodulation distortion is measured and compared with the noise level outside of the slot, as diagrammed in Figure 3. The ratio of noise powers is called the *Noise Power Ratio,* and provides a good indication of system performance, since all aspects of equipment performance are taken into account.

Although the Noise Power Ratio provides a good relative indication of system performance, it is more customary to rate equipment in terms of the quality of a communications channel, usually in dba. This allows the radio channels to be compared with any other kind of channel, such as those transmitted over wire, cable, or any other medium.

Per-channel noise (expressed as either signal-to-noise ratio or dba, since dba = 82 − S/N) can be derived from the NPR by relating the noise in the slot to a reference level and then applying a weighting factor. (For a detailed

World Radio History

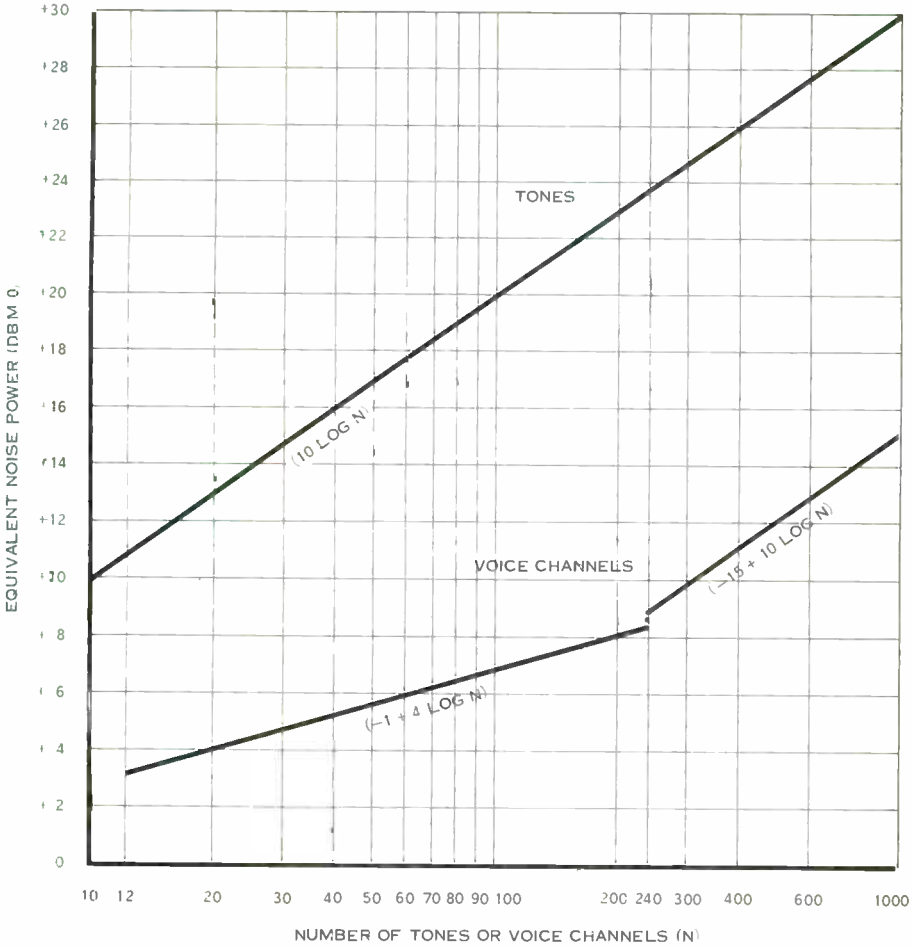


Figure 4. Random noise power recommended by C.C.I.R. for simulating the load presented by voice channels or tone signals. The two slopes representing voice channels approximate the way in which many individual channels average their power. These values also allow for the presence of some signaling and data tones.

discussion, see DEMODULATOR, *May*, 1961.) However, most commercially available noise measuring test equipment is calibrated to express noise in terms of picowatts or signal-to-noise ratio.

Note that in measuring intermodulation distortion, both the transmitter and receiver participate in the test, and the resulting noise power ratio pertains to both operating together. It is not proper to assign a noise power ratio (or NPR)

World Radio History

figure separately to receiver and transmitter, even though the number for each then appears to be 3 db better (higher) than the NPR for the transmitter-receiver combination. For instance, if a manufacturer indicates that his transmitter has an NPR of 50 db, and the receiver also has a 50-db NPR, the actual value for the system is 47 db, twice as noisy as suggested by the "split" figures.

System Channel Capacity

Like many other fields of activity, microwave design tends to reflect the needs of the moment, or at least those qualities that appear to be desirable. With the dramatic growth of microwave, and its many new uses, operators of microwave systems are tending to look ahead and plan their systems to meet future needs. Consequently, most new microwave systems have much higher capacity than systems produced only a few years ago. For instance, in 1956, most light- and medium-capacity systems were designed to accommodate 120 channels. By 1958-59, systems with a capability of 240 channels were being promoted. Within the last year or two, the magic words have become "600 channels."

Actually, these arbitrary numbers have little meaning unless they are expressed in terms of noise performance under given conditions. Thus, a system capable of just meeting certain noise standards when transmitting 120 channels over six "hops" (five repeaters), may be quite capable of handling 400 channels over only a single hop. With 400 channels, an additional repeater might raise the noise to an unacceptable level. Conversely, the same equipment might provide very acceptable noise performance for many *additional* repeater sections if the system were loaded with fewer than 120 channels.

This trade-off does not necessarily apply equally to all systems, however. Equipment with very little idle noise, but which is poor in intermodulation distortion, might be very suitable for transmitting only a few channels (or the equivalent) for great distances. Another system, with somewhat greater idle noise but much better intermodulation characteristics, while performing relatively poorly with a light load, might be outstanding when loaded with large numbers of channels.

Accordingly, a microwave system should be evaluated under the full load for which it was designed, or that it might be called on to carry. Even better is a family of curves taken at various load levels, such as shown in Figure 6 for the Lenkurt Type 76A microwave system. Note that although the 76A is called a "600 channel" system, this particular equipment produced less than 20 dba0 in the worst channel when loaded for 960 channels. This would go up to 23 dba0 for two hops, and 26 dba0 for four.

Bandwidth. Every signal, whether it be one or more voice channels, a television signal, or a train of data pulses, occupies a certain finite bandwidth. The greater the information content of the signal, the greater the bandwidth that is required to accommodate it. Since bandwidth can place a limit on the amount of information transmitted (number of channels, for instance), the mistaken assumption has spread that bandwidth alone determines the capacity of the system. This is not so.

World Radio History

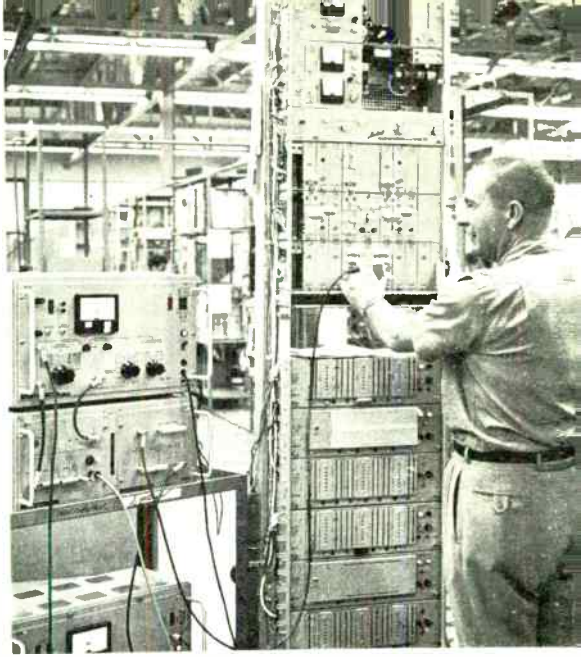


Figure 5. Production line testing of microwave equipment for noise performance. Two terminals are connected "back-to-back" through microwave attenuator. Measurements of intermodulation plus idle noise are made at both ends and middle of the baseband, using equipment arrangement shown in Figure 3.

Although bandwidth must be adequate, it is only one of several factors which determine capacity. Other important factors are the ability of the modulators and demodulators to operate over the increased bandwidth without distortion, and the linearity of baseband and modulating amplifiers.

Actually, bandwidth in excess of the load requirements is detrimental, since "front end" thermal noise increases in direct proportion to bandwidth, as indicated in the relationship shown on page 4. In the past, many types of microwave equipment have been manufactured which employed greater bandwidth than necessary for the number of channels that could be handled. This helped reduce phase distortion of the signal and also permitted the equipment to be used for diverse applications like television studio-transmitter links, thereby increasing the market.

The penalty paid for this versatility is a degradation of the equipment noise threshold. Although this does not affect performance when transmission is good and the signal is strong, it makes the system more vulnerable to fading, unless additional signal strength is obtained by larger antennas or shorter spacing between repeaters.

Transmitter Bandwidth. Strictly speaking, there is no arbitrary limit on the bandwidth of the transmitter output. By increasing frequency deviation, the bandwidth of the transmitted signal can be made as great as desired. Unfortunately, the non-linearity of the klystron or modulator will increase distortion rapidly as deviation becomes greater. The transmit waveguide filter used in all systems "cleans up" the radiated signal, and prevents high-order sidebands or other undesired modulation products from interfering with other services on adjacent frequencies. The bandwidth of this filter,

World Radio History

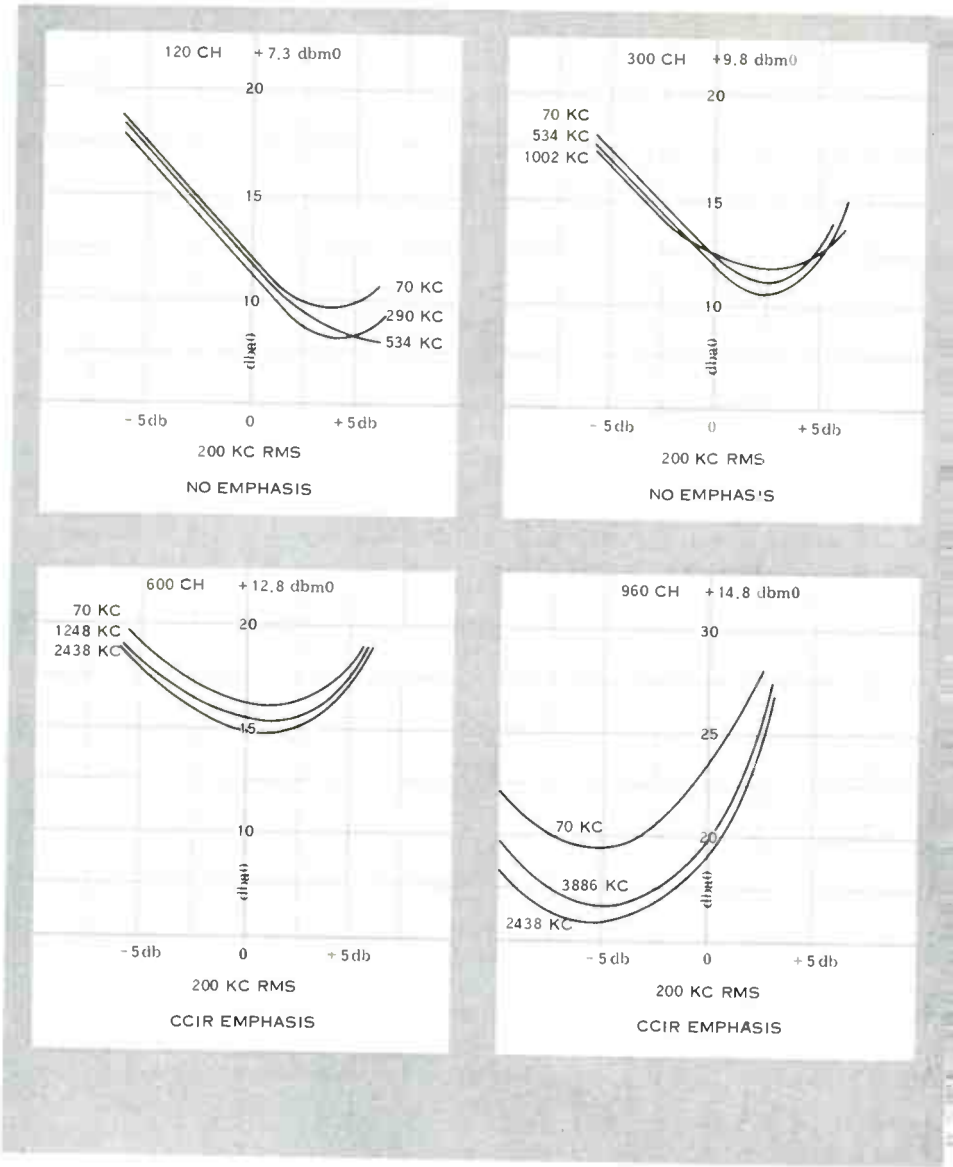


Figure 6. Family of curves showing noise performance of Lenkurt 76A microwave equipment under various load and deviation conditions. Curves for 120 and 300 channel loading were measured without emphasis, those for 600 and 960 channels with C.C.I.R. message circuit emphasis. The same broad IF passband filter was used in all cases. Note that optimum noise performance is obtained by increasing deviation for light loads, decreasing it for 960 channels. Nominal 200 kc-per-channel deviation is just right for 600-channel load, which represents the design objective of the equipment. Such performance curves vary slightly from equipment to equipment due to variations in components, equipment linearity and line-up.

World Radio History

however, does not determine the effective bandwidth of the system.

Receiver Preselector Bandwidth. Most microwave equipment uses a very selective bandpass filter at the receiver input, but before the mixer or first detector, to reject signals from adjacent bands. In addition, energy from noise and other interference is reduced. These filters, like those in the transmit waveguide consist of a series of tuned waveguide cavities. Generally, the more cavities used, the greater the selectivity of the filter. Since the preselector filter pass-band is always considerably greater than the effective bandwidth of the receiver, this may cause confusion when it is stated in equipment specifications. The principal value of stating preselector bandwidth is in estimating the minimum frequency spacing between adjacent systems, and the likelihood of receiving spurious signals due to local oscillator "images." It is of little use in determining receiver threshold, except that the threshold is raised by the amount of loss introduced by the preselector filter.

IF Bandwidth. The IF or intermediate frequency amplifier or its associated filters provides the ultimate selectivity for the receiver, and therefore, the entire system. It is in the IF amplifier that the signal undergoes the greatest amplification, and it is important that all spurious signals and interference be sharply rejected. Normally, IF bandwidth is measured at the "3 db points", or those points on the frequency response characteristic where the signal is attenuated 3 db. This is also known as the "half power" point. Occasionally IF bandwidth may also be given at the 0.1 db points, to indicate the full IF bandwidth that can be used with a minimum of phase or envelope delay distortion. Like other technical features, this is relatively academic, since baseband frequency response is hardly affected by the response of a properly designed IF amplifier. Although variations in IF amplifier response may introduce some phase distortion, this is normally eliminated by the phase equalizer present in most modern broad-band microwave equipment.

More Questions?

In preparing this article, based on typical inquiries received from time to time, it became obvious that all the semantic distortions and ambiguous expressions often found in microwave literature could not be discussed adequately in this limited space. Accordingly, another article on this subject is planned for the near future.

Rather than depend on our own feelings and ideas about which items to include, we invite you, the readers, to help shape the article. Your comments and questions about confusing terminology will be welcome; questions and suggestions will be acknowledged or answered individually, and those most typical will be discussed in the DEMODULATOR. •

World Radio History

World Radio History

the Lenkurt Demodulator

VOL. 13 NO. 5 MAY, 1964

NOISE PERFORMANCE in Industrial Microwave Systems

Part One

The performance of a communications system should be evaluated by how well it meets the requirements of the user. This sounds simple, but often there is confusion about how to establish performance criteria, and then about how to measure the actual performance. This article discusses noise performance of FM microwave systems using single-sideband, suppressed-carrier multiplex equipment. Sources of noise in such systems are discussed in terms of their effect on signal-to-noise ratios in the derived voice channels. Methods of calculating and measuring noise are considered and specific noise-performance recommendations are made.

Fundamentally, the purpose of a communications system is to transfer some form of intelligence, or "signal," from one point to another. An ideal system would deliver at the receiving end a signal identical in every detail to the signal applied at the transmitting end — with nothing altered and nothing added.

In a real communications system, this ideal performance is never completely achieved. In such a system every characteristic of the signal is altered to some degree, and there is always something

EDITOR'S NOTE

This is the first part of a two-part article written by Robert F. White, Lenkurt Transmission Engineer, in an effort to establish some guidelines for calculating and measuring noise in industrial microwave systems. Because of the exceptional clarity of the discussion, and because most of the article applies to all classes of microwave users, it is being reprinted here for the benefit of all DEMODULATOR *readers. The second part will appear next month.*

World Radio History

added along the way. Thus, the received signal is always a somewhat less than faithful reproduction of the signal applied at the transmitting end, plus some other elements which are mostly unrelated to the original signal and which may be present even when the signal is completely absent.

Performance of a communications system is measured by how closely the received signal resembles the transmitted signal and by how free it is of these other elements. The definition and measurement of the performance thus falls into two natural categories. In the first category there would be considered technical characteristics which define accuracy or fidelity of the reproduced signal: amplitude-frequency response, level stability, phase response, delay distortion, etc. These characteristics are, more or less, under the control of the equipment designer and may be held to almost any desired value.

In the second category there would be considered all the extraneous elements appearing at the channel output which were not a part of the input signal. It is these elements, usually lumped together in a single category called "noise," with which this article deals. The discussion is in terms of the noise as it appears in the derived voice channels. There are good reasons for taking noise in a voice channel as a criterion, even though present day systems usually carry telegraph and data as well as voice. The basic voice channel is familiar to all, is reasonably well standardized, and there is a large body of experience to draw on. Furthermore, the majority of equipments used for modern telegraph and data service are designed to operate over such a carrier-derived voice channel, or some fraction or multiple of it, and it is not difficult to evaluate the effect on a data system of a particular level of noise in the 3-kc band.

Noise Sources

The noise which appears in a voice channel of a microwave system comes from a number of different sources, some of which vary in a rather complex manner. It is useful to consider three general types of noise, classified in accordance with how they vary.

One is the thermal noise generated in the antenna and in the "front-end" circuits of the receiver: this noise in an FM system varies in inverse relation to the strength of the RF level at the receiver input, and is therefore affected by fading. It is not affected by system loading.

A second type of noise, also thermal in nature, is that developed in the electronic circuitry of the transmitter and in certain portions of the receiver. This type of noise, often called "idle" or "intrinsic" noise, is not affected by the RF input level, nor is it affected by system loading.

The third type of noise consists of spurious signals created by intermodulation between the various frequency components of the total composite signal. Such intermodulation is produced by every non-linearity through which the signal passes. The spurious products include the sums and differences of every frequency and its harmonics present in the modulating signal and all of the other frequencies and their harmonics.

Since the baseband spectrum of a multichannel system is extraordinarily complex, the number of intermodulation products produced in such a system ap-

World Radio History

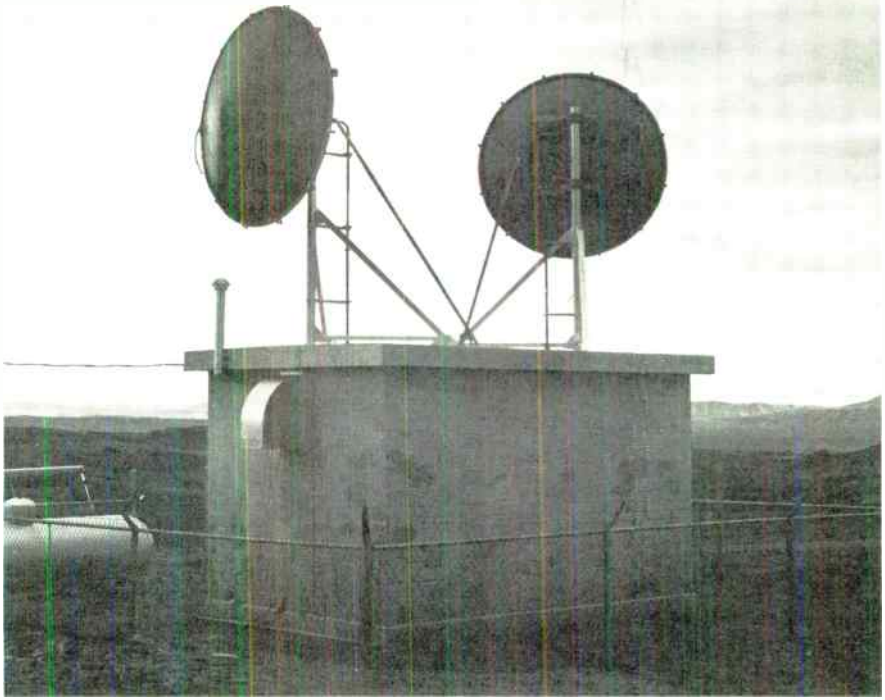


Figure 1. Thermal noise in the receiver adds to the earth noise temperature seen by the antenna.

proaches infinity. Statistically this noise becomes very similar to the thermal and idle noise. Intermodulation noise is affected by system loading, increasing as the loading increases, but it is not directly affected by the RF input level.

Each of the three kinds of noise described above affects system operation in a different way, as can be shown from Figure 2. This graph shows noise performance for one hop of a high-quality microwave system, and is a plot of typical per-channel noise as a function of receiver input level and system loading. Noise is shown at the left as unweighted signal-to-noise ratio in a 3-kc voice chan-

World Radio History

nel, and at the right in dba, F1A weighted, at a 0 transmission level point. The curve is typical for the top channel (in which noise is usually greatest) of a 300-channel system using CCIR deviation and CCIR busy-hour loading. A small allowance for antenna distortion is included.

The effect of the receiver front-end noise on the channel signal-to-noise ratio is shown by the long line starting at the lower left-hand corner and running to the upper right-hand corner. It is evident that this noise is controlling when the RF input is lower than about −40 dbm. The noise at threshold is almost entirely of this type.

At high receiver input levels idle noise becomes controlling and limits the signal-to-noise ratio available, as shown by the bend in the upper line at the upper right-hand corner. This noise sets an upper limit to the channel signal-to-noise ratio when the system is in an idle or unloaded condition.

The effect of intermodulation noise is shown by the lower branch line. This noise sets the limit to the channel signal-

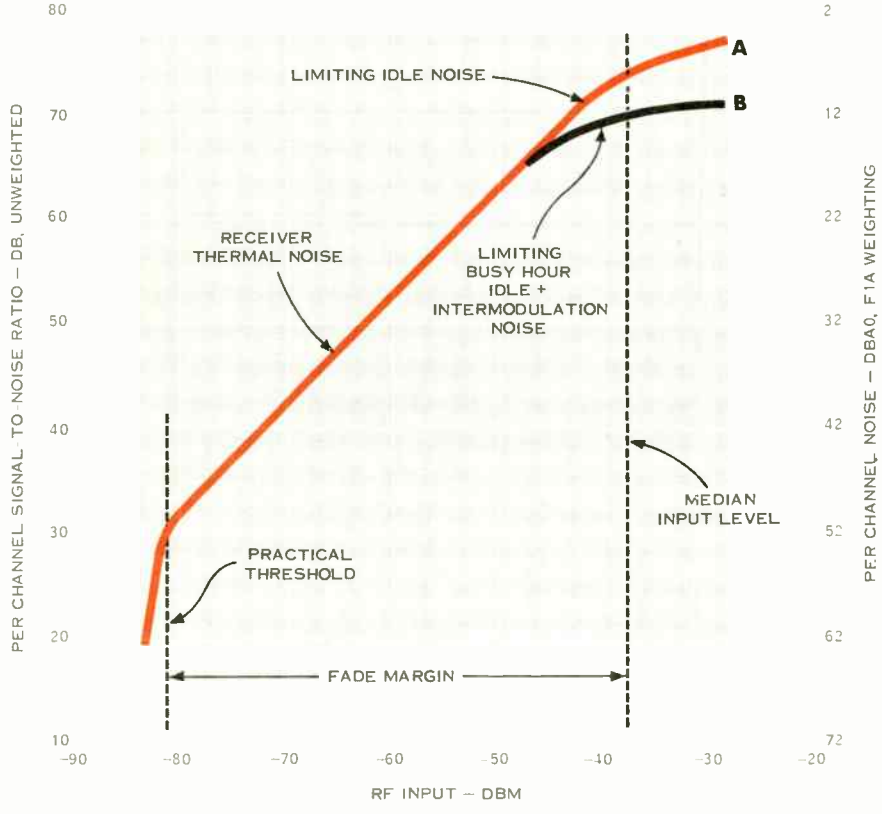


Figure 2. Noise performance of one hop of a high-quality microwave system is shown in this example of typical worst-channel noise plotted as a function of receiver input level and system loading.

World Radio History

to-noise ratio when the system is loaded to simulate busy hour conditions.

A noise characteristic curve such as Figure 2 is a good aid to understanding microwave noise performance, since it includes essentially all of the noise effects and shows them under all conditions of operation. The three most significant bits of information to be derived from the curve are the noise level at the practical threshold point, the noise level at the point of normal RF receiver input level under busy-hour loading conditions, and the fade margin.

A microwave system of the type specified by Figure 2 is usually engineered to have a median RF input level which is somewhere between −30 and −40 dbm. Such a level makes it possible to have very high signal-to-noise ratios during periods of no fading or very little fading, a condition which exists for all but a very small percentage of the time, and to have a fade margin which permits the RF input level to drop by at least 40 db (about one ten-thousandth of normal) before the signal-to-noise ratio becomes objectionable.

With a typical median input level of about −37 dbm, as shown in Figure 2, the signal-to-noise ratio for this system during non-fading periods will be very high, approaching Curve A during periods of light loading and dropping a few db towards Curve B during the heavy loading periods of the busy hour. Only after the input signal has faded several db does the signal-to-noise ratio begin to drop significantly as the receiver thermal noise begins to exceed the other noises. Over the straight line portion of the curve the signal-to-noise ratio varies db for db with the receiver input level and is determined only by the noise figure of the receiver and the deviation ratio used for the particular channel. Over this portion of the curve the unweighted signal-to-noise ratio in db in the derived 3-kc voice channel can be calculated as:

$$S/N \text{ (in db)} = C + 136 - NF + 20 \log D$$

where

C = *receiver input level in dbm,*

NF = *receiver noise figure in db,*

and

D = *deviation ratio, or peak deviation for the channel divided by the carrier frequency of the channel.*

Signal-to-thermal noise ratio can be improved in three ways: by increasing the input level with higher transmitting power or bigger antennas, lowering the noise figure of the receiver, or increasing the deviation ratio. In practice, equipment and system designers raise the effective power and lower the receiver noise as far as is economically practicable. The effect of increasing the deviation ratio is not so simple; it improves the signal-to-noise ratio for the thermal and idle noise but degrades it for intermodulation noise. For this reason the equipment designer must choose a deviation ratio which provides an optimum balance between the different types of noise.

It is worth noting that the IF bandwidth of the microwave system does not affect the signal-to-thermal noise ratio as long as the receiver input level is above threshold. It does affect the "noise performance" in two ways: it determines the point at which the "knee" of the noise characteristic occurs, often called the FM improvement threshold,

World Radio History

and it has a significant effect on the intermodulation characteristics.

The FM Improvement Threshold in dbm for a microwave receiver can be calculated as:

$$T_{FMI} = -104 + NF + 10 \log B_{mc}$$

where

NF = *receiver noise figure in db,*

and

B_{mc} = *receiver IF bandwidth at the 3-db points.*

Changing the deviation ratio does not change the point at which threshold occurs, but it does change the value of signal-to-noise ratio at that point. Increasing the IF bandwidth raises the threshold point by admitting more noise into the system, thus reducing the available fade margin, but it makes possible a reduction in the intermodulation noise. Again, the equipment designer must attempt to achieve an optimum balance between these two conflicting factors. His choices are further affected by the fact that there are legal restrictions on the total bandwidth and deviation which can be used for a microwave channel. The FCC applies somewhat tighter restrictions to industrial users than it does to the common carrier users.

Although the FM Improvement Threshold represents the practical working threshold for a microwave system, there are other definitions of threshold which do not. This has caused a certain amount of confusion among microwave engineers. In order to avoid this confusion, it is now a common practice to specify threshold as the RF input level which will produce a specific minimum acceptable signal-to-noise ratio in the worst voice channel. Since threshold is based on an arbitrary channel signal-to-noise ratio, it may be different from commonly accepted values for either noise threshold or FM improvement threshold. An unweighted signal-to-noise ratio of 30 db is widely used as the minimum acceptable both by telephone and industrial users.

When the receiver input signal becomes very high, a point is reached where the signal-to-thermal noise ratio is no longer directly dependent on the receiver input level. This effect is indicated by the bend in the upper right branch of the curve. Here the thermal noise produced in the transmitter circuits and in those portions of the receiver circuits which are not affected by the automatic gain controls provides an upper limit to the signal-to-noise ratio under non-loaded conditions. This portion of the curve, though of some interest, is not really significant from an operational point of view since the signal-to-noise ratio makes little difference if the system is not being used.

In this area of high receiver input level, the lower branch is the *significant* operational curve. This gives the signal-to-total-noise ratio since it includes thermal noise, idle noise, and intermodulation noise under loaded conditions.

Intermodulation noise is produced in many different ways in a microwave system; reducing it to the extremely low levels required to meet present day noise standards imposes stiff requirements on many areas of equipment design. The FM modulation and demodulation processes, the passbands of the filters, and the characteristics of the baseband amplifiers must be highly linear, both in amplitude and phase, over a wide dynamic range. Impedance mismatches in

World Radio History

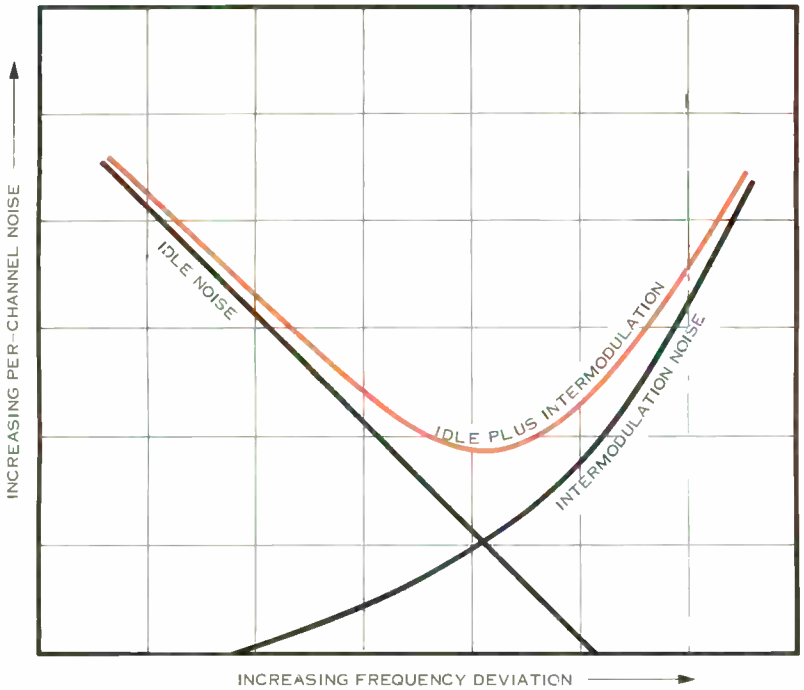


Figure 3. "Front end" idle noise is reduced in direct proportion to increase in frequency deviation due to signal level or system load. However, beyond "break point" of equipment, increasing intermodulation noise rapidly overcomes and reverses this advantage. The level at which intermodulation noise becomes dominant varies from equipment to equipment.

the equipment, and also in the associated waveguide and antenna systems, must be kept as low as possible in order to reduce intermodulation effects caused by reflections.

Multihop Performance

A noise performance curve such as the one shown in Figure 2 applies only on one hop of a microwave system, and it does *not* include the noise contribution of the associated multiplex equipment. The multiplex noise must be added to the microwave noise to get the system signal-to-noise ratio. The multiplex noise under loaded conditions usually runs about 20 to 23 dba0 for a pair of carrier terminals; this is considerably higher than the noise shown for the single microwave hop, and for a one or two hop system the over-all noise is mainly that of the multiplex. But for a long microwave system, in which the multiplex noise appears only once and the per-hop microwave noise many times, the latter becomes controlling.

World Radio History

The curve of Figure 2 shows what the noise performance will be as a function of RF input level, but it does not give any information as to the time distribution of this input level, except by the circumstance that the input will equal or exceed the median level for 50% of the time. For the purposes of this article, it is sufficient to state that the period of time during which a microwave hop will experience a fade of the order of 40 db is extremely small, and the probability that more than one hop of even a very long microwave system will be at or near the threshold level at a given moment is negligible. This means that the noise level at the threshold point on a microwave system does not increase as more hops are added.

When one hop fades down near threshold, the noise contributed by that hop becomes so much greater than the combined noise of all the other hops that the full system signal-to-noise ratio is essentially that of the one faded hop. For systems with high fade margins, then, it is seen that the threshold noise level is the same regardless of the number of hops in the system. Increasing the number of hops does not change the noise level at threshold, but it does increase the amount of time during which the system will reach threshold and, hence, directly reduces system propagation reliability, since the reliability is defined in terms of the percentage of time during which every hop of the complete system is at or above threshold.

Although the threshold noise level does not change as more hops are added to a microwave system, the noise level during average or non-faded conditions does change. To a first approximation, the multihop noise under these conditions is the sum of all the individual hop noises added together on a power basis.

This is strictly true for thermal and idle noise and even-order intermodulation noise products, which are fully incoherent even on tandem hops. But certain odd-order intermodulation products, even though they are essentially random in a single hop, have some coherency on tandem hops and may, therefore, add on a *voltage* rather than a *power* basis. In this case the system noise power can be greater than the sum of all the per-hop noise powers. Odd-order intermodulation products are normally considerably lower than even-order products, so this effect does not significantly affect the noise addition until the system becomes fairly long. For the longer systems it is important that the equipment design be such as to give special attention to the reduction of odd-order intermodulation effects. If this is done, the noise on even long systems has a summation pattern very close to that of power addition.

Noise Specifying Methods

So far in this discussion channel noise has been considered in terms of signal-to-noise ratio, expressed in db, with the "signal" being understood to be a 1,000-cycle test tone with a power of 0 dbm at a 0 transmission level point, and the "noise" being the unweighted noise in a 3-kc bandwidth. Without belaboring the point, the "signal" in signal-to-noise ratio really means "standard signal," which must be taken as test tone level. This way of defining noise, which has been adopted as a standard by the E.I.A., is perhaps the most meaningful for the industrial user.

Although conceptually, signal-to-noise ratio is the significant end result,

World Radio History

it is considerably more convenient for purposes of calculation to have the channel noise expressed in some absolute form. One such way, developed by the Bell System and widely used for many years in this country, is in terms of a unit identified as *dba, F1A-weighted*. The reference level, or 0 dba, is equivalent to a 1,000-cycle tone with a power of −85 dbm or of a 3-kc white-noise band with a power of −82 dbm.

(Bell System has recently introduced a new weighting characteristic and a new unit, the dbrnc. The reference level, or 0 dbrnc, is the equivalent of a 1,000-cycle tone with a power of −90 dbm, or of a 3-kc band of white noise with a power of −88.5 dbm, usually rounded off to −88 dbm.)

A second way of expressing noise, developed by CCITT and CCIR, is in terms of picowatts, psophometrically weighted. The reference level, 1 pw_p, is the equivalent of an 800-cycle tone with a power of −90 dbm, a 1,000-cycle tone with a power of −91 dbm, or a 3-kc band of white noise with a power of approximately −88 dbm. The shapes of the F1A weighting curve and the psophometric curve are essentially identical, and dba can be converted to picowatts, or vice versa, by the formula

$$dba = -6 + 10 \log_{10} pw_p.$$

Since the dba and the picowatt are both absolute units, it is necessary to associate them with some specific transmission level before they have any real significance. In recent years it has become quite common to do this by adding a zero to the unit to indicate that it is

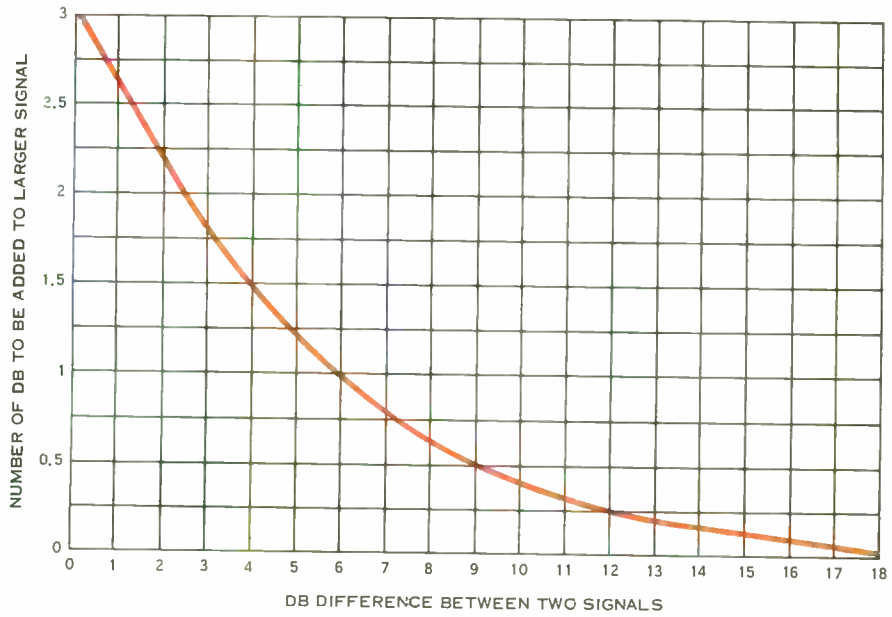


Figure 4. Addition of powers expressed in logarithmic units is simplified by the use of curve.

World Radio History

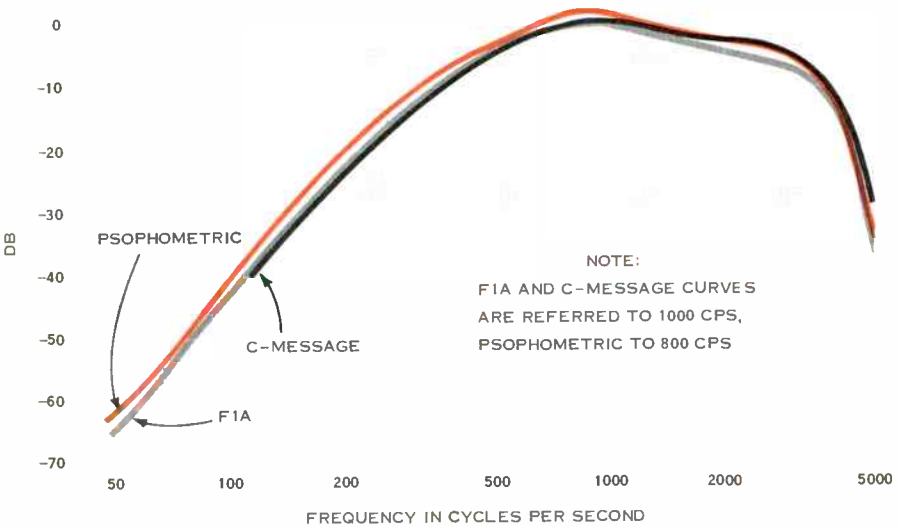


Figure 5. Weighting curves in common use indicate how noise at the band edges affects unweighted measurements out of proportion to its actual interfering effect.

referred to a 0 transmission level point. The resulting units, written as dba0 and pw_p0, can be converted to signal-to-noise ratios as defined earlier by the formulas

$$S/N = 82 - dba0$$

$$S/N = 88 - 10 \log_{10} pw_p 0.$$

These relations are correct only if the noise is essentially white noise. The noise produced in multichannel microwave systems is almost entirely of this type, so the correlations are valid for microwave noise.

The dba and the psophometric picowatt are equally valid absolute noise units, but differ somewhat in application because the one is logarithmic and the other linear. The linear unit, the picowatt, has the advantage that addition of noise powers becomes a matter of simple arithmetical addition of the picowatts. Addition of powers expressed in logarithmic units such as the dba is not quite so simple but can be done relatively easily by the use of a chart such as Figure 4.

Many people feel that the mathematical convenience of the picowatt is more than offset by the fact that the effects being measured are essentially logarithmic themselves and, consequently, a logarithmic unit is much more meaningful than a linear unit. A change of 1 db in signal-to-noise ratio has essentially the same meaning regardless of where it occurs, but a change of 100 picowatts might mean a change of 20 db or more if it occurs at a very low level; or it might mean no detectable change at all if it occurs at a high level.

World Radio History

Weighted or Unweighted?

When only the noise generated in the microwave system itself is considered, along with its measurement at the radio baseband output point, it really makes little difference whether weighting is used. For noise of this kind the effect of weighting is simply to reduce the noise by a fixed, known amount (3 db in the case of F1A weighting, approximately 2 db in the case of psophometric weighting). Using weighting simply changes the numerical value of a noise reading by that fixed amount. It is equivalent to changing the noise unit itself. A noise level giving a flat signal-to-noise ratio of 50 db will give an F1A weighted ratio of 53 db; this sounds better, but the noise is just the same.

But when noise in the complete system and its measurement at the output of the voice channels themselves is considered, with all of the multiplex equipment and drop equipment connected and functioning, weighting is far more significant. In this case there may be substantial amounts of noise which are not random in nature. Much of this noise may be at very low or very high frequencies where the effect on measurements of noise power is far out of proportion to the effect on actual transmission quality. For this reason telephone practice is invariably to use weighted noise units. Even though the weighting is based strictly on voice transmission, it is quite possible that for data transmission systems designed to operate over a voice channel, a weighted measurement may be as good a criterion as an unweighted one or perhaps even better.

Because of phase effects near the band edges, such data circuits are usually located in the interior part of the band — an area where the weighting characteristic introduces the least change. The C-message weighting, for example, is reasonably flat over most of the portion of the band which is usable for data.

The most important thing about noise units is not so much whether they are weighted or unweighted. Rather, it is that they be precisely defined so that the meaning is unmistakably clear. ●

BIBLIOGRAPHY

1. H. A. Lewis, R. S. Tucker, G. H. Lovell, and J. M. Fraser, "System Design for the North Atlantic Link," *The Bell System Technical Journal;* January, 1957.
2. R. H. Franklin and J. F. Bampton, "Coordination of British and American Transmission Techniques," *The Post Office Electrical Engineers' Journal;* January, 1957.
3. T. A. Combellick and M. E. Ferguson, "Noise Considerations on Toll Telephone Microwave Radio Systems," *Electrical Engineering;* April, 1957.
4. A. J. Aikens and D. A. Lewinski, "Evaluation of Message Circuit Noise," *Bell System Technical Journal,* July, 1960.
5. "Microwave Intermodulation Distortion — and how it is measured," *The Lenkurt Demodulator;* December, 1960.
6. Red Book Vol. III, Rec. G. 212, G. 222, *CCITT II Plenary Assembly;* New Delhi, 1960.
7. "How to Evaluate Radio and Carrier Noise Performance," *The Lenkurt Demodulator;* May, 1961.
8. "dba and Other Logarithmic Units," *The Lenkurt Demodulator;* November, 1961.
9. "Take the Mystery out of Microwave Literature," *The Lenkurt Demodulator;* July, 1962.
10. "Levels and Powers in a Carrier System," *The Lenkurt Demodulator;* September, 1963.
11. *CCIR — Documents of the X Plenary Assembly;* Geneva, 1963.
12. "Point to Point Radio Specifications (Microwave)," *REA Form 397d.*

World Radio History

World Radio History

the Lenkurt Demodulator

VOL. 13, NO. 6 JUNE, 1964

NOISE PERFORMANCE in Industrial Microwave Systems

Part Two

This is the second part of a two-part article by Robert F. White, Lenkurt Transmission Engineer. The first part, which discussed such factors as noise sources, thresholds, loading, and noise-specifying methods, appeared last month.

The term, "noise power ratio," which came into the language with the advent of white noise testing methods, has had a certain vogue as a somehow more "fundamental" quantity than the signal-to-noise ratio in the voice channel. However noise power ratio is only an intermediate step in a particular method of making noise-loaded measurements. The most modern noise loading test sets do not even use this step but go directly to the significant end-result, noise in the channel. For a given set of measuring conditions, a correcting factor can be calculated which, when added to the noise power ratio, will give the signal-to-noise ratio, but this factor is not always the same. So a given noise power ratio sometimes means one thing, sometimes another. The fairly common practice of specifying noise performance both ways — as a noise power ratio and as a signal-to-noise ratio or noise power in the voice channel — is particularly undesirable. It gives results which are either redundant if they come out the same or contradictory if they do not.

Lenkurt's practice is to specify and calculate microwave system noise performance in dba0, F1A weighted. The

World Radio History

end result is usually given both in dba0 and in signal-to-noise ratio. Conversion from one to the other, or to other noise units, is easily made. Figure 1 correlates signal-to-noise ratio, dba, and picowatts for noise which is essentially random, and Figure 2 gives dba versus picowatts in graphical form.

Column 1 in Figure 1 gives flat signal-to-noise ratio in a 3-kc voice band; column 2 gives the equivalent in dba, F1A weighted, in a 3-kc voice band; and column 3 gives the equivalent in psophometrically-weighted picowatts.

The table is applicable to signal-to-noise ratio conversion only if the noise is of the random, or "white noise" type. The dba/picowatt conversion is based on the following correlation which was established by Bell System and British Post Office engineers in connection with the transatlantic cables — it is valid for any kind of noise:

$$dba = -6 + 10 \log_{10} pw_p$$

System Noise

Despite the complexity of the problem, it turns out that it is necessary to define, calculate and measure only three significant parameters in order to determine with adequate precision the limits of noise performance which a microwave system will have under actual operating conditions, even taking into account the effects both of fading and busy hour loading.

These parameters are:

1. The receiver input level at which the noise in the worst derived voice channel reaches 52 dba0.
2. The required fade margin in db. This affects the noise performance since it determines what the receiver median input level (corresponding to the non-faded condition) must be. Adding the fade margin to the threshold level gives the median input level.
3. The noise in the worst derived

S/N	dba0	pw_p0	S/N	dba0	pw_p0	S/N	dba0	pw_p0
28	54	1,000,000	48	34	10,000	68	14	100.0
29	53	794,000	49	33	7,940	69	13	79.4
30	52	631,000	50	32	6,310	70	12	63.1
31	51	502,000	51	31	5,020	71	11	50.2
32	50	398,000	52	30	3,980	72	10	39.8
33	49	316,000	53	29	3,160	73	9	31.6
34	48	252,000	54	28	2,520	74	8	25.2
35	47	200,000	55	27	2,000	75	7	20.0
36	46	159,000	56	26	1,590	76	6	15.9
37	45	126,000	57	25	1,260	77	5	12.6
38	44	100,000	58	24	1,000	78	4	10.0
39	43	79,400	59	23	794	79	3	7.9
40	42	63,100	60	22	631	80	2	6.3
41	41	50,200	61	21	502	81	1	5.0
42	40	39,800	62	20	398	82	0	4.0
43	39	31,600	63	19	316	83	−1	3.0
44	38	25,200	64	18	252	84	−2	2.5
45	37	20,000	65	17	200	85	−3	2.0
46	36	15,900	66	16	159	86	−4	1.6
47	35	12,600	67	15	126	87	−5	1.3
48	34	10,000	68	14	100	88	−6	1.0

Figure 1. Comparison of noise performance units: flat signal-to-noise ratio in a 3-kc band; dba0, F1A weighted; and psophometrically weighted picowatts.

World Radio History

voice channel with median input level to the receiver (or receivers if it is a multihop system), measured with the radio baseband loaded with white noise power equivalent to the busy hour load for full rated channel capacity of the system.

For the purposes of this article, the first two parameters can be disposed of rather simply. There is nothing very controversial about the choice of 52 dba0 as the point at which a voice channel should be taken out of service, even though the channel would still be usable at even higher noise levels. But for present day requirements, 52 dba0 is a quite reasonable figure for determining the threshold. The choice of a figure for fade margin is considerably more complex but not really controversial either. There is no question that fade margins at microwave frequencies must be high, the only question is *how* high. The decision is a familiar one: economics versus reliability. In the 6-Gc band, which at the present time is the "work horse" for the industry, fade margins are now almost always at least 35 db and often as high as 45 db or more on "problem" paths; 40 db is a quite typical value easily achievable with conventional microwave equipment and antenna sizes. System reliability is not the subject of this article and it has been brought into the discussion only because it affects the choice of fade margin and fade margin affects noise performance.

These two parameters, threshold and fade margin, are characteristics of the individual hops rather than of the system. Their measurement is simple and straightforward. It is done in the field only to determine whether the equipment is operating properly and whether the receiver input level is at or very near the value which has previously been calculated for the path.

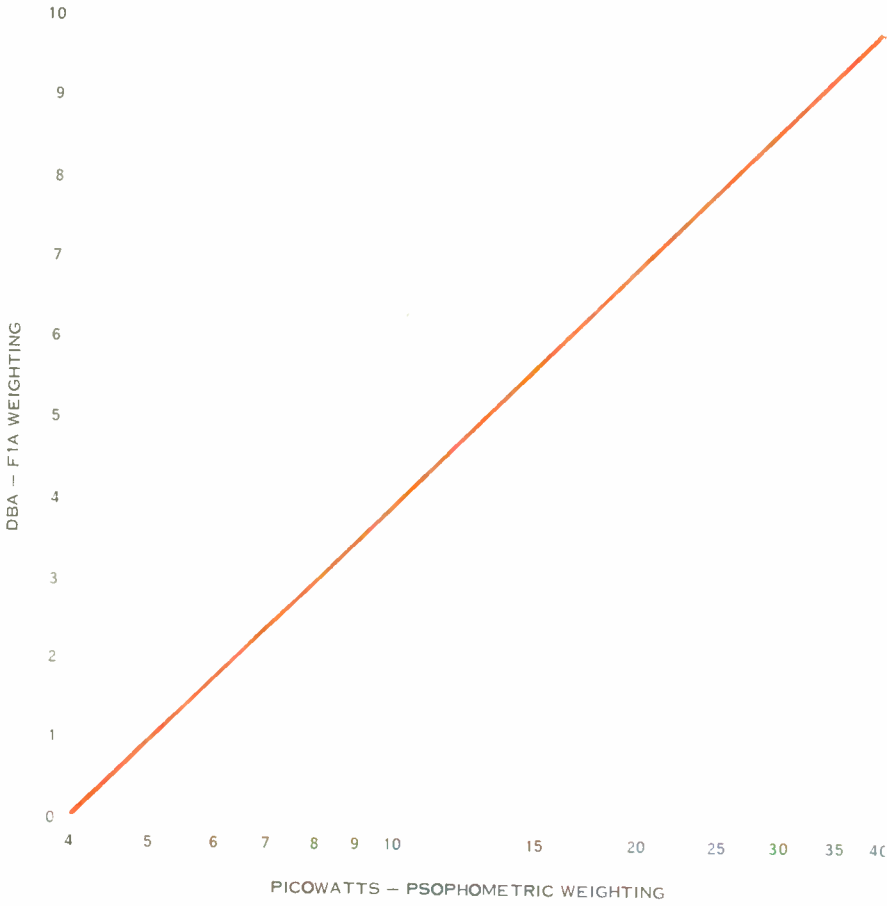
All the rest of this article will be devoted to a discussion of the third parameter, which is perhaps the most basic one, since it is the one which describes the day-in, day-out noise performance of the system. Threshold noise occurs only for fleeting instants and at very rare intervals, but *this* noise is there all the time (though it may drop a db or so during non-busy periods).

A most important contributor to this noise parameter is intermodulation distortion. The importance of intermodulation characteristics in determining the load handling capability of a microwave system has not always been fully appreciated. Until a few years ago, there was no convenient way of calculating or measuring loaded performance, and microwave systems were often judged on the basis of idle noise alone, which could be a very inadequate way of judging true performance.

Now there is a suitable way of measuring loaded performance. The discovery a few years ago that the statistical properties of a multichannel telephone load occupying a given baseband spectrum were very similar to those of a continuous noise load occupying the same spectrum, and the development of noise loading test sets based on this principle, have made it possible to determine with a fairly high degree of accuracy the noise performance of microwave equipment or systems under conditions which are quite similar to those which are encountered in service.

The method allows the microwave noise contribution to be measured separately without reference to the multiplex equipment. Measurements can be made over each hop individually and over the complete system end to end. The microwave system can be measured at its full rated channel capacity, even when it is not equipped initially with a full complement of channels.

World Radio History



$$DBA = -6 + 10 \log_{10} PWP$$

$$PWP = \text{ANTILOG} \frac{DBA + 6}{10}$$

(VALID FOR ANY KIND OF NOISE)

OTHER RANGES:

DBA	PWP
+10	X10
+20	$X10^2$
+30	$X10^3$
+40	$X10^4$
+50	$X10^5$
+60	$X10^6$
+70	$X10^7$
+80	$X10^8$

Figure 2. Curve for converting dba, F1A weighted, to picowatts, psophometrically weighted. The formula on which this curve is based is not limited to white noise, but is valid for any kind of noise.

World Radio History

Noise loading test sets are now readily available, and there is a generally accepted standard for calculating the noise load power to be used:

(−15 + 10 log N) dbm0, where N = 240 or more channels;

(−1 + 4 log N) dbm0, where N is between 12 and 240 channels.

These formulas include an allowance for the power of signaling tones and a number of telegraph channel tones as well as for the speech currents themselves.

Measuring sets are most readily available for the standardized CCIR microwave channel capabilities of 60, 120, 300, 600, 960, 1800, or 2700 channels. Thus, the third parameter can be measured with considerable accuracy using the proper test equipment — although such equipment is quite expensive.

The Real Objectives

What should the objective be for this most significant noise parameter? This can be an extremely important decision to user and manufacturer alike, since it seriously affects the cost of the system to the user and the technical problems which must be solved by the manufacturer.

Most communications users look to the practices of the telephone industry for guidance in this respect, since telephone people are in the business of selling communications and, consequently, can usually be relied on to look for a good balance between performance and cost. Telephone practice makes a rather sharp distinction between short-haul and long-haul systems, with the dividing point at about 200 miles. For long-haul standards, CCIR and Bell System are the best sources. For the short-haul systems Bell and REA provide good guidelines.

Bell System and CCIR both treat the allowable noise for systems longer than about 200 miles as directly proportional to length, and their standards turn out to be almost identical. CCIR simply allows 3 picowatts per kilometer for the microwave system contribution and 1 picowatt for the carrier contribution, making a total of 4 picowatts per kilometer for the complete system. Converting these to miles yields 4.8 picowatts per mile for the microwave alone and 6.4 for the microwave plus multiplex. The Bell System objective of 38 dba0 for 4,000 miles includes multiplex as well as microwave contributions. Thirty-eight dba is equivalent to 25,200 picowatts, which is equal to 6.3 picowatts per mile, as against CCIR's 6.4 picowatts per mile.

If three-fourths of the total noise is allotted as the microwave contribution, as does CCIR, the result is 4.7 picowatts per mile, almost the same as CCIR's 4.8.

For short-haul systems the practice is quite different. Current practice is to specify a single value of noise for such a system regardless of the number of hops. This figure is, at present, 27 dba0 for the microwave plus multiplex noise. This is equivalent to 2,000 picowatts for the microwave plus multiplex noise, or 1,500 picowatts for the microwave alone. This is about 7.5 picowatts per mile for the maximum length system of 200 miles, 15 picowatts per mile for a 100-mile system, and even more for shorter systems.

The 7.5-picowatt-per-mile figure for the most stringent case in short haul systems is 2 db less stringent than the 4.7-picowatt-per-mile long-haul figure.

As far as the microwave equipment designer or system planner is concerned, the important thing to know is not picowatts per mile, but picowatts per hop, since microwave noise power is approximately proportional to the number of hops rather than to the number of miles. So, before it can really be determined how much strain the above-listed ob-

World Radio History

jectives put on the microwave system, it is necessary to know, or arbitrarily decide, the length of the average hop. For example, take the long-haul objective of 4.7 picowatts per mile and see what it means to three different engineers:

Mr. "Conservative" figures 25-mile hops and thus gets a requirement of 117.5 picowatts or 14.7 dba0 per hop. He thinks it can be done, but it's pretty rough.

Mr. "Middleroad" figures 30-mile hops and gets a requirement of 141 picowatts or 15.4 dba0 per hop. He isn't too unhappy about it, though he still doesn't think it's a cinch.

Mr. "Optimist" figures 40-mile hops and gets 188.0 picowatts or 16.6 dba0 per hop as the requirement. He just can't understand what those other fellows are worrying about.

Looking at the worst case short-haul requirement of 7.5 picowatts per mile in the same way, Mr. C gets 16.7 dba0, Mr. M gets 17.4 dba0, and Mr. O gets 18.6 dba0 as the per-hop requirement. These figures make all of them pretty happy, so it appears that the agonizing decisions about what performance standards to use really lie only in the area of long-haul systems.

Recommendations

Although the microwave system standards established by the telephone industry provide a basis for establishing adequate performance, they stem from needs which do not always apply to industrial users. Consider the following opinions:

1. The distinction between short-haul and long-haul requirements is valid for industrial as well as telephone users. Unless a short-haul system is eventually to become part of a longer system, there is no need to set the standard any higher than 27 dba0 for an eight-hop system, or an average value of some 17 to 18 dba0 per hop. Even this requirement could be relaxed a couple of db and the service would still be perfectly acceptable. Until very recently telephone companies usually used 31 dba0 for such systems — 4 db worse than their present practice.
2. For long-haul service the industrial user faces a more difficult choice, complicated by the fact that he uses the same system for short- and long-haul service, while the telephone companies use separate systems. It is very tempting to simply fall back on the CCIR or Bell long-haul recommendations and accept them without further consideration. After all, these requirements are only 2 db tighter than the short-haul requirements which can be met fairly easily. It turns out that those 2 db of difference push performance into an area which is much closer to the edge of the present state of the art, especially if "Mr. Conservative's" gloomy estimate of average path length is accepted. This can mean a great many thousands of dollars in the initial cost of a system of even moderate length, and also a considerable increase in maintenance costs if the high performance is to be maintained. Thus, it appears that for the industrial user the cost of these 2 db is too high. The 7.5 picowatt per mile figure used for shorter systems seems perfectly adequate for industrial systems of any length. Even for a 4,000-mile system, this would mean only about 38.7 dba0 for the microwave and about 40 dba0 for the microwave plus carrier. That still is a

World Radio History

42-db signal-to-noise ratio, better than that obtained in a call across town in many parts of the world.

3. CCIR and Bell System's long-haul criteria are based on the use of heterodyne or non-demodulating repeaters. If channel dropouts are needed only at widely separated points, the heterodyne repeater is a clear choice over the demodulating repeater using back-to-back terminals because the noise performance can be made somewhat better and level problems are greatly diminished. With recent improvements in the design of back-to-back repeaters the difference in noise performance between the two has been reduced to something on the order of 1 db.

Industrial microwave systems, unlike those of the telephone companies, are likely to require channel dropping at almost every repeater point. For this kind of service the back-to-back repeater has a very positive advantage, since the full baseband is available at every point.

At the present state of the art, the 4.7 picowatt per mile criterion for long-haul circuits can probably be met using back-to-back repeaters of the very best modern design, but not without rigid control of a great many variables. The criterion can be met a little easier, but not very much, if heterodyne repeaters are used.

If the criterion is relaxed to about 7.5 picowatts per mile, the requirements can be met fairly easily with either type of equipment and, in this case, the back-to-back type appears to be the best choice in most cases.

In summary, why not use CCIR and other similar criteria as guides, but not absolute standards, and modify them to suit the particular requirements rather than following them unquestioningly? It may save microwave users a good deal of money with no significant decrease in performance. •

BIBLIOGRAPHY

1. H. A. Lewis, R. S. Tucker, G. H. Lovell, and J. M. Fraser, "System Design for the North Atlantic Link," *The Bell System Technical Journal;* January, 1957.
2. R. H. Franklin and J. F. Bampton, "Coordination of British and American Transmission Techniques," *The Post Office Electrical Engineers' Journal;* January, 1957.
3. T. A. Combellick and M. E. Ferguson, "Noise Considerations on Toll Telephone Microwave Radio Systems," *Electrical Engineering;* April, 1957.
4. A. J. Aikens and D. A. Lewinski, "Evaluation of Message Circuit Noise," *Bell System Technical Journal,* July, 1960.
5. "Microwave Intermodulation Distortion — and how it is measured," *The Lenkurt Demodulator;* December, 1960.
6. Red Book Vol. III, Rec. G. 212, G. 222, *CCITT II Plenary Assembly;* New Delhi, 1960.
7. "How to Evaluate Radio and Carrier Noise Performance," *The Lenkurt Demodulator;* May, 1961.
8. "dba and Other Logarithmic Units," *The Lenkurt Demodulator;* November, 1961.
9. "Take the Mystery out of Microwave Literature," *The Lenkurt Demodulator;* July, 1962.
10. "Levels and Powers in a Carrier System," *The Lenkurt Demodulator;* September, 1963.
11. *CCIR — Documents of the X Plenary Assembly;* Geneva, 1963.
12. "Point to Point Radio Specifications (Microwave)," *REA Form 397d.*

World Radio History

World Radio History

the Lenkurt® Demodulator

VOL. 9 NO. 12 DECEMBER, 1960

Microwave Intermodulation Distortion —and how it is measured

Modern multi-channel microwave systems achieve their best performance when operating levels are carefully controlled. If modulating levels are increased in an FM system, background noise is reduced by the wider frequency deviation, but intermodulation distortion goes up. Equipment designers seek to minimize intermodulation distortion without making the microwave equipment so expensive as to discourage its use. This article discusses some characteristics of intermodulation distortion and how it may be measured.

Every element in a communications system tends to degrade signal quality to some extent, even in the very best equipment. Amplifiers, modulators, klystrons, and other such components are inherently non-linear. That is, over a wide band of frequencies, amplitude response or rate of phase shift will not be uniform. Such techniques as negative feedback go a long way toward improving linearity, but can never achieve perfection. As performance approaches ideal, cost of the equipment rises astronomically, so that economic considerations eventually determine the performance limits of the system.

Intermodulation distortion increases as the load to be handled approaches the capacity of the device. In amplifiers, the power-handling capability of the amplifier may be the limiting factor; in modulators, klystrons, modulation detectors, and the like, bandwidth may determine how much load the device can accommodate without appreciable non-linear distortion.

Harmonics and Intermodulation

If a single frequency is passed through a non-linear device—an amplifier, for instance—the output signal will contain not only the funda-

World Radio History

mental frequency f, but also harmonics of the frequency: $2f$, $3f$, $4f$ . . . nf.

Similarly, if more than one frequency is passed through the non-linear amplifier, harmonics of each of the fundamental frequencies will appear in the output signal. Since these harmonics were not present in the original signal, they are the result of distortion caused by the amplifier's non-linearity. The more non-linear the amplifier, the greater the signal power that is converted to distortion products.

Another characteristic of non-linear devices is that the various frequencies passing through the device modulate each other so that additional frequencies are produced. These *intermodulation products* represent not only the sum and difference of the original input frequencies, but also the sum and difference of the various harmonics and the intermodulation products themselves.

For example, assume that the input frequencies are called A, B, C, etc. Then the harmonics which appear in the amplifier output will be 2A, 3A, 4A . . . nA; 2B, 3B, 4B . . . 2C, 3C, 4C, and so forth. Actually, there are an infinite number of harmonics of each fundamental frequency, but the magnitude of these harmonics diminishes very rapidly with higher order, so that only the first few harmonics of each frequency have much significance.

Second-order intermodulation products consist of such frequencies as (A + B), (A - B), (A + C), (A - C),

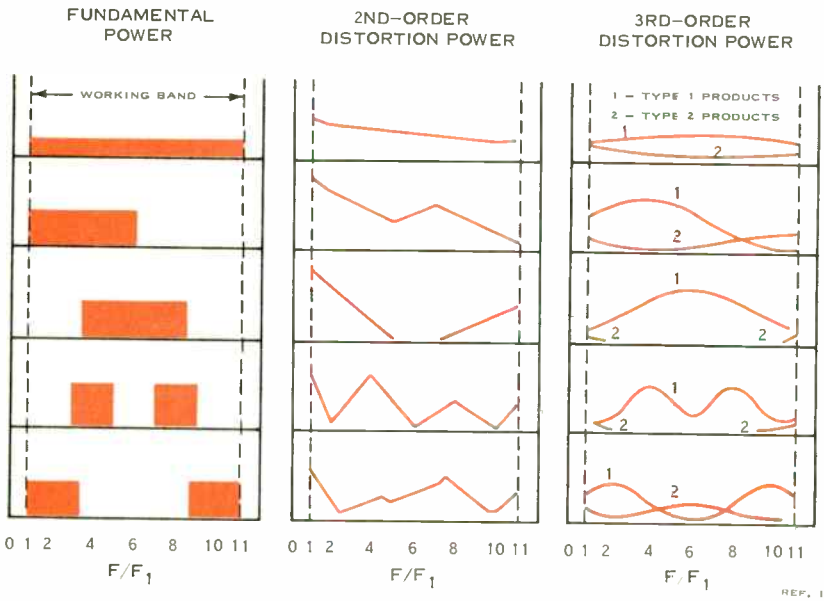


Figure 1. Each type and order of intermodulation product has a different distribution when fundamentals are distributed non-uniformly (power is the same in each example). Second-order products are greatest problem in wideband radio systems with "back-to-back" repeaters. Type 1 third-order products are most disturbing in narrow-band systems, cable and coaxial systems, and those with heterodyne repeaters.

World Radio History

(B + C), (B - C), and so forth. Third-order products are much more complex, and will typically consist of such frequencies as

(A + B + C), (A - B - C) . . ,
(2A + B), (B - 2A) . . . ,
(A + B - C) . . . ,
(2A - B) . . . .

As indicated in Figure 1, second- and third-order distortion products have different distributions across the band.

The differences result from the way in which the various intermodulation products form. Second-order intermodulation can only be (A ± B), (B ± C), and the like. Third-order products, on the other hand, may be formed from complex combinations of many frequencies, both above and below the fundamentals. The third or higher-order products are divided into two groups: Group 1 products, which add on a voltage basis at each repeater, and Group 2 products which add on a power basis (except under certain special conditions). Normally, the distortion power of Group 1 products is much greater than that of the Group 2 products, and, along with second-order products, provide the greatest distorting effect.

The various types of intermodulation products will have different effects in different types of equipment. For instance, even-order products cannot appear in narrow-band systems unless the ratio of the highest frequency to the lowest is at least 2 to 1 for second-order products, 1½ to 1 for fourth-order products, and 1¼ to 1 for eighth-order products. Odd-order products, on the other hand, appear in the band regardless of the frequency ratio. For this reason, third- and other odd-order products present the greatest problem in narrow-band systems.

Complex Signals

It is unlikely that single frequencies will be used to transmit information over a modern radio system because of the limited information capacity of such methods. Invariably, information is conveyed by a complex signal of some sort, such as produced by high-speed telegraphy, speech, or music. The number of individual frequencies in such signals is very large. When many channels are transmitted over a single system, the number of fundamental frequencies becomes extremely large.

Under such circumstances, the intermodulation products are so widespread that they resemble noise in their randomness. In a wideband radio system, intermodulation in a given channel raises noise level in other channels, rather than appearing as crosstalk.

As the signal level approaches the "break point" or overload level, intermodulation power tends to increase faster than the increase in level. In a frequency-modulated radio system, increased signal level increases the frequency deviation. This, in turn, decreases background noise in direct proportion to the increase in deviation.

The noise in a system carrying many voice channels never stays at a fixed level, but varies from instant to instant as the number of channels and the signal power of each changes. Differences in how loud or how fast the various telephone users talk may affect the total intermodulation noise at any given time (See DEMODULATOR, *August, 1959*).

Measuring Intermodulation

High-density carrier systems—those with 60 channels or more—provide a complex signal that strongly resembles "white" noise in its randomness and

World Radio History

uniform distribution across the band. When such a signal is transmitted over a modern radio system, intermodulation products appear both within the base-band and out of band. If one or more channels are left idle, it can be observed that the noise level in these channels increases when the rest of the system is heavily loaded.

This immediately suggests a way of measuring the intermodulation characteristics of the transmission system. Because the load on the actual working system is variable and cannot be controlled, a substitute load must be found which is representative of a typical signal. Thermal or "white" noise is generally used, with the noise power being adjusted to represent the power of the actual signal. Figure 2 shows a block diagram of a typical arrangement used for measuring intermodulation distortion in radio systems.

The noise source is usually a phototube, diode, or transistor operated in such manner as to provide a wide-spectrum noise signal. The noise output is amplified and passed through a band-pass filter (usually a high-pass plus a low-pass filter) which limits the noise spectrum to the bandwidth of the radio equipment.

Following the noise generator, its associated amplifiers, and level control, a band-stop or "notch" filter is inserted to *eliminate* the noise signal from a selected band. Several such band-stop filters are usually used, in order to measure intermodulation products in different parts of the band. Usually, only one band-stop filter is used at a time, and some provision is made for quickly substituting filters.

The broad-band noise signal from which the "notch" has been eliminated by the band-stop filter, is applied to the system at the same power level as the normal wide-band signal. If the system produces intermodulation distortion, the noise in the notch will increase, and this may be measured at the receiver with a frequency-selective voltmeter. In normal practice, noise in the notch is measured with the band-stop filter in the transmitting circuit and then out. The ratio (in decibels) of the noise power, with no band-stop filter in the circuit, to the noise power

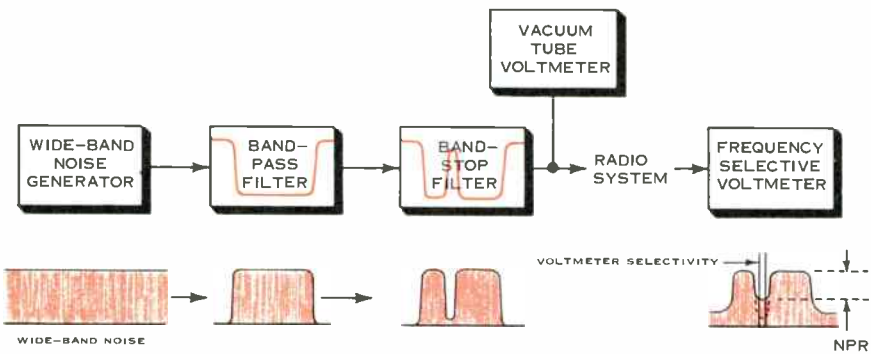


Figure 2. Block diagram of a typical arrangement for measuring intermodulation distortion in terms of Noise Power Ratio.

World Radio History

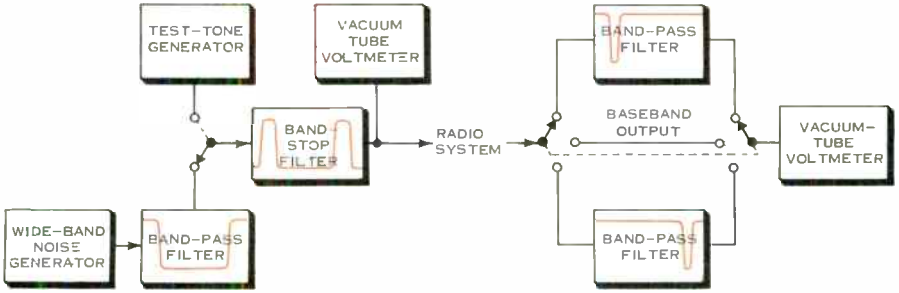


Figure 3. Block diagram of arrangement for determining signal-to-intermodulation noise ratio directly. System is calibrated with test-tone, then noise load is substituted. Total noise power from final filters provides direct signal-to-noise value.

present with the filter in the circuit is called the *Noise Power Ratio*, usually abbreviated NPR, and is frequently used to express the intermodulation distortion produced in communications equipment. NPR provides a measure of the intermodulation performance of the radio equipment over which carrier channels may be transmitted, but does not indicate how much interference will be experienced within *each* channel. Since other types of interference, such as crosstalk and background noise, are expressed in terms of their power or effect per-channel, it is desirable to express intermodulation distortion similarly, thus putting all disturbing effects on a common basis.

Intermodulation distortion may be expressed in terms of per-channel signal-to-noise ratio by adding a "correction factor" derived from the ratio of total signal power to the power appearing in one channel. Since power is distributed uniformly, this is a function of bandwidth. For example, if the total baseband is 1200 kc and the total signal load is +10 dbm, the distortion power appearing in a single 3-kc channel will be 1/400 of the +10 dbm total power. This is 26 db below +10 dbm, or −16 db relative to the channel test tone. Thus, in this example, a 16-db correction factor would have to be added to the Noise Power Ratio in order to express intermodulation distortion in terms of signal-to-noise ratio. Note that the correction factor will be different for different bandwidths and total signal power.

If it is desired to express intermodulation noise in terms of its disturbing effect, an additional 3 db must be added to the flat signal-to-noise ratio to obtain the F1A-weighted signal-to-noise ratio. This may be converted to dba by subtracting the F1A-weighted signal-to-noise ratio from −85 db. In circuits using C-message weighting, 1.5 db should be added to the flat signal-to-noise ratio to obtain C-message weighted signal-to-noise ratio. Subtracting this from −90 db yields dbrn (C-message).

An alternate way of measuring intermodulation distortion is diagrammed in Figure 3. In this method, the entire system is calibrated by transmitting a test tone through the system. The test tone frequency is selected to lie in the center of a band-pass filter at the

World Radio History

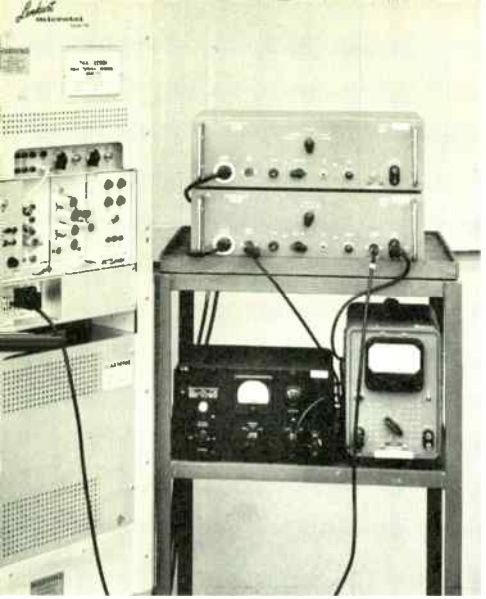


Figure 4. Lenkurt Type 5203-5204 Distortion Test Set in use. This equipment works on principle diagrammed in Figure 3.

output of the receiver. The received test tone power then becomes the reference power.

The test tone is removed and a wideband noise signal of a power which represents the baseband signal is transmitted through a band-stop filter which rejects a band of frequencies somewhat wider than the band accepted by the receiver band-pass filter. The total power measured at the output of the receiver band-pass filter is proportional to both the intermodulation distortion and the bandwidth of the receive filter. By applying a correction factor to account for the bandwidth of the filter, intermodulation distortion power may be read directly from the output of the receiver band-pass filter.

Both methods will give the same results. The latter method has the advantage of permitting special test equipment to be constructed in which filter characteristics are allowed for in calibrating the equipment, thus eliminating calculations and greatly simplifying the measurement of intermodulation characteristics. Figure 4 shows an intermodulation test set designed and manufactured by Lenkurt. This equipment requires only a noise source and ordinary laboratory RMS-indicating vacuum tube voltmeter. The instrument is calibrated to give direct signal-to-noise values (flat-weighted) for intermodulation distortion.

Intermodulation distortion measurements show the performance of the system as a whole, rather than of individual components, such as transmitter and receiver. If separate performance ratings are provided for separate components, they should be added together on a power basis to obtain a realistic evaluation of system performance. ●

BIBLIOGRAPHY

1. R. A. Brockbank and C. A. A. Wass, "Non-linear Distortion in Transmission Systems," *Institution of Electrical Engineers Journal (London)*, Vol. 92, Part 3; March, 1945.
2. R. W. White and J. S. Whyte, "Inter-Channel Crosstalk and Noise On Broadband Multi-channel Telephone Systems," *Electronics & Communications;* November, 1957.
3. M. E. Ferguson, "Intermodulation Testing of Multi-channel Radio Systems," *A.I.E.E. Conference Paper CP 59-1194;* (August, 1959).
4. S. Janson and V. Stending, "Some Problems Concerning Noise in Wide-Band Carrier Systems," *Ericcson Technics* (Stockholm), Vol. 16 (1960), No. 1.

World Radio History

the Lenkurt Demodulator

VOL. 7 NO. 3 MARCH 1958

MEASURING POWER AND FREQUENCY

At 6000 Mc

Because the wavelength is only about 2 inches, and because waveguide is used extensively in place of more conventional equipment components, the methods used for making measurements in the vicinity of 6000 Mc are considerably different from those used at lower frequencies. In carrier equipment, or in radio equipment operating at lower microwave frequencies, many measurements are made by connecting meter leads directly across points on the transmission path. This technique cannot be used with waveguide; instead, special apparatus must be employed to sample the energy passing through the waveguide.

Some of the special test instruments and measuring techniques used to measure power and frequency in waveguide circuits are discussed in this article.

Proper lineup and adjustment of microwave radio equipment requires that measurements of power and frequency be made at various points in the r-f circuits. When these measurements are made where waveguide is the transmission medium, some means must be provided to gain access to the energy passing through the circuit. Because the circumstances are not the same as at lower frequencies, microwave measurements require different types of test equipment. The individual items of test equipment utilize components which are considerably different in appearance from those used with carrier equipment. In this article, emphasis is placed on components of microwave test equipment. Among the individual components described are resonant cavities, crystal diodes, directional couplers, and bolometers.

Commercially available microwave test sets include the various components needed to make power and frequency measurements on waveguide circuits. The frequency meter portion of these test sets is normally based on the use of a built-in resonant cavity; the power meter portion is based on the use of a temperature sensitive device called a "bolometer." Test sets also normally supply a source of microwave test signal. A typical microwave test set is shown in Figure 1. Where test sets are not available, measurements can be

World Radio History

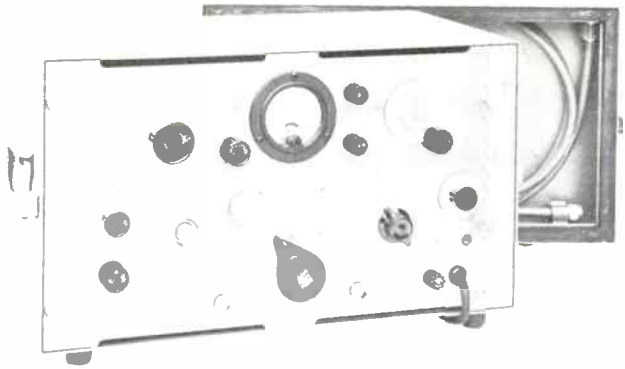


Fig. 1. A typical test set for measuring power and frequency on waveguide circuits.

made by setting up various arrangements of test equipment components.

Resonant Cavities

A typical resonant cavity of the type commonly used as a cavity wavemeter (to measure frequency) is shown in Figure 2. Is it essentially a cylindrical metal enclosure, one end of which can be moved by means of a micrometer screw adjustment. The micrometer scale normally reads in units of length. The reading is converted to frequency by reference to a calibration curve. Theoretically, a cavity wavemeter can be calibrated from its dimensions. In practice, however, the calibration is done by comparison with a standard meter. Although cavity dimensions can vary slightly with temperature, most wavemeters, once calibrated, will maintain sufficient accuracy under normal conditions to meet most field requirements. For critical applications, where extreme temperature variations are likely to be encountered, resonant cavities are made of materials (such as invar) with extremely low temperature coefficients. A crystal detector, an indicating meter, and a cavity wavemeter provide a means of measuring frequency equivalent to that provided in a standard microwave test set.

How a Resonant Cavity Works

A shorted quarter-wave transmission line is actually a resonant circuit. If two such resonant lines are connected in parallel, the resonant frequency is unchanged. In fact, connecting any number of shorted quarter-wave lines in parallel from the same two points

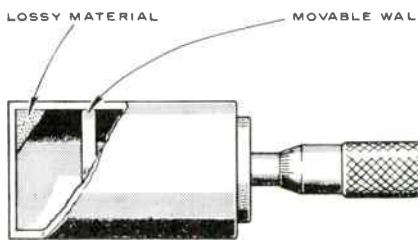


Fig. 2. A typical resonant cavity.

World Radio History

will not affect the resonant frequency. As more and more lines are connected in parallel, eventually a closed metal container will be formed—and this container is the simplest example of a resonant cavity. An example of this is shown in Figure 3.

A practical resonant cavity consists of a closed waveguide section with one dimension equal to an integral number of half-wavelengths of the resonant frequency. It can be excited in the same manner as any waveguide—i.e., by induction through a slot or from a probe inserted at the proper location.

Cavity Wavemeters

Resonant cavities used for frequency measurements are called *cavity wavemeters*. They are adjustable, and are calibrated to measure all frequencies within a certain band.

A special application of a cavity wavemeter, permanently adjusted to resonate at one frequency only, is called a *reference cavity*. In place of a micrometer dial, a reference cavity usually has only a simple screw adjustment which normally is set and locked once the cavity is tuned.

Three different types of cavity wavemeters are available—transmission, reaction, and absorption. They differ in the manner in which the resonant cavity is coupled to the waveguide. An example of each type is shown in Figure 4.

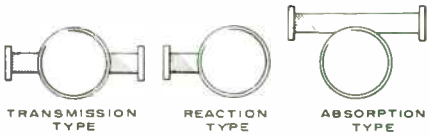


Fig. 4. Three types of cavity wavemeters.

Transmission- and absorption-type wavemeters are used with the signal source at one side of the cavity and a detector at the other. Resonance is determined by means of a microammeter connected to the detector. With the transmission type, the meter will peak at resonance, while with the absorption type, the meter will dip. Absorption-type wavemeters transmit maximum power at frequencies far from resonance. They can, therefore, be inserted directly into a transmission line and detuned when not in use. A transmission-type wavemeter, however, transmits maximum power only at resonance; it must be used with a directional coupler or some other arrangement that permits it to be removed from the transmission system when it is not in use. If left in the line, a transmission-type wavemeter would cause the r-f level to fluctuate with frequency changes.

Reaction-type wavemeters indicate resonance by a change in magnitude and phase of the reflection coefficient. The resonant frequency is best determined by instruments capable of detecting a

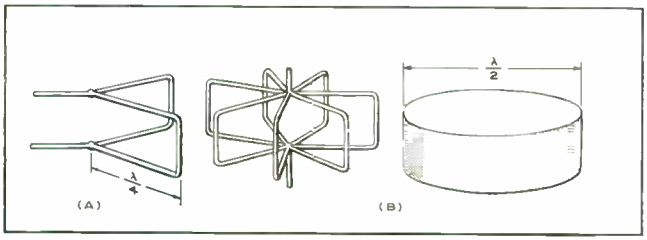


Fig. 3. Development of a Resonant Cavity.

World Radio History

change in phase. In addition to frequency measurements, the reaction-type cavity is often used as a reference cavity. The reference cavity in the transmitter afc circuit of the Type 74A Microtel system is essentially a reaction-type wavemeter. Because of its critical application, this cavity is made of invar.

Crystal Diodes

A crystal diode provides a convenient means of converting r-f energy in a waveguide into a measurable quantity. The crystal diode is connected to a probe (which may actually be a part of the crystal element) or loop which, in turn, is coupled to the electric or magnetic field in a waveguide. The sample of r-f energy intercepted by the probe is rectified by the crystal and can be measured with a microammeter.

Where the r-f energy in a waveguide is amplitude-modulated, a crystal diode can function as an AM detector. This characteristic is utilized in some types of measurements by deliberately inserting an AM signal into the transmission system. This signal can then be detected, amplified, and displayed on an oscilloscope to determine whether the desired characteristic has been obtained.

Bolometers

R-f power passing through a waveguide is normally measured by means of a *bolometer*. This device consists of a temperature-sensitive *bolometer element* in a *bolometer mount* which provides a means of connection to the waveguide The bolometer element forms one leg of a bridge circuit. Since resistance of the bolometer will vary in accordance with the amount of power absorbed, power measurements can be made by determining the degree of balance of the bridge circuit.

There are two general classes of bolometer elements: (1) *barretters,* which have positive temperature coefficients and, (2) *thermistors,* which have negative coefficients. Any short piece of fine wire is a simple barretter. Instrument fuses are often used for the purpose, although the normal type of commercially available barretter consists of a piece of extremely fine platinum wire enclosed in a suitable capsule. Thermistors are made of metallic oxide materials selected because their resistance decreases as the temperature increases.

A bolometer can be mounted to absorb power directly from a waveguide, or from a probe which samples the r-f energy in the waveguide, coupling a definite portion of it to the bolometer.

Mounts

A crystal diode mount or bolometer mount provides means for coupling the element to the r-f energy, and for matching the impedance of the element to that of the transmission system. The mounts may be shorted sections of waveguide, connected to the transmission system by means of directional couplers, or they may be coaxial, and connected to the waveguide by means of a coaxial jack and probe.

A waveguide mount is essentially a short section of waveguide closed (shorted) at one end. Waveguide mounts can either be fixed to operate over a specific band of frequencies, or they may be tuned. Tunable mounts have tuning stubs which may be adjusted to exactly match the impedance of the element to that of the guide.

World Radio History

The bolometer element is held inside of the waveguide by the waveguide mount. Crystal diodes are mounted outside the guide, with the probe projecting into the waveguide.

Coaxial crystal diode and Bolometer mounts are also available, and are often used where a coaxial jack and probe are built into the main waveguide. Coaxial connectors are provided at each end of the mount so that, when connected to suitable instruments, the r-f power can be measured.

Directional Couplers

A directional coupler is a device used to sample the r-f energy traveling in one direction in a transmission system, with a minimum of interference from the r-f energy traveling in the other direction. There are two general classifications applied to waveguide directional couplers, the *multi-hole coupler* and the *cross-guide coupler*.

Multi-hole couplers are most often used for precision laboratory measurements and are sometimes called precision directional couplers. A typical multi-hole coupler is shown in Figure 5. It consists essentially of two parallel waveguide sections that have a common wall throughout most of their length. The main section is flanged at both ends and in use is a part of the r-f transmission system. The secondary section is used for measurement purposes. For example, a bolometer or crystal diode mount may be connected to the flanged end. The secondary section has a load termination at the blind end.

Wave energy traveling through the main section is induced into the secondary section through holes in the common waveguide wall. The holes are arranged in such a manner that wave energy in the main waveguide induces a wave traveling in the same direction in the secondary waveguide. An oppo-

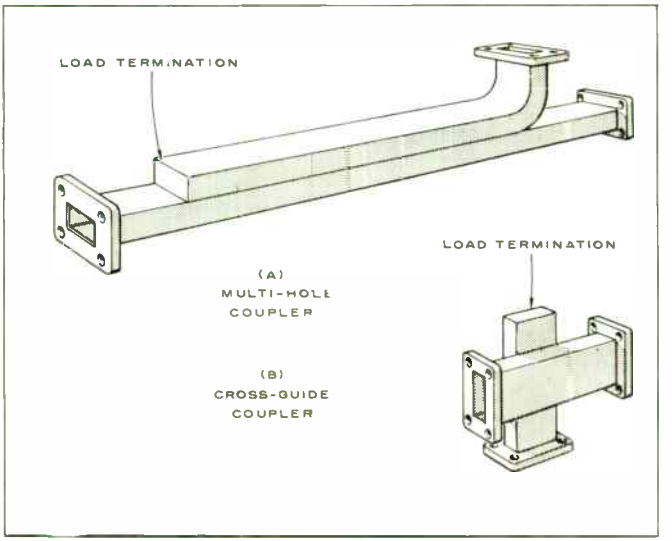


Fig. 5. Directional Couplers: At top, a typical example of a multi-hole coupler; at bottom, a typical example of a cross-guide coupler.

World Radio History

sitely directed wave is also induced in the secondary, but the energy of this wave is relatively small compared with the energy in the main waveguide.

There are two waves in the main guide. The wave carrying energy toward the load is called the preferred wave. Measurements in the secondary waveguide are normally related to the preferred wave. The preferred wave couples energy to the secondary flange, and the oppositely directed wave couples to the secondary load termination.

The power ratio between the preferred wave energy in the main guide and its component at the secondary flange is called the *coupling factor*, and is expressed in decibels.

A small amount of the power appearing at the secondary flange may be due to the energy of the oppositely-directed primary wave. The power ratio between the desired wave at the secondary flange and this undesired wave is called the *directivity*, assuming primary waves of equal magnitude. The directivity is expressed in decibels.

The definitions of coupling and directivity apply also to the cross-guide coupler, which is commonly used where there are space restrictions and where laboratory accuracy is not required. An example of a cross-guide directional coupler is shown in Figure 5. Although the axis of the waveguide sections which comprise the cross-guide coupler are at right angles, the operation is similar to that of the multi-hole coupler. Wave energy from the main guide is coupled into the secondary, and the direction of wave energy depends upon the direction of energy flow in the main guide.

Cross-guide couplers are available with flanges at both ends as well as with a flange at one end and a load termination at the other end. An advantage of a cross-guide coupler with a secondary load termination is that reflections from the termination to the secondary flange are minimized.

The secondary flanges should be covered by shorting plates or by waveguide terminations when they are not being used. Normally, shorting plates will be sufficient to prevent radiation which would cause deterioration in service. However, in some applications, matched terminations are used to provide optimum operation.

Typical values of coupling are 3 to 20 decibels for multi-hole, and 20 to 30 decibels for cross-guide couplers. Typical directivity for multi-hole couplers is 40 decibels or better, and for cross-guide couplers, 20 decibels or better.

Test Set

The components necessary for frequency and power measurements are incorporated in commercial microwave test sets. In addition to a cavity wavemeter, crystal detector, bolometer and power meter, a microwave signal generator is included in the test set, which greatly increases its versatility. The signal generator is similar to signal generators used at lower frequencies in that the oscillator may be tuned over a specific range, and the output power is adjustable. The signal generator is used where test and adjustments in a microwave system require a test signal of known value. Among the measurements which can be made are transmitter deviation, path loss and receiver sensitivity.

World Radio History

When power or frequency measurements are to be made, the waveguide energy is sampled by means of a probe or directional coupler, and is connected to the test set through a coaxial lead. The power meter includes a self balancing bridge circuit with the bolometer in one leg of the bridge. R-f power is read directly on the indicating meter. Frequency is measured in terms of the reading on the micrometer dial. To simplify field measurements, probes and coaxial jacks are often built into key spots in the waveguide system of microwave terminals and repeaters.

Frequency Measurement

The basic equipment for frequency measurements in the field include a cavity wavemeter, a crystal detector and mount, a potentiometer and a microammeter. These may be part of a microwave test set, or they may be used separately.

The sample of waveguide energy is applied to the cavity wavemeter. The crystal diode rectifies the energy so that an indication may be obtained on the microammeter. Rather than applying the rectified energy to the microammeter directly, a potentiometer is used to divide the voltage. This helps to prevent damage to this sensitive instrument. The cavity wavemeter is then adjusted to obtain the desired meter indication, either a peak or dip depending upon the type of cavity wavemeter used.

An oscilloscope can be used to facilitate the measurement. When connected to the output of the detector, the pattern on the oscilloscope (a straight line at frequencies off resonance) will jump as the resonant frequency is passed. This permits a rough frequency measurement to be done quite easily and rapidly. Once the rough micrometer setting has been made, the oscilloscope is replaced by the microammeter for the final measurement.

Power Measurement

For power measurement, a bolometer element and mount, and a microwave power meter are required. The power meter will include three legs of the bridge circuit, bridge power supply and the indicating meter. The bolometer element is the fourth leg of the bridge. R-f energy reaching the bolometer element changes its resistance and this tends to change the bridge balance. Self-balancing bridges are normally used with a power meter. The energy required to maintain the bridge balance is equal to the energy absorbed by the bolometer. The indicating instrument is direct reading and is calibrated in milliwatts or dbm.

Conclusion

While the components and measurement technique used in microwave measurements are somewhat different from those used at carrier frequencies, the quantities to be measured are not changed. In general, more precautions and a little more care must be exercised in making tests at microwave frequencies. However, they are not difficult to make, and only require an understanding of the measurement technique to be used and of the operation of the test equipment. Detailed descriptions of microwave test equipment and measurement techniques are included in the catalogs and instruction manuals of test equipment manufacturers.

World Radio History

World Radio History

the Lenkurt Demodulator

VOL. 12, NO. 5 MAY, 1963

ANTENNA SYSTEMS for microwave

Part One

An efficient antenna system is a vital part of any successful radio system. This is particularly true of point-to-point microwave, where the low power output and the high propagation losses combine to make highly directive antennas necessary. Many specialized antenna arrangements have been developed to solve specific problems, but compromises are usually necessary. Sometimes the performance of a highly specialized design must be compromised to adapt the antenna to more general usage. And a compromise between performance and cost is almost inevitable. This article discusses the characteristics of several types of microwave antennas used in communications and other applications.

Any radio system requires some sort of transmitting antenna to radiate energy into space and a receiving antenna to collect as much of this energy as possible. The efficiency of the antenna system depends on how much of the transmitted energy can be retrieved by the receiving antenna; and the amount of this received energy depends on the characteristics of both antennas.

Broadcast radio has a low efficiency because it must radiate energy more or less equally in all directions; thus, any one receiving antenna can pick up only a tiny fraction of the radiated energy. To overcome this low efficiency, the broadcast station must transmit a large amount of power. By contrast, a point-to-point microwave system radiates only a small amount of power — but it uses a directional transmitting antenna to concentrate the power into a narrow beam directed toward the receiving antenna. The receiving antenna must also be highly directional to enable it to collect as much of the incoming signal as possible and to reject radiation from other directions.

This directional property of the antenna is of vital importance to the sys-

World Radio History

tems engineer and is measured in terms of *gain*. Antenna "gain" results from the directivity of the antenna and is used as a figure of merit for the antenna. It is usually defined as the ratio of the maximum radiation intensity in a given direction to the maximum radiation intensity in the same direction from an *isotropic radiator*. (An isotropic radiator is an "ideal antenna" which radiates equally in all directions; it cannot be realized in practice, but it serves as a convenient performance reference).

Antenna gain increases the effective power of a transmission just as surely as does amplifier gain. A distant observer located along the beam would receive as much signal power from an antenna with a gain of 30 db, radiating 1 watt, as he would from an isotropic radiator at the same distance with an output of 1,000 watts. The effectiveness of the transmission is similarly increased by the gain of the receiving antenna. In fact, all the properties of an antenna are the same whether it is used for transmitting or for receiving.

Closely related to the gain of an antenna is its *beamwidth*, usually defined as the angle between the "half-power points," the points where the radiated power is 3 db down from maximum. The beamwidth may be specified both vertically and horizontally to describe the beam shape in three dimensions. Many different beam shapes are used for various purposes. For example, some radars use a "fan-shaped" beam to increase the volume of the sky covered with each scan of the antenna. Point-to-point communication systems normally use "pencil-shaped" beams to concentrate the power as much as possible toward a single point. Figure 1 illustrates typical radiation patterns of several types of antennas.

Directional Antennas

No directional antenna concentrates *all* its radiated power into the desired beam. For various reasons (such as scattering from feed or supporting structures, diffraction at the antenna edge, and the finite number of wavelengths across the aperture diameter), energy is robbed from the main beam to produce secondary beams or "lobes" which radiate in various directions away from the desired axis — even to the rear. These side lobes and back

Figure 1. Idealized representation of radiation patterns produced by isotropic, dipole, and parabolic reflector antennas.

World Radio History

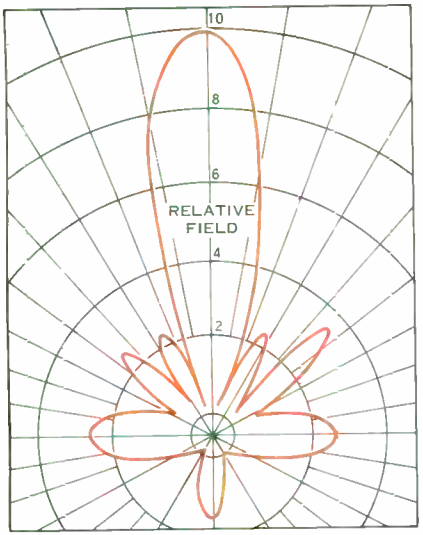


Figure 2. Typical radiation pattern of a parabolic reflector antenna. Majority of the power is concentrated in the main beam, but some power is wasted in the unwanted back and side lobes.

lobes represent wasted power and may be a source of interference with other services.

One method of providing antenna directivity is to use arrays of radiating elements. These are complex arrangements in which the characteristics of many simple radiators add up to achieve directivity. The radiation pattern of an array is controlled by its shape and size (in terms of wavelengths), and by the phase relationship between the individual elements. With careful design, gains of over 30 db are achieved with arrays, but at the expense of very complex distribution arrangements to serve the many elements required.

Arrays are not generally used for point-to-point microwave communication because of the complexity necessary to achieve the required gain and because the mechanical tolerances become difficult to attain at frequencies above 2 gc (2000 mc). Arrays composed of radiating slots cut in a plate or waveguide, rather than of radiating conductors, have recently achieved more prominence. This type of array is sometimes used as a "feed" arrangement to illuminate a microwave reflector.

One of the most attractive properties of microwaves, from the antenna designer's viewpoint, is that they follow many of the rules of optics. Microwaves can be focused into a narrow beam by various lens or reflector arrangements. In both, the lens or reflector must be illuminated by a primary radiator, in the same way that an optical lens or reflector must be illuminated by a light source. In a microwave system this primary illumination is most often accomplished by an electromagnetic horn — a flared end on a waveguide (which itself is an efficient — but sometimes cumbersome — antenna). Waveguide arrays are also used to provide illumination, and, at lower frequencies, dipoles may be used.

A microwave lens is a device for collecting and focusing divergent radiation into a parallel beam by refraction (bending), in much the same way that an optical lens focuses a light beam. Microwave lenses may take a variety of shapes including disks, cylinders, and spheres. Lenses are often constructed of dielectric materials such as polystyrene, but they may also be made of metal plates set parallel (the so-called "waveguide" lens). The main advantage of a lens is that it is fed from the back; thus, there is no front feed structure with its attendant mechanical problems and aperture blocking. A high "front-to-back ratio" (the ratio of power in the beam to the power scattered to the rear) can be achieved with a lens because the rear feed radiates en-

World Radio History

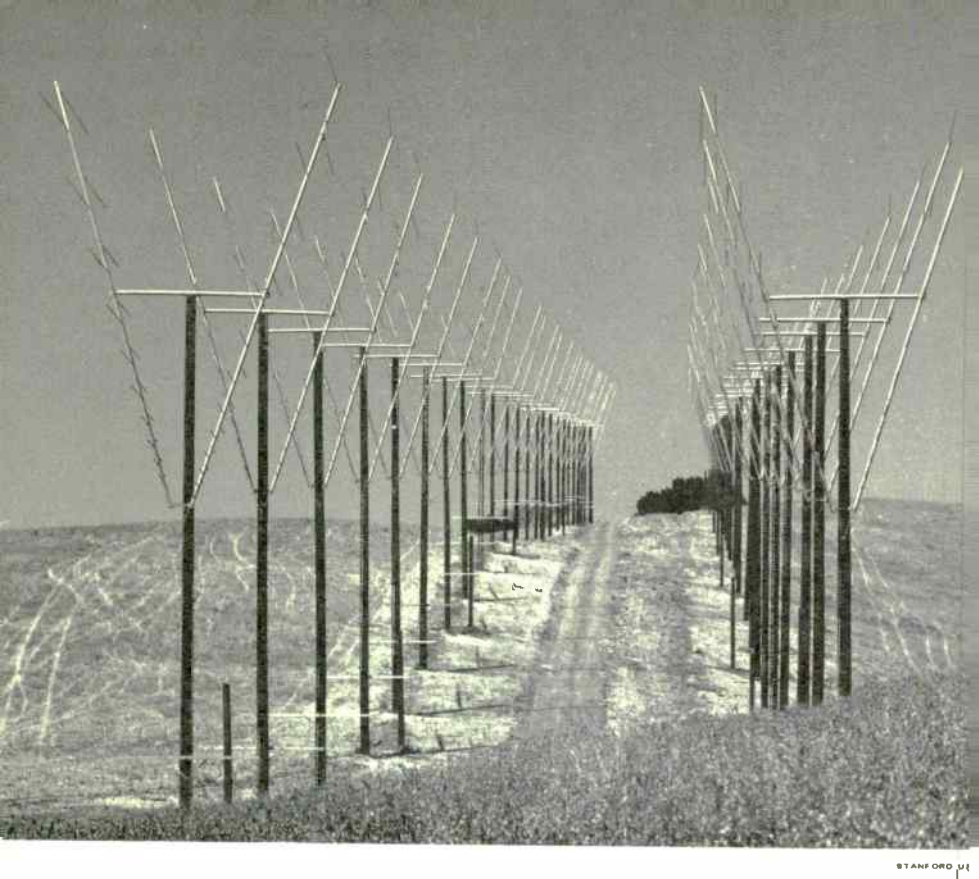


Figure 3. An array of 50 log-periodic antennas used for radar explorations of the sun, moon, and planets.

ergy in the same direction as the lens. However, the lens inevitably has reflections at the surfaces and losses in the dielectric material, resulting in an insertion loss of perhaps 1 to 3 db. Thus, the gain is a little lower than that of a reflector and the side lobes of the lens are usually larger. The lens is also more difficult to design, although once designed, it can tolerate greater surface errors than the reflector.

Reflectors

Radio beams can be formed by reflection as well as by refraction. A shaped reflector is used in the majority of communications applications where a high-gain antenna is required; and by far the most widely used reflector shape is the parabola. Many parabolic reflector arrangements are possible, but probably the most common is the paraboloid of revolution illuminated by a center-feed waveguide horn located at the focus of the paraboloid.

The gain of an antenna depends on its size, measured in wavelengths. Gain increases as the wavelength becomes shorter or the antenna becomes larger, according to the relationship

$$Gain = k\left(\pi \frac{diameter}{wavelength}\right)^2,$$

World Radio History

where k is an efficiency factor, usually about 55% in microwave applications. Thus, gain increases very rapidly as the dish diameter is increased or the wavelength is shortened. The practical significance of this is that higher frequencies must be used to attain high gain if the antenna size must be kept small. As an example, assume a 6-foot reflector and an operating frequency of 6 gc (6000 mc). The wavelength at this frequency is 0.164 ft. Then,

$$G = 0.55 \left(\pi \frac{6}{0.164}\right)^2$$

$$= 7270.$$

Expressed in db, this power ratio is 38.6 db. At 1 gc a 36-foot reflector would be required to achieve the same antenna gain.

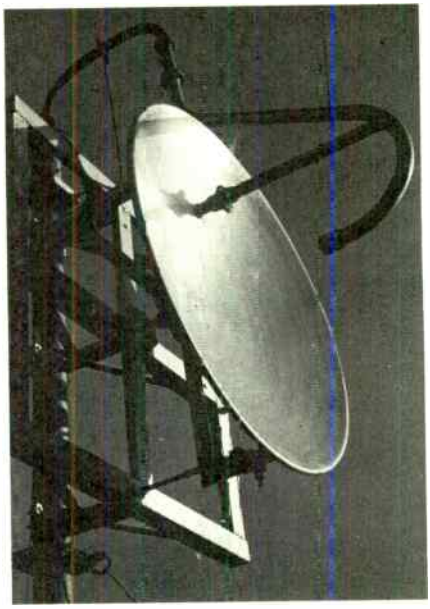


Figure 5. The center-fed paraboloid of revolution is the antenna most commonly used for point-to-point microwave transmission.

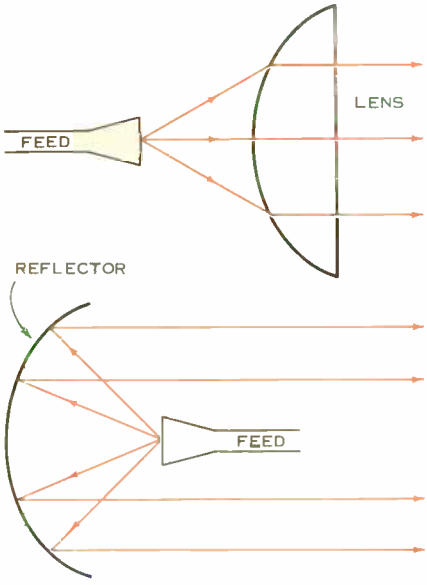


Figure 4. Microwaves follow most of the rules of optics. Lens is illuminated from behind, eliminating aperture blocking by the feed, but energy is lost in passing through the lens.

Since antenna gain is really a result of its directivity, beamwidth can be calculated using a similar formula. Antenna beamwidth, in degrees between the 3-db points, is given by the approximate relationship

$$\text{beamwidth} = \frac{(70^\circ)\,(\text{wavelength})}{\text{diameter}}.$$

For the same 6-foot antenna and the same 6-gc frequency,

$$\text{beamwidth} = \frac{(70^\circ)\,(0.164)}{6}$$

$$= 1.9 \text{ degrees}.$$

The antenna designer strives for a low side lobe level as part of his effort to design an efficient directional antenna. The side lobe levels of a parabolic

World Radio History

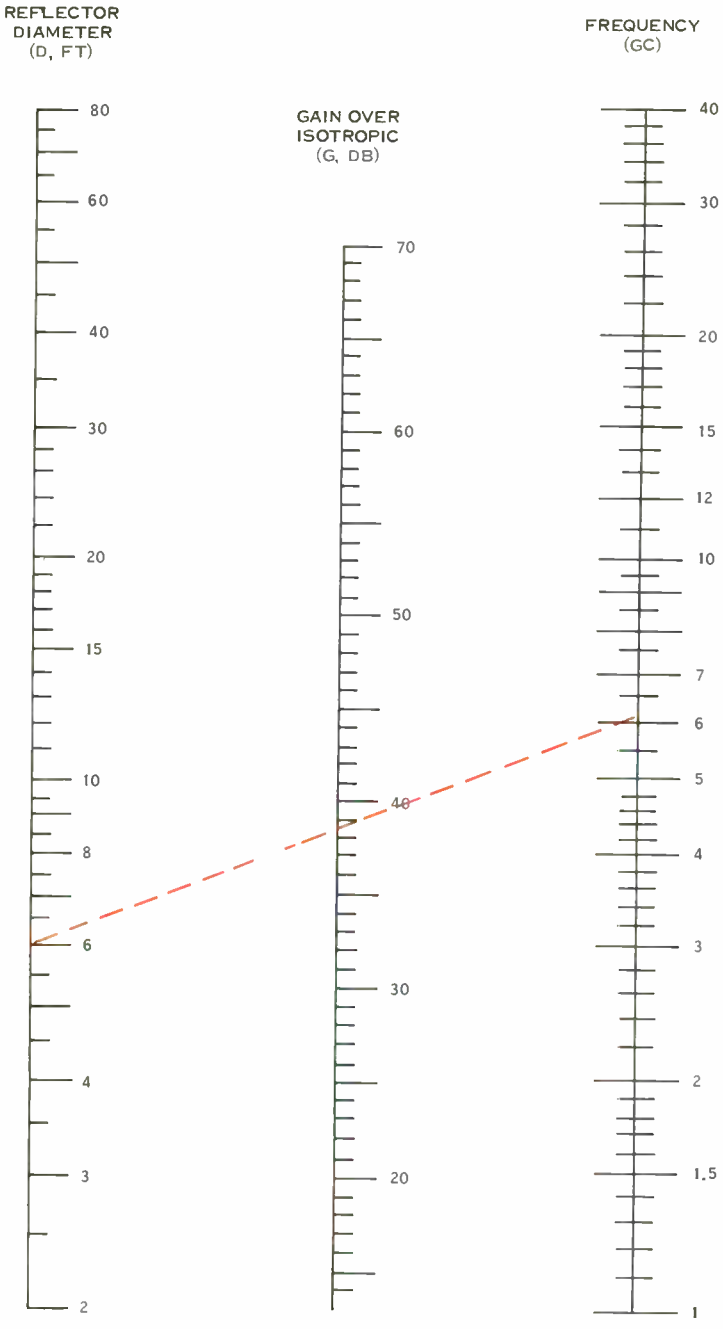


Figure 6. Gain of a parabolic reflector antenna with reference to an isotropic antenna (assumed efficiency of 55%).

World Radio History

reflector are controlled mainly by the feed pattern, not by the dish itself. One method used to reduce the side lobe level is to taper the illumination. That is, the illumination power at the outside edges of the reflector is made lower (usually about 10 db) than that at the center. This reduces the energy scattered to the side and rear by the dish edge, thus improving the front-to-back ratio. Other factors which affect the side lobe level are the scattering produced by the supporting structure, and the shape of the feed horn itself.

There are two main problems which arise from the use of a center-fed horn to illuminate a reflector: (1) the feed horn structure itself blocks a portion of

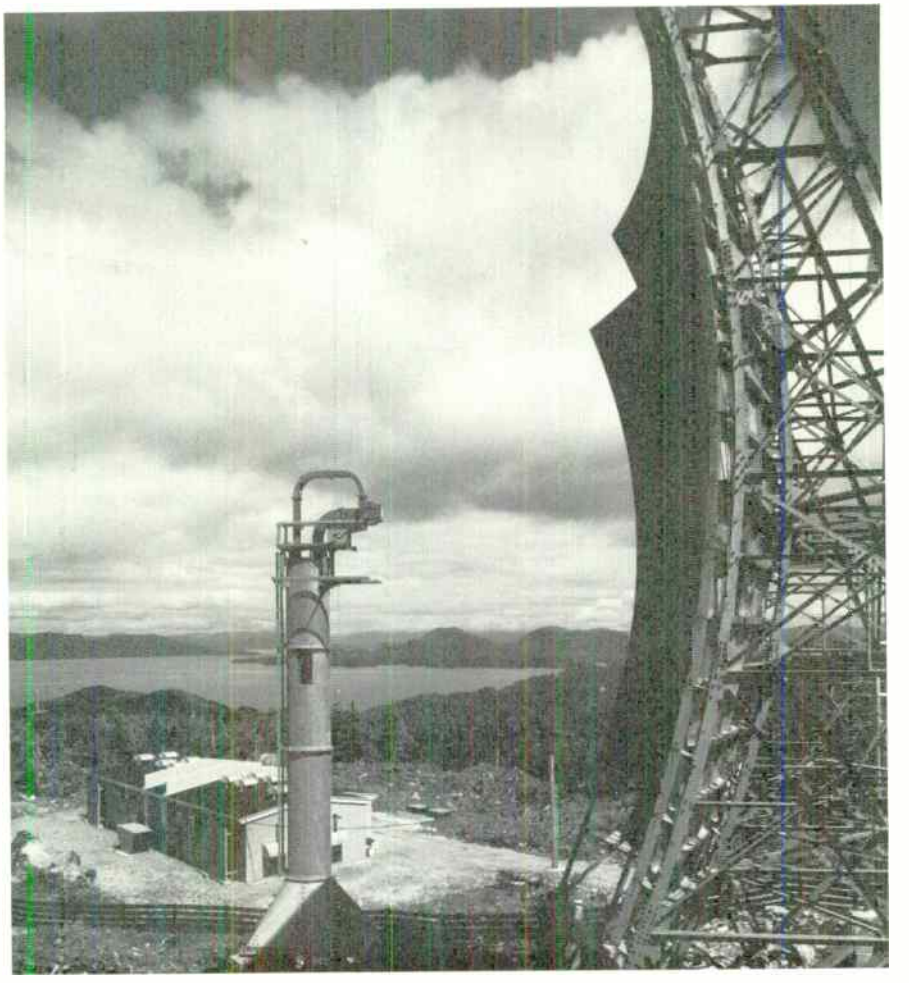


Figure 7. This 60-foot reflector is part of a tropospheric scatter system installed by Lenkurt at Trutch Island, British Columbia. Such antennas, with a gain of 39 db at 900 mc, permit "hops" far longer than line-of-sight distance by using the troposphere as an intermediate reflector.

World Radio History

the reflector aperture, and (2) some of the energy is reflected directly back into the feed horn to create a standing wave, which causes distortion and loss of efficiency. Both these problems, of course, depend on the relative size of the feed horn. One way to eliminate such problems is to use a parabolic section, rather than a paraboloid of revolution, for the reflector. This allows the feed to be located at the focus of the parabola, but offset from the aperture of the antenna. This eliminates both the aperture blocking and the direct reflection of energy back into the feed.

An extreme example of the use of such a parabolic section is becoming increasingly popular for microwave communication systems. This is the horn-reflector antenna (sometimes called the "cornucopia"). In this antenna, the feed horn and the reflector are fabricated as one structure, but it can be considered as simply a horn feeding a parabolic section. One system uses horn-reflectors for the simultaneous transmission of signals in the 4-gc, 6-gc, and 11-gc bands, with gain ranging from about 39 db at the lowest frequency to more than 47 db at the highest frequency. Because the sides of the feed horn extend to the reflector, edge "spill-over" is reduced; this results in lower side lobes and a much higher front-to-back ratio than ordinary parabolic reflectors. This is perhaps the biggest advantage of the horn-reflector; front-to-back ratios of better than 70 db have been reported.

In some applications, the waveguide and feed structure of the antenna system becomes large and complicated. A good example of this occurs in space communications, where low-noise maser or parametric preamplifiers must be placed quite close to the antenna feed in order to avoid the losses which occur in a long waveguide run. These preamplifiers often require elaborate cooling to keep the noise level low. This usually requires a rather large feed structure at the focus of the paraboloid, which may cause excessive scattering and blocking of the beam. It also complicates the mechanical arrangements for running and cooling the amplifiers.

As a way around this problem, antenna designers have borrowed a technique from optical telescope design. This technique, known as the *Cassegrain* design, permits the antenna to be fed from behind the dish, with the horn protruding through the center of the reflector to illuminate the convex side of a hyperbolic subreflector, as shown in Figure 9. The reflection from the

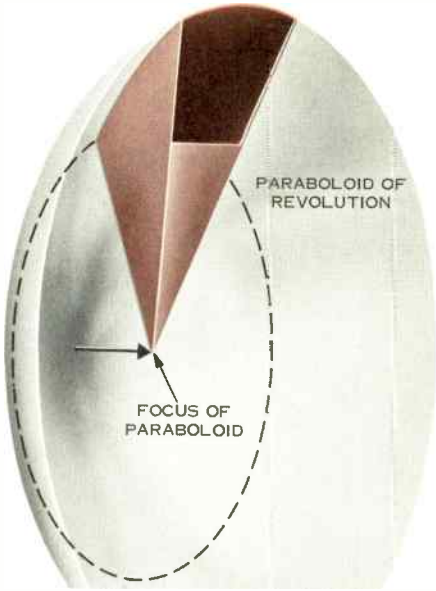
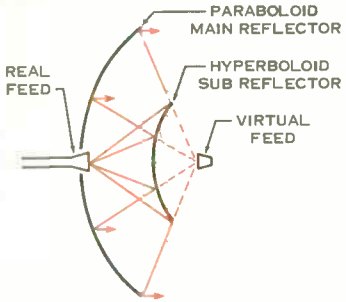


Figure 8. The horn-reflector antenna consists of a parabolic section fed from the focus of the paraboloid. Because the horn extends to the reflector edge, "spill-over" is reduced, side lobe levels are low, and a high front-to-back ratio is achieved.

World Radio History

Figure 9. The Cassegrain antenna system is used primarily in space communications where low-noise preamplifiers are often placed close to the feed. Cassegrain system permits the antenna to be fed from (and the preamps to be mounted) behind the main reflector.



subreflector illuminates the main dish, which focuses the energy into a beam. With proper design of both the hyperbolic subreflector and the parabolic main reflector, the main dish sees a "virtual feed" located at its focus, and all path lengths are the same from the feed horn to a distant point in space.

The biggest disadvantage of the Cassegrain feed system is the amount of aperture blocking introduced by the subreflector. This can be minimized by enlarging the feed horn and reducing the size of the subreflector. The feed horn is then extended forward so that equal shadows are cast by the feed and the subreflector.

The Mechanics of Antennas

As might be expected from an analogy with light, surface irregularities in the reflector tend to defocus the beam — side lobe levels are increased, gain is reduced, and the beamwidth is increased. The magnitude of these effects depends primarily on the size of the surface errors — in terms of wavelengths at the frequency of interest. Of course, there are limits to machining tolerances, particularly with larger antennas, and some compromising is necessary between performance and economics. As a general rule surface errors should not exceed 1/16 the wavelength. This tolerance is not difficult to achieve

World Radio History

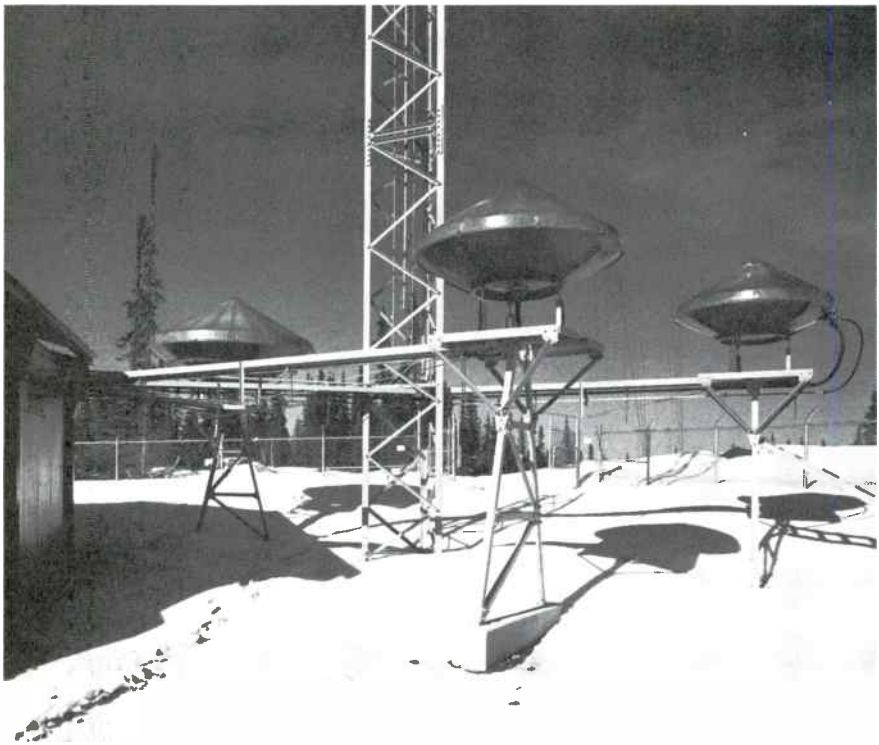


Figure 10. Radomes are sometimes used to protect antennas from weather and debris. The problem is particularly severe where reflectors are mounted horizontally, illuminating a passive reflector which redirects the beam horizontally to a distant site.

at low frequencies or with small reflectors, but it becomes a real problem with large reflectors at high frequencies. For example, at 6 gc the tolerable surface error is of the order of ±0.12 inch, while at 11 gc it is about ±0.065 inch. Achievable error tolerance is approximately proportional to the reflector diameter. Thus, a 60-foot reflector (such as might be used for space communications) would be expected to have surface irregularities about 10 times the size of those on a 6-foot dish, even though both were constructed as carefully as possible. Because increased gain achieved by increasing the reflector size is offset to some extent by the gain loss caused by increased surface error, a practical limit on gain is eventually reached. With present techniques, this limit seems to be about 70 db. Of course, this limit might be extended by improved construction techniques.

One such technique has been adapted from telescope-making experiments of half a century ago. In 1908, astronomers experimented with liquid mercury as a reflector. When rotated on a turntable, the liquid was formed by centrifugal force into a "natural" paraboloid,

World Radio History

producing an accurate mirror. Although the idea was old, modern chemistry was required to adapt it to antenna making. In modern practice, epoxy resin, catalyzed to harden slowly at room temperature, is placed in a shaped container and spun on a turntable at the correct speed to produce a parabola of the desired focal length. The spinning continues until the material is hard. It is sputtered or sprayed with a thin metallic coating to form the reflector. Very good results have been obtained with small reflectors made this way, and a recent report indicated that a 28-foot "spun" reflector had maximum surface errors of less than 0.02 inch and rms surface errors of less than 0.008 inch. Such close tolerances allowed the use of quite high frequencies. At a test frequency of 35.2 gc (wavelength = 0.336 in.), the reported gain was 67.4 db and the beamwidth was 4.4 minutes of arc.

In some locations antennas must be covered to protect them from the weather or from falling leaves and other debris. Such protection, of course, must be accomplished without impairing the electrical characteristics of the antenna. Ideally, this could be provided by a covering of low-loss dielectric material having the same propagation characteristics as free space. This ideal cannot be realized in practice, but actual coverings, or "radomes," approach it closely. Heating of the radome is sometimes required to prevent formation of ice which would degrade antenna performance.

Point-to-point Systems

Because microwave transmission follows essentially a straight line, path length — or the distance between terminals — tends to be limited by the curvature of the earth and by obstructions such as trees, buildings, and mountains. Actually, microwaves require more than just an optical line-of-sight clearance, to allow for both fading and for certain interference characteristics of the beam itself at microwave frequencies.

In order to obtain adequate clearance, it is often desirable to locate the antennas on buildings or towers. In many cases this may be impractical due to the high signal loss and expense contributed by long waveguide runs. Accordingly, it is common practice to employ antennas located conveniently close to the microwave equipment, and some arrangement of passive reflectors which serve to redirect the radio beam in the desired direction, much as in the case of an optical periscope.

Many variations of these compound antenna systems have been developed to overcome various transmission problems imposed by terrain. However, the engineering of such systems may be more complicated than is apparent from the analogy between microwave transmission and light. Some of the problems encountered in designing such compound antenna-reflector combinations will be discussed in the second article in this series, to appear in the July, 1963 issue of THE DEMODULATOR. •

BIBLIOGRAPHY

1. Henry Jasik, Editor, *Antenna Engineering Handbook;* McGraw-Hill, New York, 1961.
2. S. P. Carter and L. Solomon, "Modern Microwaves," *Electronics;* June 24, 1960.
3. R. W. Friis and A. S. May, "A New Broad-Band Microwave Antenna System," *Communications and Electronics;* March, 1958.
4. J. W. Dawson, "28-Ft Liquid-Spun Radio Reflector for Millimeter Wavelengths," *Proceedings of the IRE;* June, 1962.

World Radio History

World Radio History

the Lenkurt Demodulator

VOL. 12, NO. 7 JULY, 1963

ANTENNA SYSTEMS for microwave

Part Two

Without highly directional antennas, modern microwave systems could not achieve their high propagation reliability with such low power outputs. However, just as some antennas are more efficient than others, compound arrangements of antenna elements often provide a more effective and less expensive means of linking the transmitter to the receiver than do simple antennas. Some of the factors involved in the use of these compound systems are considered in this article.

Because microwave transmission follows essentially a straight line, the path between transmitter and receiver must clear any intervening obstructions. The limiting factor may be the curvature of the earth or it may be an obstruction such as a mountain. In either case, one way to achieve additional path clearance is simply to raise the antenna systems at the ends of the path. This can be done in several ways, but they all cost money. If the antenna is placed at the top of a tower, the expense and the transmission impairments of a long waveguide run must be considered, in addition to the cost of the tower.

Of course, if the terminal is located on a mountain top it may not require a tower. Usually, however, the cost is still there — it is no less real because it takes the form of an access road instead of a steel tower. Thus, terminals are placed for convenient access to roads and power lines, and the antenna system must get the signal over or around all obstructions.

This does not mean, however, that the path between transmitter and re-

World Radio History

Figure 1. A "periscope" antenna system can provide path clearance without the expense and transmission impairments of a long section of waveguide.

ceiver must be a single straight line. Since microwaves follow most of the rules of conventional optics, a system of "mirrors," similar to an optical mirror system, can be constructed to reflect the beam over or around an obstruction. Such a system often provides a solution to the antenna system problem, which is essentially one of balancing path clearance, directivity, and transmission loss against economic factors.

"Periscope" Antenna Systems

The simplest and most common reflector system consists of a parabolic antenna mounted at ground level and directed vertically to illuminate a passive reflector at the top of a tower. This

World Radio History

reflector, inclined at 45°, redirects the beam horizontally to a distant site, where a similar "periscope" system may be used to reflect the signal back to ground level. The vertical separation between antenna and reflector may be provided by a specially constructed tower or the reflector may be mounted on a building or other convenient structure, providing two factors are considered: (1) the structure must be high enough to provide adequate path clearance; and (2) there should be an optimum separation between antenna and reflector — signal strength is impaired by either too small or too great a separation. This second point is not obvious, but it has an important effect on the efficiency of the system. In fact, a properly laid out antenna-reflector combination can produce more gain than the antenna alone, even when reflector losses and scattering are considered. ("Gain" results from increased directivity — a "sharper" beam).

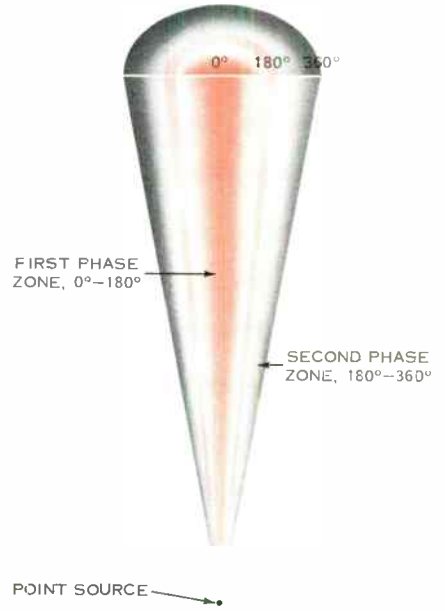


Figure 2. When a plane intersects radiation from a point source, phase addition occurs throughout first phase zone (red). When second zone energy (gray) is intercepted, phase cancellation occurs.

This seemingly paradoxical situation occurs because of a complex interrelationship between the size and shape of both antenna and reflector, the distance between them, and the operating frequency. These are the factors which control the phase relationship of the energy striking the reflector — and this phase relationship is the key to an effective system.

The phase variation at the reflector is illustrated in Figure 2, which shows a flat plane intersecting the radiation emitted by a point source. As the distance from the center of the plane increases, so does the distance from the source, until some point on the plane is reached where the distance from the source is a half wavelength longer than the distance from the source to the center of the plane. At this point the wavefront is 180° out of phase with the energy at the center of the beam. A line joining all such points describes a circle about the center, and all the energy within this circle has an in-phase component. That is, all the energy between 0° and 180° has a common in-phase component at some intermediate angle. As the phase difference exceeds 180°, all the energy has a component which is out of phase with the energy in the first phase zone (0° to 180°). This second phase zone, consisting of out-of-phase energy, extends from 180° to 360°. Additional zones exist, but most of the energy is concentrated in the first two.

Effectively, this means that the reflector will have maximum gain when it intercepts all the energy in the first phase zone, because phase *addition*

World Radio History

occurs throughout this zone. If, however, some of the energy from the second zone is also intercepted, phase *cancellation* occurs and the effective signal power is reduced. Because the maximum in-phase reflection occurs when all of the energy in the first zone — but none of the second-zone energy — is reflected, the optimum reflector size corresponds to the size of the first phase zone. A smaller reflector will not intercept all the in-phase energy, and a larger one will intercept some of the out-of-phase energy, producing cancellation.

But choosing the proper reflector size is not the whole problem. The first phase zone expands as the distance from the antenna increases. (Its diameter is approximately proportional to the square root of the distance.) This means that if antenna and reflector size are fixed, there is an optimum separation. Increasing this separation has the same effect as reducing the size of the reflector.

Furthermore, antenna size also affects the performance. The preceeding discussion considered the antenna as a point source, but in reality the area of the antenna may be a sizeable fraction of the reflector area. The significance of this is that the point source becomes an apparent source somewhere behind the antenna, as shown in Figure 4. This is the point from which the radiation appears to be emanating. Because a larger antenna produces a more nearly parallel beam, this point is farther back and the first phase zone does not expand as rapidly. In other words, for any given antenna-reflector separation, a larger antenna provides a better focus, and a smaller reflector must be used to achieve maximum gain.

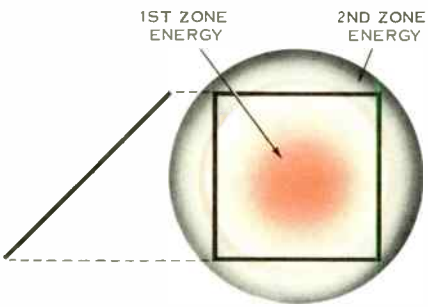


Figure 3. Rectangular reflector illumination seen from antenna. A small reflector misses some first zone energy, while corners of larger reflector catch some second zone energy, thus reducing gain.

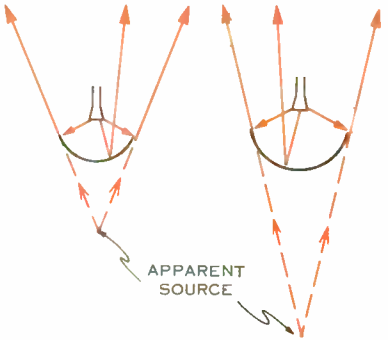


Figure 4. A smaller antenna produces a more divergent beam, causing phase zones to spread more. Hence, a smaller antenna requires a larger reflector to intercept entire first phase zone for maximum gain.

Another aspect of the problem is the *shape* of the reflector. Since the desired first phase zone is round, the reflector should present a circular area to the antenna; corners may project into the second phase zone, catching some out-of-phase energy, and straight sides may miss some of the in-phase energy of the first zone. Because a reflector is normally inclined at 45°, it must be elliptical to project a circular area. In actual prac-

World Radio History

tice, however, reflectors are seldom elliptical because rectangular ones are less expensive. A rectangular reflector is proportioned to project a square area both vertically and horizontally. The inefficiency of this square shape can cost 1 or 2 db of gain if the corners extend into the second phase zone. In practice, however, a reflector usually does not occupy the entire first zone, and the corners increase the effective reflecting area, thereby slightly increasing the gain.

Square corners also have an adverse effect on the side lobe level. An elliptical reflector has a high side lobe level, but the corners of a rectangular reflector increase the scattering, producing an even higher side lobe level. Because of these undesirable effects, it is common practice to remove the corners, forming an octagonal projected area — or two corners only may be removed to form the familiar "fly swatter."

A properly designed antenna-reflector system can produce 2 or 3 db more gain than the antenna alone. This comes about not only because just the in-phase energy is reflected, but also because of another factor. The surface of the re-

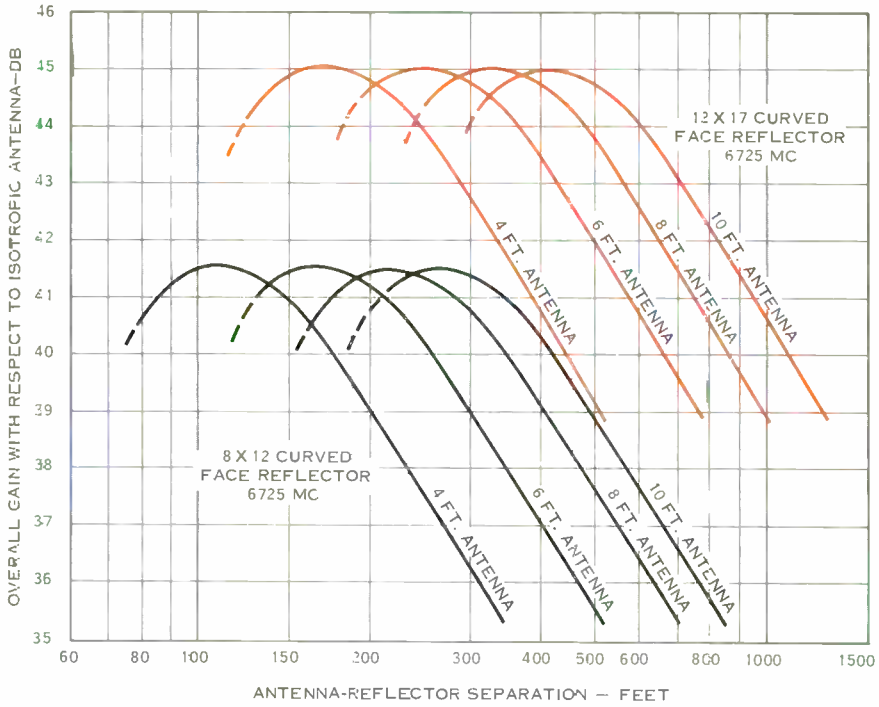


Figure 5. Gain variation of curved-face reflectors with respect to separation between antenna and reflector. The red curves are for various sizes of antennas used with a 12 by 17 foot reflector. Black curves are for an 8 by 12 foot reflector. Larger antenna requires smaller reflector or higher tower for maximum gain.

World Radio History

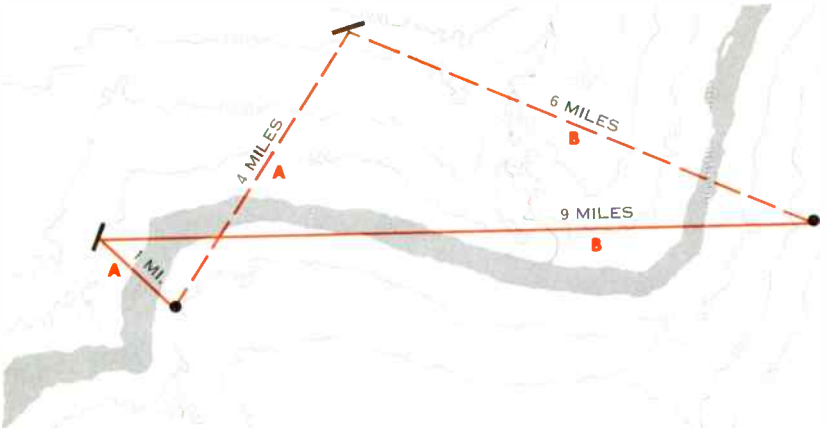


Figure 6. Topographical map showing two possible locations for a passive repeater. The total path length is the same for both locations, but placing the reflector much closer to one terminal results in 7.5 db less path attenuation.

flector is larger than that of the antenna (in most cases) so it acts effectively as an antenna of larger aperture; that is, it produces a "sharper" beam at a distant point. Replacing the reflector with an antenna of similar aperture will provide just as much gain — but the antenna and the long waveguide run would cost much more and would also introduce additional waveguide losses and reflections.

One way to realize more gain from a compound antenna system is to curve the reflector to provide additional focusing. Regardless of the size of the antenna, the beam striking the reflector is always divergent. A reflector with a parabolic curvature acts much like an extension of the antenna — it makes the beam more nearly parallel. Such a reflector would be a small section of a very large paraboloid, with the antenna feeding it from the focus. In actual practice, however, the cost of carefully fabricating and installing a parabolic section of the required size is likely to be prohibitive. One effective technique is to install a flat, but flexible, reflector. Once this is in place, and the system is operating, the reflector is curved by pulling the center of it back with tension fittings. An ideal curvature can be approximated by experimental adjustment to produce maximum gain.

The effect of antenna and reflector sizes and separations on system gain is indicated by the curves of Figure 5. Such curves may be plotted in terms of the relative gain with respect to the antenna alone, but a more useful method is to plot them to show the gain of the combination at various separations, relative to the gain of an isotropic antenna. The antenna and reflector together can then be treated as one large antenna. Figure 5 indicates that for an

World Radio History

8 by 12 foot curved reflector operating at 6725 mc, a maximum gain of 41.5 db can be obtained by using a 4-foot antenna separated from the reflector by 120 feet. The same gain can also be achieved with a 10-foot antenna at 265 feet or with other sizes at various distances.

For comparison, a similar series of curves is shown for a 12 by 17 foot reflector at the same frequency. These curves indicate that the maximum gain of 45 db occurs with a 4-foot antenna at 160 feet or with a 10-foot antenna at 405 feet. Thus, the higher gain available with the larger reflector must be paid for in additional tower height. Or, conversely, if path clearance demands a higher tower, a larger reflector is required for the best performance.

Passive Repeaters

Sometimes a tower cannot provide clearance over an obstruction. For example, if two sites are separated by a mountain, the microwave beam may have to be redirected at one or more intermediate points to get it around or over the mountain. Although repeater stations could be used at these points to amplify and retransmit the signal, *passive repeaters* may be used to merely change the path direction without amplification. These passive repeaters contribute no signal amplification, but they require no power and very little maintenance, so they can be located in places where access is very difficult.

One type of passive repeater uses two parabolic antennas connected back-to-back through a short length of wave-

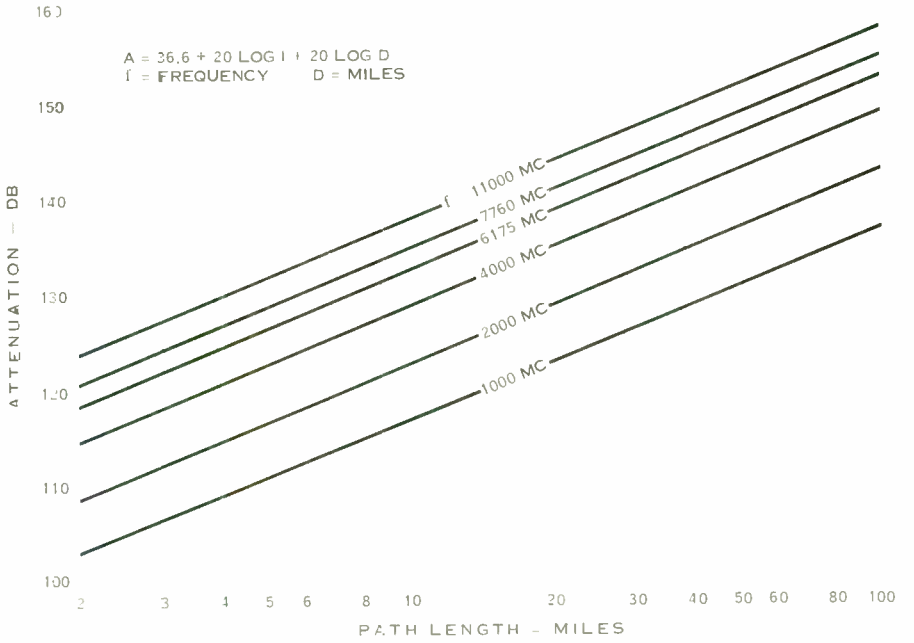


Figure 7. Free space attenuation between isotropic antennas. Attenuation increases by 6 db when path length is doubled, regardless of actual distance increase.

World Radio History

guide. With this arrangement, the beam can be redirected in virtually any direction simply by using an appropriately curved waveguide section. However, this type passive repeater is not widely used because of the losses encountered. The efficiency of each parabolic antenna is typically only about 55 percent, and the waveguide inevitably contributes some loss and reflection, resulting in considerable signal impairment.

A more common type of passive repeater for situations like this is the so-called "billboard" — a large, flat surface which acts simply as a reflector. In a typical system, a billboard repeater might be located at a turn in a valley, effectively "bending" the beam to follow the valley. Such reflectors may be illuminated by a periscope antenna system and may reflect the beam to another periscope system, forming, in effect, an arrangement which resembles a huge mirror system.

The size of a billboard reflector is not subject to the same limitations as the size of the reflector in a periscope system because the passive repeater is usually located far enough from the transmitter so that all the energy intercepted by any manageable size reflector is essentially in phase. Typical reflector sizes range from 6 by 8 feet to 24 by 30 feet.

In contrast to open wire or cable, where the attenuation per mile is con-

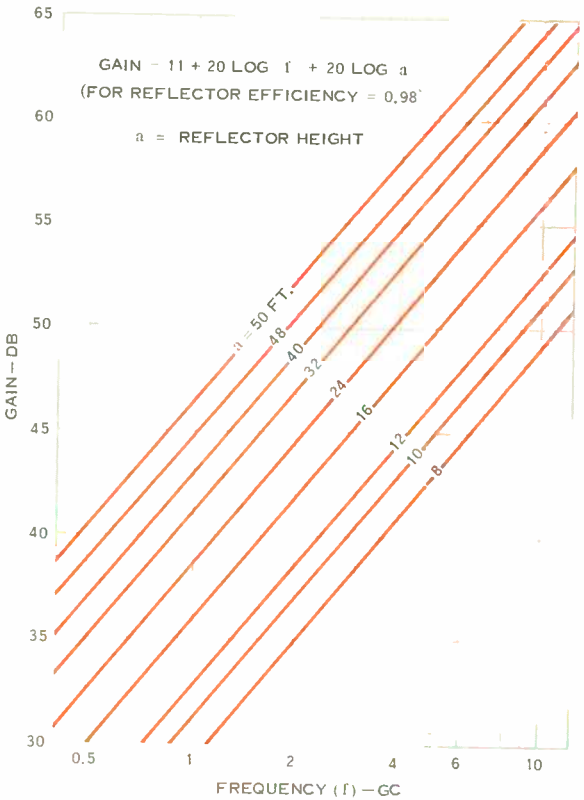


Figure 8. Passive reflector produces gain with respect to an isotropic antenna. Gain depends on size of reflector as well as on frequency of operation.

World Radio History

stant regardless of the number of miles, the attenuation per mile of a radio signal depends on the path length. Signal strength is inversely proportional to the square of the distance from the transmitter. Thus, each time the path length is doubled, the signal strength is reduced to 1/4 of its previous value. This means that path attenuation increases by 6 db each time the path length doubles, whether the actual distance increase is 1 mile or 10 miles.

Because the attenuation is greatest in the first few feet of the path, the relative lengths of the paths have an important effect on the location of a reflector. As an example, consider a passive repeater placed as shown in Figure 6 so that path A is 1 mile long and path B is 9 miles, for a total length of 10 miles. If the repeater is then relocated to give path A a length of 4 miles and path B a length of 6 miles, the total is still 10 miles — but the loss in path A has been increased by a factor of 16, while the loss in path B has been reduced only to 4/9 of the original value. The effect of this is to make it highly desirable to place a passive repeater much closer to one terminal than to the other because the total attenuation for the two paths is highest when they are of equal length.

The same example might be calculated in terms of db, using the formula

$$A = 36.6 + 20 \log F + 20 \log D,$$

where

A = *free space attenuation between isotropic antennas in db,*

F = *frequency in mc,*

and

D = *path length in miles.*

For an operating frequency of 6,000 mc, path A would originally have an attenuation of 112.2 db and path B would have 131.3 db, making a total attenuation of 243.5 db. After the relocation, path A would have 124.2 db and path B would have 127.8 db, for a total of 251.0 db. Thus, moving the reflector 3 miles closer to the midpoint has increased the total path loss by 7.5 db, even though the total path length has not changed. This means that the power reaching the receiver has been reduced by 82 percent.

If the choice between these two locations for the passive repeater is made entirely on the basis of path loss, the reflector will be located to make the two paths 1 mile and 9 miles long, respectively. The path loss of 243.5 db represents the loss between two isotropic antennas, without considering the directivity of either of the antennas at the terminals or of the billboard reflector. If the terminal antennas are 6-foot parabolas, each will have a gain of about 38 db, thus reducing the path loss by a total of 76 db. From the curves of Figure 8, a passive reflector with a projected area of 14 by 14 feet is shown to have a gain of 50 db. This effect occurs because the reflector is considered to receive and retransmit the energy. Since the gain is a measure of directivity, the billboard is 50 db better than an isotropic antenna at receiving and 50 db better at transmitting. Thus, the reflector subtracts another 100 db from the path loss. The result, 243.5 db minus 76 db minus 100 db, is 67.5 db — the actual difference between the transmitting level and the receiving level.

Such path loss calculations really amount to an "accounting" procedure — a convenient way of calculating the performance of a reflector in a transmission path. Another way to look at is to consider the entire distance as one path. In this case the reflector is *not* considered as a receiver and retrans-

World Radio History

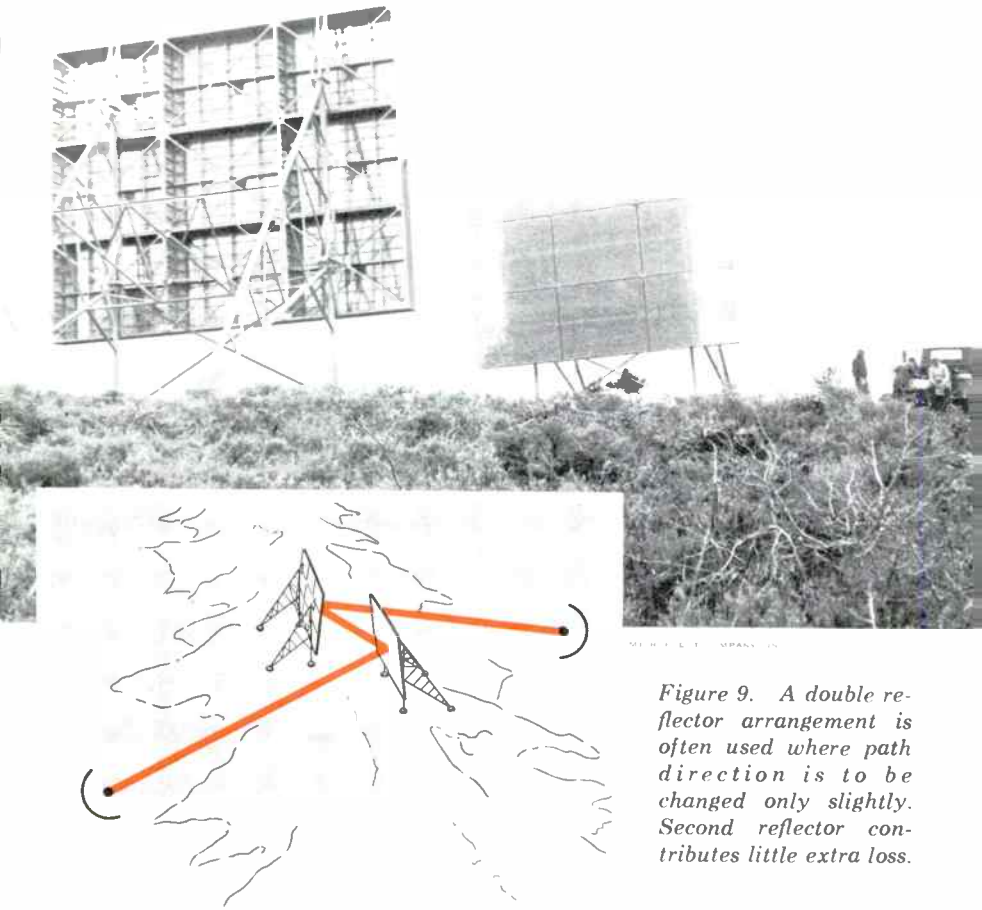


Figure 9. A double reflector arrangement is often used where path direction is to be changed only slightly. Second reflector contributes little extra loss.

mitter. For a 10-mile path, the isotropic attenuation is 132.2 db, only about 1 db more than for a 9-mile path. Subtracting the gain of the antennas (76 db), the transmission loss is 56.2 db. Because this method does not consider the reflector, a comparison with the two-path method gives the loss contributed by the reflector: 56.2 db subtracted from 67.5 db indicates an actual reflector loss of 11.3 db.

Since passive reflectors have a high efficiency (typically about 98%), the high loss in this one is almost certainly due simply to its small size. It may not reflect enough energy for some applications. A 24-foot reflector would contribute about 9 db more gain (or less loss, depending on the viewpoint), reducing the reflector loss to a more tolerable 2.3 db.

For some applications, normally where the microwave path requires only a slight bend, a double reflector installation such as that shown in Figure 9 is used. In this case a single reflector cannot be used because the beam would strike it at an extreme angle, resulting

World Radio History

in almost an "end-on shot," and severely limiting the effective reflecting area. In such an arrangement, the two reflectors are installed nearly parallel and quite close together so that virtually all the energy reflected by one is intercepted by the other. But since they are not 100% efficient, inevitably they introduce slightly more loss than a single reflector.

The Future of Reflectors

The antenna system is becoming more and more important to the performance of modern microwave equipment. In a high-performance system such as the Lenkurt Type 76, the antenna and waveguide system is the limiting factor in intermodulation distortion. The reflections caused by a long waveguide run can raise the intermodulation level considerably, even with high-quality waveguide, unless the installation is done with extreme care. Here a reflector system has the advantage because it requires only a short waveguide run.

Reflectors however, are used primarily to save money. Nearly any job that can be done by a reflector system can also be done by other means — but usually at a higher cost. This does not mean that reflectors are inferior. It simply means that where there is a choice of acceptable methods the least expensive one is usually chosen. In antenna systems this often proves to be a reflector.

But reflectors have their limitations. High side lobe levels, low front-to-back ratios, and cross illumination (unwanted illumination of a reflector by the antenna of another system), for example, make periscope antenna systems especially vulnerable to crosstalk. This makes it particularly necessary to use different frequencies in each direction of transmission from a repeater station. This is a satisfactory solution when the frequencies are available. But economic considerations are being overridden more and more by the fact that frequencies are becoming scarcer. Using the same frequency in both directions usually demands performance which cannot be provided by a reflector, or even by a simple parabolic antenna.

The passive repeater's big advantage is its suitability for remote and inaccessible locations. And the very remoteness of the installations usually makes the scattering from a billboard relatively unimportant because it is not likely to cause mutual interference with other services.

Reflectors are definitely here to stay, but they undoubtedly will be used more selectively in the future as more situations arise where the controlling factor is not one of economics. •

BIBLIOGRAPHY

1. W. C. Jakes, "A Theoretical Study of an Antenna-Reflector Problem," *Proceedings of the IRE*; February, 1953.
2. D. R. Crosby, "Theoretical Gain of Flat Microwave Reflectors," *1954 IRE Convention Record, Part 1, Antennas and Propagation.*
3. E. Bedrosian, "The Curved Passive Reflector," *IRE Transactions on Antennas and Propagation*; October, 1955.
4. *Microwave Path Engineering Considerations — 6000-8000 Mc.* Lenkurt Electric Co., Inc., San Carlos, California; September, 1961.

World Radio History

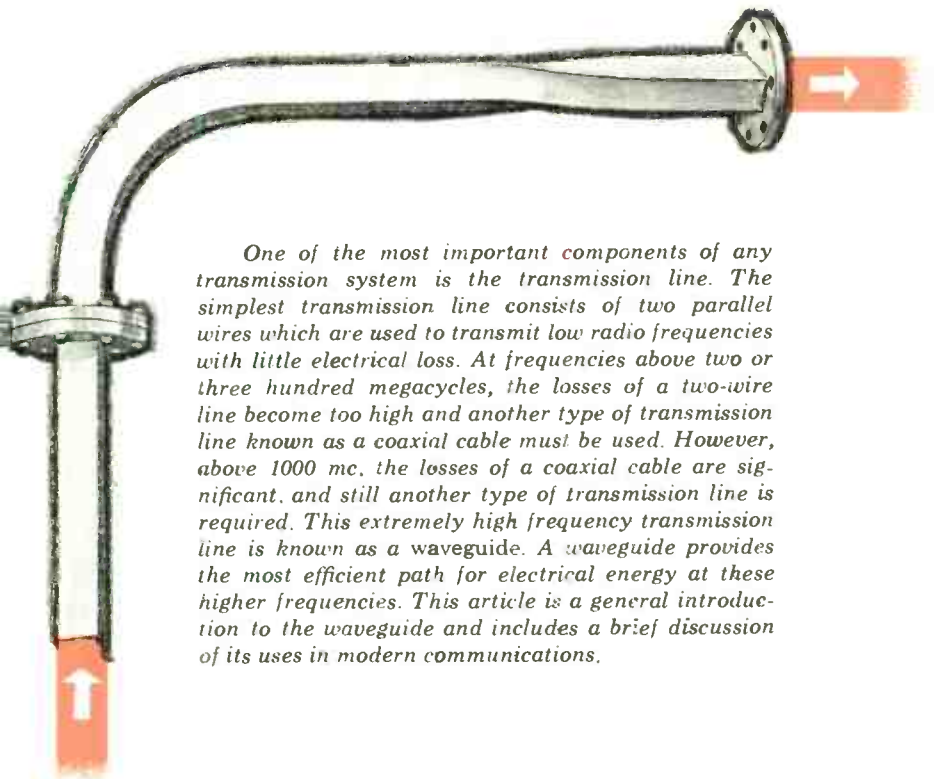
World Radio History

the Lenkurt Demodulator

VOL. 14, NO. 10 OCTOBER, 1965

Characteristics of

WAVEGUIDES



One of the most important components of any transmission system is the transmission line. The simplest transmission line consists of two parallel wires which are used to transmit low radio frequencies with little electrical loss. At frequencies above two or three hundred megacycles, the losses of a two-wire line become too high and another type of transmission line known as a coaxial cable must be used. However, above 1000 mc, the losses of a coaxial cable are significant, and still another type of transmission line is required. This extremely high frequency transmission line is known as a waveguide. *A waveguide provides the most efficient path for electrical energy at these higher frequencies. This article is a general introduction to the waveguide and includes a brief discussion of its uses in modern communications.*

World Radio History

The number of signals which may be transmitted over a given facility is directly proportional to the available frequency bandwidth. As is generally known, it is possible to group many independent message signals together and transmit them simultaneously over a single transmission facility. Broadband radio facilities occupying bandwidths of several megacycles are in common use today. The *waveguide* is a significant component of such systems.

The pioneer in the early experimental phase of the microwave field was Heinrich Hertz. Though his experiments were confined to relatively low microwave frequencies, his findings were outstanding accomplishments for the late 19th century. In the middle 1930's the practical applications of waveguides as transmission systems for microwaves were discovered. But it was not until the beginning of World War II that extensive exploitation of the microwave region was initiated. The benefits of microwave for high-frequency radar, such as improved directivity and resolution, brought forth the microwave spectrum as a very useful medium. After the war, microwave radio systems were developed that could transmit many hundreds of voice messages, and even television signals. Only because of the reach for higher and higher frequencies and the associated broad bandwidth capabilities did the waveguide find its ultimate and most efficient use as a low-loss transmission line.

Description

A waveguide is simply a single hollow metallic conductor, either rigid or flexible, which transfers electrical energy from one point to another. That it is possible to transmit electrical energy through a single conducting medium may seem strange to those accustomed to thinking in terms of currents and voltages along a two-wire transmission line. When considering energy transfer in a waveguide, it is easier to think of electromagnetic waves propagating through the waveguide in the same manner that radio waves propagate in space. A two-wire transmission line can, however, be analyzed using the electromagnetic field concepts, but it is more convenient to treat it in terms of voltages and currents. With a waveguide it is impossible to measure absolute values of voltage and current.

The shape of a waveguide may be rectangular, circular, or elliptical. By far the most common is the rectangular waveguide. It might be speculated that circular waveguides are preferable to rectangular waveguides, just as circular pipes find common application in carrying fluids. But the waves in a circular waveguide tend to *twist* as they travel through the structure. However, the circular type waveguide does find use with rotating antennas which require a rotating joint input. Elliptical waveguides are normally flexible, and are used for bends, and to reduce the close physical tolerances required of a rigid waveguide system. This article is concerned mainly with the characteristics of the more commonly used rectangular-shaped waveguide.

Modes

A waveguide is capable of transmitting microwave energy in a number of different electric and magnetic field configurations. The configuration in which energy propagates through a waveguide is referred to as the *mode*. A particular mode depends on the operating frequency and the physical dimensions of the waveguide.

Generally, there are two fundamental classes of modes that may appear in a waveguide. In one class the electric field is always perpendicular to the direction

World Radio History

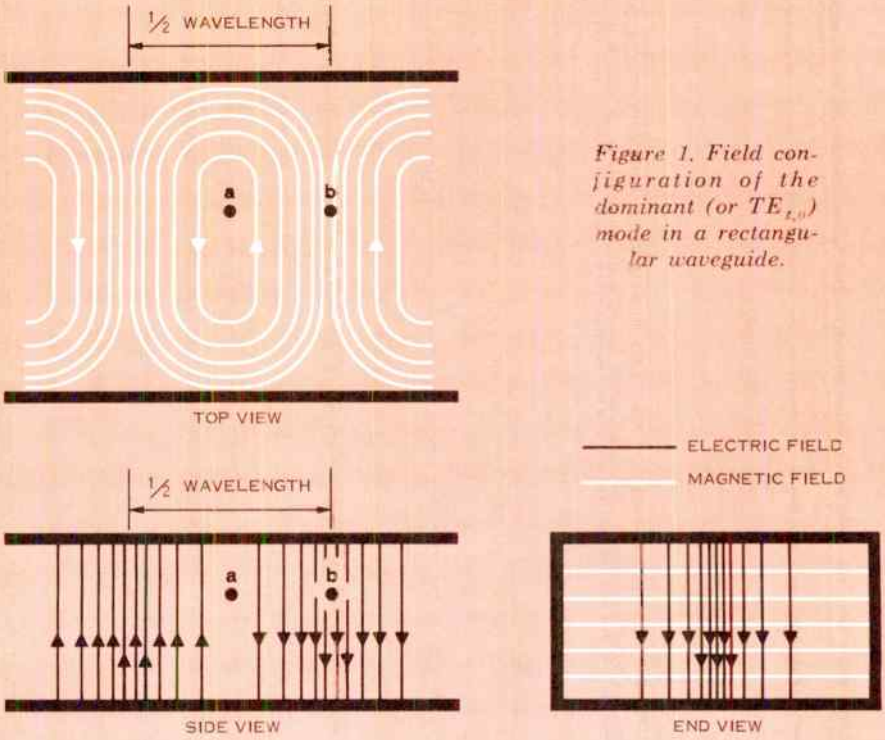


Figure 1. Field configuration of the dominant (or $TE_{1,0}$) mode in a rectangular waveguide.

of propagation. This class of modes is known as the TE or *transverse electric* class. In the second class of modes, the magnetic field is always perpendicular to the direction of propagation. This class is known as the TM or *transverse magnetic* class. The two fields are mutually perpendicular to each other and oriented at right angles to the direction of propagation.

The fields in the waveguide which make up these modes obey certain physical laws. Also, each mode has a *cutoff frequency*. This is the lowest frequency that will propagate through a waveguide while operating in a particular mode. Energy at frequencies below the cutoff frequency is greatly attenuated, while energy above the cutoff frequency is transmitted with very little attenuation.

The simplest or lowest order mode in a waveguide is called the *dominant mode*, and is the one most often used. Figure 1 shows the field pattern of the dominant mode in a rectangular waveguide. The black lines are voltage lines and indicate the direction of the electric field; the white lines indicate the magnetic field. In the end view, there is an electric field intensity between the narrow dimension sides of the waveguide which is maximum at the center. This is indicated by more voltage lines appearing in the center of the waveguide than toward its edges. The magnetic field, shown by the white lines, consists of closed loops of magnetic flux.

World Radio History

It should be noted that as the wave propagates along the waveguide, the electric and magnetic fields move together. Figure 1 represents the electromagnetic field as it exists at one instant of time. Although the amplitude at position *a* has zero intensity, a quarter cycle later the amplitude at position *a* will be the same as the present amplitude at position *b*.

The particular mode in each class is designated by two subscripts (for example, $TE_{1,0}$). The first subscript (1) indicates the number of half-wave variations of the electric field intensity across the wide dimension of the waveguide. The second subscript (0) denotes the number of half-wave variations across the narrow dimension. In Figure 1, the voltage intensity varies from zero to a maximum and back to zero across the wide dimension, which is one-half wavelength. Across the narrow dimension there is no variation in voltage intensity. Thus, in the transverse electric mode (TE) the subscripts 1 and 0 are added. The $TE_{1,0}$ mode is the dominant mode in a rectangular waveguide. Other subscripts designate higher order modes.

Many higher order modes other than the dominant mode can exist in a waveguide. But the common practice is to design the waveguide to propagate the dominant mode and suppress all others. The width of the usual rectangular waveguide is greater than one-half wavelength but less than one wavelength of the operating frequency. The height of the waveguide is then made about one-half the width. These dimensions are small enough to prevent higher order modes from forming, and give a cutoff frequency which is sufficiently below the operating frequency.

The dominant mode is more commonly used because it is an *exclusive* mode and, in its frequency range, prevents other higher order modes from forming, and because it gives the lowest cutoff frequency for a particular waveguide. Also, it provides low power dissipation, and requires smaller and less expensive waveguiding structures. Further, it gives the simplest field pattern, and is not as susceptible to impedance mismatches and reflections as the more complex higher order modes.

Characteristic Impedance

The waveguide has a characteristic impedance that is similar to the characteristic impedance of a two-wire transmission line. Therefore, the waveguide must be terminated by an impedance

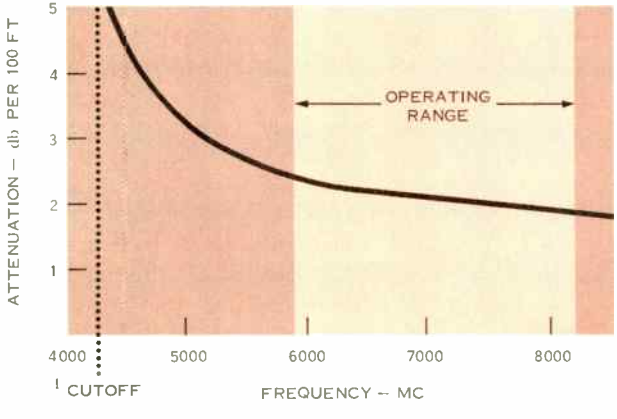


Figure 2. Attenuation versus frequency characteristics for a typical rectangular waveguide.

World Radio History

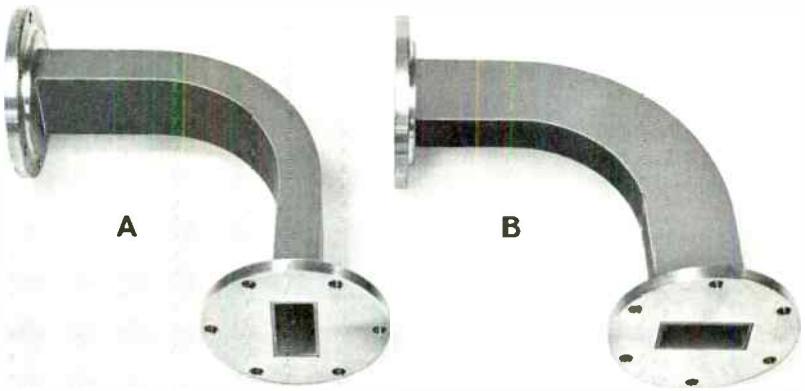


Figure 3. Two types of 90° bends for rectangular waveguide. Because the bend of waveguide A is in the plane of the electric field, it is referred to as an E bend. Waveguide B is referred to as an H bend because the bend is in the plane of the magnetic field.

equal to its characteristic impedance to prevent reflected waves. In a two-wire transmission line, the characteristic impedance is determined by the structure of the line and is relatively insensitive to frequency over a wide range. In contrast, the waveguide impedance is a direct function of frequency. The wave impedance of a rectangular waveguide in the dominant mode is given by the equation:

$$Z_0 = \frac{120\pi}{\sqrt{1 - \left(\frac{f_c}{f}\right)^2}} \text{ ohms}$$

where

f_c = *cutoff frequency, and*
f = *operating frequency.*

As can be seen from the equation, when the operating frequency is equal to the cutoff frequency, the waveguide impedance in infinite. Only by having the operating frequency greater than the cutoff frequency does the waveguide impedance achieve a finite value.

One method of achieving a good match between the waveguide and the load is, of course, to design the waveguide so that the load impedance completely absorbs all of the energy in the forward or *incident* wave. Another approach, however, is to set up a wave near the load that is equal in magnitude but opposite in phase from the wave reflected by the load. In this way, the injected wave and the wave reflected from the load cancel each other. This can be done only over a relatively narrow band of frequencies.

The voltage standing wave ratio (vswr) is an important parameter to consider when dealing with waveguides. The value of vswr determines whether or not the waveguide is correctly terminated in its characteristic impedance. As with a two-wire transmission line, the optimum vswr is 1:1 and this occurs

World Radio History

when the load completely absorbs all the energy in the incident wave.

Losses

The propagation of energy through a waveguide is accompanied by a certain amount of attenuation as a result of the current induced in the walls of the waveguide. The resistivity that the induced current encounters is the result of the skin effect losses in the waveguide, and is proportional to frequency and the resistance of the waveguide material.

The energy loss is conveniently expressed in decibels of attenuation per unit length. Figure 2 shows the attenuation versus frequency characteristics of a typical waveguide. A standard size waveguide for 6000 to 8000 mc microwave systems has a loss of approximately 2½ db per hundred feet.

To minimize resistivity, the inner walls of the waveguide can be plated with a layer of silver. Because of skin effect, most of the energy flows near the surface of the waveguide walls, and only a very thin plating is needed to reduce the resistance of the waveguide walls.

Physical Characteristics

The dimensions of a waveguide are inversely proportional to the lowest frequency which it can propagate. The larger the waveguide, the lower the cutoff frequency and, conversely, the higher the frequency, the smaller the waveguide.

For a rectangular waveguide, the maximum wavelength of a transmitted wave is equal to twice the width of the waveguide. To transmit energy through a waveguide at 1000 mc, the width must be about 6 inches. It is only at frequencies higher than this that the required dimensions become small enough to make the use of waveguides practicable.

Because of the possibility of unwanted higher order modes appearing, it is common practice to operate waveguides over a relatively narrow frequency range. By properly selecting the frequency range, it is possible to operate far enough from the cutoff frequency, but within the frequency region where modes other than the dominant mode can be prevented.

In rectangular waveguides, higher order mode suppression is most effective when the ratio of width to height is 2 to 1. In contrast, if the waveguide were square there would be no frequency range over which only a single mode could propagate. The 2 to 1 ratio gives the best mode separation of all possible physical proportions.

The table on page 7 lists the characteristics of standard rectangular waveguide types used between 1700 and 15000 megacycles. The values given are for the dominant (or $TE_{1,0}$) mode. The Radio-Electronics Television Manufacturers Association (RETMA) and Army-Navy designations are also given, since they are the common means used to identify particular rectangular waveguides which have become standard.

At frequencies below about 3000 mc, the size of the waveguide has to be so large to propagate the dominant mode that weight and costs become excessive. Consequently, coaxial transmission lines are more often used in the microwave region below 3000 mc. Above 3000 mc the desirable dimensions of the waveguide become sufficiently small—and the losses of a coaxial line become extremely high—so that the waveguide is more practical and economical to use.

Most waveguides are constructed from oxygen-free high-conductivity copper, because this material has excellent inherent electrical characteristics. In some cases, brass is used for unusual waveguide shapes, because machining

World Radio History

TABLE OF STANDARD RECTANGULAR WAVEGUIDES (1700 TO 15000 MC)

Frequency Range (megacycles)	Cutoff Frequency (megacycles)	Outside Dimensions (inches)	Wall Thickness (inches)	RETMA Designation	Army-Navy Type No.
1700-2600	1375	4.460 × 2.310	0.080	WR430	RG-104/U
2200-3300	1735	3.560 × 1.860	0.080	WR340	- - -
2600-3950	2080	3.000 × 1.500	0.080	WR284	RG-48/U
3300-4900	2590	2.418 × 1.273	0.064	WR229	- - -
3950-5850	3160	2.000 × 1.000	0.064	WR187	RG-49/U
4900-7050	3710	1.718 × 0.923	0.064	WR159	- - -
5850-8200	4290	1.500 × 0.750	0.064	WR137	RG-50/U
7050-10000	5260	1.250 × 0.625	0.064	WR112	RG-51/U
8200-12400	6560	1.000 × 0.500	0.050	WR90	RG-52/U
10000-15000	7880	0.850 × 0.475	0.050	WR75	- - -

and fabrication are much easier to accomplish with this material.

The Waveguide Run

The physical path of the *waveguide run* between the radio equipment and the microwave antenna cannot always be a straight line, but is usually made up of a combination of straight, curved, and flexible sections of waveguide. Two typical 90° curved sections of waveguide are shown in Figure 3. Curved sections tend to introduce reflections and contribute power loss, but these can be kept to a minimum if the radius of the curved section is never less than two wavelengths of the transmitted signal.

Other losses can be attributed to a number of factors. Size differences, misalignment, twists or bows, dents, and scratches all contribute losses and reflections. Since the flanges of adjoining waveguide sections are difficult to keep smooth and square with the waveguide axis, extreme judgment and care must be exercised in establishing a waveguide system so that a minimum number of sections are used.

Residual moisture appearing on the inside walls of a waveguide causes corrosion and increases attenuation. Therefore, an outdoor waveguide run is normally pressurized with dry air or nitrogen to eliminate this problem.

Conclusion

This article has covered some of the advantages of waveguides over other types of transmission lines. To sum up, the waveguide is capable of transmitting much higher powers than other types of transmission lines and at higher frequencies.

In waveguides of the types commonly employed, and for ordinary runs, attenuation is negligible. Waveguides with extremely low losses and reflections are produced by modern manufacturing processes, and with careful installation excellent performance can be achieved.

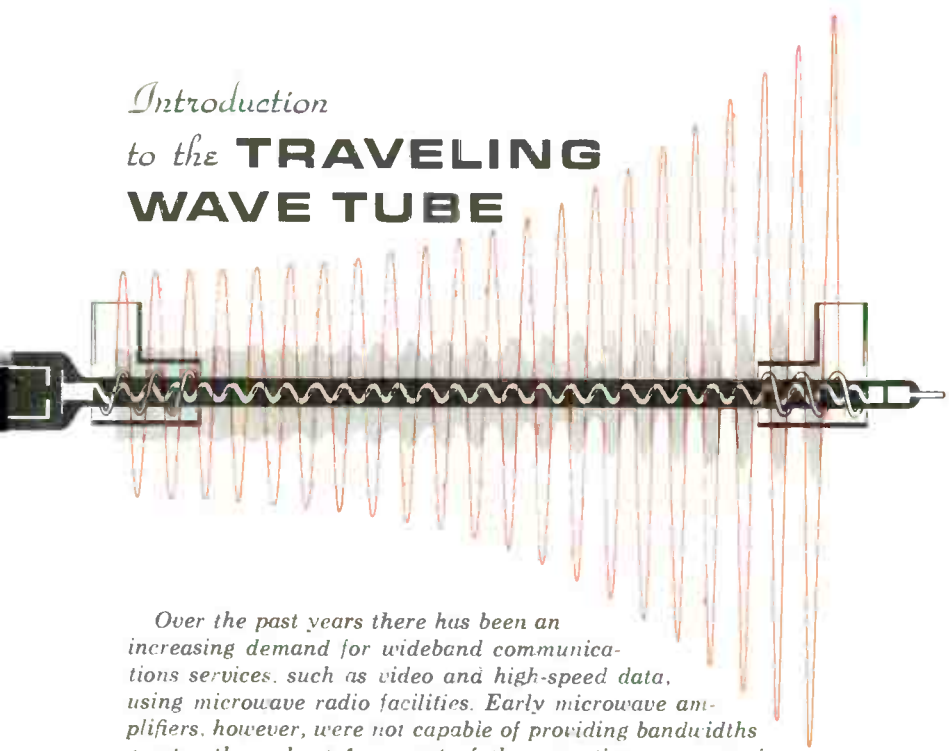
World Radio History

World Radio History

the Lenkurt Demodulator

VOL. 14, NO. 8 AUGUST, 1965

Introduction to the TRAVELING WAVE TUBE



Over the past years there has been an increasing demand for wideband communications services, such as video and high-speed data, using microwave radio facilities. Early microwave amplifiers, however, were not capable of providing bandwidths greater than about 1 percent of the operating or center frequency. A broadband high-gain microwave amplifier, therefore, was seriously needed. To fulfill this need, a microwave tube was developed that provided high gain over bandwidths greater than 10 percent of the operating frequency. This remarkable device became known as the traveling wave tube.

This article briefly describes how a traveling wave tube works and discusses some of its important operating characteristics.

World Radio History

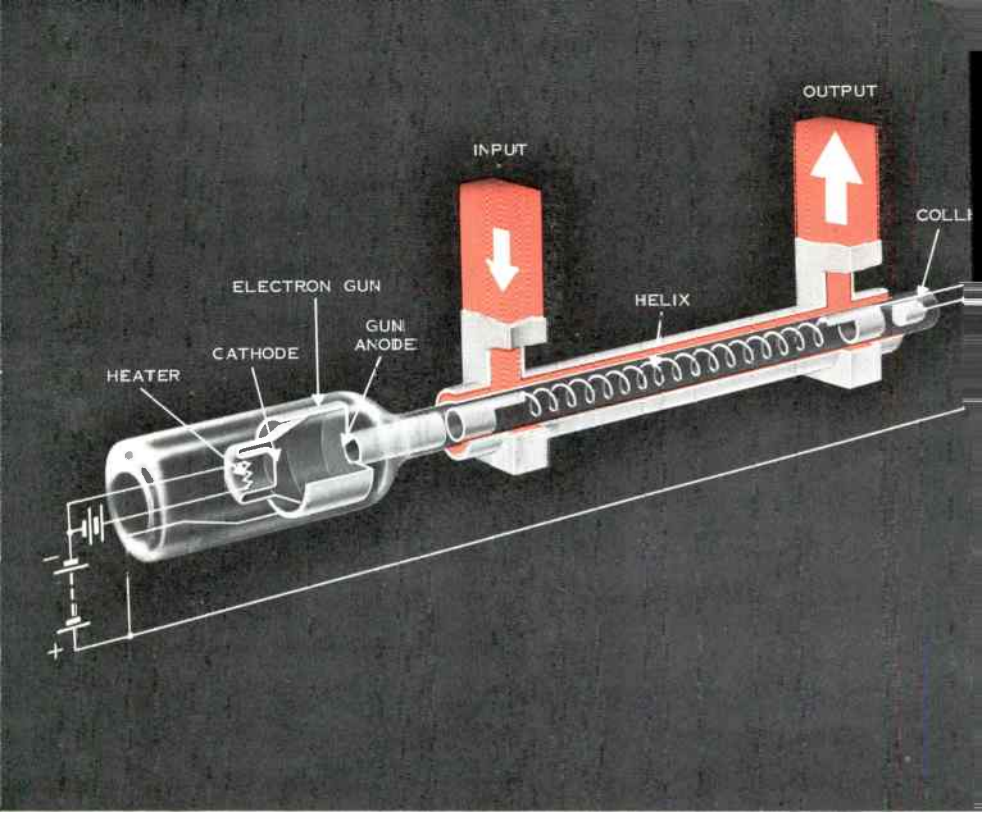


Figure 1. Construction of a typical traveling wave tube, showing the electron gun, the helix, and the collector.

The ever-increasing demand for more and more communications services soon crowds the available radio frequency bands. To acquire more radio channels, it has been necessary over the years to develop communications equipment capable of operating at higher and higher frequencies. Such progress upwards within the radio frequency spectrum was carried out successfully using so-called conventional electron devices — until the microwave region was reached. Here, things changed.

Conventional electron tubes, for example, cannot operate satisfactorily at frequencies above a few hundred megacycles. The most serious limitation of these tubes at higher frequencies is caused by a phenomenon known as the *transit time effect*. This phenomenon occurs when the period of one cycle of a signal applied to the control grid of such a tube is *less* than the time it takes an electron to travel from the cathode to the plate. When this happens, the voltage of the alternating signal applied to the control grid may completely reverse itself before the electrons in transit have had time to reach the plate. As a result, these electrons are not able to follow the signal variations exactly. This, of course, distorts the signal at the output of the tube, and also decreases gain.

World Radio History

Some method of controlling the flow of electrons had to be devised that could overcome the transit time limitation of ordinary electron tubes. Oddly enough, the first electron tubes that were developed to generate and handle microwave frequencies actually made use of the transit time effect. These tubes, known as klystrons, became quite successful primarily as oscillators, but are also capable of providing considerable gain at microwave frequencies. Unfortunately klystrons are inherently narrow-band resonant devices having a useful bandwidth of only tens of megacycles. However, they were the beginning of a family of microwave electron tubes which also includes the magnetron. Later, a very useful device known as the *traveling wave tube* (TWT) was added to this family. Traveling wave tubes are non-resonant devices and proved to be capable of amplifying microwave signals with enormous bandwidths.

The traveling wave tube was invented in England in 1943 and first used in a television relay link which began operating in March, 1952. Since then, it has found wide application in communications systems, primarily because of its broad bandwidth capability.

Description

The construction of a traveling wave tube is relatively simple. It consists essentially of an *electron gun*, a wire *helix*, usually encased in a glass envelope, and a *collector*, as shown in Figure 1. The electron gun, attached to one end of the helix, produces a focused beam of electrons which are directed through the center of the helix. The helix is simply a wire conductor that has been formed into a uniform spiral in order to slow down the forward progress of an RF signal fed into the tube for amplification. The collector, attached to the opposite end of the helix, is an electrode which receives the spent electrons that have traveled through the tube.

The focused electrons emanating from the gun are held in a tight beam usually by some type of magnetic field

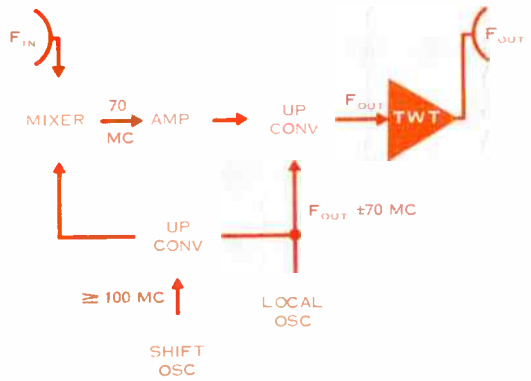


Figure 2. Circuit diagram of a typical I-F heterodyne repeater using a traveling wave tube.

World Radio History

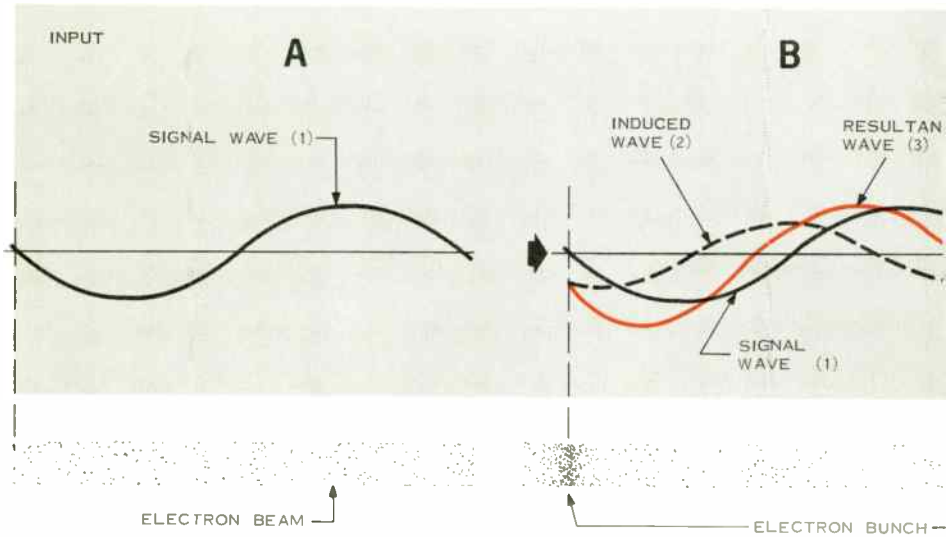
surrounding the helix. There are several methods used to prevent the beam from spreading, such as the electrostatic method, the electromagnetic method, and the periodic permanent magnet method. The periodic permanent magnet method is used most commonly in traveling wave tubes. In this method, permanent magnets are arranged in some periodic manner along the length of the tube.

How It Works

When an electron is in motion, it has a certain amount of kinetic energy. Increasing the velocity of the electron increases its energy, while decreasing its velocity decreases its energy. The energy given up by an electron whose velocity has been decreased *must go somewhere*. This basic principle underlies the operation of the traveling wave tube amplifier.

Operation of the traveling wave tube is similar to other types of microwave tubes in that the transfer of power from a direct current source to an RF signal is accomplished by *velocity modulating* a beam of electrons. This type of modulation causes the electrons in the beam to form into periodic groups or *bunches*. To accomplish this bunching, some electrons in the beam are accelerated while others are slowed down or retarded. If more electrons are retarded than are accelerated, the excess kinetic energy given up by the retarding electrons is transferred to the modulating RF signal. The bunching of the beam electrons by the modulating signal and the transfer of energy from the retarding electrons to the signal is called *interaction*. Interaction in a traveling wave tube is continuous and cumulative and is the basis for amplification.

The electron beam in a traveling wave tube is generated in the same manner as the electron beam in a typical cathode ray tube. Electrons are emitted from an indirectly heated cathode, controlled and initially focused by a grid, and accelerated by the electron gun



World Radio History

anode. As they leave the anode, the density of the electrons is uniform. The beam electrons are accelerated, by a positive voltage potential on the helix, to a velocity U_o, which is proportional to the square root of the helix voltage, V, as shown in the following equation.

$$U_o = 5.93 \times 10^5 \sqrt{V} \quad \text{meters/second} \qquad (1)$$

This equation assumes that the initial electron velocity is zero and the relationship is, therefore, approximate.

The signal to be amplified by the tube is coupled into the gun end of the helix. This RF signal travels as a surface wave around the turns of the helix, toward the collector, at about the velocity of light. The forward or axial velocity of the signal is slower, of course, because of the pitch and diameter of the helix. This forward movement of the wave is analogous to the travel of a finely threaded screw where many turns are required to drive it into position. The signal wave generates an axial electric field which travels with it along the longitudinal axis of the helix. This alternating electric field interacts or *velocity modulates* the electrons in the beam.

The relationship between the initial velocity of the beam electrons given in equation (1) and the velocity of the axial electric field is very important. In order to achieve continuous and cumulative interaction, the initial beam velocity must be slightly greater than the axial velocity of the alternating electric field.

When the electrons in the beam enter the helix they are accelerated or decelerated by the alternating electric field associated with the RF signal wave. Electrons acted upon by a positive electric field take energy from the field, and therefore, are accelerated. Electrons acted upon by a negative electric field give up energy to the field and, therefore, are decelerated or retarded. Since the initial velocity of the beam electrons is slightly greater than the axial velocity

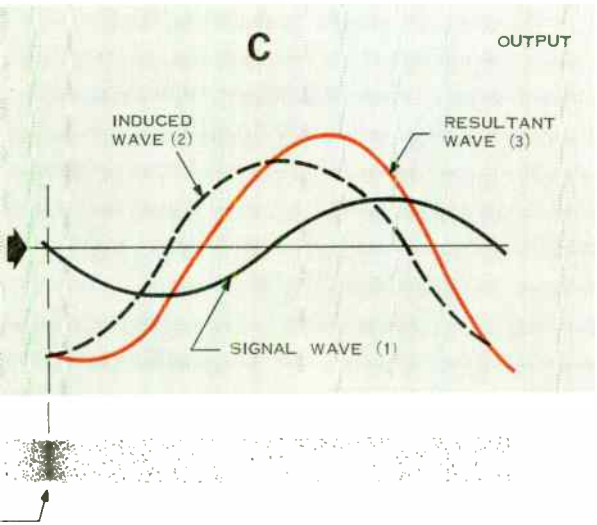


Figure 3. As the signal wave travels down the tube, electron bunches induce a second wave on the helix. The resultant of the signal wave and the growing induced wave increases exponentially.

World Radio History

of the electric field, more electrons are retarded than are accelerated. This means that more energy is transferred to the field than it gives up. Because this interaction is continuous and cumulative, the amplitude of the RF signal wave *grows* as it travels down the helix.

Consider a group of beam electrons that have just entered the helix in the presence of a signal wave (1) as shown in Figure 3A. Electrons at the point of zero field intensity are neither accelerated nor retarded. However, electrons just to the left of the zero point are accelerated by the positive electric field and therefore catch up with the electrons at the zero point. Electrons just beyond the zero point are retarded by the negative electric field and are overtaken by the electrons nearer the zero point. As shown in Figure 3B, this action causes the electrons to become more and more dense at the point of zero field strength, as they travel down the tube.

The electrons in this dense bunch induce a second wave (2) on the helix. This second wave produces an associated axial electric field that lags behind the first electric field by a quarter wavelength. As the electron bunch travels down the tube it accumulates more and more retarding electrons which are giving off energy. As more retarding electrons are accumulated, the amount of energy transferred to the induced wave increases, thereby causing it to *grow* in amplitude. The resultant wave (3) caused by the addition of the signal wave (1) and the growing wave (2) increases in amplitude exponentially as it travels down the helix until it reaches the output of the tube, as shown in Figure 3C.

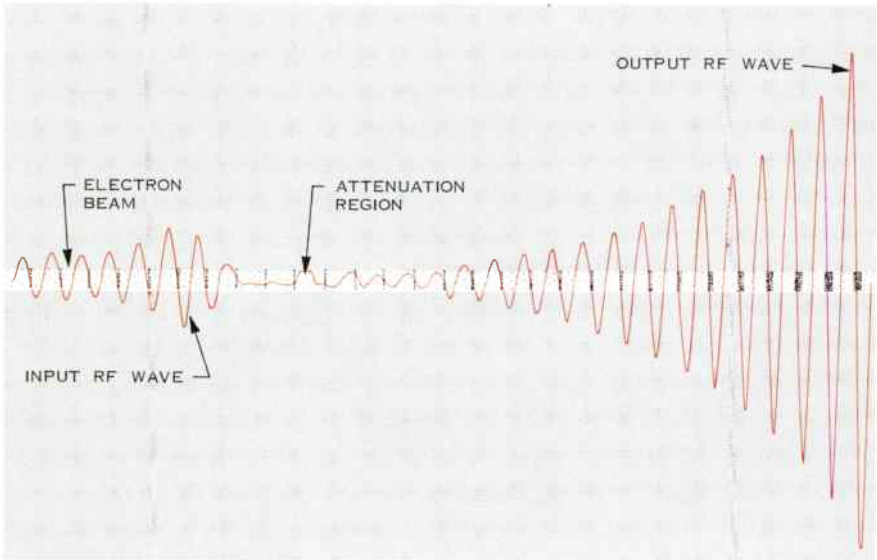


Figure 4. The growing signal wave is reduced by the attenuator but reappears at the output of the attenuator and begins to grow again.

World Radio History

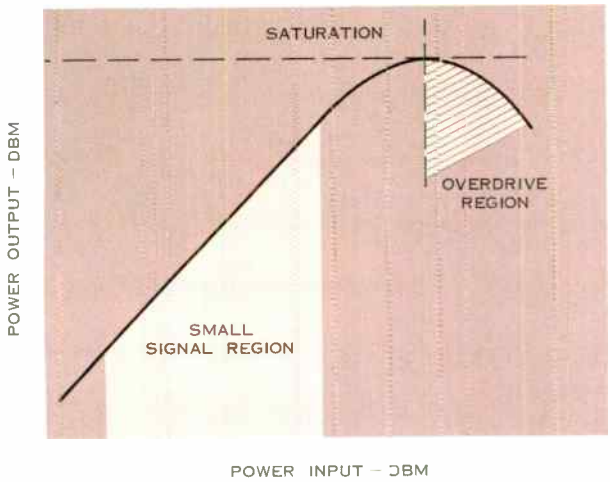


Figure 5. Power output vs. power input.

The net effect of this cumulative interaction is to reduce the *average* velocity of the electrons in the beam. The kinetic energy, given up by this reduction in electron velocity is transferred to the signal wave. Thus, amplification has occurred through the interchange of energy from an electron beam to an RF wave.

It is necessary to attenuate reflected waves on the helix to prevent spurious oscillations. This is accomplished by placing an attenuator between the gun-end and the collector of the helix. In ordinary low-power tubes, the attenuator consists of a resistive material that is sprayed around the helix.

The attenuator reduces the waves on the helix to practically zero while the electron bunches are essentially unaffected. When the electron bunches emerge from the attenuator, the waves begin their growing process all over again.

Wave Analysis

A mathematical analysis of the wave propagation along the helix of a traveling wave tube results in the following equation for the instantaneous axial component, E_z, of the electric field strength.

$$E_z = E_{max} (\cos\beta_p) (U_p t - z) \qquad (2)$$

where

E_{max} = *the crest value of the wave*
β_p = *the axial phase shift*
U_p = *the axial wave velocity*
t = *time in seconds*
z = *the distance along the helix at any instant*

When an electron passes through an electric field, a force, F, is exerted on the electron that is equal to the electric field strength, E, times the electron charge, e.

$$F = Ee \text{ newtons} \qquad (3)$$

The electric field strength exerts a sinusoidal force on the electrons within the beam. Since the electrons are free to respond to this force, their velocity will fluctuate sinusoidally. The magnitude of this sinusoidal velocity fluctuation, $|U_p|$, is

World Radio History

$$|U_e| = \frac{1.76 \times 10^{11} E_{max}}{\beta_p} \times \frac{1}{(U_p - U_o)} \quad \text{meters/second} \qquad (4)$$

The axial wave velocity, U_p, is fixed by the helix while the initial beam velocity, U_o, is proportional to the helix voltage as shown in equation (1). The quantity $|Ue|$ is a measure of the *perturbation* of the electron velocity by the wave and, therefore, gives an indication of the kinetic energy that is available for transfer to the signal wave. For maximum perturbation, hence maximum energy available, the *average* velocity of the electrons should be approximately the same as the axial wave velocity, U_p. As stated previously, the initial velocity, U_o, of the beam electrons must be slightly greater than the axial velocity of the wave. This is necessary so that the beam electrons can give up more energy than they receive and still have an average velocity approximately the same as the axial wave velocity.

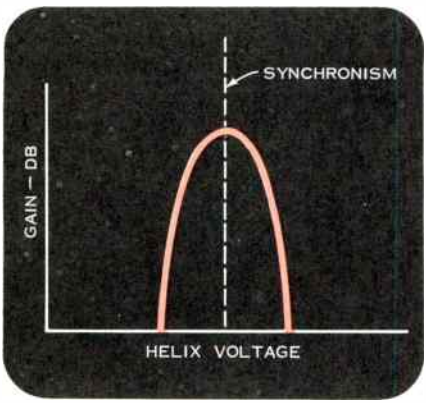


Figure 6. Gain vs. helix voltage.

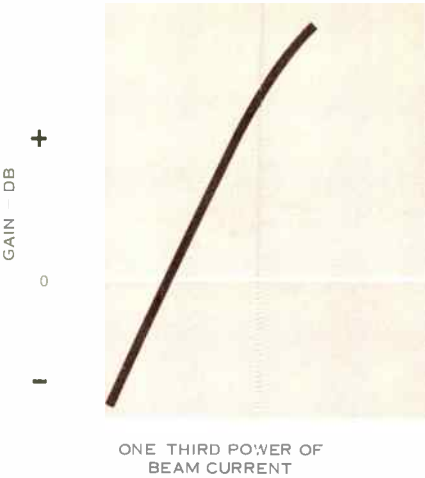


Figure 7. Gain vs. beam current.

The signal wave will grow exponentially when the traveling wave tube is operating in what is known as its *small signal region*. However, a point of maximum net energy transfer, called *saturation*, is eventually reached. This point is determined essentially by the dc operating parameters of the tube. Beyond the saturation point, the gain of the tube begins to decrease. Here the tube is operating in a condition or region known as *overdrive*. The small signal region, the saturation point, and the overdrive region, representing the dynamic characteristics of a traveling wave tube, are shown in Figure 5. Traveling wave tube amplifiers are operated, of course, somewhere in the small signal region.

Gain and Output Power

Gain and output power of a traveling wave tube are a function of the helix voltage and beam current. A par-

World Radio History

ticular helix voltage will synchronize the electron velocity with the phase velocity of the signal wave. During synchronization, interaction efficiency is at its highest and maximum gain is obtained. If the helix voltage varies from this point of synchronization, the gain decreases as shown by the curve in Figure 6. Operation above synchronous voltage extends the small signal region and increases the available output power at the expense of gain.

The gain at any frequency is essentially a function of the one-third power of the beam current and decreases as beam current is decreased, as shown in Figure 7. By operating the grid of a traveling wave tube more and more negatively with respect to the cathode, a point is reached where the gain of the tube becomes negative and it acts like an attenuator. At higher current levels, gain, saturation power, and power in overdrive all increase. The electrical and thermal characteristics of a particular traveling wave tube tend to limit the allowable beam current. Like any electronic device, traveling wave tubes are designed to operate below certain limits of input power. Operating a tube above these limits almost always degrades its performance.

Traveling wave tubes achieve gains ranging from 10 to 60 db, with continuous wave (cw) output powers above 1 kilowatt using an ordinary helix.

AM/PM Conversion

When the RF signal applied to a traveling wave tube approaches a high value of amplification near the saturation region, the average beam velocity decreases making the circuit appear *longer* to the beam electrons. The result is that the electrical length of the helix becomes proportional to the RF energy level and is indicated as a relative phase change. The mathematical slope of the RF signal level versus relative phase shift curve is a parameter called amplitude modulation/phase modulation conversion, or more commonly, AM/PM conversion. AM/PM conversion, expressed in degrees/db, is an indication of the amount of phase modulation of the RF signal contributed by the traveling wave tube. The effects of AM/PM conversion usually result in a compromise between gain, power levels, and allowable noise contributions. Traveling wave tubes used in communication systems generally are required to exhibit AM/PM conversion figures less than 2 degrees/db.

Bandwidth and Noise

The useful bandwidth of a traveling wave tube is determined by the range of frequencies over which the phase velocity of the wave is essentially constant. Bandwidth is limited at the low frequency end of the band by *dispersion* (change in phase velocity with fre-

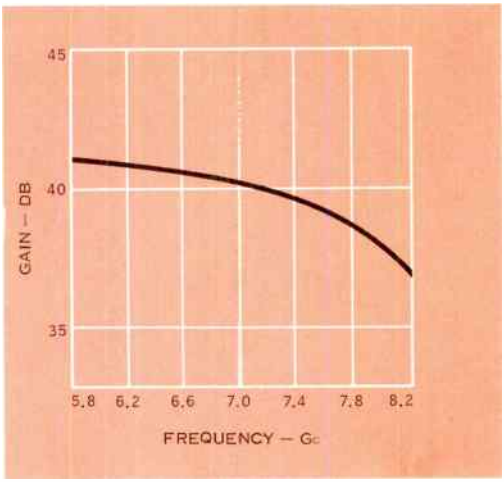


Figure 8. Typical gain vs. frequency.

World Radio History

quency) and at the upper end by reduction of field intensity with an increase in radial distance from axis to helix.

The flattest gain-bandwidth characteristic is obtained when the helix voltage optimizes gain at or near the high frequency end of the band. Under this condition, synchronization with the wave at the high end compensates for the radial decay of the RF field with frequency. At the same time, operation away from synchronization at midband lowers the maximum gain which occurs at the band center. These two effects work to minimize gain variation.

The bandwidth of a traveling wave tube is limited more by the input and output couplers than by the propagating characteristics of the helix itself. However, bandwidths greater than one octave have been achieved.

The types of noise associated with traveling wave tubes are *thermal noise* which is caused by random currents in the conductors, *shot noise* resulting from the random nature of emission characteristics and electron flow in the beam, and *partition noise* which develops as electrons are intercepted at grids and on the helix. Thermal noise is a function of temperature and bandwidth. Shot noise can be reduced by designing the cathode to operate at the lowest possible temperature. Partition noise is reduced by maintaining the electron beam in a tightly focused envelope.

Uneven emission from cathode surfaces can increase noise by producing a nonsymmetrical beam. Thus, as a traveling wave tube ages, it may become noisy unless the emission decreases evenly over the entire surface of the cathode. Partition noise increases when cathode emission becomes non-uniform as the tube ages and manifests itself as an increase in helix current at rated collector current. Normally, the useful

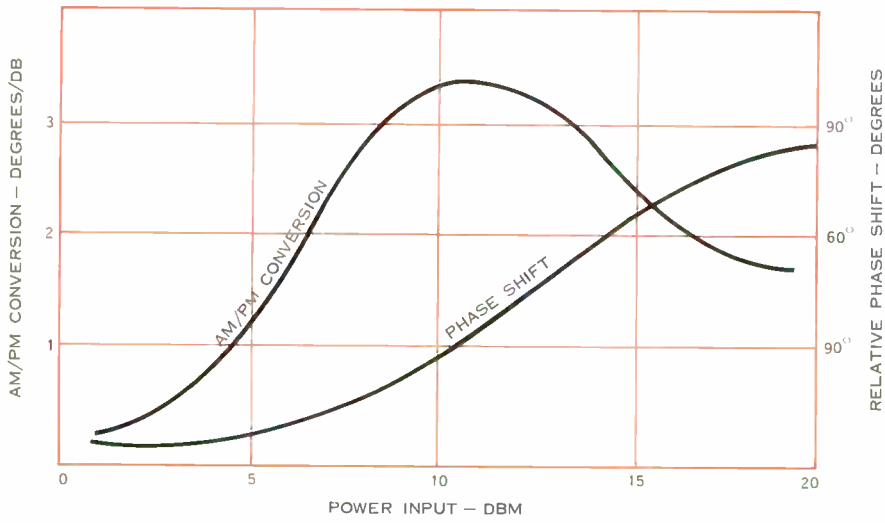


Figure 9. RF signal level vs. phase shift.

World Radio History

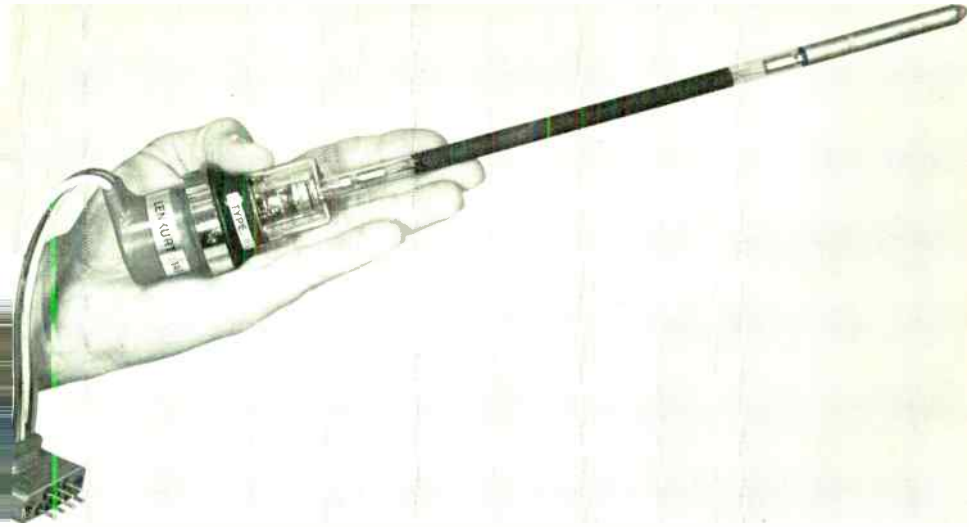


Figure 10. Photograph of typical traveling wave tube used in a microwave heterodyne repeater system. Tube operates in the 6000 megacycle band.

life of a typical communications traveling wave tube is in excess of 7500 hours.

Conclusion

The traveling wave tube is indeed a remarkable device. It overcame the bandwidth limitation of other microwave tubes and has proved to be an extremely valuable tube in modern communications systems. In fact, traveling wave tubes have been launched into space aboard certain communications satellites that are in orbit today.

Before the traveling wave tube was developed, microwave amplifiers were limited to bandwidths of about one percent of the center frequency. Traveling wave tubes have achieved bandwidths greater than 10 percent of the center frequency because of the cumulative nonresonant interaction between the electron beam and the RF signal.

This tremendous broadband feature has made the traveling wave tube especially useful in the development of *heterodyne* or non-demodulating repeaters for use in wideband transmission systems. With the expanding need for more and more wideband communications services such as high speed data and video or television, the traveling wave tube will certainly provide a significant contribution.

World Radio History

World Radio History

the Lenkurt Demodulator

Vol. 8 NO. 12 DECEMBER, 1959

The Very Important Klystron

Almost single-handedly, the klystron has revolutionized much of the technology and even the economics of communication. By expanding the usable radio frequency spectrum hundreds of times beyond its previous maximum limits, the klystron has proved to be one of the most important inventions since DeForest invented the vacuum tube amplifier. This article traces some of the progress made in klystron technology since the device was first conceived 22 years ago.

The klystron, like many other developments that have become important to our economy and way of life, resulted not from "practical" research, nor from the search for a good commercial product, but as a by-product of pure academic research into the nature of matter.

In 1933, Professors Russell Varian and W. W. Hansen of Stanford University were seeking economical ways of obtaining higher energy X-rays to support work in nuclear physics. The large high - voltage "atom - smashers" which were the subject of so many jokes and cartoons at that time, were not available, primarily due to budget limitations. Very high voltages were required to produce X-rays of the desired energy, but the means of achieving these voltages were too expensive.

Dr. Hansen believed that the desired high voltages might be achieved in a resonant circuit if the efficiency of the resonator could be improved sufficiently.

About this same time, Sigurd Varian, an airline pilot and brother of Russell Varian, was concerned about the alarming growth of Hitler's air force, and felt that electronics might provide a suitable defense against the growing

World Radio History

threat. Other researchers had reported being able to detect the presence of ships offshore, using radio waves at a frequency of about 60 megacycles. After studying the problem, the Varians concluded that frequencies higher than any then possible would be required — primarily because frequencies large enough to concentrate low frequency radio energy into useful beams would require antennas of an impractical size.

By 1937, Dr. Hansen had developed the cavity resonator or "rhumbatron" and had gone on to prove mathematically that it could be far more efficient than any other resonant devices then known. The cavity resonator was the perfect starting point from which to begin a search for a new high frequency oscillator. Accordingly, the Varian brothers and Dr. Hansen teamed up to develop such a device.

In the conventional vacuum tubes of the day, the top limit of operating frequency was set by the time required for electrons to travel from the tube's cathode to its anode or plate. The tube could no longer amplify or oscillate at frequencies where one cycle required less time than required by electrons to go from cathode to plate. Accelerating the electrons by increasing the voltage on the plate interfered with the tube's functioning. Narrowing the distance between tube elements increased capacitance — also interfering with the tube's high-frequency performance. Furthermore, reducing the tube's physical size was limited by the difficulty of manufacturing the microscopic structures that would be required.

Obviously, some method of controlling the flow of electrons had to be devised that would not be penalized by the relatively long transit time of the electrons between tube elements.

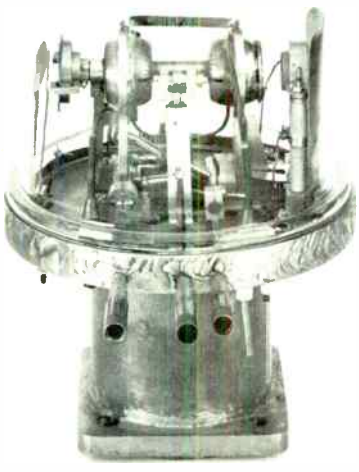
The solution to the problem occurred to Dr. Varian while he was attempting to systematically classify all known methods of controlling electrons. His solution was to allow the characteristics of the cavity resonator itself to control the electrons as they flowed toward the tube anode. He reasoned that an alternating field on a short portion of the electron stream would retard some electrons and accelerate others. The different electron velocities would cause the electrons to gather in groups or "bunches." The speed of the electron stream and the time interval between bunches would determine the frequency of oscillation. Instead of requiring ever-shorter transit times between cathode and anode for higher frequencies, this scheme actually worked better if the transit time was increased!

The team's first design consisted of two resonators linked by a "drift" tube, and suitably equipped with an electron gun and grids for forming a beam. The "drift" tube was to allow sufficient time for electrons to group in bunches.

Before the researchers could be sure that their device worked, they had to solve the basic problem of detecting the high-frequency oscillations. No detector known at the time would be able to function at microwave frequencies. Without a detector it was very unlikely that anyone would know if the machine operated successfully. This problem was solved by providing a small hole in the last resonator so that during oscillation, a portion of the electron beam used to drive the klystron would be deflected through the hole and hit a fluorescent screen. If oscillations were produced,

World Radio History

the screen would glow. This detection method was easy to incorporate in the experimental model because the entire device was operated in an evacuated bell jar. The third model constructed, shown in Figure 1, provided reproducible oscillations at a frequency of about 2300 megacycles.



HANSEN MICROWAVE LABORATORY, STANFORD UNIVERSITY

Fig. 1. First successful klystron produced microwave oscillations at about 2300 megacycles. Device incorporated water cooling and an ingenious fluorescent radio frequency detector. Shafts on front are micrometer controls for adjusting resonator positions and coupling between resonators during operation.

Wartime Urgency

The growing power of the German Luftwaffe, and the beginning of the European war in 1939, stimulated urgent research on the klystron in America, Britain, and France. Working under wartime pressure, the Bell Telephone Laboratories developed a line of klystrons which included prototypes of some still used today. The famous type 2K25 solved the problem of tuning the device over a wide frequency range.

Tuning had become a serious problem. Electrical tuning was possible by varying the voltage on tube elements. However, this permitted tuning ranges of only a few megacycles. Slight variations in the size of the resonator gap produced large changes in frequency, but were very difficult to control. Frequency changed 200 megacycles for every thousandth of an inch change in spacing between cavity grids. Since it was necessary to adjust frequency to within a megacycle, the tuning mechanism had to be able to position the grids accurately within five-millionths of an inch!

The Bell design is shown in Figure 2. Changes in gap spacing and cavity dimensions were made possible by a flexible diaphragm. This diaphragm, functioning like a bellows, permitted the upper portion of the tube to be moved in relation to the main body. The tuning device consisted of a fixed strut on one side of the tube, and an adjustable tuning bow on the other. The fixed strut provided enough hinge action to allow the diaphragm to flex,

World Radio History

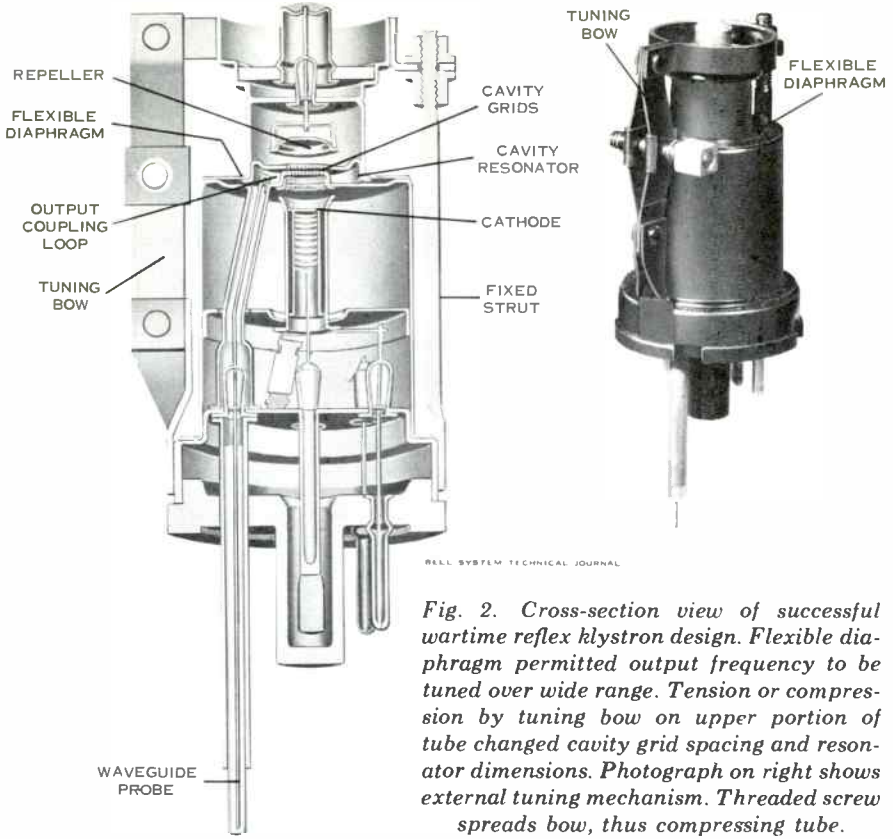


Fig. 2. Cross-section view of successful wartime reflex klystron design. Flexible diaphragm permitted output frequency to be tuned over wide range. Tension or compression by tuning bow on upper portion of tube changed cavity grid spacing and resonator dimensions. Photograph on right shows external tuning mechanism. Threaded screw spreads bow, thus compressing tube.

and the tuning bow provided tension. When the two members of the tuning bow were spread, the bow was shortened, thus reducing the gap between the cavity grids. This arrangement provided the very high leverage required to hold the grids to the desired spacing, and permitted slow motion tuning.

The design was easy to produce, and by the end of the war, thousands of klystrons of this type were in use all over the world. Even today, some tubes of this type are still in use — a tribute to the excellence of the wartime engineering effort.

After the war, a new type of communication came into being. Point-to-point microwave circuits, using klystron oscillators as transmitting tubes, permitted interference-free radio circuits with enough bandwidth to carry hundreds of telephone channels. The high frequencies opened up by the klystron increased the efficiency of antennas to such a degree that transmitter powers of only a watt, or even less, could provide reliable communication across many miles.

Under the seemingly hum-drum peacetime conditions, certain characteristics of the wartime klystron became evident that hadn't mattered too much in the military applications where life expectancy of the equipment was short and maintenance was plentiful.

Small factors, such as temperature change, affected frequency stability. Because of the flexible diaphragm in the

World Radio History

gap-tuned klystron, even changes in barometric pressure tended to shift operating frequency. Stresses on the diaphragm resulted in metal fatigue and "creep," thus producing a gradual shift in operating frequency which required periodic readjusting.

The biggest problem, however, proved to be temperature compensation. In this type of tube, cavity dimensions and gap spacing have a profound effect on operating frequency. Both the cavity itself and the external tuning mechanism require very careful temperature compensation. Since they are physically separated, long warm-up periods are required to stabilize operating frequency. Legal and practical requirements for frequency stability make it essential that this type of klystron be operated in some sort of temperature-controlled chamber or "oven." Al-

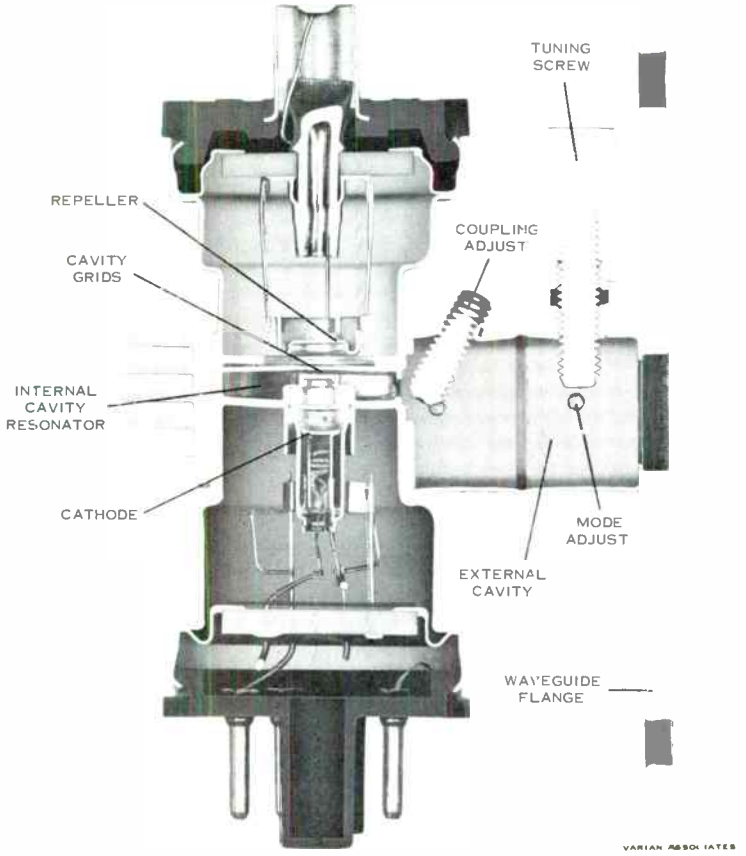


Fig. 3. Cross-section view of external-cavity reflex klystron. Internal resonator cavity is shaped at time of manufacture, has no moving parts within vacuum. Over-coupling between internal and external cavities permits tuning output frequency by simple mechanical screws in outer cavity. Waveguide output is under-coupled to load, preventing frequency "pulling" by changing load impedance.

World Radio History

though this helps the situation considerably, even a brief loss of power, or a malfunction of the thermostat or heater will throw the klystron off frequency for a considerable time.

To cope with these and other problems, many new types of klystrons have been devised. One type finding wide acceptance in microwave communications is shown in cut-away form in Figure 3. This tube avoids the troubles of the older design by having no variable or moving parts within the vacuum. Unlike the gap-tuned klystron, the shape and dimensions of the internal cavity are fixed at the time of manufacture and are not changed.

Tuning is provided by an external cavity electrically coupled to the internal cavity by means of an iris or window between the two. A vacuum seal of ceramic or mica preserves the vacuum in the inner cavity. The two cavities are strongly over-coupled so that changes in the electrical characteristics of the outer cavity will affect the fields set up in the internal resonator, thus permitting the tube to be tuned. Simple mechanical screws entering the outer cavity permit adjustment of coupling, frequency, and mode selection.

The output aperture for this type of klystron is designed to under-couple the tube to its waveguide load in order to prevent differences in load impedance from affecting the output frequency. This feature provides the additional benefit of holding output power constant in the face of reduced cathode emission as the tube ages. The "reserve against age" thus obtained reduces the amount of maintenance required, a valuable consideration in microwave communications systems where some stations are rather inaccessible.

Fig. 4. Ballistic missile early warning radar at the Lincoln Laboratory of M.I.T. Using the klystron shown in Figure 5, this installation made radar contact with planet Venus.

Temperature compensation is achieved by a proper choice of materials used in the internal cavity. Because of the proximity of the cavity to the electron gun and the stream of electrons, this type of klystron achieves its final operating frequency only a few seconds after the filament reaches operating temperature and oscillation begins. Since a temperature-compensated oven is not required, the tube operates at lower temperatures, thus increasing life expectancy.

Not only is this new-found frequency stability useful in the communications industry, but it has permitted the klystron to enjoy greater usefulness in the missile and space development field. The external-cavity klystron has proved especially valuable where high levels of vibration and rapidly changing pressure

World Radio History

Fig. 5. World's largest electron tube, this amplifier klystron was designed for ballistic missile early warning radar. Shown here without external resonators and focusing coils, this tube provides 1¼ million watts peak output power.

make the gap-tuned klystron undesirable or unsuitable.

In the early days of radar, the klystron was overshadowed by the magnetron as a source of high power radar energy. The klystron was relegated to use as the local oscillator in the radar receiver. At that time, frequency stability and the ability to amplify were not as desirable as high-efficiency and extreme compactness.

In recent years, the klystron has begun to displace the magnetron. Modern radar systems obtain considerable information from the phase characteristics of the returning echo, but require accurate phase control of the outgoing pulse. The magnetron, which is only an oscillator and cannot amplify, is generally unable to satisfy this criterion. The klystron, which can amplify, has no such limitation. Modern high-power klystron amplifiers are even approaching the magnetron in efficiency, often achieving 40 to 45% efficiency.

The klystron amplifier shown in Figure 5 is believed to be the world's largest electron tube. This tube was developed for use in a large early warning radar to detect ballistic missiles, and can deliver a peak power output of 1¼ million watts, or sustain a continuous power output of 100 kilowatts!

Impressive as these figures are, other klystrons have been developed which can deliver even greater peak power. At Stanford University, where the klystron was invented as a tool to help explore the atom, 21 extremely high power klystrons are used in the world's largest linear accelerator. These klystrons, each having a peak power capability of 30 million watts, provide the power for accelerating particles to speeds very close to the speed of light. Particles achieve energy levels approaching a billion electron volts, and reach velocities so close to the speed of light that their mass increases 2,000 times!

It is very fitting that the klystron has contributed so well to the basic research for which it was invented. Not only

World Radio History

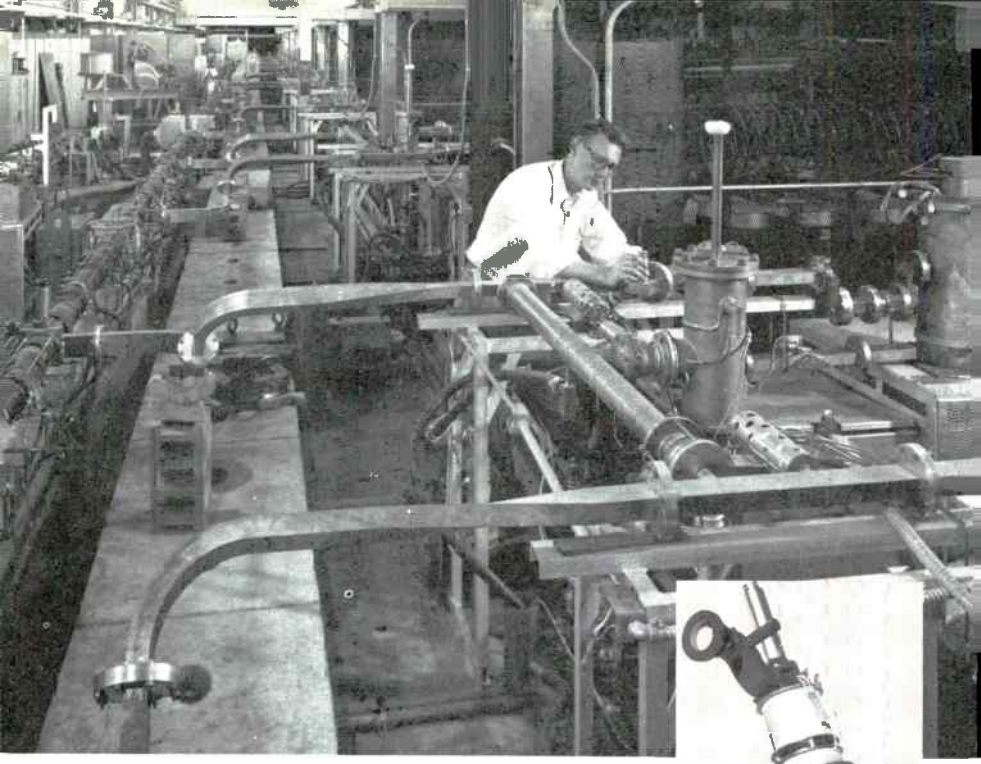


Fig. 6. World's most powerful electron tube. Twenty-one of these klystrons are used in the Stanford linear accelerator. Researchers say that although tubes are each capable of 30 million watts peak power, they are usually operated at "only 20 million watts." Microwave energy from klystrons is used to accelerate particles in evacuated tube on left to speeds approaching that of light. Screened area in background contains pulse network for one klystron.

has it fulfilled its original purpose, but has produced tremendous social and economic benefits as well. Large klystrons make possible tropospheric "scatter" communication sytsems which link remote communities, which never knew a telephone, to the rest of the world. Similar facilities enable early warning radar stations (which also depend on the klystron) to maintain constant communication with control centers in central locations. Tomorrow, a klystron may be the means of talking back to earth from outer space or the planets!

References

1. Russell H. Varian, "The Invention and Development of the Klystron," *Military Automation*, Vol. 1, No. 5; September-October, 1957.
2. J. R. Pierce and W. G. Shepherd, "Reflex Oscillators," *The Bell System Technical Journal*. Vol. 26, pp. 462-681; July, 1947.
3. Theodore Moreno, "The Klystron." *Military Automation*. Vol. 1, No. 6; November-December, 1957.
4. Edward L. Ginzton, "The Klystron." *Scientific American*: March, 1954.

World Radio History

SECTION III
DIGITAL DATA TRANSMISSION

World Radio History

World Radio History

the Lenkurt Demodulator

VOL. 14, NO. 4 APRIL, 1965

INFORMATION THEORY and CODING

Part 1

Information theory began as an academic mathematical inquiry into factors which limit communication. The fantastic growth of man's ability to communicate in the past several decades has confirmed the importance of the inquiry and added to pressures of many important practical problems. Billions of dollars have been spent on communications in recent years and will be spent again to keep up with public demand. Improving the efficiency of communication facilities could channel these funds into providing even greater advances. This article is the first of two which discuss some of the highlights of information theory and how it is used to improve communications.

Transmitting information by various communications techniques is an important part of everyday life. Certainly everyone has used the phrase "a lot of information," but few people regard the fact that it is possible to measure information quantitatively. However, information has been given a numerical value that is very useful in the study of communications.

What is the importance of information theory to a communications engineer? It seems quite reasonable that an engineer who is responsible for selecting solid-state radio equipment for a communications system should have a knowledge of the theory of transistors. It is very unlikely that such an engineer will have anything to do with the design of transistors, but a knowledge of basic transistor theory is nevertheless required for the engineer to be adequately qualified to do his job. In other words, a great deal of fundamental background knowledge is necessary for the broad understanding re-

World Radio History

quired by a competent engineer. For the communications engineers, information theory is becoming an increasingly important part of this background knowledge.

The accelerating growth of data transmission creates an obvious requirement for a broader understanding of information theory. In addition to other uses, information theory provides the fundamental principle for analyzing and comparing existing and future data transmission systems.

Unfortunately, many of the improvements in communications systems suggested by information theory are rather complicated and expensive, thus preventing them from being readily put into use. Nevertheless, as technology progresses, cheaper means of performing the complicated operations suggested by the theory will undoubtedly be found and information theory can be expected to play a more important role in practical communication systems of the future.

Meaning of Information

Information can perhaps be explained as *choice* or *uncertainty*. The effect of the information in a message is to change the probability concerning a situation, as far as the receiver of the message is concerned, from its value before the message is received to what is usually a larger value after the message is received. If an event is certain to occur, the mathematical probability of its occurrence is, by definition, one. If the event is certain not to occur, the mathematical probability is zero. The mathematical probability of the occurrence of any event whose occurrence or nonoccurrence cannot be predicted with certainty lies somewhere between zero and one.

One of the first steps in determining the exact nature of information was the selection of a unit, or yardstick, by which information could be measured. This unit had to be such that it could easily be determined and did not depend upon the *importance* of the message, since a message's importance is difficult to evaluate mathematically.

It turned out that the simplest and most basic unit was the amount of information necessary for a receiver (person or machine) to make the correct choice between two equally possible messages. This choice may be between the messages yes-or-no, on-or-off, A-or-B, 0-or-1, black-or-white, and so on. Since the two possible messages correspond to the two symbols in the binary number system, a unit of information based on two symbols (messages) came to be called a *binary digit* and was abbreviated bit.

A message consisting of one simple electrical pulse has the informational value of one bit because the presence or absence of the pulse permits the receiver to choose the correct message from a set of two. As shown in Figure 1, transmitting two pulses permits the receiver to select the correct message from four equally possible messages. Three pulses, or bits, will enable the correct selection from a set of eight. This selection process gives the average amount of information which must be transmitted to specify a message from a set of equal possibilities.

For instance, suppose a message is to be sent indicating a choice in an election among eight candidates. In the form of communication, the sender has a certain definite limited choice as to what message to send. The receiver of the message will have some uncertainty as to what it will say but he knows that there are only eight possible choices. In order for the sender to indicate to the receiver which candidate he has chosen, he must use some sort of signal.

World Radio History

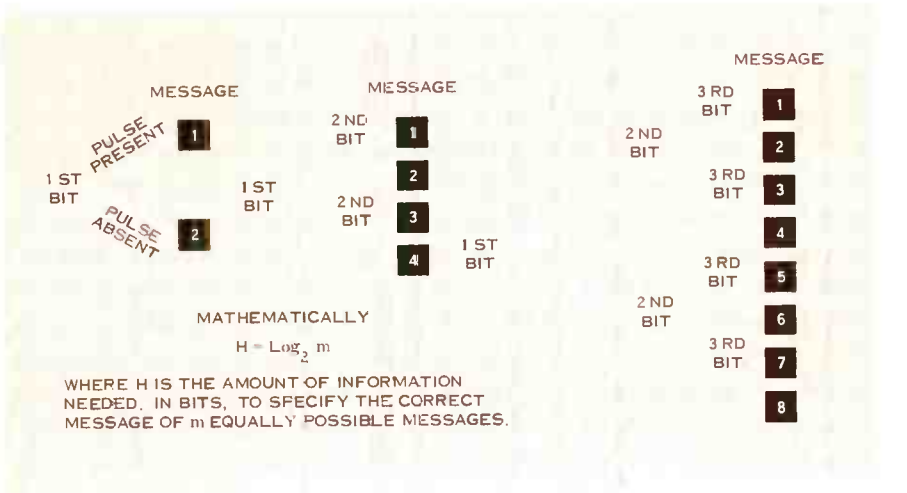


Figure 1. Amount of information required to specify a message from a set of equal possibilities.

One form of signal might be a series of pulses or absence of pulses (for convenience, a *one* can be used to represent a pulse and a *zero* to represent the absence of a pulse). Since there are eight unique series of three pulses or no-pulses possible, any one of the messages can be designated by one series of three pulses or no-pulses. For the receiver, three elementary decisions decide which message among eight was intended.

The information content of a message, expressed in bits, is determined by the formula:

$$H = \log_2 m$$

where

H = number of bits of information
m = number of equally likely choices

If m is 8, the first bit corresponds to a choice of which half of the 8 possibilities is chosen, the second bit to a choice between the first and second pair of the selected half, and the last bit to a choice between the first or second member of the chosen pair. Thus 3 (or $\log_2 8$) bits of information determine the selection, and this is the amount of information acquired by the receiver.

In the example cited, each pulse provides one bit of information. In binary code each code element may be either of two distinct kinds of values: for example, the presence or absence of a pulse. In a ternary or three level code, each code element may be any of three distinct kinds or values; in an N'ary code, each code element may be any one of N distinct kinds or values.

With a simple N'ary code, if all values are equally probable and the probability of any code element is independent of preceding code elements, the amount of information (H) is proportional to the number of code ele-

World Radio History

ments (n) comprising the message multiplied by the logarithm of N.

$$H = n \log N$$

The average information per symbol in any of the commonly used sets of symbols or in any language is always considerably less than its maximum possible value. In effect, this means that parts of messages usually tell things which are already partly known. Thus the intersymbol influences (including interword influences) can predict the nearby succeeding parts of a message to a considerable extent. The actual reception of the message then gives partly a verification or correction of the prediction in addition to completely new ideas. This partial or complete repetition of message content which occurs in languages is called *redundancy*. However, despite the fact that it causes a loss in the rate of transmitting information, redundancy is a very useful property of languages, for it allows individual errors in the transmission of messages to be recognized easily and corrected.

Message Sources

In information theory, message sources are classified as either *discrete* or *non-discrete*. A speaker is an example of a non-discrete message source since the values of a speech wave are drawn from a continuum of possibilities. Such non-discrete sources are very difficult to evaluate mathematically. For this reason, the application of information theory to communications systems is concerned mainly with so-called discrete sources. A discrete source produces messages which are sequences of symbols, the symbols being drawn from some finite list. The most familiar example of such messages is printed English text. The sequence of telegrams passed to the telegrapher for transmission can be thought of as such a source of English text. Other examples of discrete information sources are the input tape to a large computer or the string of symbols printed on a stock exchange ticker-tape.

In analogy with English, the different symbols of a message from any discrete source can be called *letters,* and the finite list of letters from which messages are composed are called the source *alphabet*. The number of letters in the source alphabet represent the *size* of the alphabet. The occurrence of a letter in a message is called a *character* regardless of what letter it is. For example, the word *Mississippi* contains eleven characters but only four different letters.

The first step in describing a given information source is to list its alphabet. This is far from a complete description of the source, however, for the messages produced by most sources have an elaborate statistical structure. The character being printed now is not independent of characters just produced by the source, but depends upon them in a complicated way. What the next character produced by a source will be is not certain. A good guess as to the next character, however, depends strongly on how much of the past message has already been received. If the source is a telegraph and the message "General Eisenho" is observed, then it is quite certain what letter the next character will be. If, however, only the last letter of the message, "o", is observed, what will follow is not so clear.

The statistical structure of the messages produced by a given source can be described mathematically by associating with the source a long list of probabilities. The first probabilities on this list are quantities p_i—the probability that the source will produce the ith letter of the source alphabet. These

World Radio History

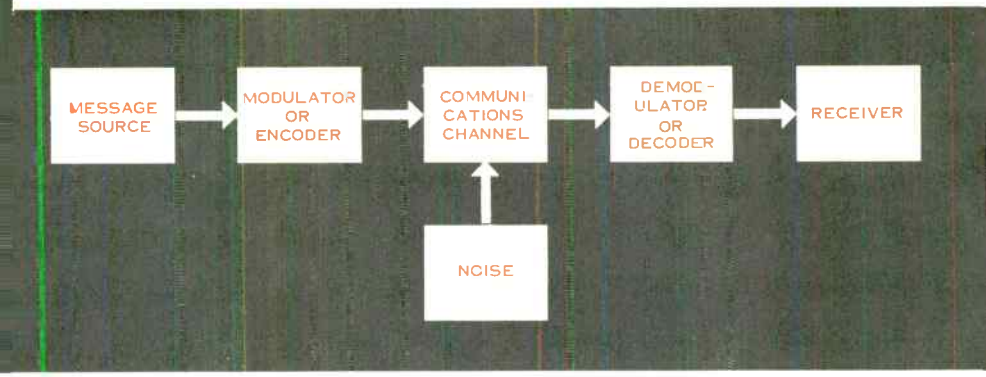


Figure 2. Generalized communications system used in the study of information theory.

quantities reflect the best guess as to what letter the source will produce when the text already produced has not been seen. The next set of characterizing numbers for the source are quantities p_{ij}—the conditional probability that the source will next produce the jth letter of the alphabet when it is known to have just produced the ith letter. Next are the quantities p_{ijk}—the probability that the source will next produce the kth letter of the alphabet when it is known that the source has just produced the ith letter followed by the jth letter. Listing such probabilities in this manner can continue indefinitely, each set giving more information about the long range structure of the messages. In the mathematical model of a message source used in *information theory*, this list of probabilities, along with the source alphabet, characterizes a particular information source.

In actual sources the infinitely remote past of a message certainly exerts no influence on characters being printed now. In fact, in many sources correlation between characters does not extend very far into the past at all. Sources may be classified then according to the number of past characters that exert an influence on the character being produced by the source at the present moment. A source in which the past characters exert no influence on the present character is called a *monogram* source; one in which only the last character produced influences the choice of the present character is called a *digram* source, and so on.

Information Content of English Text

As previously stated, information can be measured in terms of bits or informative yes's or no's, and an element of a binary code contains one bit of information. For the more complex or non-binary codes the information content of each code element increases. If there is an equal possibility of any code element appearing in a sequence, the information value of each code element is equal to the logarithm to the base two of the possible number of code elements.

World Radio History

In written English, there is a finite set composed of 27 symbols, the 26 letters of the alphabet and a word space. If the next letter in a written message appeared at random with a probability of 1/27 for each letter, then the information content of a message expressed in bits would merely be equal to the number of symbols multiplied by $\log_2 27$, using the formula previously given.

However, when the information in a message is in the form of a language, occurrence of the various symbols which comprise the language's alphabet is never completely random. Thus, the appearance of a given letter or a given word in English is subject to "constraints" which act to modify an otherwise completely random probability of occurrence.

In setting up his dot-dash code, Morse made one of the first applications of statistics to a communication problem. On the basis of type counts made in a printing shop, Morse assigned a short code to the most frequently used letters and longer codes to the less frequent. Thus, he could transmit E, the most frequent, by simply sending a dot, but for V, one of the least frequent, he had to send dot dot dot dash.

Thus, Morse would have expected to use more time transmitting letters of gibberish, which might use V as frequently as E, than sensible English in which letters appeared with their familiar frequency.

To get the most possible combinations from its alphabet, a language should allow its letters to fall with uniform probability. Constraints on where the letters fall serve to introduce substantial redundancy in the transmission of information.

As an example, consider the number of bits of information that are con-

LETTER	MORSE CODE
A	·−
B	−···
C	−·−·
D	−··
E	·
F	··−·
G	−−·
H	····
I	··
J	·−−−
K	−·−
L	·−··
M	−−
N	−·
O	−−−
P	·−−·
Q	−−·−
R	·−·
S	···
T	−
U	··−
V	···−
W	·−−
X	−··−
Y	−·−−
Z	−−··

Figure 3. The familiar Morse code was one of the first applications of statistics to a communications problem.

World Radio History

tained in a letter of the English alphabet. If any letter of a 27 letter alphabet were equally probable, the information in one letter would be the logarithm to the base of 2 of 27 or 4.76 bits. Actually, since all letters are not equally probable, when the known probabilities are applied for each letter, it develops that, on the average, each letter probably contains less than two bits. Putting it another way, an 18 letter alphabet of uniform probability could do the same job as our less efficient 27 letter alphabet. However, the redundancy of the English language permits great liberties in transmitting written messages. For example, the telegraph message

PLLESE SXND MONEZ

can easily be interpreted as

PLEASE SEND MONEY

Transmitting Information

If there were no noise to degrade transmission, there would be no limit to information transmission. By transmitting a perfectly measured voltage to represent information, for example, any desired rate of communication could be achieved. In reality, noise masks signals transmitted over communication circuits and introduces uncertainty as to their exact value. Signals tend to be converted into noise by a process of degradation and distortion. In transmitting many channels of information over a multiplex system, each channel requires a certain bandwidth in order to distinguish the signal from random noise. Thus, the greater the number of channels, the greater the bandwidth required. However, as the number of channels increases, each channel signal represents a smaller and smaller portion of the total band. As this occurs, it becomes increasingly difficult to distinguish the signals from background noise unless transmission power is increased.

Information theory studies have revealed the exact relationship between information capacity, signal power, noise, and bandwidth. While these studies have generally confirmed knowledge acquired on an experimental basis, a number of possibilities were revealed that had not been self-evident. It was well known that a smaller signal-to-noise ratio would be acceptable in communications if greater bandwidth were employed, as in FM. It was surprising, however, to discover that in principle, bandwidth could be reduced by increasing signal-to-noise ratio. Heretofore, it was firmly believed that channel bandwidth could never be less than the bandwidth of the original message.

H. Nyquist, a mathematician at the Bell Telephone Laboratories, proved mathematically that the required bandwidth for a communications channel is directly proportional to signaling speed, and that the minimum bandwidth required for transmission of a signal is essentially equal to half the number of binary pulses per second.

Nyquist showed that although there was a limit to the number of pulses per second that could be transmitted over a given communications channel, each pulse might have several distinguishable states or conditions, each of which could carry information. Thus, if amplitude were the variable conveying the information, and each pulse had four possible amplitudes, twice as much information could be transmitted as in a system where pulses had only two possible values.

Nyquist showed that the limit to the number of information-carrying states was related to the noise in the circuit. As stated previously, without noise, there would be no limit to the rate at which information could be transmit-

World Radio History

ted. In the presence of noise, however, the difference in value between two levels or states must be at least twice the value of peak noise. Otherwise, there will be uncertainty as to the value of the pulse.

The same limitation applies to continuous waveforms as well as pulse signals. Actually, there is no real difference between the two. Although a continuous wave may contain an infinite number of points which define its shape, it does not contain an infinite number of information-carrying values. In fact, periodic samples of the waveform can be used to reconstruct or define the waveform perfectly if they are taken often enough. The waveform doesn't have to be sampled very often to make a perfect reconstruction—sampling at twice the highest useful frequency in the signal will do. Thus, if 3000 cps is the highest useful frequency in a telephone channel, a series of brief samples taken at the rate of 6000 per second will precisely and exactly duplicate the telephone conversation! The samples can be as brief as desired, in fact, the shorter the better. Thus, a series of pulses can serve in lieu of a continuous waveform, with no loss whatsoever.

The 3000 cps telephone circuit is a universal communications channel, available almost anywhere in the world. Almost all general purpose communications facilities are designed to accommodate voice signals. Accordingly, this bandwidth has been taken into consideration in designing equipment used to transmit telegraph and data signals and other forms of information.

According to Nyquist's formula for maximum signaling speed, a 3000 cycle channel should be capable of carrying 6000 binary pulses per second. Translated to words per minute, and using the standard Baudot or teletypewriter code, this is approximately 8000 words per minute. Furthermore, the information capacity is considerably higher if codes other than binary are used. The relationship between bandwidth, signal power, and noise is complex and depends upon many factors such as the kind of noise present in the channel, the nature of the power limitation, the type of modulation used, and the method of encoding the information. In 1948, C. E. Shannon, also of the Bell Telephone Laboratories, devised a mathematical formula which defined the *capacity* of a communications channel or the maximum transmission rate. This formula, which relates information rate to the bandwidth and the amount of interfering noise in the system, is shown graphically in Figure 4.

Using Shannon's formula, a channel of 3000 cycles bandwidth and a signal-to-noise ratio (signal power/noise power) of 30 db has a capacity (C) of about 30,000 bits per second:

$$C = W \log_2 (1 + \frac{S}{N})$$
$$= 3000 \log 1001$$
$$= 3000 (9.96)$$
$$= 29{,}880 \text{ bits per second}$$

Of course, this is ideal, non-surpassable performance, and achievable only by the most elaborate coding. Practical communications systems cannot begin to approach that rate of transmission. To achieve such a rate, three conditions would have to be met. First, the transmission medium must be distortionless. Second, the noise power in the channel must be equal throughout the frequency band. Third, the method used to encode the signals must be so complex that no possible combination of noise impulses will ever cause errors to occur during transmission. None of these conditions can be met or even closely achieved with present-day techniques.

World Radio History

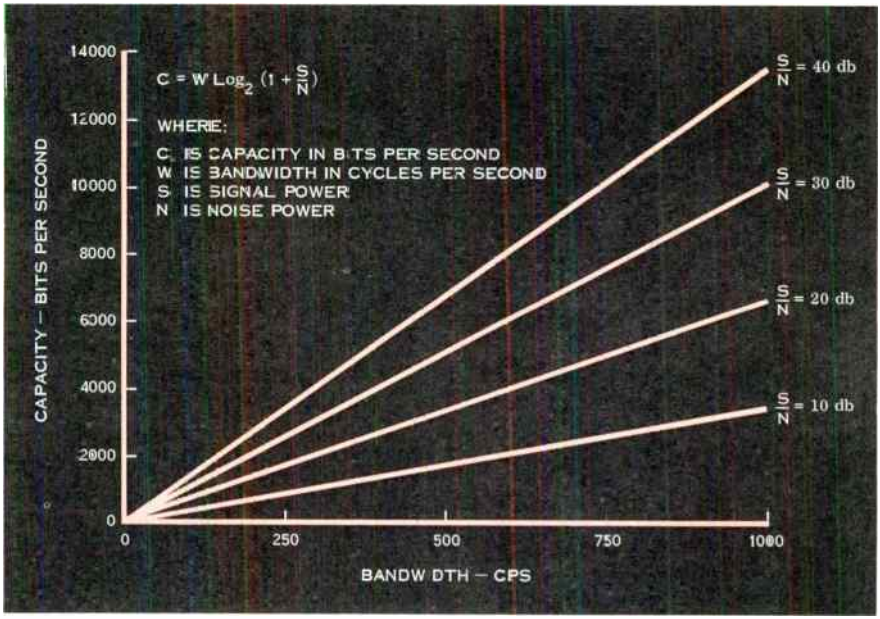


Figure 4. Graph of Shannon's formula which relates information rate to the bandwidth and the interfering noise in a communications system.

Such performance would require whole buildings-full of equipment to encode and decode the message. In addition, the time required for encoding and decoding messages would be far too great for practical needs.

Information theory studies have indicated the existence of ideal codes for transmission over the noisiest channel at rates up to the theoretical limit, and permitting as low a probability of transmission error as required. Redundancy or repetition reduces error probability. By introducing a controlled amount of redundancy in proportion to the channel noise, the desired transmission reliability can be maintained under the worst conditions, at the cost of reduced transmission rate. If reduced transmission rate cannot be tolerated, more complicated coding can theoretically be employed to keep down the probability of error. The subject of coding and error control is discussed in Part 2.

World Radio History

World Radio History

the Lenkurt Demodulator

VOL. 14, NO. 5 MAY, 1965

INFORMATION THEORY and CODING

Part 2

There is a never-ending need to overcome the various kinds of interference inherent in today's communications systems, so that information can be transmitted faster and with fewer errors. One way of providing improvements is to design special coding schemes which better organize the messages sent over communications systems. This article discusses some of the considerations in coding and decoding messages and the methods of detecting and correcting transmission errors.

COMMUNICATION requires that information be transported from one place to another and, for this purpose, must be converted into a form suitable for handling. Electrical communication requires additional conversions to prepare the words or other symbols for transmission. Sound waves are converted to a variable voltage; electrical pulses, like drum beats or smoke signals, provide the means for transmitting letters and numerical data over today's modern communications systems.

Regardless of the exact means of transmission, some form of symbolic language or code must always be used to carry information from its source to its destination, and most of these codes and languages are inherently wasteful. In language, some words are used more than others and letters occur in predictable patterns. This predictability and pattern in sounds, letters, and words make it possible to receive the meaning of a spoken or written message, even when some part of it is altered or deleted in transmission. A

World Radio History

reader's familiarity with the words and syntax of a language allow him to supply missing or incorrect letters and words in the text. The prolonged sounds of speech, and their inflection and pattern preserve the intelligibility of speech except in the presence of extreme interference.

A simple experiment will confirm how predictable language actually is. A short passage of written prose is selected, and someone is asked to guess the characters (including spaces and punctuation) one at a time. The subject continues guessing until he names each character correctly. As each character is guessed, it is written down as an aid in predicting the next character. The results of such an experiment are shown in Figure 1. The numerals show the number of guesses required for each character. Of the 109 symbols in the text, the subject guessed correctly on his first try 79 times, and was able to identify all 109 characters in 235 attempts. This is an average of only about two guesses (or information "bits") per character. Further experiments have indicated that long passages of English text have an information content of only about one bit per letter. This means that, theoretically, it should be possible to transmit text by pulses no more numerous than the letters themselves, thus enabling 24 of the 26 letters to be discarded without loss of communication. Although this ideal cannot be achieved, it provides a goal to be approached in the design of coding techniques.

Transmission codes can be made more efficient by designing them to fit the statistics of the language. Thus, letters which occur most frequently—E, T, and A, for instance—are represented by the shortest code symbols, while the least probable characters have longer symbols. Figure 2 shows such a code which has an average information content of about 4 bits per character. By contrast, the standard teletypewriter code employs 5 bits per character, not counting synchronizing pulses.

Although the additional redundant symbols and pattern in language may help overcome errors, *unsystematic* redundancy is wasteful, and merely lowers the rate of communication. It follows logically that the more redundancy removed, the more efficient the communications channel, but the greater the likelihood of error due to interference. Since interference is always present to some degree, a very efficient communications system would use a code in

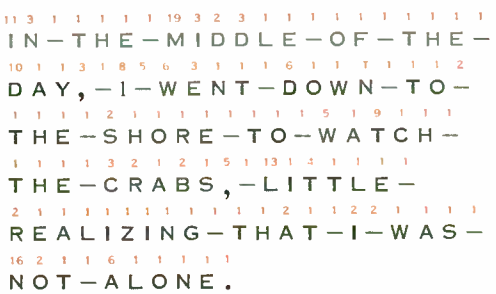


Figure 1. Predictability of language is indicated by large number of characters correctly guessed on first try. Numerals indicate number of guesses required to identify each character.

World Radio History

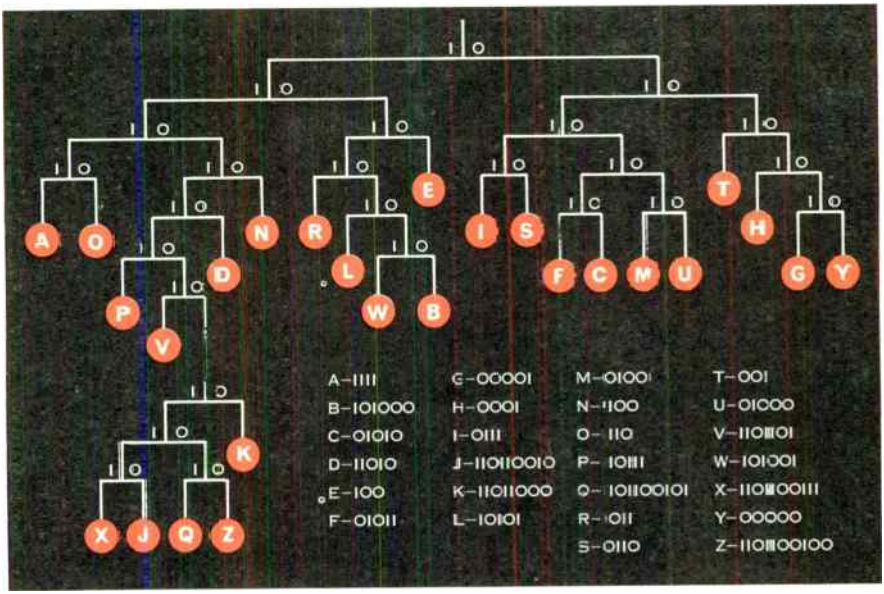


Figure 2. Efficient binary code for English requires average of only 4 bits per character by taking advantage of language statistics.

which all message redundancy was eliminated to obtain maximum information rate; then, just enough redundancy would be re-inserted to overcome the interference present in the transmission path.

Unlike the redundancy in spoken and written languages, data-type messages have no inherent redundancy. Machine-generated characters occur without pattern, and errors cannot be detected by inspection, as in the case of text. To complicate matters, data errors cannot be tolerated to the same extent as errors in text, because control operations or machine calculations may be completely ruined by a single error. Yet the high speed with which data is generated and transmitted makes the occurrence of errors more likely.

One way of overcoming errors in handling and transmitting high-speed data is to design codes which, by their very construction and organization, are able to detect or even correct errors automatically. Unfortunately, most such codes cannot be created without adding redundancy. The problem then becomes one of finding a coding method that provides maximum error-free transmission with the least possible redundancy.

Error Probability

The information capacity of a communications system with a finite bandwidth depends primarily upon the effective signal-to-noise ratio at the detector or receiver. Noise power in the system is generally considered to be completely random and adds to or sub-

World Radio History

tracts from the signal power. The addition of this noise to digital signals makes it difficult for a detector to always make a correct decision, thus causing errors.

In a typical binary system, for example, the digital codes 1 and 0 are represented by different amplitudes as shown in Figure 4. The detector must determine whether a signal pulse is a 1 or 0 by its amplitude at the time of sampling. (Signals are usually sampled at the center of the pulse.) If the signal amplitude at the time of sampling exceeds a set level, called the *decision threshold* or *slicing level,* a binary 1 will be indicated. If the signal amplitude is less than the slicing level at the time of sampling, a binary 0 will be indicated.

The amplitude of the pulse at the sampling time is proportional to the vector sum of the signal power and the noise power. If the signal at the detector is a binary 0, then the noise power would have to be of such amplitude and phase that it would *raise* the amplitude of the pulse above the slicing level to produce an error. Conversely, if the signal is a binary 1, the noise power would have to be of such amplitude and phase that it would *lower* the amplitude of the pulse below the slicing level to produce an error.

In Figure 4, the slicing level is set at half the peak signal amplitude. This means that whenever the amplitude of a signal pulse at the sampling time is distorted by an amount equal to half the peak amplitude, an error will occur because the detector will indicate the wrong binary symbol. If random noise is considered to be the only cause of signal distortion, then the chance of error is related to the probability of the noise power becoming greater than half the peak amplitude set for the signal pulse. This implies, of course, that the greater the signal-to-noise ratio, the less chance there is for error. Therefore, given the signal-to-noise ratio, the probability of random noise peaks causing errors can be estimated by using the mathematics of statistics and the so-called *normal* or *gausian distribution* values which have been well tabulated. The error rates established by this sta-

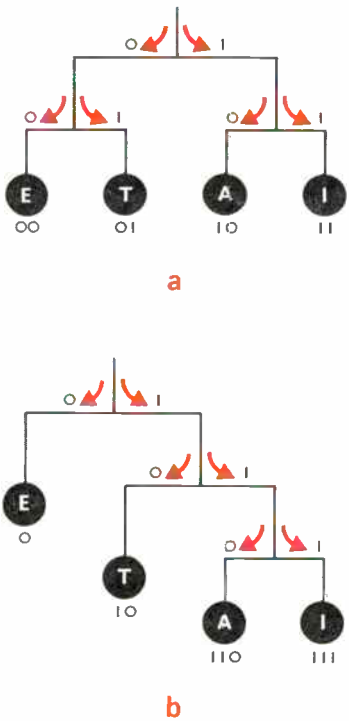


Figure 3. Two methods of coding information. In (a), characters are equally probable and two bits are required for each. In (b), characters known to appear more frequently are assigned a shorter code thus reducing the total number of bits required.

World Radio History

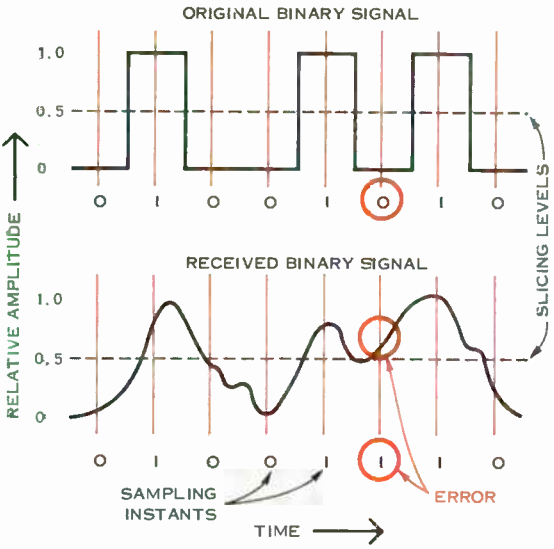


Figure 4. In a binary system, digital codes I *and* O *are represented by different amplitude levels. When a pulse is sampled at the detector, its amplitude is proportional to the vector sum of the signal power and the noise power. The addition of this noise may distort the signal to such a degree that the detector will indicate the wrong code at the time of sampling.*

tistical method provide an excellent measure of performance that is especially useful in rating digital communications systems. The error rate performance of today's communications systems ranges from one bit in 100,000 to over one bit in 1,000,000. Figure 5 shows curves of error probability versus signal-to-noise ratio (in decibels) for three types of digital systems.

It is important to note that in comparing multilevel codes, such as ternary and quaternary, with a binary code, the probability of error increases when the peak-to-peak signal range is the same. As shown in Figure 6, two slicing levels are required for a ternary signal and three slicing levels are required for a quaternary signal. Although these additional levels increase the information capacity of the signal, when compared to a binary signal, the margin against noise is reduced by a factor of $1/(n-1)$ where n equals the number of levels. Thus, for a quaternary signal, where n equals 4, the margin against noise is reduced by one-third

Error Control

Error control has become an essential part of pulse or data transmission systems since it is not practical to make circuits perfectly error free. The method adopted depends on whether or not the circuit provides one-way or two-way transmission and its error performance – that is, the type and distribution of errors. The use of error detecting and error correcting techniques can increase the overall accuracy of the transmission within the capacity of the channel to any accuracy required but at the expense of equipment complexity.

Shannon's formula described in Part 1, arrives at the remarkable conclusion that even a noisy channel has a definite

World Radio History

errorless capacity. No matter how low the error rate must be, it can be achieved while still transmitting a signal over the channel at the desired rate provided that the rate of information does not exceed the channel capacity. However, as the error requirements become more stringent, it becomes more difficult to transmit the signal but only because the

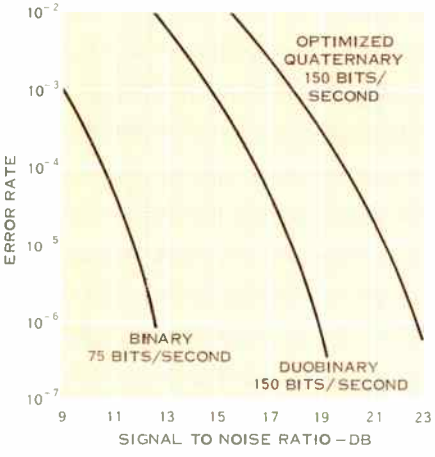


Figure 5. Error rates versus signal-to-noise ratio for three different types of digital systems.

encoding becomes very complex and imposes a long transmission delay in coding and decoding. Practical considerations necessitate using a simple encoding alphabet, thus wasting a good fraction of a channel's capacity. Shannon showed that the information capacity of a channel expressed in bits per second is:

$$C = W \log_2 (1 + \frac{S}{N})$$

where

W = *Bandwidth in cycles per second*

S = *Average signal power*

N = *rms noise power in a one cycle band*

Take for example a typical telephone-type voice channel:

W = *3000 cps*

$\frac{S}{N}$ = *20 db (signal S at −15 dbm0 and noise N at −35 dbm0)*

C = *20,000 bits per second*

In existing systems only a small fraction of this maximum possible channel capacity is obtained. Indeed the equation gives no indication of how such ideal encoding of the message may be realized. In any practical system yet proposed there will be a finite probability of error for a finite transmission rate. The sophistication needed to reach the channel capacity of Shannon's formula would result in an extremely complex system.

Error control systems today are either error detecting or error correcting. One of the simplest methods of reducing errors is to repeat the message several times. A more elaborate approach is to use some form of coding which enables a block of characters to be tested for errors. If no errors are found, a feedback signal is sent which acknowledges receipt of the block and asks for the next block to be transmitted. If no acknowledgment is received, the original block is retransmitted. These systems achieve more accuracy at the price of a slower rate. The transmission efficiency of this type of system can be expressed as the number of information bits per block divided by the sum of the information bits plus the redundancy bits, plus the number of bits that could have been sent in the waiting time for an-

World Radio History

swer-back signals together with the average additional time for repeated transmission.

Thus:

$$E = \frac{Bi}{Bi + Bh + Bw}$$

where:

E = *Transmission efficiency*

Bi = *Information bits per block*

Bh = *Redundant bits per block*

Bw = *Waiting bits per block*

Figure 7 shows how waiting time effects the transmission efficiency of the system. It also indicates that adding redundant bits seriously limits the overall rate of transmission.

It is interesting to note that the coding system based on the Lenkurt developed duobinary technique (described in the February 1963 DEMODULATOR) gives a degree of error detection without adding redundant digits. This capability is achieved by increasing the amount of information per digit. The duobinary code is a more powerful error detection system than the simple *parity check* and has the additional advantage that the data does not have to be processed before the error becomes apparent.

Parity Checks

A widely used error detection scheme is the so-called parity check. An extra digit is added to the regular binary code group so that there will always be an even (or odd) number of 1's in each group. A single error will cause an odd number of 1's to appear at the receiver, indicating an error. A single parity check will detect all odd numbers of errors, but will not detect double errors or other even-count errors, since the count of 1's will still provide the required even number. By adding an additional parity check for every other digit, all odd numbers of errors and about half the even number of errors can be detected. A third parity check added for the remaining digits will further reduce the undetectable errors.

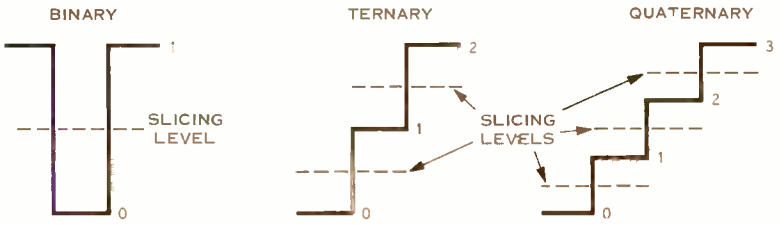


Figure 6. Multilevel codes such as ternary and quaternary increase the information capacity of the signal. However, when the peak-to-peak amplitude is the same as for a binary signal the margin against noise is reduced.

World Radio History

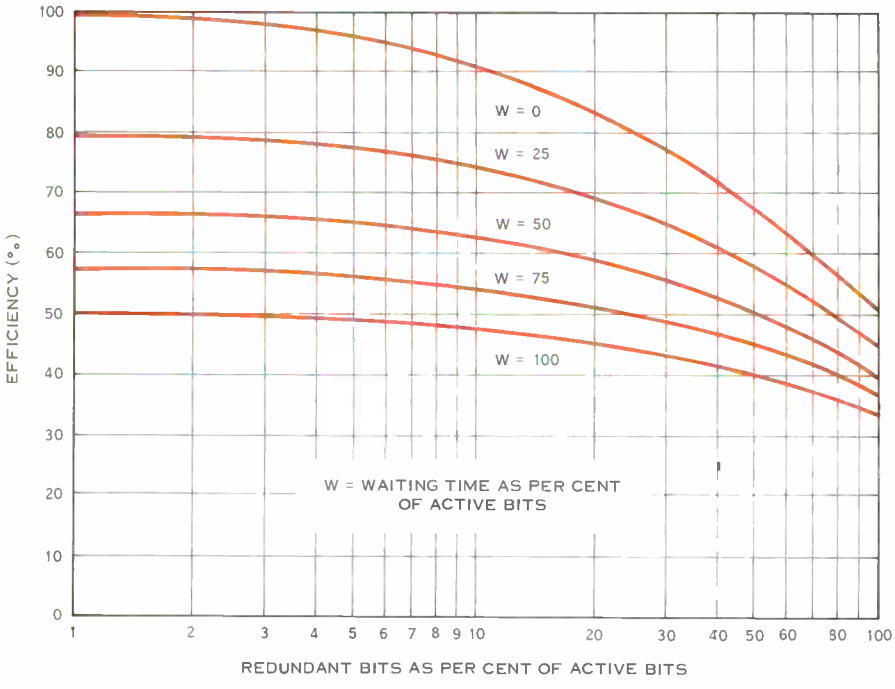


Figure 7. Effect of waiting time on transmission efficiency for one type of error-control system where a block of characters are tested for errors. Curves show that adding redundant bits limits the overall rate of transmission.

Parity checks provide some protection against errors, but like all redundancy, they slow down the transmission of the message. If a single parity check is used with each five-digit code group, as shown in Figure 8, the message will contain about 16% redundancy. This can be reduced by increasing the number of information digits for each check digit, but this increases the probability of undetectable errors occurring.

Many types of parity check systems exist. Where parallel transmission is used (tape-to-tape computer data, for instance), parity checks may be used in both the horizontal and vertical directions, in order to reduce the chance of data errors going undetected.

A related approach to error detection uses a fixed ratio of marks and spaces for all code characters. When designed to reduce the likelihood of compensating errors, this code can be very effective in detecting most errors.

Essentially, error detection coding and retransmission make an excellent system for reliable communication if the transmission channel introduces only a few scattered errors, and if a high quality return channel is available. The system deteriorates rapidly as the error rate increases.

World Radio History

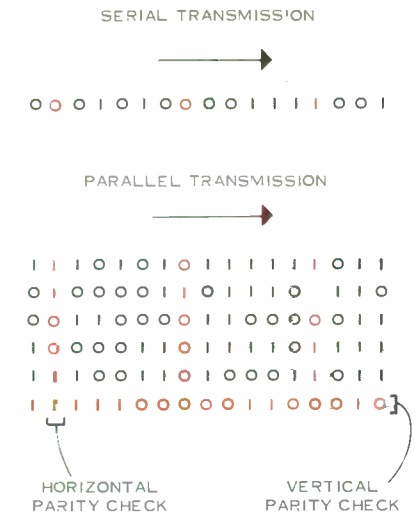


Figure 8. Typical parity checks for serial and parallel transmission. Check digit is chosen so that sum of marks *in each block is always even (or odd) — wrong count reveals error.*

The fraction of errors actually detected in transmission will depend on the characteristics of channel noise. Adding more parity bits will increase the protection afforded by the code. If a block of bits of arbitrary length includes parity bits, the number of undetected errors will amount to the fraction 1/2p of all possible errors if the error detecting code is efficient.

Thus:

$$d = \left(1 - \frac{1}{2p}\right) \times 100$$

where

p = *Number of parity bits*

d = *Percent of errors detected.*

Figure 9 illustrates the dependency of detection efficiency on the number of parity bits.

When properly arranged, parity checking by block is a very powerful

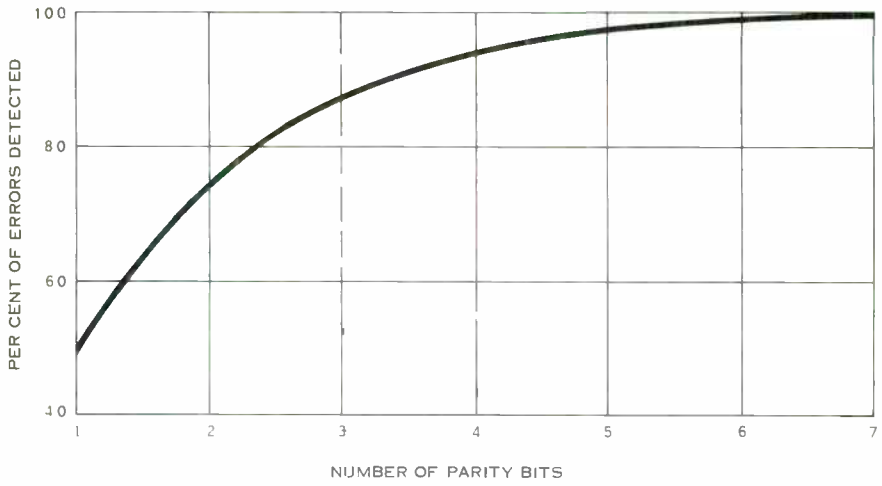


Figure 9. Efficiency of parity check error correction system depends on the number of parity bits.

World Radio History

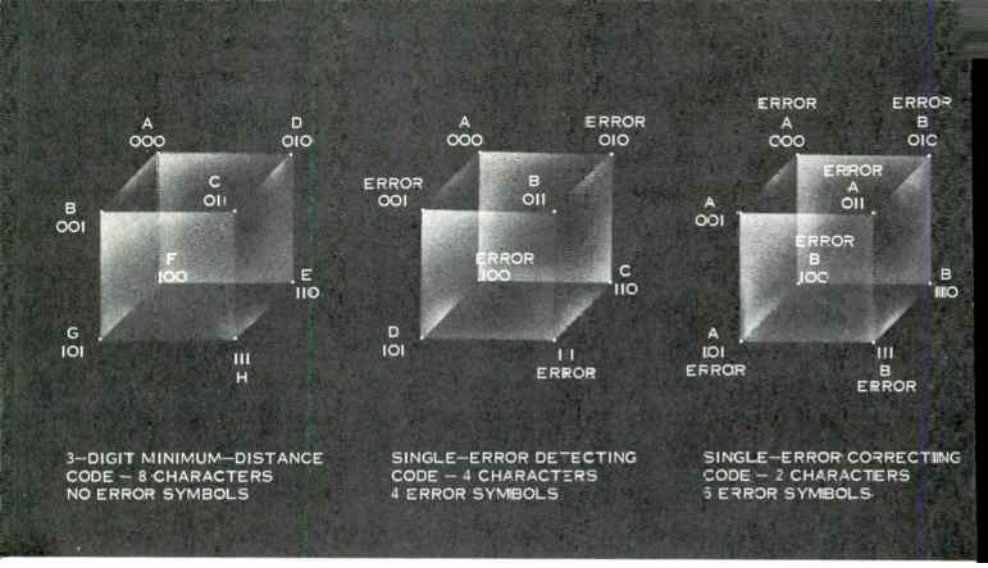


Figure 10. The efficiency and certain other characteristics of complicated codes are analyzed easily by mathematicians through the use of solid geometry. Simple 3-digit code shown in the example requires three-dimensional figure. Each vertex represents a code combination assignable to characters or errors.

device. Consider, for example, a block of 80 6-bit characters used to transmit the information contained on a standard 80-column punched card. A single parity check on each character is only 50 per cent efficient and requires 80 parity bits to a card. By contrast in a block detection system, 9 parity bits could be 99.8 per cent efficient and detect any burst of errors up to 8 bits long. The redundancy will be less than 2 per cent.

Error Correction

It is not enough merely to identify the existence of errors. Some means of correcting the message is required in order to complete the transmission or control function. One basic form of error correction is to transmit the message several times in the hope that errors will not destroy identical portions of each mesage. A similar approach would be to transmit each digit several times and count the bits received. A majority count would presumably reveal the correct digit. Obviously, this method fails if more than half the digits are in errors.

Error-correcting codes which do not require retransmission have been devised, using principles similar to those used in the code of Figure 1. Error correction is obtained by adding additional redundant digits so that an erroneous code group still most nearly resembles the intended group, despite changes occurring in one (or more) binary digits. Obviously, the redundancy is greatly increased.

Mathematicians specializing in information theory and advanced coding techniques find it useful to describe codes in terms of geometry, so that each character in the code is located at a "corner" or vertex of a geometrical

World Radio History

figure. Thus, a code having 2 digits could be described by a square with all four combinations located at the four corners. A code with three digits would require a three-dimensional figure for the eight possible combinations, and a four-digit code requires a solid having four dimensions to adequately describe its properties. Although it is difficult or impossible to diagram multi-dimensional figures accurately on paper, they are relatively easy to handle mathematically.

Since each code combination, whether an error or a correct symbol, lies at a vertex of the solid figure, a change in one digit represents the difference between one vertex and an adjacent one. Two changes move it two places, and so forth. The ideal code, then, will use the least possible number of code combinations, but separates all *valid* (non-error) code groups by as many locations as possible. The less the "distance" between correct symbols, the lower the redundancy. If additional "distance" is placed between valid characters, the code can either detect multiple errors or correct single errors, depending on how the code is set up. Figure 10 diagrams how a geometric figure can be used to express "distance" between symbols, and shows how efficiency or information capacity can be traded for error correction or detection capability.

Conclusion

Information theory has provided the designers of communications systems with new insight into the intangible commodity with which they work. By providing engineers with a specific measuring stick, information theory enables them to measure the efficiency of their communications apparatus and make improvements. The theory is important not only to conventional communications media, but it has important implications in computers, control systems, and data systems where machines communicate directly with machines.

Information theory studies have revealed two basic approaches to improving communications. One is based on improved coding of the signal to be transmitted, the other stems from new knowledge of the relationship between signal power, noise, and bandwidth. There are indications that both approaches lead to a common goal; the most efficient coding method will possibly be the most efficient way of compressing bandwidth and overcoming noise in a transmission channel.

BIBLIOGRAPHY

1. Shannon, C. E., and W. Weaver, *The Mathematical Theory of Communications*, University of Illinois Press, Urbana, Illinois, 1949.
2. "Information Theory and Modern Communications," *The Lenkurt Demodulator*, June, 1959.
3. "Methods for Transmitting Data Faster," *The Lenkurt Demodulator*, March, 1960.
4. Reza, Fazlollak M., *An Introduction to Information Theory*, McGraw-Hill Book Co., Inc., New York, 1961.
5. "Special Code Techniques for Improved Data Transmission," *The Lenkurt Demodulator*, June, 1961.
6. Abramson, Norman, *Information Theory and Coding*, McGraw-Hill Book Co., Inc., New York, 1963.
7. Bennett, W. R., and J. R. Davey, *Data Transmission*, McGraw-Hill Book Co., Inc., New York, 1965.

World Radio History

World Radio History

the Lenkurt Demodulator

VOL. 9 NO. 3 MARCH, 1960

Methods for TRANSMITTING DATA FASTER

Increased use of data transmission between distant business machines, computers, and other information-handling devices is focusing attention on the need for reliable data transmission at higher speeds. This article discusses some of the methods being developed to increase the speed of reliable data transmission over telephone circuits.

High speed data transmission presents no problem if there are no restrictions on the cost of the system or the kind of transmission to be used. The great bandwidth of microwave radio links, for instance, permits millions of bits per second to be transmitted. Such means of transmission may be relatively expensive, however.

Although many special applications require high-capacity data transmission channels, most needs may be satisfied by systems operating within the capacity of a standard telephone voice circuit. These telephone circuits are available throughout the inhabited world. Standard communications facilities are usually divided into such channels, and these channels almost always have a useful bandwidth of 3000 cycles or more. In some cases, inductive loading or other physical characteristics of the transmission line may reduce its ability to handle high speed data, but this quality of circuit is becoming more the exception than the rule.

It is only logical, therefore, that communications engineers have concentrated on data transmission systems designed for voice circuits. The prob-

World Radio History

lem is to make the most efficient use of existing facilities.

Transmission Methods

Information has to be converted into some sort of code or set of symbols for transmission. These may be letters, words, or numbers — or they may be the dots which make up a television or facsimile image. The symbols must undergo further transformation in the form of modulation to adapt them to the transmission medium. For instance, a tone or a direct current may be keyed to convey code symbols; a subcarrier frequency may be shifted in phase or frequency in a systematic manner. The choice of modulation method is very important in determining the actual performance of a high speed transmission system. There are advantages and disadvantages to nearly all the methods now being used or seriously considered.

Amplitude Modulation

Amplitude modulation (AM) methods are historically related to direct-current telegraphy. In d-c telegraphy, a battery or other source of direct current is keyed on and off. At the receiving end, the signals are detected by some sort of magnetic receiving device. In AM telegraphy, the process is similar. Instead of direct current, a tone or subcarrier is keyed, so that the two binary states, "mark" and "space", are indicated by the presence or absence of the tone.

This method has several disadvantages: it does not use bandwidth efficiently, since two sidebands of the carrier are produced. Unlike single-sideband voice communications methods, the carrier and one sideband can not be completely eliminated and still do a satisfactory job.

Sidebands are produced when the modulating wave causes the carrier to change from one value or state t another. In the case of voice communi cations, the modulating waveform i continuous, thus causing modulatio products (sidebands) to be forme continuously. If the carrier and on sideband are eliminated, the other side band remains to convey the modulatin intelligence.

In the case of telegraphy, where on off pulses are the modulating signal modulation products are formed onl during the transition from "on" t "off", and from "off" to "on." Thes modulation products are transient whose bandwidth is a function of the keying or switching rate. Except when a pulse is started or ended, no modulation products can appear in the trans-

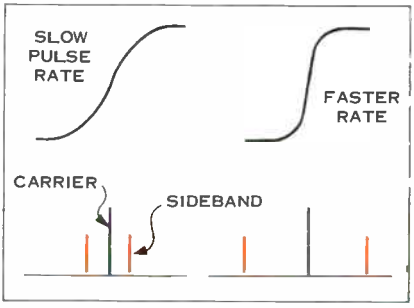


Figure 1. Sidebands are farther from carrier at high keying rates. Round-cornered pulses require less bandwidth than square-cornered pulses.

mission path. Thus, it would be impossible to continuously transmit a steady mark or space. It is possible to design a system with "memory" so that only changes are transmitted. In such a system, after receiving a mark, the receiver holds a marking condition until a signal is received indicating a space. Such systems usually must be quite complex to offset the problems of ambiguity and errors due to interference.

World Radio History

The information-carrying characteristic of an AM signal is its amplitude. For this reason, AM is particularly vulnerable to impulse noise and changes in transmission level. Impulse noise is particularly disturbing. Noise pulses caused by electrical storms, switching transients, and similar disturbances, may equal or exceed the information pulses in amplitude and duration. Under severe conditions, impulse noise may completely obliterate an AM information pulse.

Vestigial Sideband Transmission

Vestigial sideband systems result from an effort to reduce the bandwidth requirements of AM transmission. As we saw above, it isn't practical to eliminate one entire sideband from a telegraph or data transmission channel, despite the reduction in bandwidth. Instead, one entire sideband and only a vestige (10% to 15%) of the other sideband are transmitted. This is done by filtering the telegraph channel so that the output signal barely includes the edge of the sideband to be reduced, as shown in Figure 2. Although this reduces the total bandwidth to be transmitted, another difficulty appears immediately. Because of the lack of symmetry of the signal obtained from the filter, a *quadrature component* — a wave component 90° out of phase with the basic signal — is produced and combined with the signal. This distorts the pulse, making it more difficult to identify or reconstruct at the receiving end, if normal envelope detection is used. This difficulty adds to the inherent AM vulnerability to changes in level, and makes such a system even more sensitive to interference.

To reduce quadrature component effects, the transmitting modulation index is reduced. The amplitude of the "space" may be increased from the normal value of zero to 30% or 40% of the amplitude of the "mark." In other words, the *difference* in amplitude between the mark and the space is reduced, thus making vestigial sideband transmission even more sensitive to noise.

The sensitivity of vestigial sideband transmission to noise has been confirmed by extensive experience in the SAGE data systems. Since most SAGE

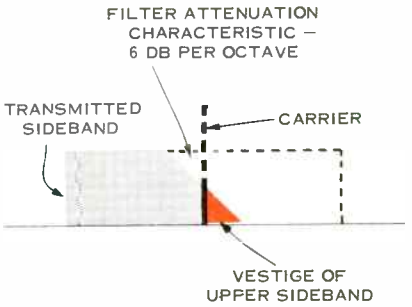


Figure 2. Vestigial Sideband is produced by special filter which attenuates carrier and second sideband at 6 db-per-octave rate.

channels are operated through regular telephone offices, impulse noise caused by dialing pulses and other switching transients normally found on telephone circuits are present in large amounts. All of these types of impulse noise reduce the accuracy or reliability of vestigial sideband transmission. In addition, vestigial sideband is so sensitive to disturbance that in SAGE systems many errors occur which have no readily apparent cause. The actual cause is probably a combination of minor line disturbances, each of which is too small to be very noticeable.

On the good side, vestigial sideband methods are believed to require less

World Radio History

correction for delay distortion than most other approaches to high speed transmission. Also, it is relatively uncomplicated and easy to maintain.

Phase Modulation Methods

From a *theoretical* viewpoint, phase modulation (PM) is perhaps the most attractive approach to higher speed data transmission. In this method, the *phase* of a carrier is shifted a certain number of degrees ahead of or behind the normal sine wave of the carrier. For instance, "mark" might be signified by momentarily advancing the phase of the carrier by 90° or 180°. A "space" might be represented by retarding the carrier phase a like amount.

In practical PM systems, as much phase difference as possible is used between channels (or between mark and space conditions of one channel) by employing as few phases of a single carrier as possible. The most rudimentary form of phase modulation uses the unmodulated carrier to signify a space, and reverses the phase of the carrier (shifts it 180°) to signify each mark. Thus, a series of consecutive marks would be indicated by continuous phase reversals, such as shown in Figure 4.

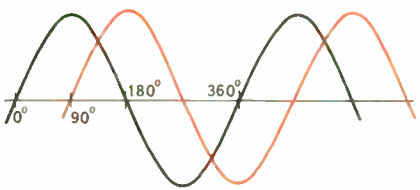


Figure 3. Phase Modulation systems are based on shifting the phase of a carrier from the normal or "zero" phase. Red waveform is shown advanced 90° in phase.

Phase modulation has the advantage of being insensitive to level variations, and being able to transmit low modulating frequencies, including zero frequency (a continuous steady-state condition). Theoretically, PM makes the best use of bandwidth for a given transmission speed.

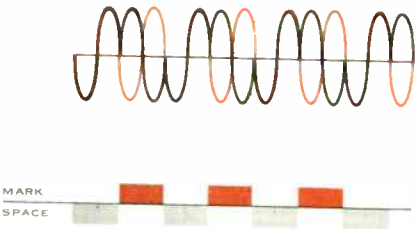


Figure 4. Phase reversal indicates marking condition. Red waveform indicates zero phase.

On the negative side, phase modulation methods have very great difficulty achieving or approaching their theoretical superiority over other methods except at the cost of extremely complex and sophisticated methods. Since the information-bearing characteristic of the received signal is its phase, the quality of any phase-modulated transmission depends on the accuracy with which the phase shift of the transmission path is stabilized. All actual communications paths show some phase drift. This is caused by temperature changes in cable or wire, atmospheric effects on radio paths, momentary variations in transmission equipment performance, and so forth.

Another major problem of phase modulation methods is *phase ambiguity*. Actual transmission paths are subject to many kinds of disturbance, both major and minor. Impulse noise may

World Radio History

occur because of man-made disturbances, or because of electrical storms. If the signal is momentarily interrupted by any cause, the receiver has no way of knowing whether the first data pulse received after the interruption is the reference (or zero) phase, or whether it represents a mark or a space. If the receiver makes the wrong decision, a mark or space might be interpreted as the reference phase. All other signals transmitted after the interruption would then be misinterpreted, resulting in a thoroughly garbled message.

Most approaches to solving this ambiguity problem require the transmission of a pilot tone or other signal which carriers no message information. Some systems transmit timing pulses in addition to the pilot tone, to make sure that there is no ambiguity. Such techniques, while quite necessary for reliable performance, reduce the efficiency of the system. Although the ideal phase modulation system should enjoy about a 2-db advantage over FM, 7 db over AM, and 16 db over vestigial sideband transmission, much of the advantage is lost in preventing phase ambiguity. Like vestigial sideband transmission (which reduces modulation index to minimize quadrature component distortion of the transmitted wave), phase modulation methods sacrifice some of their basic advantage in solving related problems.

Any data transmission system that achieves the full theoretical advantage offered by phase modulation must do so at the cost of extreme complexity — with attendant high cost and difficulty of maintenance and adjustment — unless the signal path is very short or the signal power is unusually high.

Frequency Modulation

Frequency modulation, sometimes called *frequency shift keying* when applied to telegraphy or data transmission, has long been used for transmission of telegraph signals by radio or carrier systems. Like phase modulation, FM is insensitive to variations in level, and has a six-to-sixteen db signal-to-noise advantage over amplitude modulation methods. Because of the way in which FM trades bandwidth for freedom from interference, it doesn't enjoy as much freedom from noise interference as PM, for the same bandwidth.

In normal practice, binary signals are transmitted by allowing one frequency within the transmission passband to represent mark and another frequency to represent the space. Since the mark and space signals are represented by different frequencies of equal strength, amplitude variations have no effect on the signal unless the signal has the same or less amplitude than the noise. This contrasts strongly with amplitude modulation where a mark is represented by the presence of carrier and a space is indicated by a lack of carrier. Level changes due to fading, noise, and other interference have a strong effect on AM signals. FM systems can tolerate level changes of about 40 to 50 db, and are

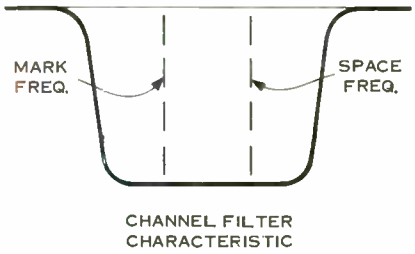


Figure 5. Single oscillator alternates between two frequencies to indicate binary mark and space in FM data system. Sidebands occupy remaining bandwidth.

World Radio History

about 12 db less sensitive to impulse noise than AM systems.

Conventional FM telegraphy is accomplished by alternately transmitting two frequencies representing mark and space. A diode keyer in the tank circuit of an oscillator changes the tank circuit so as to shift the tone back and forth between the two frequencies. Such frequency shifting does not occur instantaneously, however. The inherent resonance of the tank circuit causes the resulting waveform to change smoothly from one frequency to the other.

In an FM system, information is conveyed by the instantaneous frequency of the waveform. This instantaneous frequency determines the exact time at which the waveform crosses zero. The shorter the time, the higher the frequency. The only way that impulse noise or interference can affect an FM transmission is to change the instantaneous frequency. The random pulses of energy which comprise impulse noise have relatively little effect on an FM signal except when they occur near a zero crossing. Where noise impulses combine with the signal to hasten or retard a zero crossing, pulse distortion results. This may or may not produce an error, depending on how much the pulse was distorted.

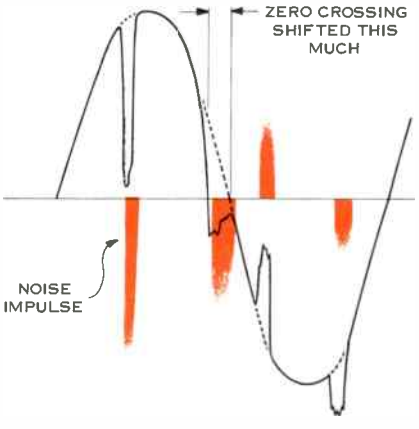


Figure 6. Impulse noise distorts received FM pulse only by speeding or delaying zero crossing of waveform.

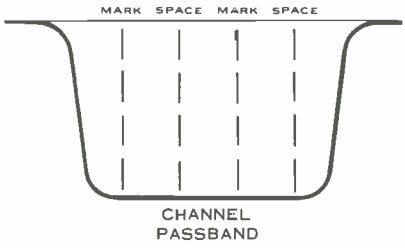


Figure 7. Four tones in channel bandwidth permit information rate to be doubled.

Band Compression

Phase modulation systems can transmit more information in a given band of frequencies by displacing a carrier fewer and fewer degrees of phase between channels. A similar increase in information rate can be obtained in AM systems by transmitting more levels. This is undesirable because noise sensitivity — which is already great — becomes much worse. An equivalent method of band compression is available to FM, based on the manner in which the modulating signal is encoded. To make the comparison clearer, let us first review the operations involved in phase modulation.

In a system where binary mark and space are indicated by 180 degrees phase separation, the voltage appears as "on" and "off." If an additional channel is obtained by displacing the mark and space conditions 90 degrees instead of 180 degrees, the phase detector output will consist of four different voltage levels — two for each channel. Since the amplitude difference between mark and space has been re-

World Radio History

duced to half its previous value, the system is 6 db more sensitive to noise interference.

In FM systems, the same type of band compression may be obtained by using an increased number of significant frequencies to obtain additional channels. To obtain two channels, four instead of two frequencies might be employed. A single oscillator could be deviated to one frequency to indicate a mark, and to another frequency to indicate a space. The remaining two frequencies could represent the mark and space of the other channel. If bandwidth were increased to keep the same frequency spacing as in ordinary binary telegraphy, there would be no noise disadvantage. Where bandwidth is limited, however, the frequencies must be more closely spaced. This is analogous to the more closely spaced phases used in the phase modulation method.

A Practical System

An interesting variation of this band compressed FM system was recently installed at Cape Canaveral to transmit data between a tracking radar and a computer at Patrick Air Force Base. The radar set transmits data concerning the position, velocity, and heading of missiles, to an impact prediction computer at the launching site. The computer constantly analyzes the data and predicts the point where the missile would land if thrust were cut off at any given instant. The radar information requires a transmission rate of 6720 bits per second. This is transmitted to the computer over two cable channels.

Each channel is supplied data at the rate of 3360 bits per second. Because the FM system is necessarily double sideband, a *quaternary* or four-tone system is used to double the information rate without requiring more bandwidth.

Four frequencies within the voice

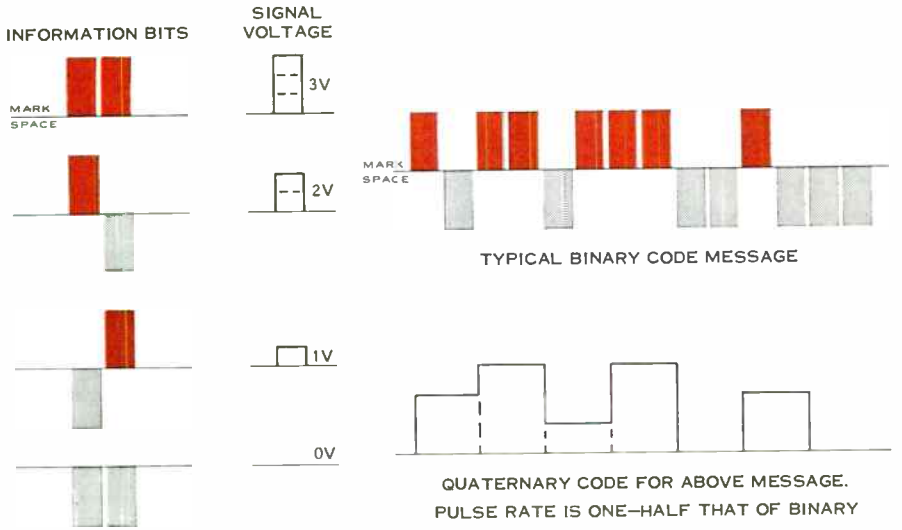


Figure 8. Quaternary (four-level) code adapts binary information for transmission by four tone system shown in Figure 7. Each pair of information bits is represented by one tone or level. System requires transmission system with good tolerance to noise and level changes.

World Radio History

band are used to convey the information. Instead of using two of the four tones for mark and two for space, thus yielding two channels, the data signal is encoded so that each transmitted pulse carries two bits of information. Since there are four possible combinations of mark and space taken two at a time, one tone is assigned to each of these combinations. Figure 8 shows the encoding scheme by which information rate is doubled. The actual transmitted pulse rate is 1680 pulses per second, and the information rate is 3360 bits per second. At the receiver, the tones representing the four combinations appear at the discriminator as four voltage levels — the same as in the phase modulation method. This straight-forward approach avoids many of the problems of complexity inherent in the more elaborate phase modulation methods.

Synchronization

Frequency modulation does not require synchronization, unlike phase modulation systems. In phase modulation, failure to maintain perfect synchronization between the two ends of a transmission path results in failure of the system, since information is carried only by the phase of the transmitted wave.

FM systems can achieve additional signal-to-noise advantage if they are operated synchronously. This makes it possible to mute the receiver until the exact moment when a signal pulse is strongest. Instead of receiving everything that comes over the line, including noise, the receiver samples the incoming signal at the optimum moment. Figure 9 shows a simplified block diagram of a typical synchronous detector which yields about 4 db signal-to-noise advantage, compared to non-synchronous detection.

Pulse Integration

A useful technique for obtaining even more signal-to-noise advantage is to integrate the received signal over a short period of time. Since background noise is perfectly random, there is no coherence to the integrated noise. Positive and negative pulses tend to offset each other on a statistical basis. Signal pulses, on the other hand, are coherent, and tend to build up by integration.

Extreme examples of this principle are the historic radar contacts made with the planet Venus in 1958, and with the Sun in 1959. In the case of

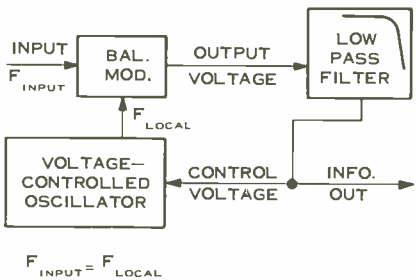


Figure 9. Block diagram of typical synchronous or homodyne detector. This is mandatory for PM, optional for FM.

the Venus contact, the planet was about 28 million miles away, requiring a round trip of about five minutes for the radar signals. The radar transmission consisted of a series of 2-millisecond pulses transmitted every 33.3 milliseconds. The transmitter was operated for about 4½ minutes, turned off, and the return signal was recorded on magnetic tape for five minutes. The estimated signal-to-noise ratio was -10 db. That is, background noise was 10 db greater than the radar echo.

The recorded signal was passed through a shaping filter and quantized to 64 values of amplitude. A digital

World Radio History

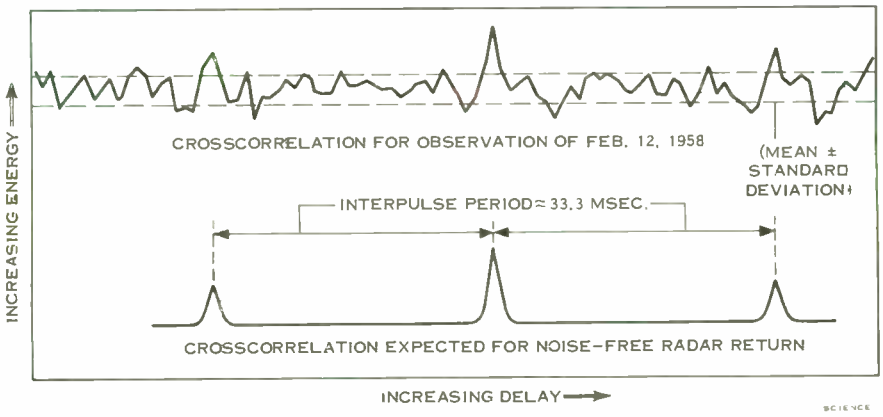


Figure 10. Use of pulse integration and statistical correlation techniques made possible radar contact with planet Venus and Sun. Similar technique can be used to improve high speed data transmission, particularly with longer pulses.

computer integrated each one-millisecond period of signal over a time span equal to several thousand pulses, in an attempt to obtain correlation between the transmitted pulses and the received signal. Figure 10 compares the cross-correlation that would be expected in the absence of noise, and that which was actually obtained.

In data transmission, each pulse may be integrated over its own length, and the integrated value sampled at the end of each pulse. This will normally yield about six db additional freedom from noise interference, but will also require synchronous operation. While such techniques are not mandatory in FM data systems, they are available when circuit conditions are noisy, or reliability requirements are so stringent as to justify the additional complexity and expense.

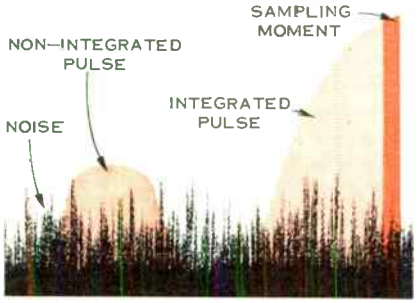


Figure 11. Bit integration builds up pulse amplitude. Noise fails to build up because of statistical cancellation of individual noise pulses.

At the present state of the art, how close a system comes to ideal performance seems to depend on how much the budget can stand. In practical terms, a very few decibels of signal-to-noise tolerance may cost a surprisingly large number of dollars. A review of various available methods of high speed data transmission would indicate that the cost of one terminal capable of transmitting 3000 bits per second in a 3000-cycle channel goes up at a rate which rapidly exceeds $1,000 for *each* db drop in signal-to-noise ratio! The choice of the methods available for higher transmission speeds might very well be decided on the basis of the quality of circuits available. •

World Radio History

World Radio History

the Lenkurt Demodulator

VOL. 13, NO. 2 FEBRUARY, 1964

DATA TIMING

For HF Radio Transmission

Data transmission, by its very nature, involves a compromise between speed and error rate. Higher speed demands a shorter time interval for each information bit; and reduced bit length usually means more errors, particularly in high-frequency radio transmission, where path conditions may vary from moment to moment. This article discusses some of the problems inherent in the transmission of data over HF (3 to 30 Mc) radio, and describes how high transmission speed can be attained while maintaining the relatively long bit length necessary for a low error rate.

The transmission of information in digital form has grown tremendously in recent years, due in large part to the increased use of computers, but also stimulated by the increase in communications between people. Part of this additional communication goes over telephone circuits as analog signals; but a large portion is transmitted in digital form. This includes various types of printed messages, as well as facsimile. Furthermore, digital voice transmission is being used more widely, particularly for "secure" communications, where privacy is vital.

The result is increased demand for transmission facilities for long-distance digital data communication. For overland links, there is likely to be a variety of facilities available, but when an ocean must be spanned the present choice is either submarine cable or HF radio. Submarine cables are usually both crowded and expensive, but often their biggest disadvantage is that they don't go where the prospective user needs them. If no

World Radio History

transmission facilities exist between the points to be linked, and new ones must be installed, radio is almost certain to be the choice, whether it is judged by cost or by installation time. Thus, radio is quite attractive.

Historically, data transmission on HF radio has been either slow or unreliable. Often it has been both. Recent efforts aimed at increasing both the speed and reliability of HF data transmission have shown considerable success, but many problems which are unique to HF radio have been encountered.

The HF Transmission Problem

For distances longer than about a hundred miles, HF radio transmission depends entirely upon the ionosphere, a region of charged particles high in the earth's atmosphere. Unless the radio beam is refracted (bent) by the ionosphere, little or no propagation path exists beyond the earth's curvature and the signal will not reach a distant receiver. For this reason, the quality of HF radio transmission is intimately involved with the behavior of the ionosphere.

The ability of the ionosphere to refract or absorb radio waves is directly dependent on the density of the negatively charged particles, the free electrons. The free electron density is controlled by several factors, but by far the most important is solar activity. In fact, the action of the sunlight in ionizing the rarified air is primarily responsible for the existence of the ionosphere. Hence, the free electron density is much greater during daylight than it is at night.

Although the return of radio energy to the earth by the ionosphere is commonly called "reflection," it is actually a form of refraction which is controlled by both the frequency of the signal and the density of the electrons. At higher frequencies the radio wave is refracted less than at lower frequencies. For any given frequency, the greater the electron density the greater the refraction or bending of the wave. In other words, the *index of refraction* of the ionized medium, which determines the amount of deflection of the ray, is a function of both the electron density and the operating frequency. The operating frequency is important because if the frequency is too high, the radiated energy goes right on through the ionosphere. Conversely, if the frequency is too low, the beam is effectively absorbed by the ionized layer, and little or no energy reaches the receiver.

One of the main complications in using the ionosphere for radio transmission is the fact that it is nonhomogeneous. In some places the electrons are grouped together in bunches, while in other places there are "holes" in the ionosphere. The effect is much like a layer of broken or scattered clouds. To further complicate matters, the whole ionosphere is constantly shifting. Thus, where at one moment the free electron density is high in a given spot, at the next moment the density may be so low that a radio signal can pass straight through and be lost in space.

This erratic behavior of the ionosphere has several important effects on HF radio transmission. One of the more important of these is called the "multipath effect." Multipath effects occur when two transmitted signals take a slightly different path from the transmitter to the receiver. The result, of course, is that the two signals arrive at the receiver in a slightly different time relationship, due to the different lengths of their propagation paths. One of the most common ways for this to happen is for one signal to take a "one-hop"

World Radio History

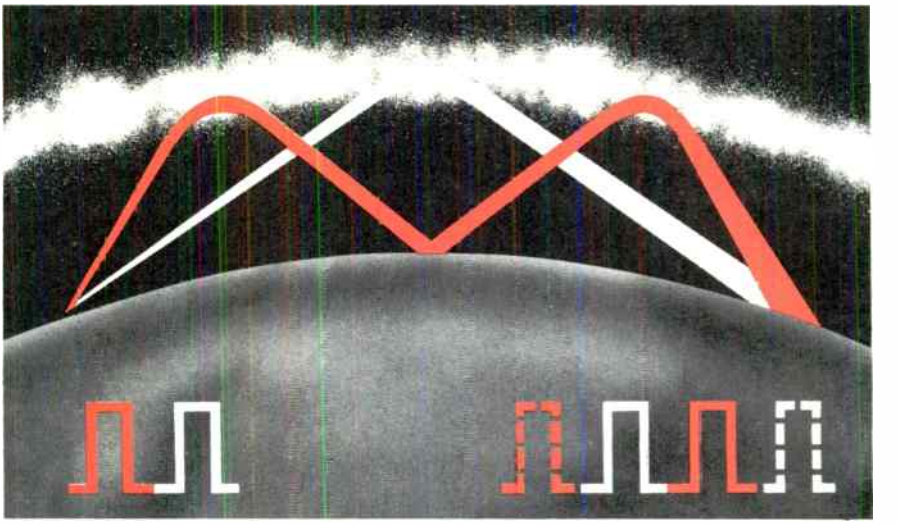


Figure 1. Multipath effect occurs when signals take different paths from transmitter to receiver, arriving in a different time relationship.

path while the other signal takes a "two-hop path" as illustrated in Figure 1.

This simply means that one signal is deflected only once by the ionosphere while the other signal is deflected three times—twice by the ionosphere and once by the earth. In the illustration, pulse two is transmitted after pulse one, but because of the difference in path length, pulse two arrives before pulse one. Or, if the difference in transmission time is not so great, the two pulses may arrive simultaneously. When this effect occurs in a high-speed data stream, several pulses may be overlapped or "smeared." The same effect may occur when the same pulse takes both paths simultaneously. The difference in propagation time makes the receiver see two pulses, often overlapped. This smearing effect is what limits the minimum acceptable pulse length for data transmission on HF radio.

Solving the Bit-Length Problem

From a qualitative viewpoint, reducing the speed of data transmission can be likened to introducing redundancy. Holding a mark or a space for a longer period of time has somewhat the same effect as sending the same information bit more than once. In either case there is a better chance that it will be correctly identified.

To avoid the garbling caused by the smearing which occurs in the HF path, however, it is often necessary to use such a long bit length that transmission speed is severely restricted. For example, consider an HF path of 4,000 miles, with a propagation time of about 22 milliseconds. A bit length of perhaps 2 to 4 milliseconds would be required to counteract the multipath effects. Thus, the data transmission speed over a single channel is limited to 250 to 500 bits per second, *regardless of the bandwidth*

World Radio History

of the channel. The same channel on cable (assuming a 3-kc voice channel) would have a data capacity of perhaps 2,000 to 3,000 bps (probably 2,400 bps, a common data-transmission speed over a 3-kc channel).

One solution to this problem is to divide the single serial data stream into several parallel streams which together provide the same total transmission capacity. This way, bit length can be increased without reducing speed or using more bandwidth. For example, suppose that a standard 3-kc HF radio channel is available for the transmission of 2,400 bits per second. The bit length at this speed is 0.42 millisecond. If, however, the 2,400-bps serial stream is split into 16 150-bps channels, the bit length in each of these slower channels is 6.67 milliseconds — 16 times as long as it would be in the single channel. Since these 16 channels operate at much lower speeds, they require less bandwidth, and can be frequency-division multiplexed into the single 3-kc voice channel.

This arrangement effectively solves the bit-length problem, but it introduces another complication. In order to reconstruct the 2,400-bps serial data stream, the 16 parallel channels must be recombined at the receiver. If the information bits which make up a single word are distributed among several channels, it is apparent that they must be rearranged in the same order at the receiving end.

But because of the nature of HF radio transmission, each of the sixteen channels is likely to have a slightly different propagation path. The channels are also likely to have different fading characteristics because they are at slightly different frequencies, and are not affected identically by the ionosphere. The usual practice in HF radio is to use *space-diversity reception* to counteract fading. In such a diversity arrangement, two receivers are used with antennas separated by several wavelengths. This minimizes the possibility that the signals received by the two receivers will fade simultaneously. The signals from the two receivers are put together in a combiner.

Because the sixteen channels are all at different frequencies and have different fading characteristics, it is entirely possible that at a single instant some channels will be stronger at receiver A,

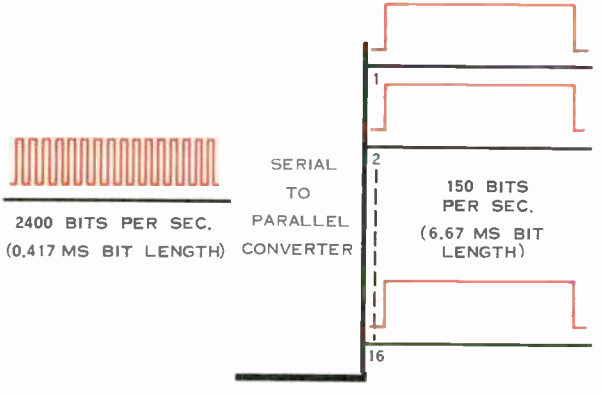


Figure 2. Use of 16 150-bps channels rather than one 2400-bps channel permits bit length 16 times as long while requiring no more bandwidth.

World Radio History

while the other channels will produce a stronger signal at receiver B. This further complicates the recombination because the filters used in the two receivers may have different delay characteristics. This delay variation, added to the difference in propagation time between the various channels, produces a real timing

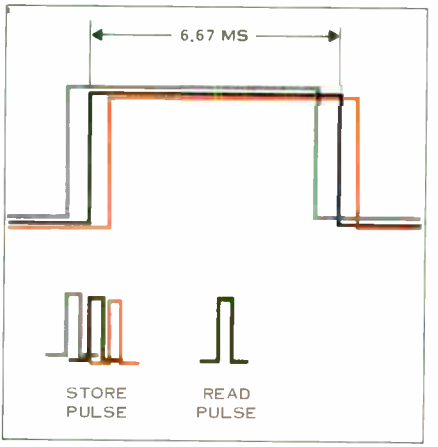


Figure 3. Optional synchronizer has 6.67-millisecond storage capacity. Read-out rate is fixed, while storage is controlled by phase-corrected pulses from phase resolver.

problem at the receiving end. The channels cannot simply be brought together and combined into a serial stream, because it is likely that the information from one channel would be placed "on top of" the information from another channel. In effect, this would be multipath smearing again.

Furthermore, because of rapid shifts in the ionosphere, the total path length changes with time. The result is that the number of information bits "in the air" at any instant varies. In other words, even though the data is being transmitted at a constant rate, it is received in bunches or groups of information bits. The long-term output rate is equal to the transmisison rate, but the short-term output rate varies. For some purposes, this non-uniform data stream is perfectly acceptable, while for other uses it must be retimed. For example, if the equipment is being used with vocoders for digital voice transmission, the non-uniformity of the data stream will probably have little effect. On the other hand, for some computer applications the data stream must be phase locked to a highly stable clock.

Retiming

A good example of how the serial stream may be reconstructed is provided by Lenkurt's Type 27A Data Terminal. Since the 27A uses Lenkurt's Duobinary Coding technique (see THE LENKURT DEMODULATOR, February, 1963), which effectively halves the required bandwidth for a particular bit rate, its operation at 2,400 bps is much like that of conventional binary equipment operating at 1,200 bps. The retiming problems, however, are similar regardless of the coding techniques used.

The receive timing system of the 27A has two automatic adjustments. One follows the short-term (fractions of a second) phase variations between the 16 channels to provide a single sampling instant for all channels. The other varies the long-term (several minutes) data read-out rate to match the transmission rate.

Figure 4 is a simplified block diagram of the data timing system used in the 27A receiver. An adjustable crystal oscillator generates a timing signal at the same frequency (nominally 2400 cps) as the data-transmission rate. This "clock" signal is fed to a phase resolver

World Radio History

which adjusts the phase of the clock signal to follow the short-term phase variations in the received data. The phase-corrected output from the phase resolver goes to the sampling circuits for the 16 150-bps channels (after being counted down to 150 cps), causing them all to sample their incoming signals at the same instant.

This phase-corrected signal also goes to the phase detector, where it is compared to a composite signal from the 16 incoming channels. This composite signal indicates the *mean* phase of the channels, thus providing a reference for correcting the clock signal. The phase detector produces an output which is proportional to the phase difference between the corrected clock signal and the composite received-data signal. The error signal goes to a servo which reduces the error signal to zero by adjusting the phase resolver.

The servo is also connected to the crystal oscillator. The oscillator, however, has a long response time, and is unaffected by rapid fluctuations. Its function is to adjust slowly to long-term variations in the transmitted data rate.

The oscillator output may also be used to control the read-out rate from an optional synchronizer, which provides a completely uniform data stream for some applications. The synchronization is accomplished by a "buffer" shift register in each of the 16 channels. Its purpose is to store extra data when the input is temporarily too fast, and to provide a "reservoir" for the read-out to draw on when the input is temporarily too slow.

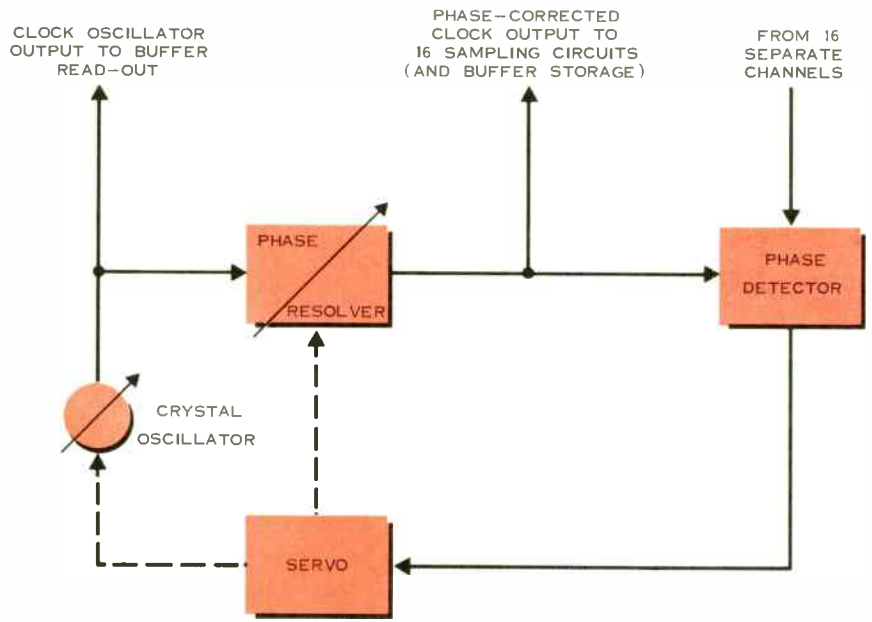


Figure 4. Simplified block diagram of timing system used in Lenkurt's Type 27A Data Terminal. Dotted lines indicate mechanical links from servo to phase resolver and oscillator.

World Radio History

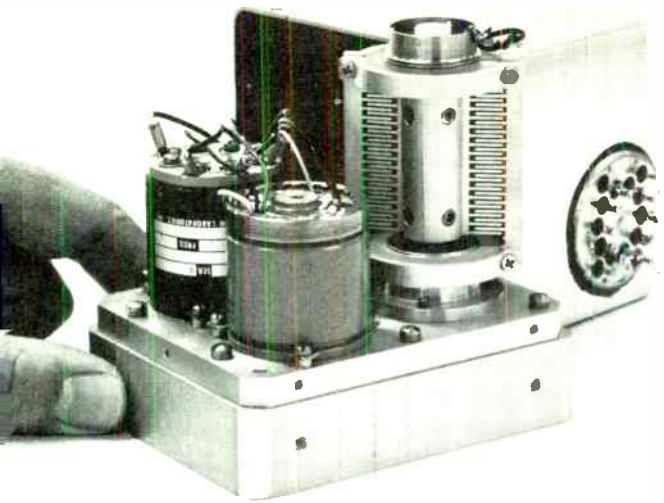


Figure 5. Mechanical action in 27A timing system is provided by motor (left) which drives phase resolver and variable capacitor. Phase resolver follows short-term variations in data rate, while capacitor permits oscillator frequency adjustment up to plus or minus 20 parts per million.

The shift register has a storage capacity of one bit length — 6.67 milliseconds. (The same effect could be obtained by putting a shift register with a 16-bit capacity in the serial stream. Again, the storage capacity would be 6.67 milliseconds.)

Figure 3 illustrates the operation of the synchronizer. The read-out rate is fixed, controlled by regular pulses from the oscillator. However, the input rate to the shift register is controlled by a non-uniform series of phase-corrected "store" pulses from the phase resolver. The nominal read-out point is the center of the pulse, but it can vary across the whole width of the pulse. This provides a tolerance of plus or minus 3.33 milliseconds to allow for rapid variations in path length, and hence in propagation time. In the highly unlikely event that the read-out were to "run off the end," it is automatically reset to the center with a loss of only 8 serial bits.

Conclusion

The problems posed by the transmission of data over HF radio are somewhat different from those encountered in transmitting voice over similar facilities. One of the main differences is the lack of redundancy in data, whereas speech may include 75 percent redundant information. When a bit is lost from a data stream, an error results. But when a portion of a word is lost from a sentence, the meaning is usually clear from the context. This is one reason why data transmission puts more severe requirements on HF radio than does voice transmission.

One of the more pressing of the problems has been that of timing the data to achieve both speed and reliability. The solution has not been simple, but it has proven practical. The result is that HF radio is now taking its place as a reliable, as well as economical, medium for high-speed data transmission. •

World Radio History

World Radio History

the Lenkurt Demodulator

VOL. 12, NO. 2 FEBRUARY, 1963

Major Breakthrough in Data Transmission ...

"DUOBINARY CODING"

An ingenious method of encoding digital signals has overcome a theoretical limit to the maximum rate at which signals can be transmitted through a communications channel. This technique, called "Duobinary Coding," doubles the speed of transmission through ordinary binary channels, but at a far lower cost, in terms of vulnerability to interference and equipment complexity, than conventional multi-phase or multi-level transmission systems. As a sort of "bonus," Duobinary Coding permits automatic detection of most transmission errors, although no redundant information is added to the transmitted data. This article reviews some of the laws that restrict the transmission of information, and how Duobinary Coding helps bypass them.

Almost immediately after the first electrical telegraph circuits were put in service a century or more ago, telegraphers discovered a very real limit to the rate at which telegraph symbols could be transmitted. The symbols tended to spread out and "melt down" in transit, but still could be recovered and identified if they were not too close together. As the transmission rate increased, however, the telegraph pulses necessarily fell closer and closer together; above a certain rate, they would interfere with each other so badly that they became garbled and unidentifiable.

This situation still prevails, of course. However, the concept of circuit bandwidth and how it determines the maximum rate of transmission is now well understood. We know how to create circuits of enormous bandwidth, capable of handling vast amounts of information. Knowing how is one thing, but doing, another.

In most applications, bandwidth must be limited, either because of its high cost or because of the practical need to be compatible with other portions of a wide-spread network. Most of the world's communication facilities consist of telephone channels limited

World Radio History

to about 3000 cycles of bandwidth. Because this type of facility is so universal, most data communications equipment is designed to operate over such voice channels, even though other types might be preferred. The transmission of digital data and other pulse signals is growing very rapidly, thus forcing system designers to seek ways of increasing the capacity of these ill-suited standard channels which go nearly everywhere.

Theoretical Limits

In two classic studies of the ultimate capacity of a communication channel, Harry Nyquist, a Bell Laboratories scientist, showed that no more than $2B$ binary pulses per second can be sent through a channel having a bandwidth of B cycles per second. Thus, in a 3000-cycle channel, up to 6000 binary pulses per second theoretically can be transmitted. In order to achieve this ideal performance, however, some rather stringent (and unattainable) conditions would have to be met. The channel would have to have the characteristics of an ideal low-pass filter, transmitting all frequencies between zero (direct current) and the cutoff frequency, B, equally well. This requirement is necessary in order to avoid the waste of two sidebands, as would usually be the case in a channel having band-pass characteristics; i. e., not transmitting direct current. In addition, our ideal channel would have to have perfect phase linearity so as not to introduce any distortion.

Although the term *bit* (for binary digit), meaning a fundamental, irreducible unit of information, had not been invented when Nyquist made his studies, he recognized that each pulse or *signal element* could be made to carry additional information if one or more of its characteristics such as amplitude or phase were varied in a prescribed manner. Thus, not only the presence or absence of a pulse, but also its level, could be made to carry information. It should be understood that all references to signal "levels" apply equally to modulation or transmission methods in which phase, frequency, or amplitude of the signal may be varied. By increasing the number of signal levels, the information capacity of the signal could thus be enlarged indefinitely.

Nyquist showed that the number of bits that can be transmitted through a channel in a given time is proportional to the *logarithm* of the number of levels that the signal can assume. Since the binary digit has only two values, being either present or absent, on or off, the logarithm necessarily must be to the base 2. Thus, according to Nyquist, the maximum amount of information, C (or bits per second), that can be transmitted through his ideal channel is $C = 2B \log_2 m$, where B is bandwidth in cycles per second and m is the number of levels. In a binary system, $\log_2 2 = 1$ and $C = 2B$. In a four-level or quaternary system, $\log_2 4 = 2$, and $C = 4B$.

Practical Limits

The trouble with any such multi-level system is that as the number of signal levels increases, it becomes harder and harder to distinguish between and accurately identify each level or phase value in the presence of noise or distortion. In order to avoid errors caused by noise peaks, the *difference* in value between each signal level must be at least twice as great as noise peaks, as shown in Figure 1.

An even greater impairment is the obscuring effect caused by *intersymbol interference*, an often overlooked factor. As the signals travel through the system, various types of distortion cause some of the signal energy to be

World Radio History

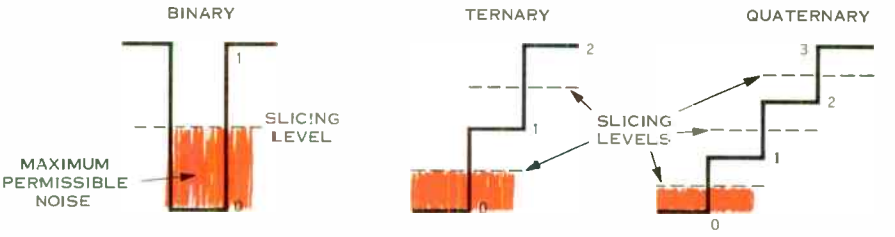


Figure 1. Additional "levels" (which may be phase, frequency, or amplitude) increases information capacity of signal, but reduces tolerance to interference. Peak noise must not exceed "slicing" or decision level (half the distance between signal levels).

displaced in time or frequency so that it is no longer associated with its own signal element. Delay distortion, for instance, slows some frequency components of a signal pulse more than others. The delayed portions of the signal may be overtaken by the leading portions of the following pulse, thus changing the value of both. The uncertainty that results makes the pulse even more vulnerable to noise, and tends to increase transmission errors.

Delay distortion is a particularly great problem when transmitting data through telephone channels designed primarily for speech. Although a channel may occupy a bandwidth of 3000 cycles, the *effective bandwidth* is always much less.

Although it is possible to partially correct or equalize the delay characteristics of channels, this is expensive and rarely practicable for channels in the dial telephone network. Even when channels are carefully equalized, only about 2000 cycles of bandwidth can be used. This, plus the general use of double-sideband transmission (which "wastes" half the available bandwidth) always restricts actual transmission rates to a value much lower than allowed by theory.

Although most calculations of the interfering effect of noise are based on random thermal noise, pulse or data transmission systems are more vulnerable to various types of *impulse noise*, particularly those created by telephone-type switching. Switching impulses are much more likely than thermal noise to cause errors, providing a reasonable signal-to-noise ratio is maintained. Voltage "spikes" or transients may have a much greater amplitude than the signal, thus making them difficult to cope with. As a result, data transmission through dial-up or switched networks usually requires binary signals and relatively low speeds in order to minimize errors.

Unfavorable Trade

A multi-level signal experiences a very unfavorable trade between information capacity and freedom from interference. Each time the number of levels is doubled, the signal-to-noise ratio must be increased 20 $\log_{10}(m\text{-}1)$ db in order to maintain a given freedom from error. Thus, a 64-level system could transmit information six times as fast ($64=2^6$) as a binary system, but would require a signal-to-noise ratio 36 db higher in order to maintain the same error rate. In other words, to increase information capacity six-fold would require that transmitter power be increased 4000 times. Actually, more than 36 db improvement in signal-to-noise ratio would be required.

World Radio History

since intersymbol interference becomes much greater as the number of signal levels is increased.

This is an expensive way of increasing capacity and is the main reason why few systems use more than four levels in signaling over commercial communications circuits. Most data transmission systems now used over "switched" telephone circuits employ binary transmission, and are limited to about 1200 bits per second. Where better circuits are available, a quaternary or four-level transmission method is often used, thus permitting data transmission speeds up to about 2400 bits per second with reasonable error rates (one error in 10^5 bits, on the average). Higher speeds are possible in a given channel, but only at the cost of rapidly increasing error rates. Transmission speed and accuracy are strongly related. Speed alone provides no basis for judging or rating a transmission method—the error rate must also be shown.

Another characteristic of multi-level transmission systems is the inherent delay that is introduced by the encoding and decoding processes as the number of levels is increased. For instance, in a quaternary system, each signal element must represent two information bits. The encoder must wait until it has received *both* before it can determine the value of the signal element to be transmitted. At the receiver, a similar delay is encountered as the decoder interprets the received signal and then produces the information bits in their original sequence. As the number of signal levels is increased, encoding delay goes up accordingly.

Signaling at the Nyquist Rate — and faster

Nyquist showed that $2B$ signal elements were the most that could be sent through an ideal low-pass transmission

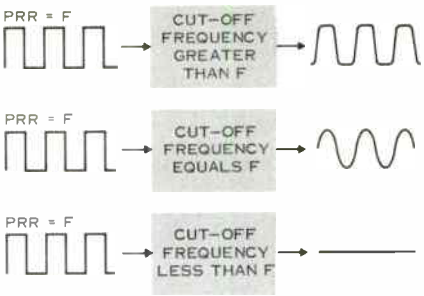


Figure 2. Square-wave pulses are always rounded somewhat by band-pass filter. When filter cut-off frequency equals pulse repetition rate, only sine wave component can pass. When pulse rate exceeds filter cut-off frequency, no energy passes.

path of bandwidth B, and still be recovered intact at the receiver. Signals transmitted at a higher rate would be hopelessly garbled. In other words, the criterion for "signaling" is the ability to decode and recover the original data at the receiver. Nyquist thus implied that a channel cannot possibly accommodate more than $2B$ bits of information except by paying the price of signal delay and a very rapid increase in vulnerability to interference.

An exception to these requirements has been recently developed at Lenkurt. The new technique, called "Duobinary Coding," consists of systematically encoding the data stream in a specific way to achieve a "compression" of the bandwidth required for transmission, thereby increasing the permissible transmission rate. Unlike conventional multi-level techniques, there is no inherent encoding and decoding delay, nor is the information content restricted only to the *log* of the number of levels, as in the Nyquist rule. The encoding method changes the pattern of the data stream and "redesigns" the signal so that it can pass through the

World Radio History

channel at twice the maximum binary rate without loss of information.

Filter Response

Communications channels behave like filters, and therefore can be analyzed and discussed as though they were filters. When a square-wave pulse is passed through a filter (or channel), it is altered or distorted by the filter's limited bandwidth. Square waves can be shown to consist of a very large number of sine wave components. If, for instance, the signal pulses are repeated 1000 times a second, the signal spectrum will consist of a 1000 cps sine wave and an infinite number of odd harmonics: 3rd, 5th, 7th, and so

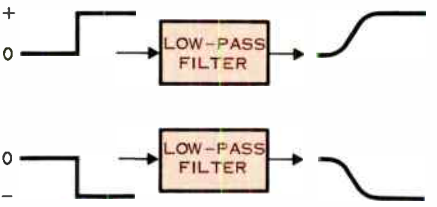


Figure 3. When an abrupt voltage "step" is applied to a low-pass filter, output voltage response assumes form of a sine wave of the filter cut-off frequency.

on. When bandwidth is restricted, the higher harmonics are attenuated, with the result that the pulse is no longer perfectly square, but somewhat rounded. As bandwidth is further reduced, the signal pulse becomes more and more rounded. When the filter bandwidth just equals the pulse frequency, only the sine wave 1000 cps fundamental frequency can pass.

If the bandwidth of the filter is *less* than the pulse frequency, no energy at all passes through the filter (assuming that it is an ideal filter). Practical filters do not achieve this, but attenuate out-of-band frequency components greatly. For simplicity, the discussion and most illustrations in this article will assume ideal filter characteristics.

If a positive or negative voltage is suddenly applied to such a filter, its output will respond gradually, as though it were part of a sine wave of the filter cut-off frequency. When the voltage is removed, the output voltage drops off at the same sine-wave rate, as indicated in Figure 3. It is this inability of the wave to change value at a faster rate which limits its information content in a purely binary system. Since only two variables remain in each cycle—a positive and a negative half-cycle, additional information can be carried by the wave only by varying phase, frequency, or amplitude.

Duobinary Coding

The Duobinary Coding technique is believed to be the first method for transmitting more information through a given channel than specified by the Nyquist relationship. Figure 4 shows one of the ways by which it can be achieved. Binary data pulses are applied to one input of a simple AND gate, and a stream of binary "clock" pulses of the same repetition rate as the data pulses is applied to the other input. Whenever a data space occurs, the AND gate yields an output pulse which is applied to a "flip-flop" circuit or bi-stable multivibrator. The multivibrator has the characteristic of changing its output voltage from zero to plus or plus to zero, whenever it is triggered by the AND gate. Thus, for every data space, the output of the flip-flop changes state. Whenever a data mark is received, no change occurs.

The binary waveform from the flip-flop could be amplified and applied directly to the line for transmission, but this would merely waste some of

World Radio History

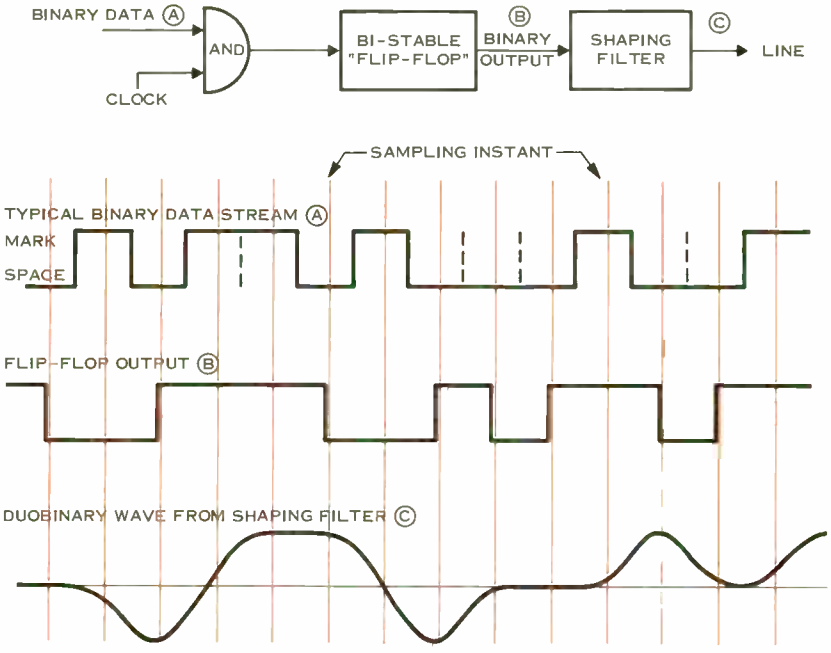


Figure 4. One form of Duobinary encoder and its resulting waveforms. Coincidence of "clock" pulse and data space causes bi-stable multivibrator ("flip-flop") to change state. When passed through transmitter wave-shaping filter, the three-level Duobinary wave results.

the transmission power and increase crossfire into other circuits. Instead, it is normally passed through a low-pass shaping filter. It should be noted that the output signal from the flip-flop is a binary waveform which, although of a different pattern, has a direct correlation with the original data, and from which the original data can be recovered.

Because of the encoding process, the signal now has several unique properties. As it passes through the channel or shaping filter (which must have a cut-off frequency ¼ the data rate), the waveform is altered so that it tends to occupy three amplitude levels. Superficially, this three-level signal resembles a so-called "ternary" (three-level) signal or a bipolar signal, both of which have been used in telegraphy for many years. However, there are significant differences: Ternary and other conventional multi-level signals must be able to go from any given level to any other level in order to transmit information at a rate faster than binary systems. By contrast, the Duobinary signal has the effect of forbidding the signal from going directly from one extreme level to the opposite extreme within one bit-interval.

A continuous series of input data *spaces* causes the flip-flop to change state repeatedly, thus creating a series of binary pulses at the data rate. This is too fast to pass through the filter, and a "zero" level signal results. A continuous series of *marks* results in the flip-flop remaining in whichever state

World Radio History

it was in at the beginning of the sequence. The single voltage, tone, or phase by which this marking condition is transmitted will pass through the channel or filter. Since either plus or minus represents mark, the signal will not change for consecutive marks, but only when a space follows a mark. A second space holds the output in the zero condition, but a mark following the space returns the signal to a mark condition, as shown in Figure 4.

At the receiver, the data stream is particularly easy to decode and recover. Marks are represented by *either* of the two extreme values— positive or negative—that the signal can assume. A simple full-wave rectifier will reassemble the marks and spaces in their correct sequence, as shown in Figure 5.

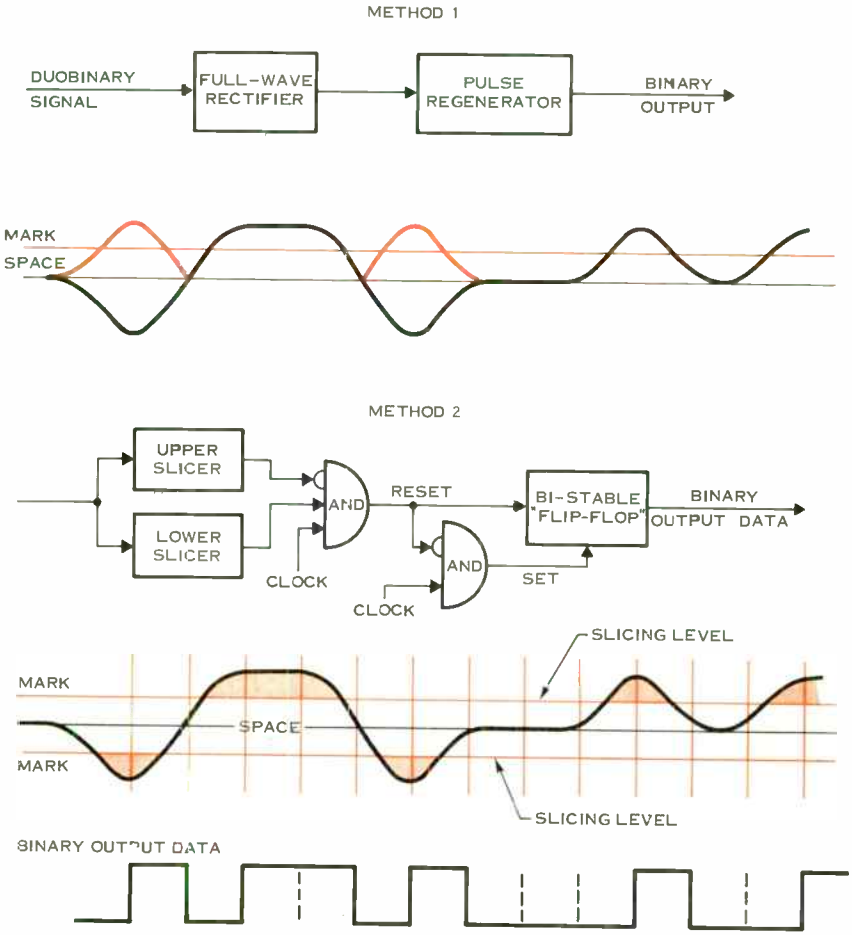


Figure 5. Two forms of Duobinary decoders. Simplest and most versatile consists only of full-wave rectifier and pulse regenerator. Pulse retiming is optional. More elaborate version below use "slicers" to determine whether Duobinary wave is in "mark" or "space" condition at sampling instant.

World Radio History

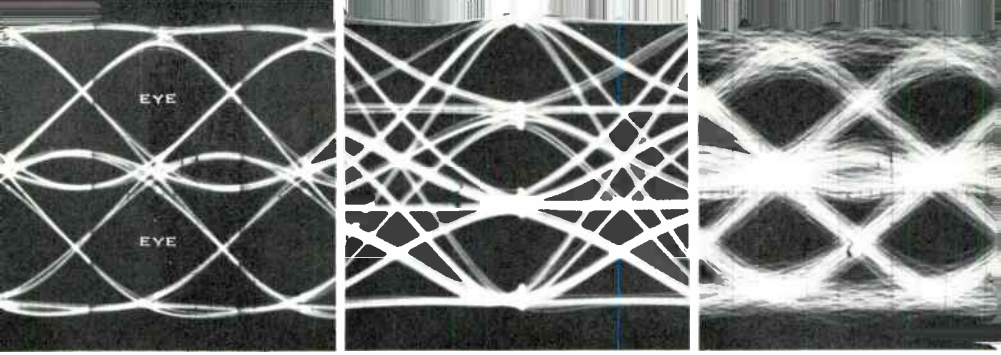


Figure 6. Typical "eye" patterns of Duobinary, quaternary, and noisy Duobinary signals. Notice how noise tends to close in the eyes of the Duobinary signal. Duobinary signal with noise (S/N = 16 db) still has clearer eye than optimized quaternary signal.

Synchronization of receiver and transmitter, an absolute necessity in multiphase transmission methods, may be used but is not required.

Another method of decoding Duobinary signals uses a synchronized "clock" signal at the receiver to sample the waveform at the center of each bit interval, thus reducing the likelihood of noise impulses causing transmission errors. Two binary "slicers" determine at the sampling moment, whether the signal is in the central or "neutral" region, thus indicating a space, or whether it is outside the central region in either the positive or negative region, thus indicating a mark.

Performance

An effective way of evaluating the performance of a data transmission system is by displaying so-called waveform "eye patterns" on an oscilloscope. The eye pattern is formed by applying random data signals to the vertical deflection plates and triggering the horizontal sweep at the data bit rate. The resulting pattern graphically demonstrates the relative vulnerability of the signal to the effects of noise, envelope delay, and "timing jitter." Figure 6 shows typical eye patterns for binary, quaternary, and Duobinary signals. The "eye" is the dark area between the levels; the larger this area, the easier it is to distinguish between levels and to transmit without error.

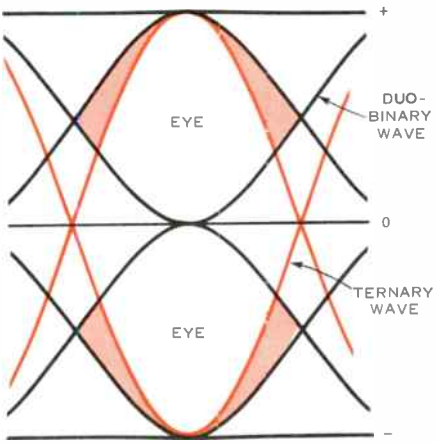


Figure 7. Comparison of conventional ternary signal (red) and Duobinary signal. Red-tinted area represents the eye area lost by the need for the ternary signal to go from any level to any other. Lost area increases inter-symbol interference, thus making signal more vulnerable to noise and increasing error rate.

World Radio History

Noise, timing jitter, and various forms of distortion tend to fill in the eye, thus reducing the ability of the receiver to distinguish between levels. In Figure 6, note that the eyes of the Duobinary signal to which noise has been added are still larger and more open than a noise-free quaternary signal of the same data capacity.

Figure 7 compares the eye patterns of a ternary signal and a Duobinary signal. Since the ternary signal may go from the lowest level to the highest value, or vice versa, the required path cuts across the eye, reducing its area and making the ternary transmission much more vulnerable to intersymbol interference than a Duobinary signal.

Error Detection

An outstanding feature of the Duobinary technique is its ability to provide "free" error detection; that is, error detection which is not dependent on the addition of redundant bits or data code symbols to the transmitted data. Error detection is very important in most digital data transmissions because the message may frequently consist of unrelated numbers or characters suitable for direct use by business machines or computers. Single errors may have the effect of invalidating or "spoiling" a much larger block of information. A good example of this might be in the transmission of information from punched cards. The information in such transmission has little or no inherent redundancy which might help reveal errors. By contrast, written text or spoken words are very redundant; occasional errors are usually very conspicuous and thus easily corrected by the recipient.

Heretofore, all digital transmission methods have been able to achieve error detection by the deliberate addition of redundant symbols to the transmitted information, thus "diluting" it and lowering the effective transmission rate.

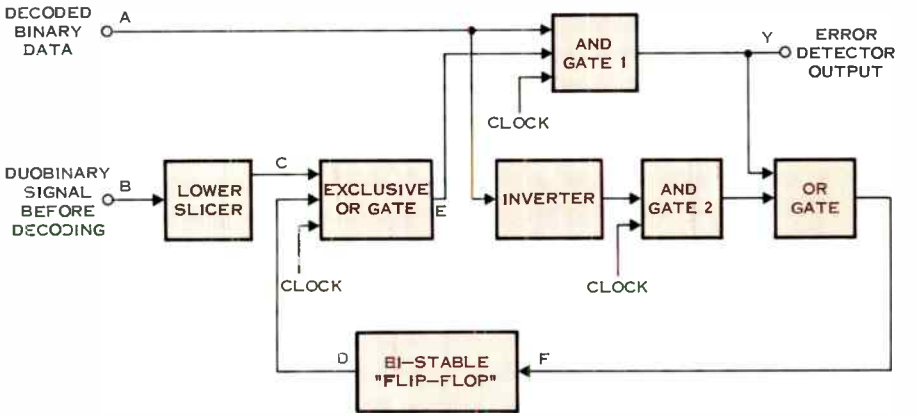


Figure 8. One form of automatic error detector. In absence of transmission errors, proper polarity of data marks is determined by the number of intervening spaces. The flip-flop acts as memory for polarity of preceding mark and counts intervening spaces. Points C and D should have opposite polarity, causing output at E, only during data spaces. If output at E coincides with mark at point A, an error has occured, causing indication at point Y. Error signal resets flip-flop to "correct" polarity to avoid subsequent false error indications.

World Radio History

(For a fuller discussion of conventional error detection techniques, see DEMODULATOR, June, 1961.) Duobinary Coding permits error detection without redundancy because the encoding process follows a systematic pattern which can be constantly monitored by error detecting circuits. Any violation of the pattern results in an error indication. This pattern can be summarized as follows:

Space is always represented by "zero" or neutral.

Mark is represented by either + or —, depending on the number of spaces between marks.

The polarity of a mark remains the *same* as the previous mark if the number of intervening spaces is *even*, but is *reversed* if the number of intervening spaces is *odd*.

Since any single error changes the number of spaces between marks, the encoding plan would be violated, and the error can be detected automatically. The error detector diagrammed in Figure 8 follows the pattern of the received signal and indicates any violation of the encoding rules. As a result, *all single errors, all odd numbers of errors, and those even numbers of errors which also result in a violation of the encoding rules can be detected without the addition of redundancy*. In order to achieve this same capability with conventional parity-check methods, a large portion of the transmitted data would have to consist of redundant check bits.

Transmission Method

Duobinary Coding can be used with any type of modulation—FM, AM, vestigal sideband, phase-shift keying, or others. This permits selection of the best method for any given application. One of the best for use over HF ("high frequency") radio circuits, which are notoriously subject to multi-path fading and phase instabilities, is frequency shift-keying, a form of FM. Unlike the phase-shift methods which are often used for high speed transmission, neither synchronization between transmitter and receiver nor between the data digits and the modulated carrier is mandatory. The FM receiver is little affected by the phase instability of HF transmission or by fading or other amplitude disturbances which may be disastrous to AM or vestigial sideband transmissions. The additional signaling speed made possible by Duobinary Coding matches or exceeds that available with these more vulnerable modulation methods.

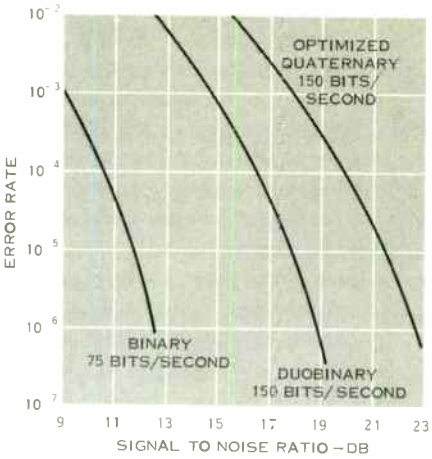


Figure 9. Comparison of error rates versus signal-to-noise ratio. Difference between binary and Duobinary becomes about 3 db when transmission speed is put on same basis ("normalized").

Duobinary Coding has been tested over actual HF radio circuits between Europe and the United States with very satisfactory results. A conventional binary system and a Duobinary system operating at twice the speed of the binary system were operated side by side and transmitted over the same radio circuit. The relative error rates

World Radio History

of the binary and Duobinary system were as predicted by theory.

One of the applications to which the Duobinary Coding technique is particularly well suited by virtue of its lack of complexity and freedom from the need for synchronization between data bits and carrier is in the transmission of several parallel data streams over a single narrow radio channel. These data signals might consist of a number of teletypewriter channels or voice signals which have been converted to digital form for encrypting.

Summary

Duobinary Coding is, strictly speaking a three-level code, and therefore must pay the price of increased vulnerability to noise, although this price is much less than that paid by quaternary systems of the same capacity. In general, the Duobinary system has the same sensitivity to noise as a ternary system, and the same susceptibility to intersymbol interference as a binary system.

Figure 9 shows a typical comparison of noise performance, based on experimental data. Duobinary Coding is about 4 db less sensitive to noise than a quaternary system of the same capacity, and less than 3 db more sensitive than a binary system. Although the chart shows a 6 db difference, this is based on a binary system of only half the capacity of the Duobinary. When this difference is eliminated, a differential of less than 3 db remains. In order to transmit at the same rate as the Duobinary system, however, the binary system must use twice the bandwidth as the price for its noise advantage.

In addition to providing obvious improvements in the speed and efficiency of many types of data and other digital communication methods, the Duobinary coding technique has an exciting future in a complete new range of other applications which will prove to be of substantial value in both commercial and military communications. •

A rigorous and more detailed treatment of the Duobinary technique is outside the format and intended scope of the DEMODULATOR. *However, Lenkurt Monograph No.* 181 *is being prepared for readers who wish such a definitive treatment. The Monograph is a reprint of the paper on the Duobinary Coding Technique presented by Mr. Adam Lender, inventor of Duobinary Coding, before the Institute of Electrical and Electronic Engineers on February 1, 1963 in New York. Mr. Lender is a Senior Staff Engineer in Lenkurt's Advanced Development Group. Copies of the Monograph are available on request from* EDITOR, THE LENKURT DEMODULATOR, *in care of Lenkurt or any Lenkurt Field Office.*

BIBLIOGRAPHY

1. H. Nyquist, "Certain Factors Affecting Telegraph Speed," *Bell System Technical Journal;* April, 1924, p. 324.
2. H. Nyquist, "Certain Topics in Telegraph Transmission Theory," *Transactions of the AIEE;* April, 1928, p. 617.
3. A. B. Glenn and G. Lieberman, "Performance of Digital Communications Systems in an Arbitrary Fading Rate and Jamming Environments," *Paper 19.2 WESCON;* August, 1962.
4. A. Lender, "The Duobinary Technique for High Speed Data Transmission," *Paper 63-283, IEEE Winter General Meeting:* February 1, 1963.

World Radio History

World Radio History

the Lenkurt® Demodulator

VOL. 10 NO. 8 AUGUST, 1961

Novel Uses of datatel

Data transmission has become one of the most dynamic efforts in the field of telecommunications. Because of growing industrial and military requirements, each month sees the introduction of new equipment for transmitting and handling data. Yet, the needs and new applications continue to grow at such a pace that even the equipment manufacturers are sometimes taken aback at the diverse uses to which their products are put. While much publicity has been given the new, special-purpose systems, little mention has been made of the ingenious ways in which similar needs have been satisfied by more conventional data systems. This article describes some of these new uses for standard telegraph carrier.

There is a growing tendency to identify *data transmission* almost exclusively with information for business machines and computers. However, data transmission takes in a much broader spectrum which includes telemetry and remote control, as well as digital data for business machines.

Telemetry enables a train dispatcher to "observe" the location of distant trains; it reports on the electrical load at an outlying power substation; it may state the pressure developed by a remote pump or monitor the viscosity of the fluid being pumped.

Remote control adds a new dimension by allowing a central operator to alter the distant conditions reported by telemetry. When telemetry and remote control are used together, complex "organisms" of vast efficiency can be created. The central control point becomes the "brain" of the organism, receiving reports from the telemetry "senses" and performing distant operations with the

World Radio History

remote control "muscles." The essential key to such an industrial organism is found in the nerves which link central brain with distant limbs — the data communication system.

Tremendous efficiency can be gained by such an operation. Centralized control provides superior coordination, and decision-making is speeded. It becomes possible to harness the speed and accuracy of electronic computers to achieve a quality of management not previously available.

In many industries, such centralized control must be approached in gradual steps. Although the concept is old, the means for accomplishing it are new and change rapidly as refinements upon refinements appear. By planning carefully and using familiar, proven equipment, many such systems have been started, modestly at first, then expanded as better techniques and equipment prove themselves particularly suitable.

This philosophy is particularly valid in the field of data transmission itself. Many industries making heavy use of data processing in their normal operations are meeting their transmission requirements adequately and economically with conventional telegraph systems. By planning their operations carefully, they have avoided the need for expensive, highly-specialized transmission equipment.

This is less surprising than it might seem. From a transmission viewpoint there is no difference between telegraph signals and the digital data generated by computers and punched card machines. Although the new machines operate at high speeds within themselves, they don't have to be supplied with data at these speeds. Most industrial operations do not require high speed or "real time" handling of data as in some military systems—air defense, for instance -- in which large

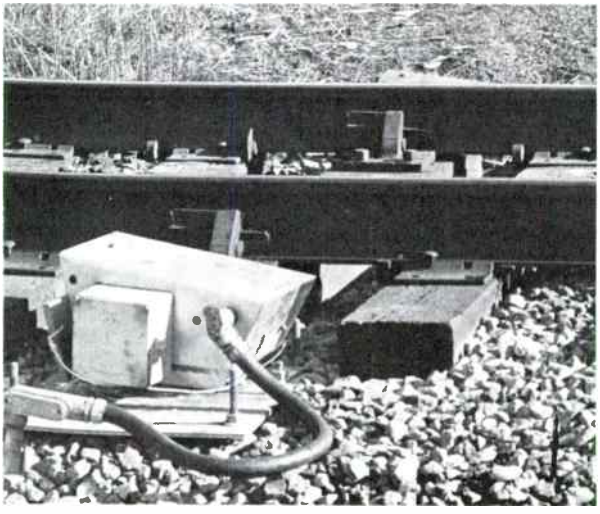


Figure 1. Typical hot-box detectors alongside railroad track. Separate units are required on each side. Detectors are sensitive enough to detect glowing cigarette several yards away.

World Radio History

quantities of data must be transmitted and processed very quickly before they become obsolete. Most of the data handled by industrial systems—payroll, inventories, and similar statistics—do not change very rapidly. The same is true of operating functions which may be monitored and controlled over data links. Almost all of these are adaptable to the 50-75 bits-per-second rate of ordinary telegraph transmission.

Lenkurt's basic telegraph carrier equipment, the 23A Datatel system, provides several excellent examples of its versatility in satisfying communications needs quite different from those for which it was designed. One of the important industries in which 23A systems have demonstrated remarkable versatility is the railroad industry. Modern railroads use the full spectrum of industrial communications facilities: telemetry, remote control, conventional voice circuits, facsimile, business machine data, and even closed circuit television. Confronted by stiff competition and ris-

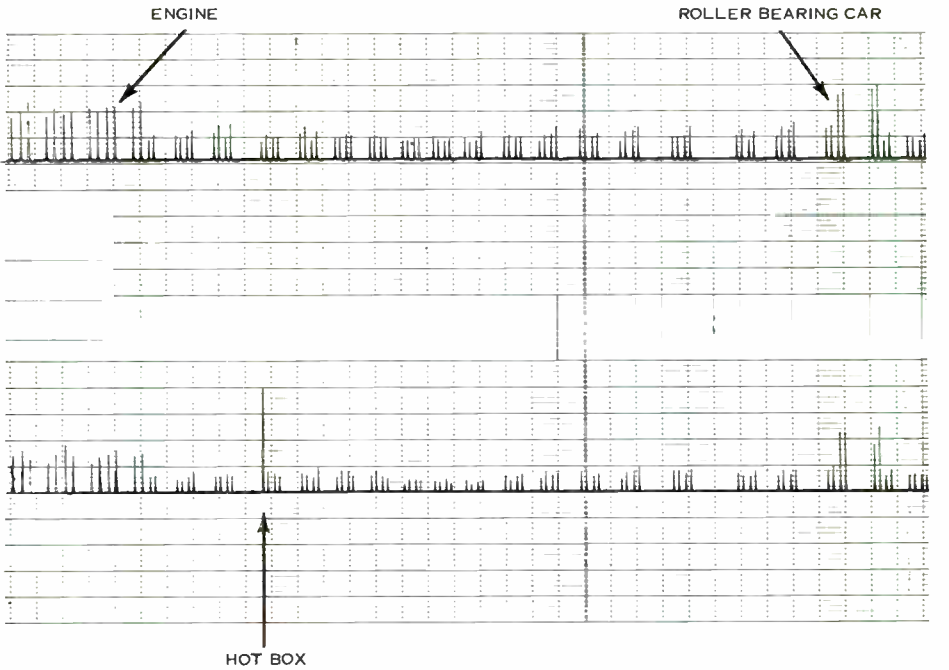


Figure 2. Typical hot-box detector transmission. Each pip shown represents a single wheel. Two recordings are made simultaneously, one for each side. Engine driver wheels, shown at left, and roller bearings always operate at higher temperatures than ordinary solid or plug-bearing journals. Note hot-box indication.

World Radio History

ing costs, railroads have had to exercise extreme diligence in making the most of existing facilities.

Hot-box Detection

One of the persistent problems of railroading is the detection of so-called "hot boxes". On cars equipped with solid bearings, overheated wheel journals can cause derailment by shearing an axle or locking the wheels of a car. If a wheel bearing becomes seriously overheated, the lint packing in a journal box will ignite, causing smoke to issue from the overheated journal box. Before the advent of automatic detection systems, the detection of hot boxes was the responsibility of train crews and personnel along the track. But visual spotting, with its universal signal of thumb and forefinger clasped over the nose, poses many problems. Considerable damage is done to the bearing before the journal box begins to smoke. On some lines with many twists and curves, the train crew is limited in the number of times that they can see the full length of a long train, and hot boxes may go undetected even after the journal box has begun to smoke.

In the past few years, efficient infrared detectors have been developed which can sense the heat radiated by the passing journal boxes. When these are installed along the track, they can detect hotter-than-average journal boxes and signal an alarm before the journal reaches damaging temperature. With the development of these systems has come the need to transmit their data somehow to the train engineer. Since various factors such as train speed, type of bearings, or air temperature do affect the temperature of the journals, in many systems the information is sent to a central location for evaluation. If a hot box is indicated, the train is halted.

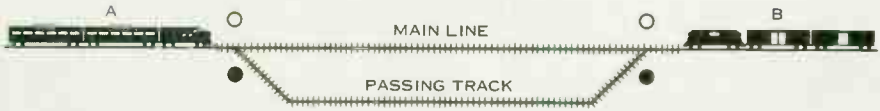
At least one railroad uses 23A Datatel to carry the digital output of hot box detectors. Since the 23A system is restricted to handling binary information, only two conditions can be transmitted. In this case, the tone representing "hot" is transmitted only when an arbitrary temperature threshold has been exceeded.

Another system, which has so far only been proposed, would dispense with transmission to a central point, but broadcast directly to the train crew by VHF radio. In this case, the detector would be connected to a logic network which would count the number of wheels (and thus the cars) *following* a hotbox, and on the basis of count, would select a suitable pre-recorded message for the engineer, telling him the exact location of the hot box. Since the broadcast message would be made up of several portions, code signals at the beginning and end of each would identify individual portions, and these would be matched with the combination ordered by the logic circuit. The identification codes would be detected by a 23A receiver and passed on to the logic circuit. In this case, the railroad already uses 23A for conventional communications, and by using it for this application, would not have to maintain a new type of equipment.

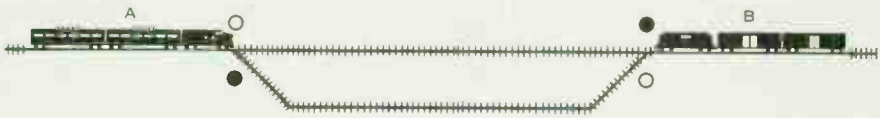
Centralized Traffic Control

Centralized Traffic Control, or *CTC*, is the most important control function in railroad data transmission systems.

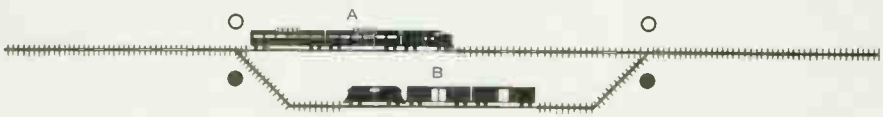
World Radio History



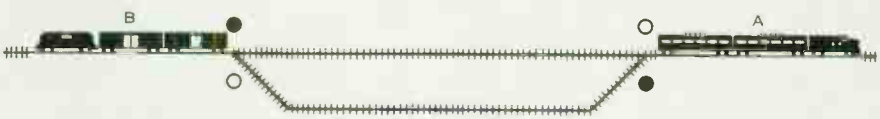
An eastbound passenger train (A) and a westbound freight train (B) approach each other on a single-track main line. The CTC dispatcher must set up a meeting of the two trains and route one of them to the passing track.



The CTC dispatcher throws the east switch of the passing track from the central control board. The new switch positions are indicated to the train crews by the track signals. This step permits the westbound freight train (B) to enter the passing track without stopping.



The westbound freight train (B) has entered the switch and proceeds on the passing track. The eastbound passenger train (A) proceeds on the main line without stopping. Trackside telemetry, such as relays, display the positions of the two trains on the control board. Based upon this information, the CTC dispatcher throws the east switch back to its mainline position to permit the eastbound passenger train (A) to proceed.



The CTC dispatcher throws the west switch of the passing track to permit the freight train (B) to re-enter the main line and proceed. When the freight train clears the switch, the dispatcher will re-set it to main line position. Both trains have met and passed without stopping.

Figure 3. The use of Centralized Traffic Control has enabled railroads to achieve great economies by better coordination of all traffic from a central point. A typical train "meet" controlled from distant CTC board is shown here.

World Radio History

From a central control panel, a CTC dispatcher can throw switches at remote points along the line, re-set signals, if needed, follow the movement of trains across his district, and monitor the results of his actions. The steps of a CTC "meet" are illustrated and described in Figure 3. CTC has become one of the most important tools used by the railroads to cut operating costs and improve service and safety.

CTC systems require rather extensive communication facilities. Coded control signals for throwing switches and setting signals must be transmitted from the control board to points along the line. When these actions are performed, centralized traffic control "telemetry" signals a verification to the control board that the function was actually performed. This same telemetry reports on the location and movement of trains over the line.

Datatel systems are used by a number of railroads in their CTC operations. The application made by one large railroad is typical of the use of carrier telegraph to consolidate data and control functions. A 600-mile mainline section of this railroad was operated with CTC dispatching. As in most of the earlier CTC systems, the 600-mile dispatching zone was broken down into five CTC districts, each having a control panel within its boundaries. Each CTC district was, in effect, a large number of d-c loops closed at the control board, and with data transmitted as coded direct-current pulses. For most of the earlier systems, the poor transmission qualities of on-off d-c pulses and the required high safety factors limited the length of a single CTC district to approximately 100 miles of line.

A few years ago, as the railroad's requirements for data transmission grew with the installation of business machine and car reporting centers, Datatel channels were added to its communications plant to provide increased capacity. A number of factors, among which were long-distance transmission reliability, ease of adding channels and compact terminal equipment, led to testing the Datatel channels for long-distance transmission of CTC data.

In the testing program, a single CTC district served in much the same fashion as a number of d-c teleprinter loops. From a control point located outside the district, frequency-shift tone signals were translated to d-c pulses and transmitted over the signaling circuits of the test district.

The tests proved successful. Subsequently, all CTC dispatching over the 600-mile section was consolidated in a central dispatching office affording greater speed, safety and economy of operations. A CTC dispatcher can establish a train meet and perform all control functions at sidings more than 500 miles away. Data requirements for this particular railroad have continued to grow. Telegraph channels at any one point in the system may carry a broad range of information including telemetry and control functions of CTC, data for business machines and computers and hot box alarm telemetry.

Pipeline Communications

Pipeline transmission systems are called upon to handle massive amounts of data because of extensive use of automation and remote control. Illustrating

World Radio History

the industrial trend toward long-distance remote control, pipeline automation has developed from control of nearby "satellite" pump stations and tank farms, to centralized remote control of points hundreds of miles away. The dispatcher's control panel of a modern pipeline is almost identical to the CTC board of a modern railroad.

Unlike railroads, the "freight trains" of the pipelines are not under the immediate control of a crew. For this reason, pipelines were among the first industries to make extensive use of telemetry, even in short-distance "satellite" operations. Effective remote control of a booster pumping station, for example, requires knowing the suction,

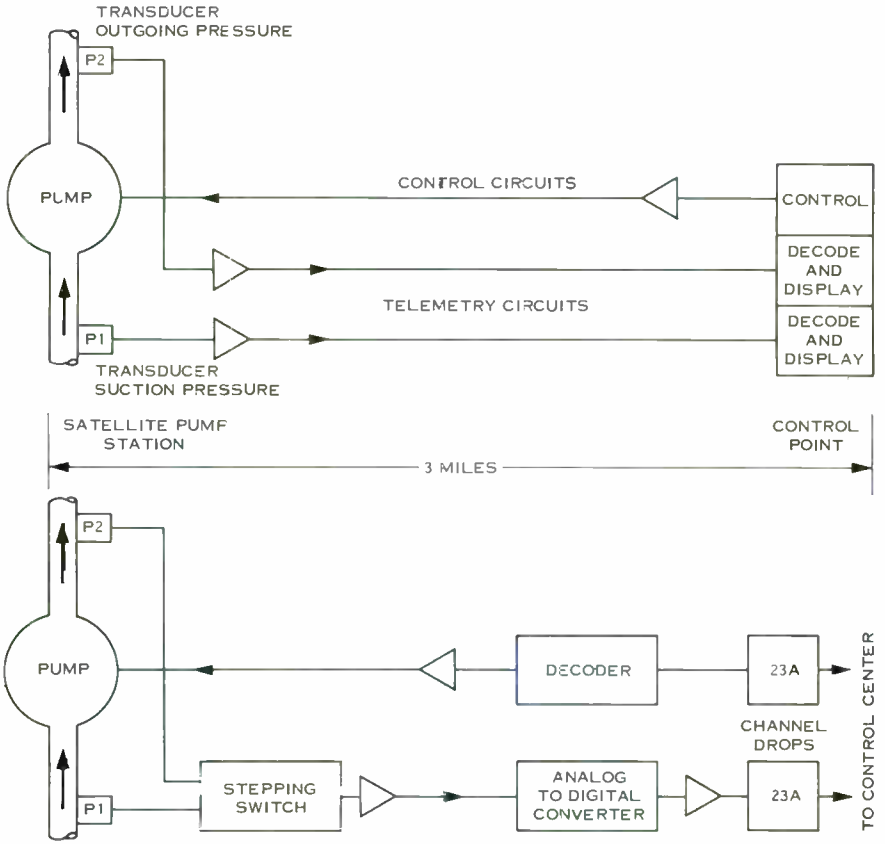


Figure 4. Oil and gas producing units and pipelines have long used relatively short-range direct-current telemetry to monitor and control satellite operations. Direct current transmission impairs accuracy of telemetry beyond a few miles. Judicious use of analog-to digital converters and modern telegraph carrier permits centralized control over far greater distances without obsoleting existing short-range equipment.

World Radio History

outgoing, and differential line pressures, the most critical of these being differential pressure. The conversion to automation and long-distance remote control calls for a greater amount of telemetry, and places an even greater reliance upon the accuracy of telemetering systems.

In changing to centralized control, pipelines have been faced with the problem of using analog telemetering systems which are adequate for satellite operations, but not so suitable for long-distance transmission. The analog equipment frequently represented a substantial capital outlay in its own right. The problem has been solved by using analog-to-digital converters and pulse-duration translators, thus permitting binary transmission via Datatel channels. Thus, pipelines have achieved centralized control without having to discard existing telemetering equipment.

In one particular example, a satellite booster pump station was controlled and monitored by analog telemetry from a tank farm office three miles away. With conversion to centralized control, it was necessary to transmit pump station telemetry over Datatel channels in the company microwave system to a control point 155 miles distant. A number of channel drops were installed for this purpose in a repeater terminal at the pump station.

With the short-haul satellite operation, pressure data were supplied by an analog telemetering system (Figure 4) consisting of two pressure transducers mounted in the pipe on either side of

Figure 5. A typical centralized control station for a pipeline. From this location are monitored such operational details as flow rate, pump pressure and suction, fluid viscosity, and the like. Distant operations can be adjusted in response to changing loads and other variables. Panels in background show profile of terrain which pipeline traverses, and which has important bearing on operating parameters. Centralization is made possible by efficient digital transmission.

World Radio History

the pump; two potentiometer amplifiers to raise the output current of the transducers; display equipment at the control point; and two physical circuits. One transducer indicated suction pressure and the other outgoing line pressure. The operator at the control point obtained the differential pressure by reading the difference between the two.

The continuously variable voltage signals of the analog system could not be transmitted over suitable long-distance channels and the factor of distance in centralized control placed a premium on the accuracy of the information. Yet, this problem was solved with maximum use of existing equipment in a manner that provided improved accuracy of information, as shown in Figure 4.

The signal leads from the two pressure transducers were coupled to a slow-speed stepping (or sampling) switch, the output of which is fed to an amplifier for transmission over a short physical circuit to the tank farm repeater station. At this point, voltage signals are translated into coded pulses by an analog-to-digital converter, and the digital output of the converter is transmitted to the control center over a Datatel channel. At the control center, the digital signals are restored to analog form and connected to suitable meters which show the suction, outgoing, and differential pressures.

Similar arrangements, using existing analog equipment at other booster stations, have yielded substantial savings in new telemetering equipment. The accuracy of the telemetry was improved. The effects of calibration drift on the critical differential pressure were minimized with the use of the common circuits, amplifiers, coding and read-out equipment. Other Datatel channels on the pipeline carry the necessary control signals for centralized operation.

TASI Signaling

The telecommunications industry itself provides an interesting example of how telegraph carrier equipment can be adapted to still another type of control function. In this application, the 23A equipment is used to provide dialing and supervision for the voice circuits handled by the TASI systems in the California-Hawaii and Florida-Puerto Rico submarine cables. Here, the cost of the cable—$37 million in the case of the Hawaiian cable—is so high, that even very expensive means of increasing circuit capacity provide great savings.

In the TASI system (which means "*T*ime *A*ssignment Speech *I*nterpolation"), extra channel capacity is obtained by using the idle time present in the speech messages transmitted over the cable. All telephone conversations have pauses, listening periods and the like, and in ordinary practice this transmission time is wasted. By salvaging these silent periods and using them to transmit additional conversations—which have their own idle periods—twice as many conversations can be accommodated as there are carrier channels to transmit them.

To accomplish this remarkable juggling feat, TASI monitors each carrier channel to determine when it is active or idle. When an "excess" circuit is connected to the system, it is instantly switched to a cable channel which is idle at that moment. When the displaced circuit becomes active, it is, in turn,

World Radio History

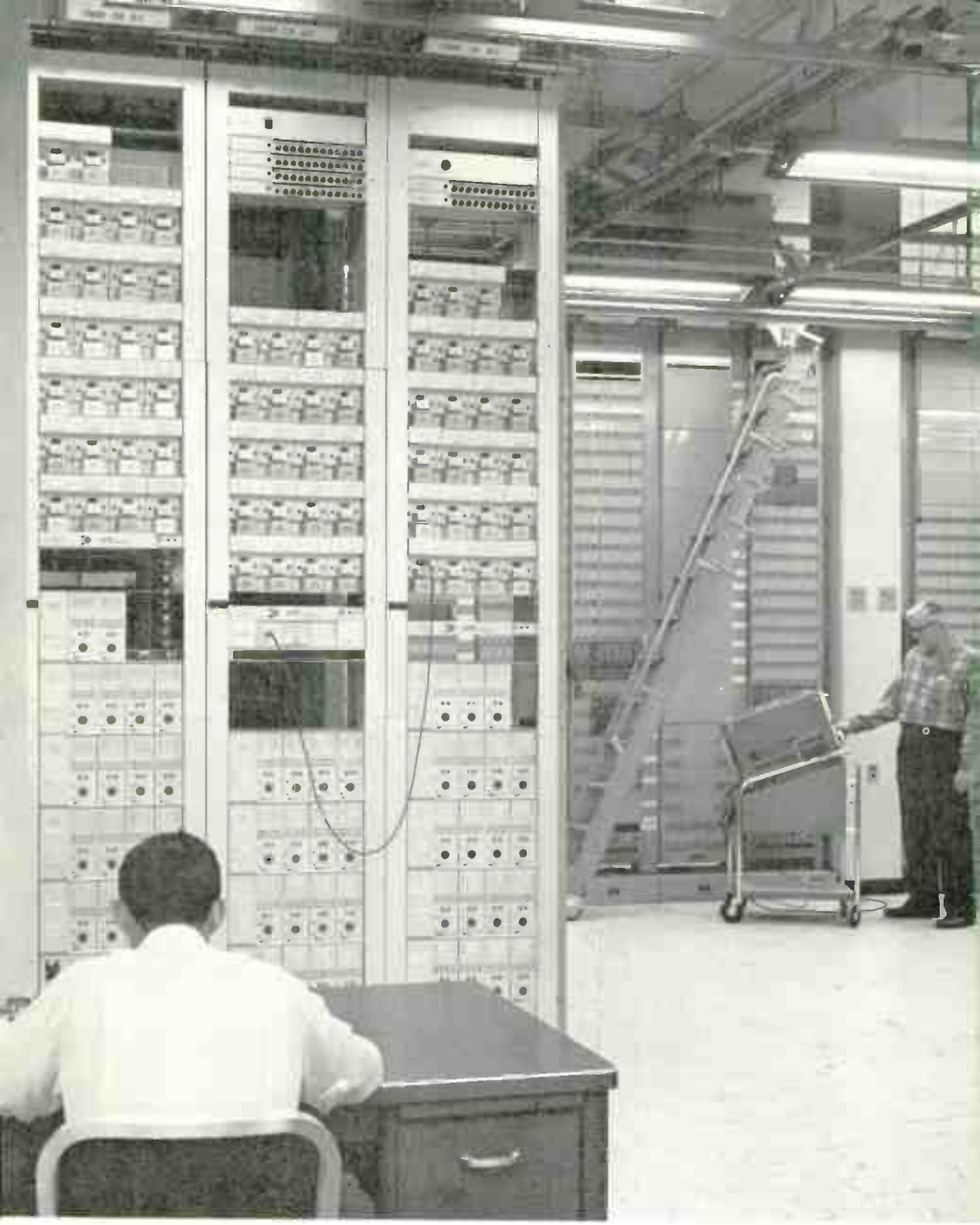


Figure 6. California end of Hawaii-mainland submarine cable. In background are some of the racks for TASI system which squeezes waste time out of conversations to obtain additional voice circuits from the limited cable bandwidth. In left foreground is the Lenkurt 23A Telegraph Carrier system which provides signaling and supervision for the TASI voice circuits. The 120-cycle channel filters permit 22 to 26 signaling channels or 100-speed telegraph channels to be sent through each voice channel.

World Radio History

connected to another channel which has become idle.

Obviously, if this were attempted with only a few circuits, there would be many moments of conflict in which all circuits would be active simultaneously, leaving no room for an additional circuit. When several dozens of conversations are used, however, the probability becomes quite high that some circuit will be idle at any given instant.

Although the high-speed switching and monitoring equipment required for this job is complicated and expensive, the extra circuits obtained in this way cost less than additional channels obtained without TASI.

Because TASI depends on interruptions or pauses in the individual messages, the presently-used single-frequency signaling tones cannot be transmitted with the individual circuits in the normal way. Since the signaling tone is present on the line when there is no connection or during dialing, TASI would interpret this idle-condition tone as "activity" and would connect it to the first available cable channel. This, of course, would result in idle circuits competing with the active circuits for the limited number of cable channels.

To get around this difficulty, signaling and supervision is separated from the message circuits and transmitted through non-TASI channels, using 23A telegraph carrier. In order to make the fullest use of the limited bandwidth available, 120-cycle spacing is used, thus permitting 23 or more telegraph (or signaling) channels to be sent over a single 3-kc voice channel, instead of the 16 possible with the more commonly-used 170-cycle spacing. This saves one of the expensive cable channels, since the cable has a capacity of only 48 three-kc channels at this time, and some of these must be reserved for data transmission, private-wire service, and the signaling channels for TASI.

Conclusions

Many of the uses made of carrier telegraph equipment have taken it far afield of the application envisioned by its designers. For Datatel systems in particular, one of the newest applications suggested involves parallel transmission with a number of channels for increased speed and direct read-out to and from the parallel characters of tape.

A review of the applications made or proposed for various types of carrier telegraph equipment suggests that users, rather than manufacturers are contributing the greater amount of imaginative applications engineering. •

BIBLIOGRAPHY

1. M. T. Nigh, "Digital Telemetering Proves Accurate and Reliable," *Pipeline Industry*; March, 1957.
2. "How to Make Non-stop Meet with CTC," *Railway Signaling and Communications*; April, 1957.
3. T. B. Collins, Jr. and J. E. Pitts. "Application of a New Carrier Telegraph System," *Communication and Electronics*; January, 1960
4. J. N. Albertson, "Voice Frequency Telegraph Carrier Systems and Centralized Dispatching on Texas and New Orleans Railroad," *Paper presented at Petroleum Industry Electrical Association Annual Meeting*; April, 1960
5. K. Bullington and J. M. Fraser, "Engineering Aspects of TASI," *The Bell System Technical Journal*; March, 1959.

World Radio History

World Radio History

the Lenkurt Demodulator

VOL. 14, NO. 6 JUNE, 1965

Delay Distortion

Within the last decade, high-speed data transmission has become one of the most vital services provided by the great communications networks. The explosive growth of electronic data processing and the construction of elaborate defense data communications networks from existing telephone facilities has placed new importance on some rather old transmission problems. One of the most important of these is delay distortion. This article reviews some of the more important aspects of delay distortion — why it occurs and how it may be overcome.

For many years, the quality of a communications channel has been judged by its ability to reproduce the moment-to-moment level or amplitude of the original signal. In judging the ability of voice communications systems to satisfy the listener, amplitude response has proved to be a good measure of performance. About 1930, when pictures were first transmitted by wire, it was discovered that a channel having excellent amplitude response might be quite unsatisfactory for picture transmission. The harmful effect, originally called *phase distortion,* and now more correctly called *delay distortion,* resulted from the phase-shift characteristics of the transmission path, and had little relation to the amplitude response or fidelity of the transmission path. Since then, other communications services have sprung up which are equally vulnerable to this type of distortion — facsimile, television, and high-speed data transmission are typical.

About this article

This article is a revision of the July, 1960 issue of The Lenkurt Demodulator. Because of the interest in this subject, resulting especially from the increase in the amount of data transmitted over telephone transmission systems, the original article has been modified slightly and issued again for the benefit of our present readers.

World Radio History

What Is Delay Distortion?

Electromagnetic waves travel 186,000 miles per second in free space. Electrical signals, however, do not travel this fast through communications channels. In fact, signals may travel over certain types of circuits as slowly as 15,000 miles per second, and will rarely travel faster than about 100,000 miles per second over a microwave radio path. These lower velocities result from the nature of the communications equipment or the transmission path.

A telephone line behaves like a low pass filter, particularly if inductive loading is used to reduce attenuation. Multiplex systems use very sharp filters to separate one channel from another, and the tuned circuits in a radio receiver serve the same purpose. All these filters and filter-like elements introduce delay.

The slowing down of a signal in its passage through a communications channel is of little importance. Delay becomes a problem only when it interferes with the ability of the receiver to understand the message. In the case of speech, delay distortion causes little interference since the ear is relatively insensitive to phase variations. Thus, it has not been necessary to correct for delay distortion in telephone systems. Facsimile, telegraph, and data signals, however, are quite vulnerable to delay distortion.

For example, if two tones, such as 1200 and 2200 cycles per second, are used to transmit binary data (shifting from one frequency to the other), it is important that these two tones experience approximately the same transmission delay in going from one end of the circuit to the other. If the data is being transmitted at 1000 bits per second, each bit will be one millisecond long. If these transmissions consist of alternate **1**'s and **0**'s, the signal will be alternately shifting between 1200 and 2200 cps. The transmission propagation time for these two tones between the two ends of a given circuit can vary considerably. For example, 60 miles of loaded telephone cable may introduce a delay to the 2200 cycle tone of 6.1 milliseconds as compared to the 5.1 milliseconds for the 1200 cycle tone, a difference of 1 millisecond. If the 2200 cycle tone is transmitted first followed by the 1200 cycle tone (each transmitted for 1 millisecond), it can be seen that they will both be received at the same time, rather than one following the other. In 120 miles they would be received in the reverse order!

In high-speed data transmission, the problem of delay distortion becomes more and more serious and troublesome as the transmission rate increases. Data bits usually originate as rectangular-shaped pulses which are used to modulate a carrier at a particular keying rate for transmission over a communications circuit. The pulses resulting from this

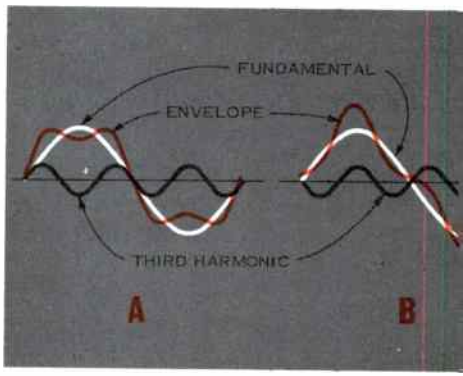


Figure 1. As shown in diagram A, the normal vectorial addition of the energy at the fundamental and third harmonic frequencies produces a slightly rectangular-shaped pulse. If the third harmonic is delayed by one-half cycle relative to the fundamental, as shown in diagram B, the pulse envelope becomes seriously distorted.

World Radio History

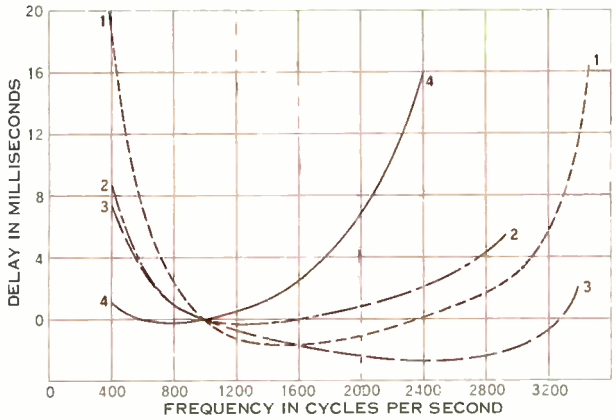


Figure 2. Comparison of the envelope delay in typical voice communications channels. Curves 1 and 3 represent the delay in several thousand miles of a toll-quality carrier system. Curve 2 shows the delay produced by 100 miles of lightly loaded cable; curve 4 shows the delay in 200 miles of heavily loaded cable.

modulation process are composed of many frequencies whose amplitudes and phases have a fixed relationship in time. The envelope of these pulses results from the energy at the fundamental and harmonic frequencies adding together vectorally. If the pulses are processed through circuit components with very non-linear phase characteristics, such as multiplex channel filters, the pulse shape can become seriously distorted. As shown in Figure 1 if the third harmonic is delayed by one-half cycle relative to the fundamental, the pulse shape is severely distorted.

Higher data speeds are achieved by increasing the rate at which the carrier is keyed, thereby shortening the width (or duration) of the signal pulses. Because of the shorter pulses, slight shifts in time or phase of the component frequencies have a greater effect in distorting the signal, with a corresponding increase in error rate.

Also, the higher the data speed, the greater the channel bandwidth required for successful transmission. The reason for the increased bandwidth lies in the nature of a high-speed pulse. When the pulse begins or ends, the rapid change causes signal energy to be distributed over a wide band of frequencies on either side of the pulse frequency. The exact amount of energy appearing at each frequency on either side of the pulse frequency depends on the nature of the pulse — its shape, rise-time, and so on. If, for any reason, some of the energy from either sideband is displaced in time or amplitude from the original value, the pulse will be distorted when it is reconstructed by the receiver. If the delay is great enough, some of the energy from a signal pulse may actually be delayed enough to interfere with the following pulse, thus destroying information carried by both pulses. It is evident, therefore, that delay introduces distortion only when various frequencies in a communications channel are delayed by different amounts of time.

Phase Shift

The phase and frequency of a signal are, by definition, inseparable. In fact, a good definition of frequency is the rate of change of phase with respect to time, or $d\phi/dt$, where ϕ is the phase shift (usually in radians — π radians equal 180°, 2π radians equal one cycle) and t is time in seconds. Thus, it follows that the more the phase of a signal is shifted in passing through a channel, the more time is required for it to get

World Radio History

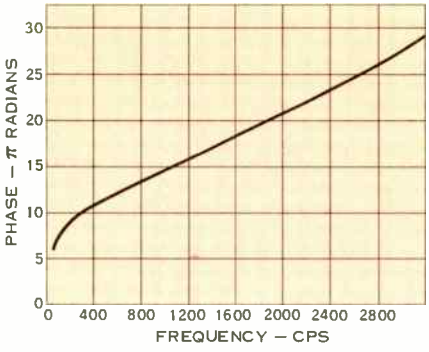


Figure 3. Phase shift characteristic of a high-quality 100-mile carrier telephone circuit.

through the channel. Where phase shift is known, the phase delay of a single frequency is

$$time = \frac{phase\ shift\ (radians)}{frequency\ (radians\ per\ sec.)}$$

This is usually expressed

$$t = \frac{\phi}{\omega}.$$

It is important to note that in practical systems, phase delay, as expressed above, is applicable only to single, steady-state frequencies.

In an ideal system, phase shift is directly proportional to frequency. All signals passing through such a system would be delayed equal amounts, regardless of their frequency. Unfortunately, phase shift in a communication channel is never linear. In a high-quality system, the overall phase shift characteristic may look like that shown in Figure 3.

Envelope Delay

Whenever a complex signal (such as a modulated or keyed carrier frequency) is transmitted, the relationship given above for phase delay no longer holds true, unless the system is perfectly distortion-free. Since phase shift is always non-linear in actual systems, some of the component frequencies undergo more phase shift than they would in a linear system. As a result, they travel through the system slightly slower than some of the other frequency components.

For simplicity, assume that the complex signal consists of only two component frequencies. Added together, the two frequencies form a beat-frequency or modulation envelope. Since the two component frequencies travel at different velocities through the channel (because of non-linear phase shift), the relationship between them constantly changes, and the modulation envelope travels through the channel at a third velocity. If phase shift were linear, both component frequencies would travel at the same velocity, and there would be no displacement of one frequency with respect to the other, and no independent delay of the modulation envelope or envelope delay.

The more non-linear the phase shift, the greater the envelope delay. In other

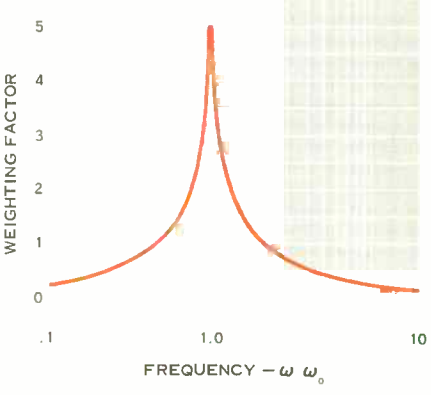


Figure 4. Weighting factor for the phase shift near the resonant or cutoff frequency of a network or line. Phase shift is very high at resonance, but falls off rapidly at higher and lower frequencies.

World Radio History

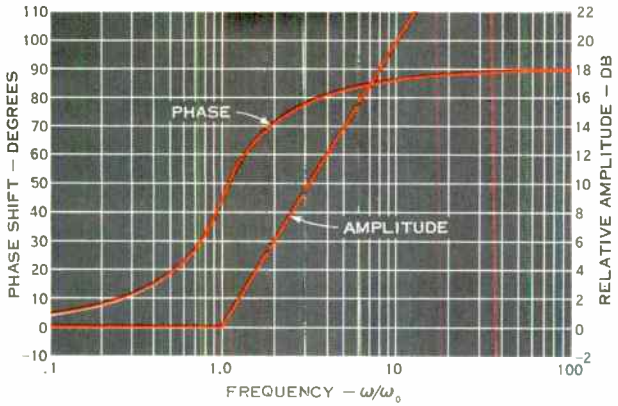


Figure 5. Fictitious amplitude characteristic of a network and the phase shift that would result. The smooth phase shift curve shows the weighting effect diagrammed in Figure 4.

words, the greater the rate of change of phase shift, the more envelope delay will result. The delay in seconds can be calculated by differentiating phase shift with respect to frequency:

$$\textit{envelope delay} = \frac{d\phi}{d}.$$

Since virtually all forms of electrical communication employ signals which require a band of frequencies for successful transmission, envelope delay is the form of delay of greatest general importance. In this article, further reference to "delay" will mean envelope delay unless other indicated.

Usually, only relative delay — the maximum range or difference in delay values in a channel — is of importance, since only the delay *difference* causes distortion in the received signal. Absolute delay—the total delay experienced by signal elements — is usually not important except where signals or parts of a signal are transmitted from one point to another over different routes and must arrive at the same time.

Delay Equalizers

Where amplitude response of a circuit is unsatisfactory, an equalizer is used to introduce a controlled amount of loss at certain frequencies to obtain the desired performance. In the case of excessive relative delay, a network which would correct the phase shift characteristics of the communications channel might very well neutralize the desired attenuation of the filters responsible for the delay. Special delay equalizers are required to overcome this problem. Ideally, a delay equalizer is a network which introduces a controlled amount of phase shift at various frequencies, but causes no signal loss at all. Practical delay equalizers, however, cannot avoid affecting amplitude response to some degree.

Although it is possible to design filters which will reduce or postpone the appearance of delay distortion (by adding special kinds of filter sections), the more common practice is to add a so-called "all-pass" network which adds delay at selected frequencies, but which adds negligible attenuation at any frequency.

In a carrier system, the channel bandpass filters which isolate individual channels from each other are the principal sources of delay distortion. These filters should have uniform amplitude response within a desired band of frequencies, but must exhibit a very rapid attenuation of all frequencies outside the desired band. Unfortunately, such

World Radio History

rapid change in the attenuation characteristic of a filter is also accompanied by rapid changes in the phase shift, as indicated in Figure 5. As a rule, phase shift is maximum at the cutoff frequency or resonant point of a circuit, but declining at other frequencies, as shown in the weighting curve, Figure 4. Thus, a circuit having flat attenuation characteristics would still have nonlinear phase shift and would introduce envelope delay distortion.

Figure 6 shows a typical all-pass network and some of the delay characteristics that may be obtained with such a network. The different delay characteristics are obtained by changing the values of components. Since the network impedance and the frequency at which the delay is obtained are also affected by component values, design of such equalizers is a complex art.

A typical equalizer will have several sections, each of which may have a different reference frequency (f_c) and a different "width factor" (B). Such equalizers may be custom designed to correct the delay characteristics of a certain type of equipment or communications path or may be continuously adjustable for universal application. Figure 7 compares the envelope delay characteristics of a carrier system channel without equalization, the amplitude response of an experimental 3-section equalizer, and the delay characteristics of the equalized channel. Note that there are three "humps" or ripples in the equalized delay characteristics—one for each section in the equalizer.

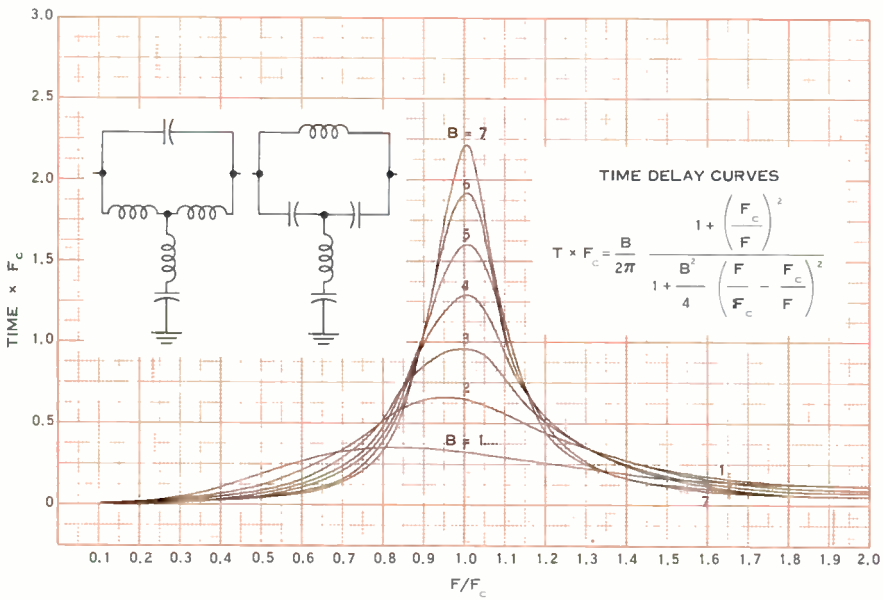


Figure 6. Two typical all-pass network configurations and a family of envelope delay curves which they might produce. Width factor (B) and critical frequency (f_c) are controlled by the values of the network components. Typical delay equalizer will consist of several sections, each designed to add a carefully-determined amount of delay to a small portion of the channel passband.

World Radio History

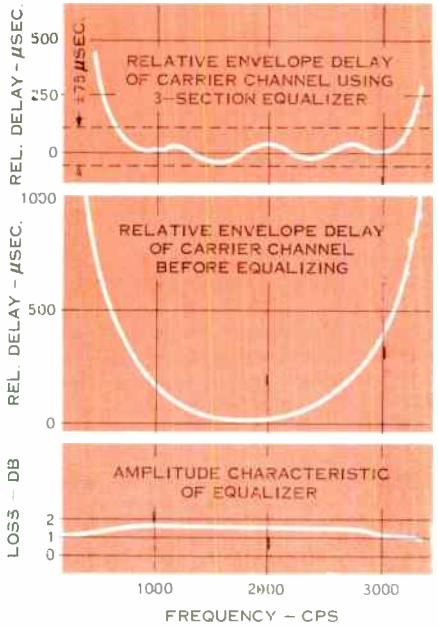


Figure 7. Center panel shows the envelope delay characteristic of a typical carrier system voice channel before equalization. Addition of a 3-section equalizer reduces relative envelope delay to ±75 microseconds over a large portion of the bandwidth, as shown in top panel. Bottom panel shows that amplitude response of the equalizer varies less than 1 db despite effect on envelope delay.

The more sections that are used in a phase equalizer, the finer the ripple in the delay characteristic. The residual delay causes distortion in the form of echoes. Where the ripples are coarse, as when few equalizer sections are used, the echoes are separated very little from the signal and may have only the effect of changing the amplitude of the signal. Where the ripples are fine, the echoes are delayed more. It has been shown that a system is considerably less tolerant of noise interference as the time separation between signal pulse and echo is increased. Thus, the fewer equalizer sections required to achieve a given relative delay tolerance, the less susceptible the system is to interference. This sensitivity does not increase, however, after the echo is delayed more than one pulsewidth for data, or about 60 picture elements in the case of facsimile.

Future Needs

The long-range trend in the communications industry is toward more and more pulse transmission, both for data and for speech. It is primarily pulse transmission that is vulnerable to delay distortion. In modern systems, a message may be transmitted from one point to another over one of several possible routes which may differ greatly from one another in their phase response. To make better use of the existing communications networks, it will be necessary to obtain uniform delay characteristics, regardless of the routes. Techniques for automatically equalizing the phase response of an entire circuit after a connection has been obtained, regardless of circuit length or the nature of the means of transmission, are now being developed. Although these techniques are now rather expensive, they will undoubtedly play an important role in the development and growth of future data transmission systems.

World Radio History

World Radio History

SECTION IV
GENERAL COMMUNICATIONS

World Radio History

World Radio History

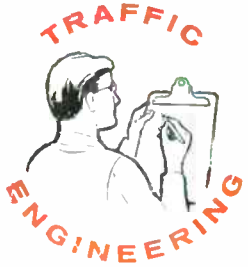
the Lenkurt Demodulator

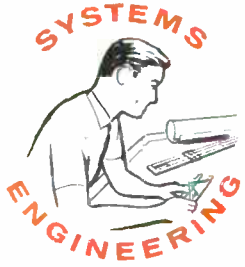
VOL. 14, NO. 7 JULY, 1965

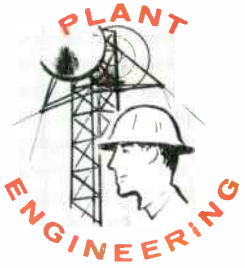
TRANSMISSION SYSTEM PLANNING

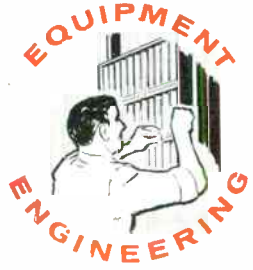
Modern communications networks are becoming more and more efficient and reliable, despite the fact that they are continually becoming more complex. These improvements undoubtedly are a result, not only of advanced engineering techniques, but more importantly, of effective planning. *However, many people in the communications industry seldom become directly involved in the intricate planning that goes into developing an efficient and reliable system.*

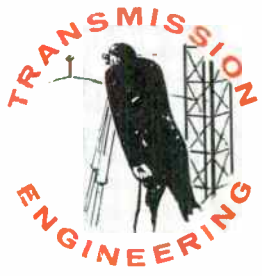
This article discusses some of the many considerations involved in the planning of multichannel transmission systems for a communications network.











World Radio History

The coordinated efforts of a number of specialized engineering functions are needed to plan and build a communications system. The size and number of such functions depend, of course, on the type of system proposed and its complexity. Certain of these functions, however, usually play an integral part in the development of all types of systems. Most significant among these is *systems engineering* — that function which has the overall responsibility for directing and coordinating the efforts of the others. These other functions include such activities as *traffic engineering, plant engineering, transmission engineering,* and *equipment engineering.* All of these functions are usually active during various stages of planning, engineering, and installation of a multichannel transmission system.

The term *communications system* is very broad. It may mean anything from a huge global complex, with thousands of miles of transmission facilities and many subscribers, down to a walkie-talkie carried by a soldier. A typical telephone communications system consists of user or *subscriber* equipment such as telephones, teletypewriters, and business or data processing machines which are connected by wire lines to a switching center. The switching center permits each subscriber line to connect with other lines of the same network. To communicate outside of the local network, however, common-use or trunk circuits must be established from the switching center to other switching centers. Such common-use circuits are typically multichannel *transmission systems* consisting of voice multiplex equipment and open-wire, cable, or broadband radio facilities with intermediate repeaters, as required.

Technical Requirements

When the need for a communications transmission system is firmly established, the technical and operational requirements that will satisfy the need must then be planned in detail. There is usually a choice of different operating arrangements and types of equipment, any of which might provide good service. The principal objectives are that the system be simple and reliable, and be practical and economical to operate. It should also be suitable for future expansion—which is almost always necessary. Accurately defining the system requirements early in the planning stage will certainly result in better overall economy and greater satisfaction in the communications services provided.

The amount of terminal equipment and the number of trunk circuits or channels required for the transmission system are determined by the intended number of users; the possible number, kind, and duration of messages sent by each of these users; and by the desired quality of service. The first task in planning the system, therefore, is to determine the volume and types of traffic that it will be expected to handle. This is usually done by a traffic engineer, whose traffic study provides the basis for estimating immediate requirements and also the nature and extent of future needs. The traffic study is essentially an analysis of the amount and type of use the proposed system will receive—in other words, the characteristic behavior of the users.

A typical transmission system must be able to handle various types of traffic,

World Radio History

such as speech with signaling, telegraph, high-speed data, and graphics. Because each type of signal imposes a different load on the system equipment, it is also necessary to determine the distribution of each possible type of traffic.

If the proposed system is simply an expansion or an improvement of an already existing communications network, a traffic study of the intended users can be made directly. However, in planning new systems there is, of course, no way to examine the behavior of the intended users. In this case, it is necessary to extrapolate information from existing systems where the requirements very nearly match those of the proposed system. Such an indirect study can be made from the substantial amount of existing traffic data, collected by various agencies, that is applicable to the design of most communications systems.

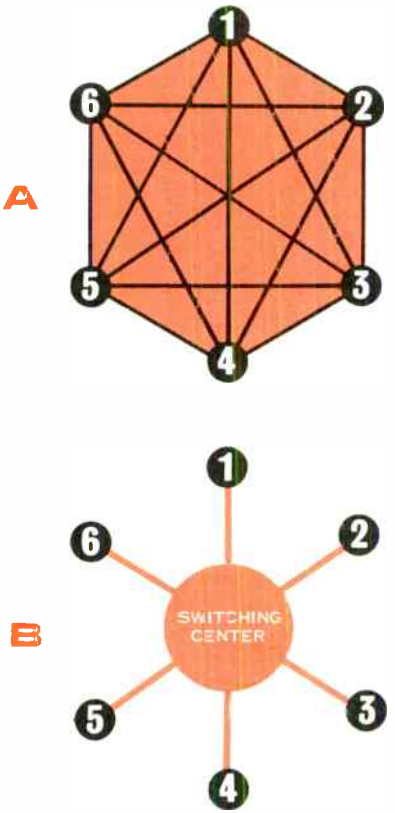


Figure 1. Example of two types of routing used in communications transmission systems. Diagram A shows six stations interconnected by direct and alternate routes. Diagram B shows the same six stations interconnected through a central switching center.

After establishing the types and distribution of traffic and the number of channels required for the new system, the technical quality and the desired operating characteristics of the circuits must be determined. The facilities that may be used to furnish transmission circuits differ greatly in such things as net loss, bandwidth, distortion, noise and load-handling capacity. To meet the specific transmission requirements, facilities must be selected that are suitable for each type of service, particularly in cases where special or unusual capabilities are needed.

When a message first enters a transmission system, it is converted into a signal whose electrical characteristics vary depending upon the particular treatment and processing it receives in the system. As the message travels through the system, its electrical characteristics are changing continuously. All along the circuit, it is being attenuated, amplified and re-amplified, shifted in frequency and phase, and even distorted, before reaching the output of the receiving terminal. So long as the signal is confined to one system its characteris-

World Radio History

tics are limited to the requirements of that system.

However, before a signal can be transferred from one system to another, its characteristics must be suitable to the technical requirements of the next system. It is desirable, of course, to be able to transfer the signal to the next system in an economical and efficient manner—that is, without the need for complicated interface equipment. This is especially true in public or common-carrier systems where long-distance direct-dialing telephone communications depend upon the cooperation and coordination of hundreds of different systems. In the United States alone, there are more than 2500 independent telephone systems, in addition to the Bell System, that must interconnect to provide complete nation-wide communications.

Standards and Practices

Inherent in the development of most transmission systems is the need to recognize and accept existing operating practices and performance standards to promote a uniform system of communications networks. This need has fostered the development of numerous written standards and practices covering almost every significant aspect of electrical communications. These standards provide the basis for establishing the performance criteria of transmission systems. The use of these standards, of course, is not obligatory, but because of the benefits to be derived, they are often accepted by general or implied consent.

The standard practices for interconnecting transmission systems in North America have been developed largely by the Bell System and generally accepted by the rest of the telephone industry. To help establish international communications, many countries, especially in Europe, use the recommendations of the CCITT (International Telegraph and Telephone Consultative Committee) and the CCIR (International Radio Consultative Committee). These two committees, which are agencies of the United Nations International Telecommunications Union, have been effective in assuring that the communications practices and performance standards of various nations are compatible.

In establishing the performance requirements of many types of special purpose communications systems, it is not always practical to use the universal standards developed for common-carrier networks. This is especially true in certain private or industrial systems and in various military systems which are not necessarily expected to interface with common-carrier networks. Often, these systems must be specially designed to meet unusual or higher-than-normal performance requirements. In these cases, relying upon universal standards and practices to establish system performance criteria may not provide adequate operational capabilities. In fact, the use of such broad standards might impose objectionable restraints on the capabilities of these systems, resulting in poor or inadequate performance.

The need for unusual performance capabilities is especially evident in many military communications systems. In common-carrier communications systems, new services requiring unusual performance capabilities can be added in a deliberate, carefully-planned manner. However, military needs are apt

World Radio History

to change radically and quickly, and often under the worst possible circumstances.

Because the circumstances under which a military system may be used are often unpredictable, it must be assumed that the traffic load will be severe. Military systems can expect to handle proportionately more facsimile, digital data, and other non-voice traffic than commercial systems. Digital transmission provides the speed required by modern military operations and, consequently, its use has become increasingly important in the central control of complex tactical, strategic, and air defense weapon systems, and in transmitting detailed intelligence reports, weather reports, and world-wide logistics information.

Such digital traffic imposes a much heavier load on a transmission system than ordinary voice traffic. For this reason, military systems must be capable of handling much greater average input signal powers than required for comparable commercial systems. In addition, tactical military communications equipment must be ruggedly built to withstand severe mechanical shock and vibration encountered in typical military exercises. Also, the equipment must be extremely reliable and easy to maintain, since well-trained personnel may not always be available to keep the system operating properly during emergency conditions.

Today's military communications systems range from small portable or mobile systems used in combat areas to support tactical operations, to the immense, highly complex Defense Communications Systems (DCS) which interconnects U.S. military and government installations located all over the world. The DCS is directed by the De-

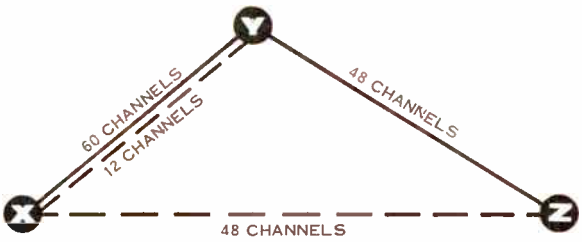
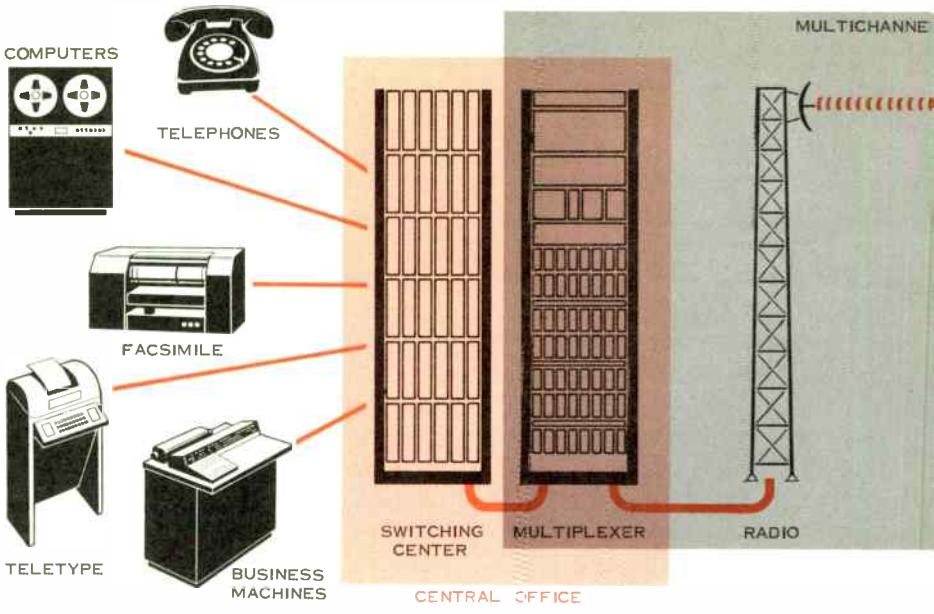


Figure 2. Possible routing for 12 channels between stations X and Y and 48 channels between stations X and Z.

World Radio History



fense Communications Agency (DCA) which issues *DCS Engineering-Installations Standards* to establish uniform performance criteria for each component of the system.

After selecting the performance criteria, the next step in planning a transmission system is to prepare the specifications that will guide its engineering and development. These specifications are more effective when manufacturers are able to bid the particular features of their equipment which will best satisfy the technical and operational requirements of the proposed system. For instance, the operating company might prescribe in its specifications that the transmission system connect several points by direct and alternate routes, as shown in Figure 1A. However, a manufacturer, bidding on the job, might determine that central switching (shown in Figure 1B), rather than direct or alternate routing, provides the most economy and the best service for his particular equipment. In other words, the system planners should take advantage of the manufacturers' knowledge of the use of their products and of their experiences in systems engineering.

Planning Ahead

The system planners must also remember that one of the criteria by which the proposed system will be judged will be the specific plans for expansion to meet future growth needs. Nearly any system can be expanded—with enough money and effort. The competent engineer, however, plans for orderly and economical growth and de-

World Radio History

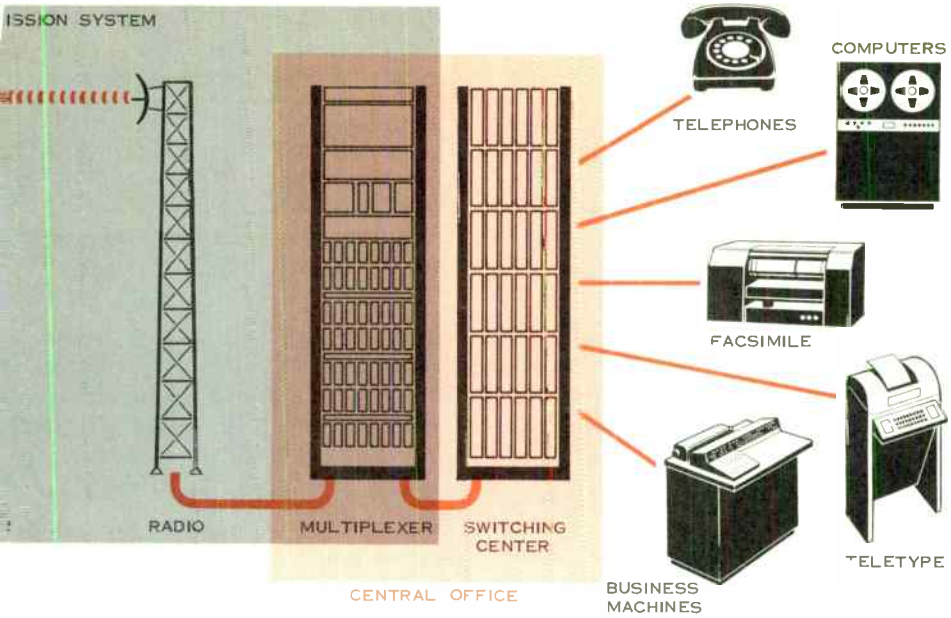


Figure 3. Typical communications network with a multichannel transmission system containing multiplex equipment, microwave radio terminals, and intermediate repeater station.

tails in his plans how it will be done and estimates how much it should cost.

When accepting bids from a communications equipment manufacturer, it is important to consider not only the initial cost, but the possible long-term costs. Under the usual pressures of doing business, it may be tempting to accept the lowest bid, while overlooking the potential limitations and costs to operate the transmission system for a long period of time. Although the initial cost of a more reliable, efficient, and capable system may seem high, it usually proves to be a real bargain when compared to the excessive costs and work required to repair and maintain an inferior system. It may also be costly later to improve the performance capabilities of a low-cost system to make it suitable for the expected increase in newer types of traffic, such as data and graphics.

The load-handling capabilities and noise performance of most of today's transmission systems are based primarily on voice traffic with allowance for the usual pilots, signaling tones, carrier leak, and a small amount of telegraph traffic. So long as these systems handle mostly voice traffic, they should certainly perform their job satisfactorily.

World Radio History

However, will such systems be able to handle the large amount of data traffic that is predicted for the future? This is a very important question that must be carefully considered when developing a system that is expected to operate economically and efficiently throughout its lifetime.

Data signals are presently transmitted over voice channels at a level somewhere between -5dbm0 and -15dbm0. It is desirable, of course, to transmit data signals at the highest possible level to prevent impulse noise from causing too many errors. Dial-operated data service, provided by the telephone industry, as well as other similar data services, presently operates at a level of -8dbm0. The signals from data equipment impose a much greater load on multichannel transmission systems than do ordinary voice signals. Because of this heavier loading, data service cannot be simply *added later* without first considering its effect on the load handling capacity and noise performance of the transmission system (see *The Lenkurt Demodulator*, March 1965).

For example, consider a typical 600-channel transmission system designed to perform in accordance with today's CCITT voice loading recommendations. If data service at -8dbm0 is later assigned to 120 of the 600 voice channels, then the remaining 480 channels must be removed from service to accommodate the data load. Even if only 60 of the 600 channels are used for

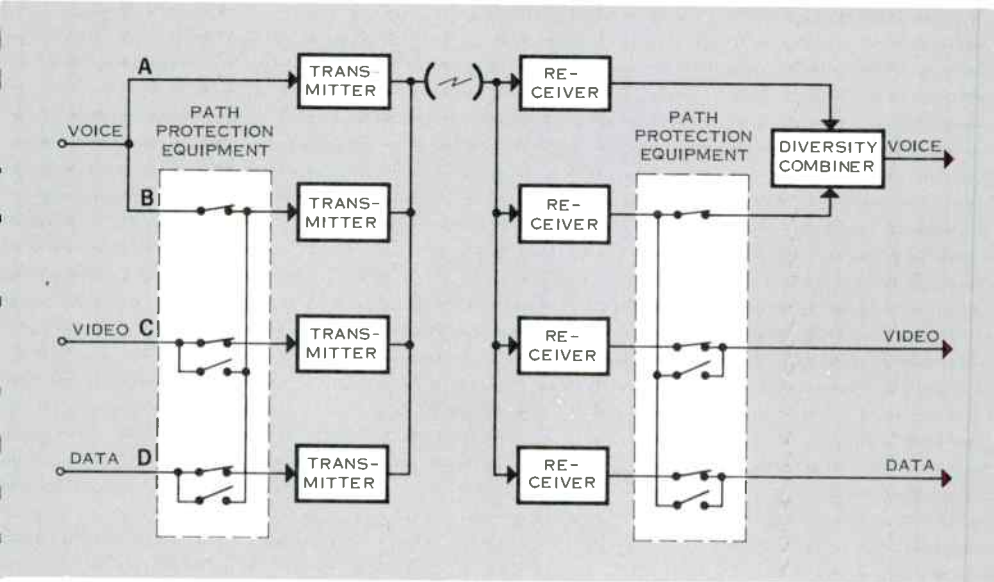


Figure 4. Functional diagram of typical one-for-three broadband radio path protection system.

World Radio History

data, 300 of the remaining 540 channels must be removed, leaving only 140 channels available for voice traffic.

It may be possible to lower the level of the data signals to keep more channels active, but this will drastically raise the data transmission error rate. Theoretically, the error rate in a typical binary system may increase by a *factor of 10* for every db that the signal level is decreased—due to the resulting lower signal-to-noise ratio. This means that if the level of the data signals were lowered from, say -8dbm0 to -13dbm0, the error rate may increase *one hundred thousand times*.

But what about the idle channels? Since they can no longer be used, but were part of the original system cost, they represent wasted investment and, even worse, are no longer able to produce revenue. Adding data service later to a transmission system designed primarily for voice traffic, therefore, may seriously decrease the efficiency of the system. To prevent such waste or to avoid having the system become quickly outmoded, particular attention must be paid to the traffic growth and to the load-handling requirements of the proposed system. In this regard, future traffic requirements must be carefully considered, especially for data where such things as signal levels, frequency stability, impulse noise, delay distortion, and intermodulation distortion are much more critical than for ordinary voice traffic.

Satisfying the Requirements

After specifying the technical requirements of the system it is then necessary to determine what type of equipment will be necessary to meet the initial requirements while providing for economical expansion to meet future growth. Of course, if the project is an addition to an existing system, or if it must tie into another system, the new equipment must be compatible with that already installed.

Before selecting the equipment, a number of specific technical questions must be considered. For example, should points A and B be connected by microwave or by cable? What is the distance involved, and what terrain problems are likely to be encountered? Would it be better to route the communications between X and Z via point Y? What about the accessibility of repeater sites, and which route will provide the necessary path reliability at the lowest cost?

For example, suppose two outlying points, Y and Z are to be connected to a central point X, as shown in Figure 2. Twelve voice channels are required between X and Y, while 48 channels must be provided between X and Z. No direct communication is necessary between Y and Z. After careful analysis it may be decided to run 60 channels on microwave radio from X to Y, drop 12 of them there, and run the remaining 48 channels to Z, since it is a simple matter to bridge off a 12-channel group without demodulating the entire 60 channels. Although the total length of the system connecting X and Z is somewhat greater this way, less equipment is required.

If a microwave system is considered to be the best way to satisfy the requirements, then a radio path survey must be accomplished. Profile maps, available for most areas, are helpful in performing such surveys. In many cases,

World Radio History

however, a transmission engineer must go out and make a direct path survey of all the proposed routes. He may also have to consult meteorological records to determine what effect the climate of the area will have on path reliability.

On the basis of the path surveys, it might be decided to use frequency-diversity operation to achieve the desired path reliability. If a fading problem exists, but dual-frequency operation cannot be authorized for this particular type of service, space-diversity operation (which provides two paths using the same frequency but with different fading characteristics) may be the answer.

Antenna sizes and heights must also be established as well as a system frequency plan which will make efficient use of the spectrum available while minimizing interference with other systems or other *hops* of the same system.

Suppose the system planners are faced with the problem of providing a microwave system for a route with growing requirements for traffic capacity. Initially, one broadband microwave channel is enough, but the need for a second is foreseen—and eventually even a third channel may be required. Furthermore, the necessary system reliability demands that each channel be protected with an alternate path.

For the initial installation, two channels operating in a frequency diversity arrangement will satisfy the traffic-handling requirements and provide the necessary reliability. But what of the future? One possible solution might be a one-for-three protection system, as illustrated in Figure 4. Such a system permits a single channel (channel B) to provide *back-up* protection for three working channels. Under normal conditions, however, this protection channel does not stand idle. Instead, it operates in a frequency-diversity arrangement with one of the working channels (channel A), as shown, to make the most effective use of the frequency spectrum when carrying message traffic.

It is not necessary initially to install all four channels. A common procedure is to install first only channels A and B. Later, as more channels are needed, channels C and D can be added (one

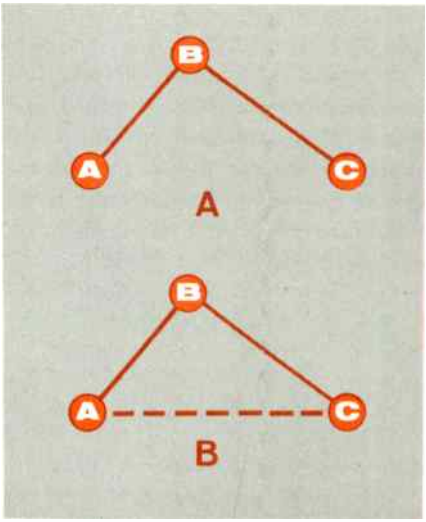


Figure 5. Diagram A shows trunk facilities interconnecting three stations without alternate routing. Diagram B shows trunk facilities interconnecting the same three terminals, but with a direct channel established between stations A and C. The additional channel between stations A and C provides an alternate route between each station in the system.

World Radio History

at a time or together). When this is done, channel B continues to operate in frequency diversity with channel A. However, if either channel C or channel D fails (because of either propagation or equipment problems), the baseband switching equipment disconnects channel B from the frequency- diversity arrangement and connects the traffic on the failed channel to channel B. Thus, a little foresight has provided a system with built-in expansion capability.

Another important consideration is the routing of traffic between the various terminal sites of the proposed system. For instance, consider a proposal for a communications system that interconnects three stations, A, B, and C, as shown in Figure 5A. In this system, traffic between stations A and C is routed through station B. If an outage occurs between B and C, there can be no communications to C from either A or B. The same problem arises if an outage occurs between A and B. However, if a direct link is installed between sites A and C, as shown in Figure 5B, an alternate route for traffic is immediately established between all stations. The effective use of alternate routing in a system containing more than two terminal sites, therefore, can result in a more reliable system, in addition to providing better service.

Planning a complex multichannel transmission system involves examining various equipment arrangements in order to achieve the best results. The system planners put together various combinations of "building blocks" and then analyze the results to see if they meet the specifications. They do this because their job is not merely to develop an adequate system, but to conceive the best possible system *at the lowest cost.*

Thus, practical economics play a large part in systems planning—where there is no substitute for experience. There is certainly no "guidebook" which details all shortcuts which can save money without degrading performance.

Conclusion

Developing a multichannel transmission system for a communications network is not a simple and easy task. There are many things to consider and many decisions to make before such a system can even begin to take form.

Today's transmission system must be capable of handling many types of message signals, each having a different effect upon the performance criteria of the system. Early systems handled only telephone and telegraph signals, while most modern systems must handle not only these, but possibly low-speed data, high-speed data, telephoto, and television signals as well.

The ever-increasing demand for these new communications services has brought about the development of new engineering techniques and equipment with extraordinary capabilities. However, new equipment and techniques alone do not make a reliable and efficient communications transmission system. Effective planning and a good experienced engineering team are surely just as important.

World Radio History

World Radio History

the Lenkurt Demodulator

VOL. 13, NO. 9 SEPTEMBER, 1964

Basic Concepts of

ENGINEERING RELIABILITY

There are many basic factors which must be considered in the design and development of communications and electronics equipment. Among them is the subject of reliability which has, in recent years, acquired a very distinct meaning. Indeed, reliability has grown into a full-fledged engineering discipline, complete with mathematics and its own special jargon. Its aim is to assure the success of a product through a scientific program of performance evaluation, statistical analysis and prediction. This article discusses some of the fundamental aspects of reliability, including such related subjects as quality control and human engineering.

The demand for high quality and reliable products has always been an important consideration in the development of communications systems. To meet this demand, most commercial manufacturers have, over the years, developed very stringent engineering standards and quality control procedures to assure reliable products. Shortly after 1950, however, the subject of reliability began to receive separate attention, especially in the aerospace industry. Since that time, the word *reliability* has acquired a very specialized meaning in respect to the quality of manufactured products.

The rapid development of highly sophisticated missile systems and manned space vehicles created some special problems for the design engineer. The failure of one essential electronic component in a manned space vehicle, for example, could result in a catastrophic failure of its mission and the loss of life and millions of dollars. Consequently, an unusually high degree

World Radio History

of reliability had to be achieved, generally within a very short development period. The need arose, therefore, for a means of *measuring* the reliability achieved in the *design* of these vital aerospace systems and *predicting* the mathematical odds of their success.

It became evident, with the growing use of computers and data processing equipment, that such a need could be partly fulfilled through statistical analysis. Emphasis was placed on compiling data relating to the causes of electronic component and system failures. Such data can be used to determine the *mean life* of components, to reveal the most prevalent causes and modes of failure, and to expose substandard parts and circuits.

This new technology, therefore, has added a statistical approach to the time-proven methods of achieving reliability, and has also given rise to a highly useful reliability rating system.

Evaluation and Prediction

What is considered satisfactory reliability? The reliability of a product is measured in relation to the mission that it is designed to accomplish. It would be ideal, of course, to accomplish this mission 100 percent of the time. Unfortunately, from a practical standpoint, the ideal is rarely possible to achieve. This can be attributed to many factors, such as design errors, material deficiencies, or cost limitations. In any event, the most important reason for considering reliability is to assure, with a *measurable* degree of confidence, that a product can accomplish its mission. Therefore, it is absolutely necessary to describe the mission clearly so that there is no doubt as to what must be achieved in the design of a product. Such a description must also include the tolerances which are to be allowed before the mission is considered a failure. When this is done, the design engineer can specify the degree of reliability in terms of the operational conditions involved.

It is significant to note the difference between *evaluating* the reliability of equipment and systems that have already been developed, and *predicting* the inherent reliability of a proposed new design. Evaluating reliability involves measuring the past performance of a product or component to determine what degree of reliability has been achieved. This is accomplished by subjecting the product to a variety of tests and by acquiring accurate reports of failures occurring during actual field use. Such information is of considerable value in evaluating the product's performance under typical operating conditions. The ultimate reason for accumulating failure reports from the field, of course, is to effect product improvement. This is usually done by analyzing the failure reports to determine the nature of the failures, and then taking steps to prevent them from occurring in the future. It is important, therefore, that these field reports be accurate so that a high degree of confidence may be placed on any conclusions derived from them.

Reliability prediction, on the other hand, involves the extrapolation and interpolation of statistical data to estimate the *inherent* reliability of a product *design,* before the product is approved for manufacture. This is done by carefully examining all pertinent engineering data and documentation, espe-

World Radio History

Figure 1. Quality assurance engineers perform complete system performance test on Type 46A carrier equipment, before shipment to customer.

cially the reliability ratings of all recommended components and parts, and then calculating the overall reliability of the proposed design using the mathematics of probability. By using such statistical techniques, designers are able to disclose design deficiencies or potential problems such as marginal circuits and misapplication of parts.

Measuring Reliability

How is reliability measured? Presently, there are a number of ways to measure reliability, many involving complex

World Radio History

statistical analyses which are beyond the scope of this article. If not properly understood, however, mathematical expressions that purport to measure reliability can be easily misleading. It is helpful, therefore, to understand some of the more prevalent mathematical expressions linked with reliability.

The three most popular expressions of reliability are the *probability function,* the *failure rate,* and the *mean-time-between-failure* or *MTBF.* (For products that are not repairable, the latter expression is referred to as the *mean-time-to-failure).* Each of these expressions can be applied to a part, component, assembly, or to an entire system, depending on the particular needs. The probability function is expressed as a decimal or a percentage and is an estimate of what the chances are that a particular device will perform its mission. The failure rate is ordinarily expressed in terms of the number of failures per unit of time, usually 1 hour, 100 hours, or 1000 hours, or as a percentage of failures per 1000 hours. The MTBF is expressed in hours and

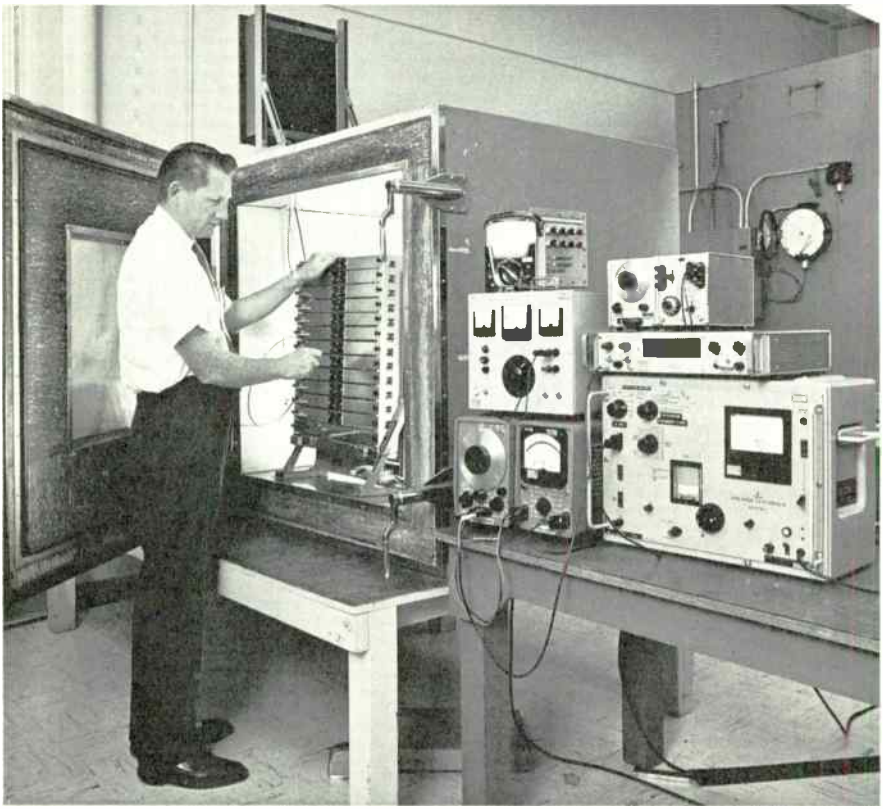


Figure 2. Type LN2 cable carrier equipment undergoes extensive temperature testing to assure reliable operation under a variety of environmental conditions.

World Radio History

is the ratio of the total test time (or operating time) of a device to the total number of failures that occur during the test period.

The probability function P can be expressed mathematically as:

$$P = \frac{a}{a + b}$$

where

a = *number of successes*
b = *number of failures*

To illustrate how this expression is applied, consider the following example. If 100 components were tested for 1000 hours and there were no failures during the test period, the probability function would be 1.0 or 100 percent.

$$P = \frac{100}{100 + 0} = 1.0$$

If, however, 10 components failed during the test, the probability function would be 0.9 or 90 percent.

$$P = \frac{90}{90 + 10} = 0.9$$

Thus, stating that a product is 90 percent reliable does not mean that it will probably operate only 90 percent of the time, but that there is a 90 percent chance that it will successfully complete its mission. It is important to note that the probability function must be *qualified* in order to be meaningful. Expressing reliability in terms of an abstract number is meaningless unless the physical conditions that prevailed when the reliability was assessed are included. It is also important to know the size of the sampling used to determine the probability function. In the previous example, it can be seen that 10 failures represented a 10 percent decrease in reliability. If the 100 components were taken from a production run of 5000, then the sampling may not be large enough to accurately predict the performance of the entire run.

The second expression is the failure rate f, which can be expressed mathematically as:

$$f = \frac{a}{b}$$

Where

a = *number of failures*
b = *duration of test, in hours*

As an example, if 100 components are tested for 1000 hours, and ten of them fail during the test, then the failure rate is:

$$f = \frac{10}{1000} = 0.01 \text{ per hour}$$

When calculating the failure rate, it is important to consider the age of the product. Failure rates of new electronic products are apt to be high because of such factors as production errors, defective parts, faulty installation, and improper alignment. After a normal *break-in* period, however, failures become less frequent and failure rates tend to remain relatively constant during the useful life of the equipment. When the product begins to wear out, the failure rate may begin to increase steadily. A typical curve of electronic equipment failure rate versus age is shown in Figure 3.

Closely associated with the failure rate is the mean-time-between-failure (MTBF). This expression is merely the average time between failures and is the reciprocal of the failure rate. Using the previous example, the MTBF would be expressed as:

$$MTBF = \frac{1}{0.01} = 100 \text{ hours}$$

World Radio History

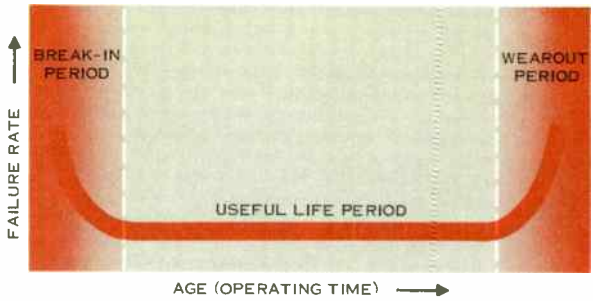


Figure 3. Curve showing typical failure pattern of electronic equipment.

Therefore, the larger the value of MTBF the greater the reliability and, inversely, the smaller the value of the failure rate, the greater the reliability.

Users of communications equipment are more concerned generally with system or equipment reliability. However, the reliability of parts, components, and circuit design provide the basis for measuring the overall reliability of communications equipment or systems. Perhaps the most important factor affecting overall reliability is the increasing number of components required in single systems. Since most system failures are actually caused by the failure of a single component, the reliability of such components must be considerably better than the required overall system reliability. This fact becomes quite evident when considering how the overall system reliability is measured.

If all the components of a system are considered to be functionally in series, and if the failure of any component results in a system failure, then the overall system reliability R is:

$$R = r^n$$

Where

r = *mean reliability (probability function) of each component*

n = *number of components in series*

The formula for calculating the overall system reliability produces some rather interesting results as seen in the following table.

n	r	R
10	*0.99*	*0.90*
100	*0.99*	*0.40*
200	*0.99*	*0.19*
500	*0.999*	*0.60*
1000	*0.999*	*0.37*

One means of improving reliability when designing a product is through simple redundancy — that is by providing an alternate means of accomplishing a given function. The probability function for redundant electronic circuits, arranged in parallel, is expressed as:

$$R = r_1 + r_2 - r_1 \times r_2$$

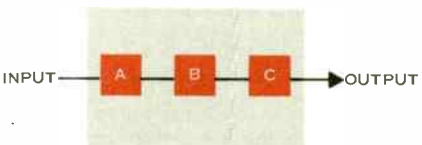
where

R = *Overall reliability*

r_1 = *Reliability of circuit 1*

r_2 = *Reliability of circuit 2*

As an example of how redundancy works, consider the following circuit containing three components, A, B, and C, each connected in series.

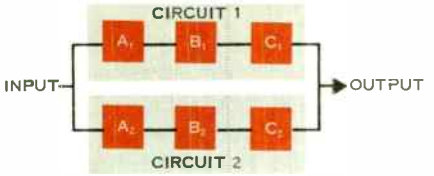


World Radio History

If the reliability of each component is 0.95, then the overall reliability for the series circuit is:

$$R = 0.95 \times 0.95 \times 0.95$$
$$R = 0.86$$

When a redundant circuit is added in parallel, as shown in the following diagram, the overall reliability increases.



Using the formula for computing the probability function of a parallel (redundant) circuit, the overall reliability becomes:

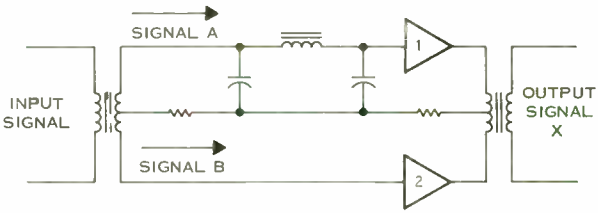
$$R = 0.86 + 0.86 - 0.86 \times 0.86$$
$$R = 0.98$$

Thus, there was a 14 percent gain in the overall reliability as a result of adding the redundant circuit. However, redundancy often requires the use of additional components, such as a switching circuit, which may lower the overall system reliability.

Testing and Quality Control

Two very important practices which affect the reliability of a product, after it is designed, are testing and quality control. The major function of quality control is to inspect the workmanship of a product to determine if it meets the level of quality proposed in the design. A good quality-control effort can detect design and manufacturing errors which otherwise might have shortened the life of a product and adversely affected its reliability. It is far better to correct errors during the manufacturing process than to make corrections later on the basis of failures that occur during actual field operation.

Electronic components must be capable of operating satisfactorily, not only under the conditions required of the equipment or system of which they are a part, but more importantly, in the *electrical environment* in which they



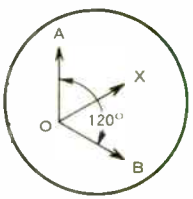


Figure 4. An example of redundancy without a switching circuit is exhibited by this amplifier circuit used in the AN/FCC-17 multiplexer set. In this circuit, there is a 120° phase difference between the output signals of the two amplifiers. Thus, the combined output level is equal to the output level of each amplifier. If one amplifier should fail, the final output level will remain constant because of the 120° phase difference. (In the vector diagram, line OX will shift toward the vector point, A or B, representing the output level of the amplifier with the higher gain.)

World Radio History

operate. To determine if any deleterious effects may occur during sustained operation of an electronic device, it must be subjected to many tests and field trials before being declared operationally suitable. The main purpose of such tests is to determine whether or not the device meets the acceptance criteria, and to provide concrete evidence of its performance capability. These tests range from environmental and temperature tests of individual components to burn-in tests and field trials of entire systems. In addition to testing the usual electrical characteristics of an electronic device, it should be subjected to a variety of physical or environmental tests. These may include such things as temperature tests, vibration tests, corrosion or salt spray tests, fungus-resistance tests, sand and dust tests, and shock tests. Failure data gathered from these tests are used to analyze component reliability as well as the overall reliability of the product or system.

Human-Factors Engineering

Any reliability effort would not be complete without considering the people who must operate and maintain the equipment. Because the performance of a communications system is determined by the human operator as well as by equipment reliability, it will certainly be improved if the mechanical component is designed to fit the human component. Such a design must consider the capabilities and limitations of the human operator and, where possible, relieve him of purely mechanical tasks. The art or science that deals with such problems is known as *human-factors engineering*.

A good example of human-factors engineering occurred in the development of the modern telephone. In earlier telephones, the letters and numbers were placed inside the dialing holes. This had the disadvantage of covering up the letter or number as the user placed his finger in the hole. This seemingly unimportant factor not only annoyed the user, but also contributed to dialing errors. To overcome this, the letters and numbers were placed outside the dialing holes. This appeared to be a fine idea, but it resulted in an increased number of dialing errors. A human-factors investigation showed that the letters and numbers, while inside the holes, had provided a natural *target* by which the user could aim his finger. By removing the letters and numbers from inside the holes, the so-called target had also been removed. By putting a mark inside each hole, however, the target was replaced, and dialing errors decreased.

Some Practical Considerations

The fact that aerospace companies have been more aggressive in developing formal reliability programs is not difficult to understand. Most aerospace contracts require that the contractor develop and build exotic and sophisticated missiles or space vehicles that have little or no precedent in their design and mission. The advanced electronic and communications equipment that supports these rather complex systems is equally as unprecedented. In addition, there is seldom enough time for these systems to mature, as they are being continually modified to take advantage of new techniques — and are soon re-

World Radio History

placed to make way for superior or more advanced systems. The relatively short life and the vital mission of these aerospace systems, therefore, does not permit reliability to be increased through a routine course of product improvement.

Under such irregular and accelerated conditions, it has become absolutely essential to be able to measure the reliability of a product and to predict its probable success before placing it into an operational environment. Thus, in the aerospace industry, reliability is recognized as a design parameter.

Unlike aerospace or military equipment, commercial communications products do not undergo radical changes and are seldom replaced simply because a newer design is on the market. This is especially true in the telephone business where equipment is generally retained as long as it performs its job. Usually, new equipment introduced into the communications industry must be compatible with existing equipment to the extent that a complete departure from earlier designs seldom occurs. This allows commercial manufacturers, who have done business with the communications industry for many years, to develop mature products and to acquire a skilled and experienced organization in which to produce reliable systems.

Most manufacturers of commercial communications equipment take special care in selecting parts and designing circuits that provide optimum reliability

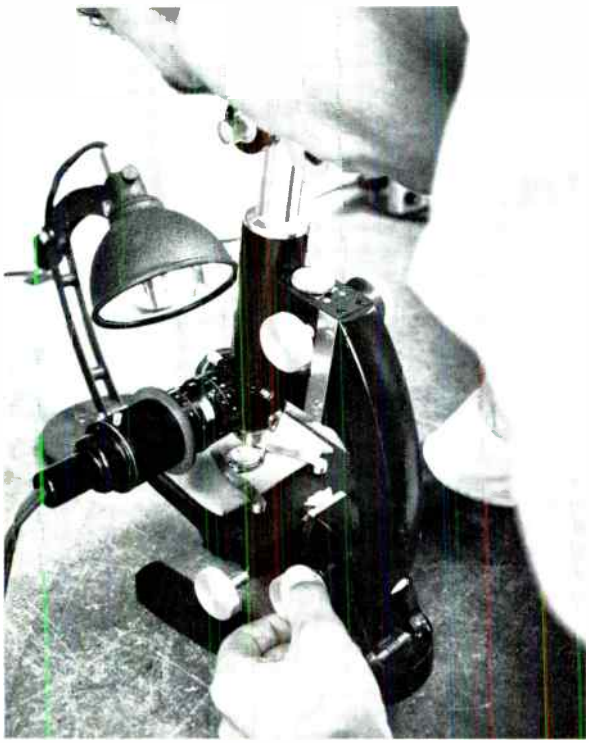


Figure 5. Transistors that fail during operation are examined and tested in Lenkurt's Material Evaluation Laboratory. If electrical tests do not reveal the cause and mode of failure, the transistor is cut open, potted, and then slowly abraded and polished until the physical defect can be seen under a microscope. Cause of failure can usually be determined by viewing the damaged area of the transistor.

World Radio History

for their equipment. Usually, this equipment is expected to operate for a useful life period of at least twenty years. Unfortunately, parts of a given type made by different manufacturers may have the same initial characteristics, but may change in different directions and degree with time and environment; usually as a result of different manufacturing processes. Because of this, only high quality parts should be selected, and used so that stress levels are safely below the manufacturer's recommended rating.

Lenkurt manufactures many of its own parts, such as quartz crystals, toroid coils, transformers, and capacitors, to assure a high degree of product and system reliability. In addition, all products are guaranteed for one year against faulty components and workmanship, which more than covers the *break-in* period shown in Figure 3. Records are maintained of all equipment and components that are returned to the factory for replacement during this warranty period. If any particular type exhibits an excessive number of failures, it then becomes the subject of a special investigation. The purpose of such an investigation is, of course, to determine the causes of the excessive failures so that steps can be taken to increase the overall reliability of the product. Accurate records are also kept of all costs incurred in support of this warranty. Such costs provide a good yardstick for measuring the quality achieved in the design and manufacture of each product.

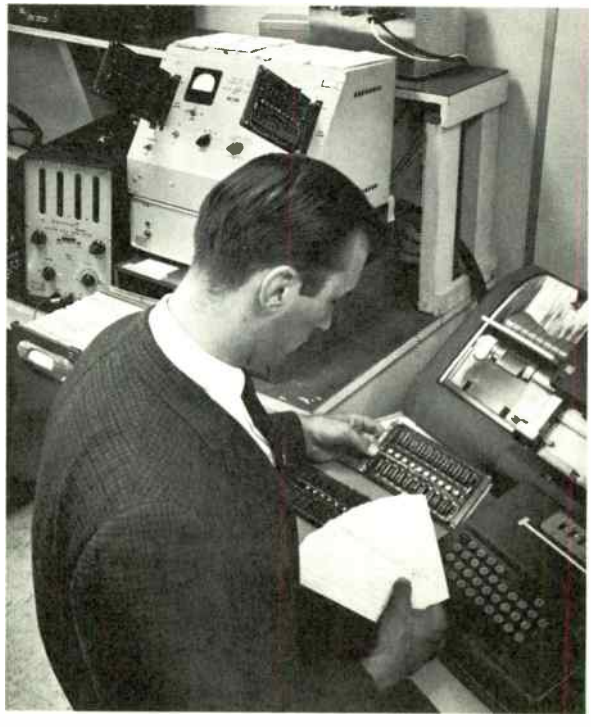


Figure 6. Transistors used in Lenkurt products are life-tested to determine their quality and operating characteristics. At certain intervals the essential parameters of each transistor are measured and recorded. Photograph shows 20 transistors being checked automatically. Analog measurements are fed into analog-to-digital converter and then into a keypunch machine which punches the data into a card for future processing. Punched cards are used in computer programs to predict transistor failure rates or MTBF, and end-of-life.

World Radio History

Conclusions

Historical data and extrapolation form the basis for predicting the reliability of a product. The concept of predicting reliability by purely statistical means is new to engineering, and it has yet to gain the respect and understanding enjoyed by certain older engineering concepts. Many view it suspiciously as an abstract numbers game that is of value and interest only to the statistician. This lack of respect may be caused, in part, by its misuse in applications ill-suited to promote its real value. Unfortunately, until such a new discipline reaches maturity, its concepts, theories, and applied methods are apt to be disorderly and easily misunderstood. Consequently, a large amount of money spent for reliability programs has undoubtedly been wasted — because users have misinterpreted the statistical data, or have compiled such data without an effective plan for its use.

Most of the present statistical concepts of reliability have been developed and put into formal practice by the aerospace industry and related government agencies. These practices have proven to be effective in guiding the development and assuring the reliability of highly sophisticated systems, especially where precedent and other engineering guidelines are lacking. In commercial practice, however, it has not yet become necessary or economical to employ elaborate statistical methods to achieve reliability This is due to the regular manner in which the needs of commercial industry evolve. Manufacturers of commercial communications equipment, for example, are able to concentrate on a particular line of long-life products and develop highly dependable equipment through tradition and experience.

The government has had to pay a high price to advance this new concept of reliability to assure the success of vital missile systems and manned space vehicles. Such costs, of course, are a real bargain if these advanced systems successfully perform their mission, especially where human lives are at stake. In time, the statistical approach to achieving reliability may also prove to be an effective tool in advancing the technical excellence of commercial products. •

BIBLIOGRAPHY

1. C. E. Leake, *A Simplified Presentation for Understanding Reliability,* United Testing Laboratories, Pasadena Lithographers, Inc., California, 1960.
2. D. N. Chorafas, *Statistical Processes and Reliability Engineering,* D. Van Nostrand Co., Inc., New York, 1960.
3. C. M. Ryerson, "The Reliability and Quality Control Field From Its Inception to the Present," *Proceedings of the IRE;* May, 1962.
4. S. R. Calabro, *Reliability Principles and Practices,* McGraw-Hill Book Co., Inc., New York, 1962.
5. Vol. R-13, Number 1, *IEEE-Transactions of Reliability,* Professional Technical Group on Reliability, March, 1964.
6. R. T. Haviland, *Engineering Reliability and Long Life Design,* D. Van Nostrand Co., Inc., New York, 1964.

World Radio History

World Radio History

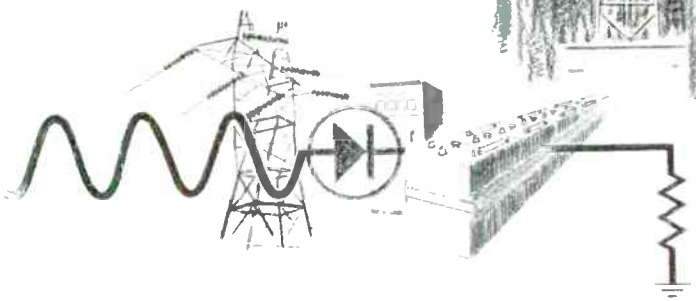
the Lenkurt Demodulator

VOL. 14, NO. 2 FEBRUARY, 1965

DC POWER PLANTS

For Communications Systems

Electric power is basic to the operation of modern communications systems. So basic, in fact, that its production and distribution undoubtedly receives far less attention than the more interesting and sophisticated communications systems that it serves. Today's modern automatic power plant performs its vital though undramatic role in such an efficient and humdrum manner that it is easily taken for granted — except by those directly concerned with its operation. However, there are many technical and economic factors associated with supplying electric power to communications systems that make the subject important. This article describes a typical communications system power plant, including the functions of such devices as power rectifiers, generators, and storage batteries.



World Radio History

A source of reliable and continuous electric power is required to operate the many communications systems which serve our ever-growing and dynamic society. The vital function performed by these systems cannot tolerate even a momentary interruption in the supply of electric power. Such an interruption in a large telephone plant, for example, would cause important circuit elements, such as relays, to de-energize, affecting perhaps thousands of important messages and long-distance circuits.

The main source of electric power for most communications systems is the commercial or public utility company serving the area in which the system is located. In the United States, electric power distributed by these companies is alternating current at 60 cycles per second, usually 230 volts three-phase and 115 volts single-phase.

A communications system requires dc power to operate most of its components such as electron tubes, transistors, telephones, and switching apparatus. Therefore, commercial or public ac power must be rectified to various dc voltages before it can be used. Certain types of communications equipment, such as radio and multiplex terminals and repeaters, employ built-in or optional power supplies which convert 60 cps power to the dc voltages required by the various circuit elements. To obtain power in such cases, it is necessary only to insert the respective power plug into an ordinary ac outlet.

In most large communications systems, however, it is not always economical or practical to use separate power supplies for each piece of equipment. Typically, a central power plant located at the central switching office is used to supply all the electric power required for the particular system.

Although the electric power supplied by most public and commercial utility companies is usually very reliable, it is subject to interruptions. Transmission lines may be damaged by storms, lightning may strike transformers, and switches and insulators may deteriorate and become faulty. It is necessary, therefore, to have an auxiliary source of electric power, ready to assume the load in the event of a primary power failure.

Emergency power is normally provided by prime mover generators, storage batteries, or both. Since an engine-driven generator requires time to start and warmup, there is unavoidable delay before it can assume the load after a failure occurs. Since even a momentary delay cannot be tolerated, some means of providing power instantly must be available. The usual practice is to use storage batteries, keeping them fully charged from the primary power source during normal operation. When a primary power failure occurs, the batteries assume the load instantly with no interruption in service.

Unfortunately, batteries do not have the capacity to supply power for long periods of time. For this reason, they are used only to assume the power load at the instant of a primary power failure, and for a short period afterwards. Because of the vital function of communications systems, they must also be protected against the possibility of long interruptions in the supply of primary power. For full protection, therefore, it is necessary to have available some type of prime mover generator in addition to the storage batteries.

Typical DC Power Plant

The ordinary dc voltages used in most communications systems today were determined by the original needs of the telephone industry. Early manual exchanges used dc power at 24 volts to operate the various telephone apparatus, while telegraph equipment required dc

World Radio History

power at 130 volts. Dial telephone exchanges were designed to operate with dc power at 48 volts, thus establishing three dc voltages which have become standard throughout the communications industry.

Since new equipment introduced into the communications industry has to be compatible with existing equipment, it must be designed to operate from the standard dc voltages. Electron tubes employed in most communications equipment, therefore, use +130 volts as the plate voltage, while transistorized equipment is designed to use 24 or 48 volts.

A typical dc power plant for a communications system uses commercial or public ac power and converts it to the standard dc voltages required by the various equipment. As previously mentioned, it must also be capable of continuous operation in the event of a primary power failure. To accomplish this, such a power plant is ordinarily equipped with an engine-driven generator and storage batteries for use during emergencies.

Power plants for large telephone communications systems, for example, usually have a separate power system for each of the required major voltages. Figure 2 illustrates the arrangement of a typical dc power plant containing a 24-volt system, a 48-volt system, and a 130-volt system.

As shown in the diagram, 60-cps power from the commercial or public utility company is fed into the power plant through a meter which measures and registers the amount of power used. A main entrance switch is normally provided to turn the power on or off. From the switch the power is fed to a distribution panel where it is divided and routed to the separate power systems.

Each of the three power systems contains a power rectifier, a power board, and a storage battery. The rectifier performs essentially two functions. Its main function is, of course, to convert the primary ac power to dc power at one of the standard voltages. In addition, the rectifier supplies energy to the battery so that it remains in a fully-charged condition during normal operation. This action is called *floating*, and is accomplished by connecting the battery in parallel with the output circuit of the rectifier.

Since the rectifier is also used to charge the battery, it is often called a *battery charger*. However, the name is slightly misleading in this application as it identifies only a secondary although important function. This misnomer also tends to create the erroneous impression that the main task of the rectifier is to charge the battery which, in turn, supplies the dc power for the communications equipment. This is not exactly true. During normal operation, substantially all of the power for the communications equipment is supplied from the commercial or public utility company through the rectifier. Electric energy is not obtained from a storage battery unless it is discharging; therefore, it does not supply power unless the rectifier power drops below a certain level or after a primary power failure. The storage battery does help to offset voltage variations in the public or commercial power and to filter out the ripples and line noises from the rectifier output, however, and so does perform an active function during normal operating periods.

When the battery does assume the load during emergencies, it begins to slowly discharge, causing the voltage in each cell to drop. To offset the effect of this dropping voltage, *emergency* or *end* cells are normally employed in a typical battery system. These end cells are arranged so that they can be

World Radio History

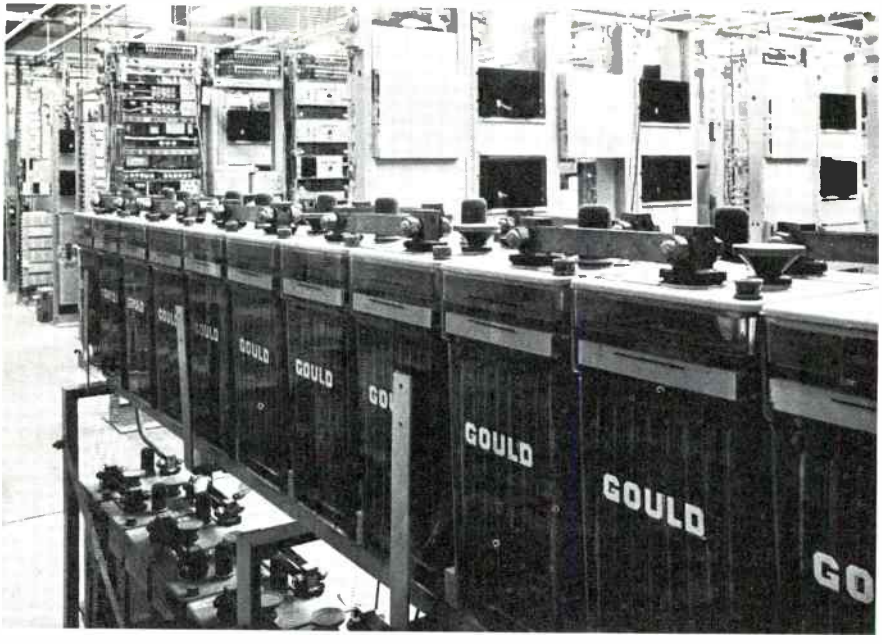


Figure 1. When the primary source of power for a communications system is interrupted, emergency dc power may be supplied by large battery cells, such as those shown in photograph.

switched into the regular battery circuit, one cell at a time, when needed to raise the battery voltage to a proper level.

The dc output from each rectifier is fed into a *power board* which functions as the nerve center for the particular power system. The power board contains various switches, meters, fuses, circuit breakers, and other devices needed to control and monitor the operation of the power system. For example, if the primary power drops below a specified safe level or fails, sensing devices in the power board automatically switch in the emergency battery system and, if necessary, start up the auxiliary generator and switch it into service when needed. The power board also switches in the end cells, one at a time, when the emergency battery voltage drops below a prescribed level. After regular service is resumed, the power board restores all circuits to normal and provides extra power to recharge the battery as quickly as possible.

In addition, the power board controls the amount of float charge necessary to keep the battery in a fully-charged condition. The power board may also include various audible and visible alarm devices which alert operating personnel of any abnormal conditions or failure in the power plant.

The main power output from the power board is fed through protective fuses to a system of large, usually copper or aluminum, busbars which are normally located overhead in the power plant room. The power input circuits of the various communications equipment

World Radio History

are connected, through cables, to the necessary busbars.

Emergency Generators

The emergency motor-driven generator remains idle during normal operation of the power plant and also during short interruptions in the primary power supply when the batteries assume the load. However, when the duration of a primary power failure extends beyond the time limits of the emergency battery system, the auxiliary generator is put into use.

In addition to supplying emergency power for the communications system, the generator also must be capable of supplying power to lights and other electric appliances such as air conditioning or heating equipment located at the power plant.

The size and capacity of these engine-driven generators depends, of course, on the emergency power requirements of the particular system. Many types of generators are used, ranging from small portable gasoline engines, to large stationary diesel engines permanently installed in the power plant.

Diesel engines are usually more reliable, durable and economical to operate than gasoline engines and are generally preferred, especially where the power requirements are relatively large. The speed at which these engines operate determines the characteristics of the output power of the generator and must be closely controlled. Such control is normally accomplished by a governor which regulates the amount of fuel supplied to the engine. Usually, emergency generators are equipped with automatic starting and stopping switches that are actuated by circuits in the emergency detection system of the power plant.

Gasoline engines have certain advantages over diesel engines, such as lighter weight, easier starting, availability of fuel, and lower initial cost. However, they are not as reliable or efficient as diesel engines and gasoline is more dangerous to handle and store than diesel fuel. For these reasons, gasoline engines are not ordinarily used except where weight and size are critical and the power requirements are relatively small.

Power Rectifiers

Rectifiers are devices that convert alternating current to direct current and, as already explained, play an important role in the typical communications system power plant. There are essentially three classes of rectifiers in general use today: gas tube rectifiers, metallic rectifiers, and solid-state rectifiers.

Gas tubes are limited to power requirements of less than about 50 amperes. The most common type of gas tube is the *Tungar* which consists of a glass bulb filled with argon gas, a tungsten filament, and a single carbon plate or anode. The operation of these tubes is essentially the same as for all other electron tubes.

Metallic rectifiers are widely used in telephone power plants, and are of two general types—copper oxide and selenium disk. The copper-oxide rectifier consists of a copper disk covered with a layer of copper oxide. This combination offers a low resistance to current flowing from the copper oxide to the copper, and a high resistance to current flowing from the copper to the copper oxide.

The selenium-disk rectifier consists of a steel or aluminum plate (electrode) coated with a thin layer of metallic selenium. This plate is in contact with a second plate (counter electrode) of conducting metal. Current flows with little resistance from the second plate to the plate coated with metallic selenium, while a high resistance to current flow exists in the opposite direction. A complete selenium-disk rectifier consists of

World Radio History

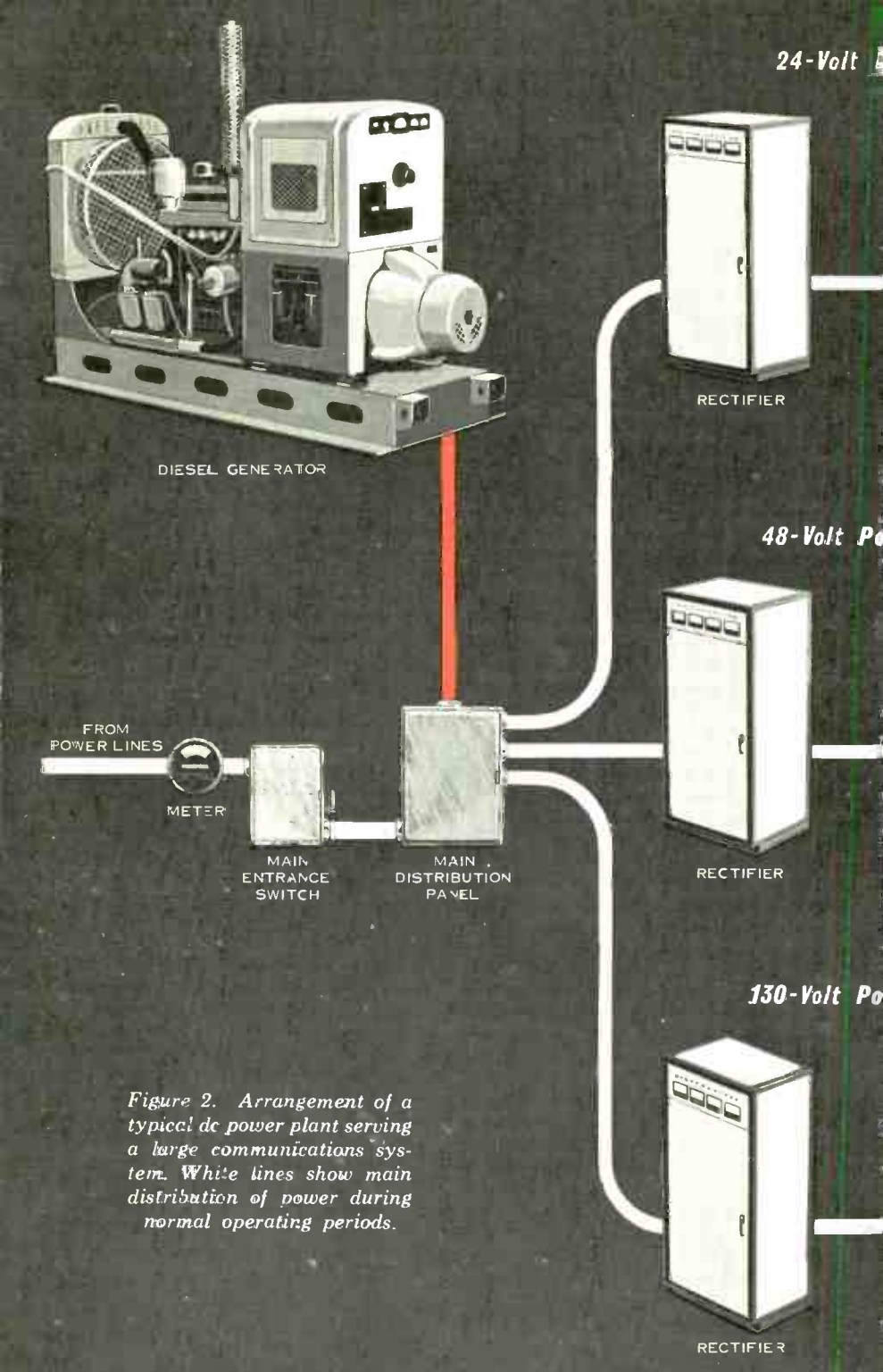
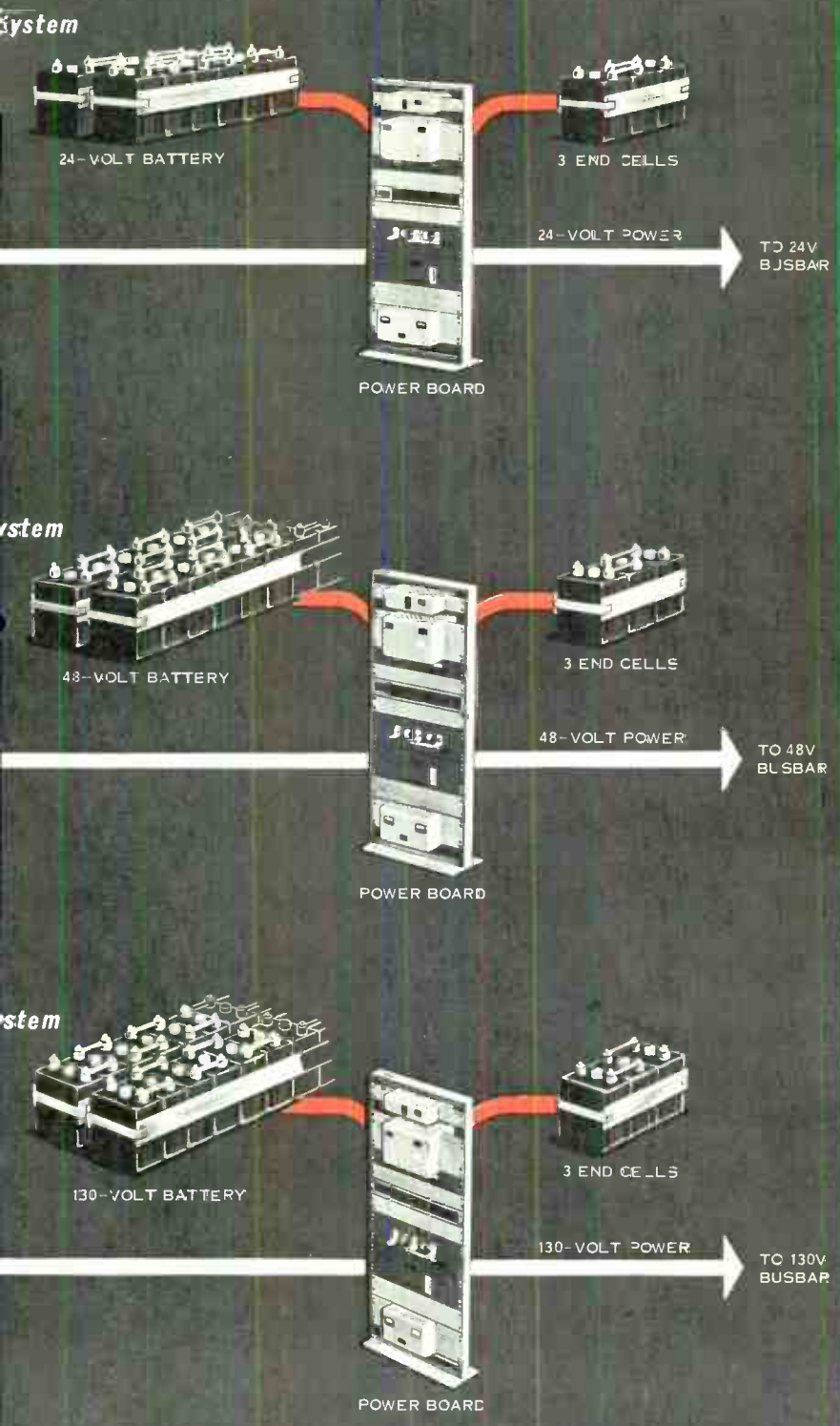


Figure 2. Arrangement of a typical dc power plant serving a large communications system. White lines show main distribution of power during normal operating periods.

World Radio History



World Radio History

a number of plates fastened together to form a *stack*.

Solid-state rectifiers are superior to the gas tube and metallic rectifiers and are replacing them in all modern power plants. They are more efficient, have a longer life, and are much more economical than the older types of rectifiers. Solid-state rectifiers are usually made of germanium or silicon diodes.

The most promising solid-state rectifier in use today is the silicon-controlled rectifier (SCR). Its PNPN semiconductor rectifying element consists of a tiny wafer of silicon with a small amount of impurity diffused into a thin layer of its surface. The junction between this *doped* layer and the pure silicon forms the rectifying barrier.

Batteries

A battery is an assembly of *cells* which convert chemical energy into electric energy. Each cell consists of a positive electrode and a negative electrode submerged in a liquid or paste-like electrolyte. There are two general classes of batteries: primary batteries and secondary batteries.

The chemical action that generates electric energy in a primary battery cannot be reversed and, therefore, once the chemical energy is expended the battery can no longer be used. For this reason, they are often referred to as *one-shot* batteries. An example of a primary battery is the ordinary single-cell flashlight battery which is discarded after use. (A single cell is also referred to as a battery.) Such batteries are seldom used in large or fixed communications plants, but are widely used in portable communications equipment.

A secondary battery is different from a primary battery in that the chemical process can be reversed by passing an electric current, from an external source, through each cell. This reverse action restores the chemical energy to the battery allowing it to be reused.

There are two classes of secondary batteries: those with *acid* electrolyte

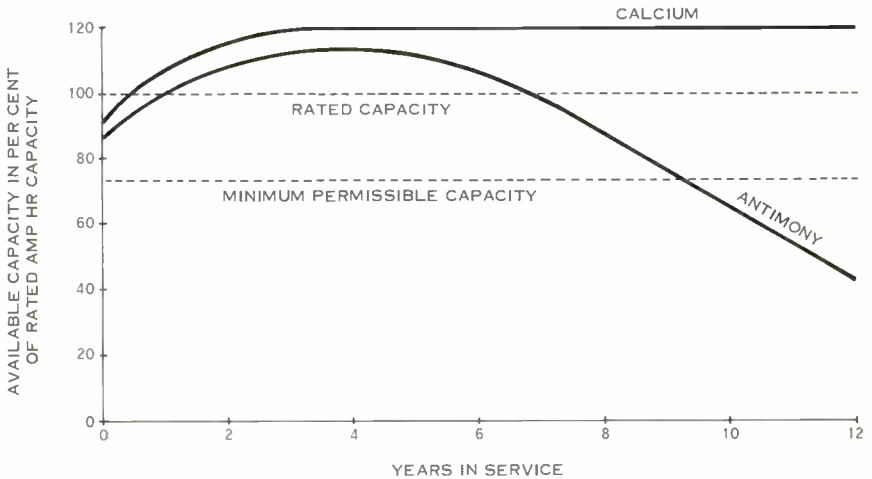


Figure 3. Life performance of lead-calcium and lead-antimony cells. (Courtesy The Warren Manufacturing Co.)

World Radio History

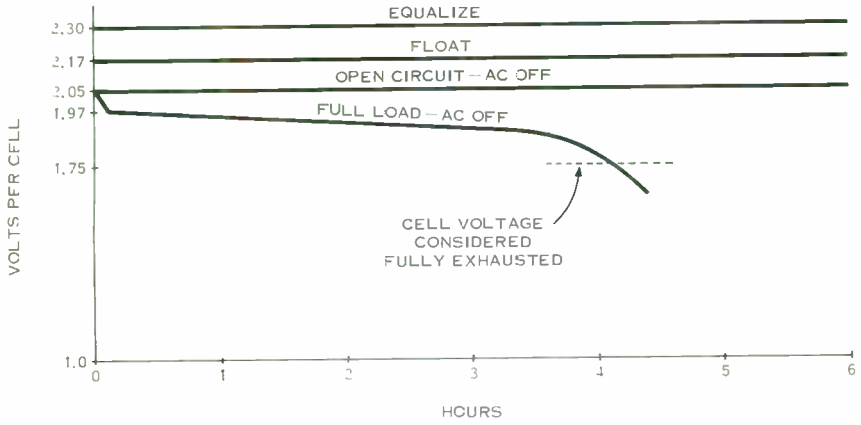


Figure 4. Voltage of lead-acid cells under various operating conditions. (Courtesy The Warren Manufacturing Co.)

(lead-acid batteries) and those with *alkaline* electrolyte (silver-zinc, nickel-iron, nickel-cadmium, and silver-cadmium batteries). The most prominent type of battery found in the communications industry is the lead-acid type which has a long service life, relatively high voltage per cell, and generally costs less than the other types of secondary batteries.

A fully-charged lead-acid cell has a positive electrode (or plate) made of lead peroxide (PbO_2) and a negative electrode made of spongy lead (Pb), both submerged in an electrolyte of dilute sulfuric acid ($H_2SO_4 + H_2O$).

When a lead-acid cell (or battery) is discharging, current passes from the positive plate, through the external circuit (load), to the negative plate, and returns to the positive plate through the electrolyte. Electrolysis occurs in the cell as a result of the electric current passing through it. During this process, the spongy lead of the negative plate is combining with the positively charged component (SO_4) of the electrolyte, forming lead sulphate ($PbSO_4$) and losing its negative charge. At the same time, the oxygen of the lead peroxide of the positive plate is combining with a part of the hydrogen in the electrolyte, forming water (H_2O), and also reducing the positive plate to pure lead (Pb). In addition, electrolysis is taking place at the positive plate, forming more water and converting some of the lead into lead sulphate.

When a lead-acid cell is charging, the chemical action is reversed, thus restoring the chemical energy released during discharge. The chemical action in a lead battery is expressed by the formula:

$$PbO_2 + Pb + 2H_2SO_4 \underset{\text{Charge}}{\overset{\text{Discharge}}{\rightleftharpoons}} 2PbSO_4 + 2H_2O$$

World Radio History

Lead used to construct the plates in a lead-acid cell is relatively soft and does not possess much structural strength. For this reason, alloys of lead are used to provide the necessary mechanical strength. The most prevalent type of alloy in lead-acid batteries used today is lead-antimony.

Recently, another type of alloy, lead-calcium, has proven to be an excellent material for constructing cell plates. Lead-calcium batteries have a much longer expected operating life than lead-antimony batteries—about 40 percent—and require less maintenance. However, they are more expensive and not necessarily the most economical battery in all applications. Since they require less maintenance and attention than the lead-antimony battery, they are very useful at remote, unattended stations.

A lead-acid battery can be maintained at full charge by placing its terminals across a dc power source. This is called *floating*. The open circuit voltage of a typical lead-acid cell that is fully charged is about 2.05 volts. To float a battery and maintain it in a fully charged condition, it is necessary to raise the float voltage above 2.05 volts to overcome the cell resistance. Under normal temperature conditions, the average voltage of the float charge is about 2.17 volts per cell.

A second type of charge, the equalizing charge, is a special charge given a battery to raise all of its cells to a uniform, equal voltage and specific gravity. Each cell in a battery has its own individual characteristics such as rate of local action (self-discharge), rate of charge, and capacity. Although differences between cells are usually very small, over a long period of time it is possible for an imbalance in cell voltages and the specific gravity of the electrolyte to become quite pronounced. The equalizing voltage is usually about 2.30 volts per cell. Equalizing charges are also used to recharge a battery after it has been discharged during emergency use.

Future Developments

With the advent of transistors, and more recently of microelectronics, the amount of power required to operate modern communications equipment is being reduced more and more. This fact has been an enormous help to the communication industry, especially in regards to supplying power to remote radio repeater sites.

Radio repeaters, for example, are often located at mountain-top sites where commercial power is not available. Conventionally, the equipment at these sites has been operated from prime mover generators and storage batteries which require frequent attention and maintenance to provide satisfactory service.

The development of missile systems and space vehicles has fostered a great deal of research into so-called *unconventional* methods of providing electric power. Among these are solar cells, chemical fuel cells, nuclear cells, and thermoelectric converters.

It is not yet practical to use these devices where much more than about 500 watts of power are required. However, they are expected to be of tremendous value in supplying power to such things as solid-state repeaters located in remote unattended sites where conventional methods of supplying electric power have, heretofore, been rather costly.

These new devices operate without moving parts, do not require frequent maintenance or attention, and generally provide uniform efficiencies over a wide range of power outputs at extremely low operating costs. Presently the most promising of these unconventional power sources, at least for the communi-

World Radio History

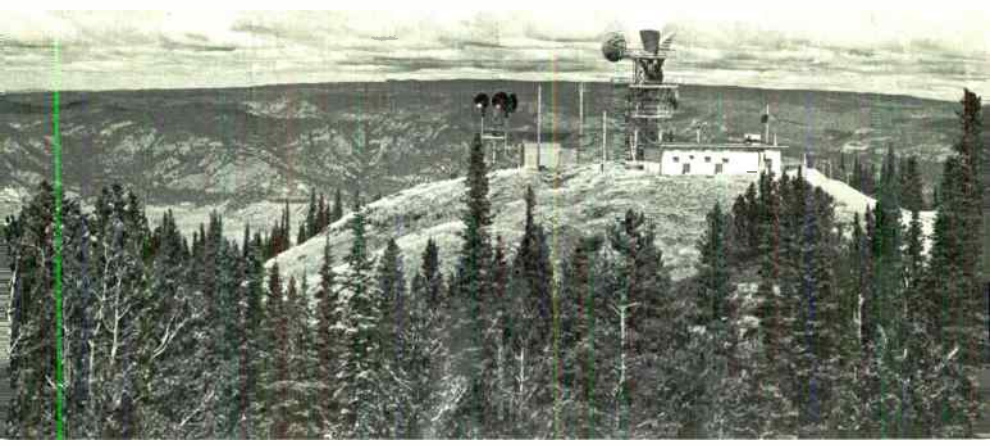


Figure 5. Often, microwave radio stations are located in isolated areas where public or commercial power is not available.

cations industry, appears to be the thermoelectric converter or generator.

Thermoelectric generators convert the heat from the flameless combustion of propane gas directly into electricity. They are capable of operating unattended for as long as the supply of propane fuel lasts, and in all types of weather extremes.

In the future, when these power sources prove to be technically and economically practicable, they should be an enormous aid to the communications industry.

Conclusion

The power equipment described in this article is typical of that used in many types of communications systems. It should be emphasized, however, that there are many different arrangements for communications power plants, each designed to meet the special needs of the particular system. For example, dc to dc converters are sometimes used to supply 24-volt power from the 48-volt power system, rather than from a separate system, or ac to dc converters might be used instead of rectifiers.

Regardless of the particular arrangement, however, each dc power system must meet certain common technical requirements. They must meet the requirements for voltage and current-carrying capacity and also the requirements for stability and regulation. Probably the most important of all requirements, however, are reliability and continuity. The system must be capable of supplying electric power continuously so that the vital mission of the communication system will not be vulnerable to disastrous and costly interruptions.

World Radio History

World Radio History

the Lenkurt Demodulator

VOL. 12, NO. 8 AUGUST, 1963

HF RADIO TRANSMISSION

Reliable global communication is increasing in importance as jets and rockets shrink the earth, and international commerce and traffic grow. The major burden of long distance communication falls on "high-frequency" radio, a method fraught with difficulties and irregularities. This article reviews the nature of HF transmission and some techniques for improving its quality and reliability.

With the exception of artificial earth satellites, which are still highly experimental, long distance radio communication around the perimeter of the earth has been practical only in the so-called high frequency band lying in the range of approximately two to thirty megacycles. Other frequencies have also been used, but are usually very limited. For instance, global transmission at very low frequencies is possible, but requires extremely high power and is limited to relatively slow transmission rates. Transmission at near-microwave frequencies requires extremely large antennas and expensive radio equipment, and even then the distance which can be spanned is sharply limited.

All of these techniques are made possible by accumulations of ions and electrons in the earth's upper atmosphere. At altitudes above about forty miles, the atmosphere is very rarified. Radiation and high energy particles from the sun constantly disassociate some of the gas molecules into free electrons and positive ions. These charged particles, which may have a relatively short life, have a pronounced effect on radio waves passing through the area, and the magnitude of the effect is directly related to the density of the electrons.

One way of regarding this effect is to assume that each electron acts as a small antenna that extracts energy from the radio wave and then reradiates the energy. As the radio wave passes through the ionized region, the electrons are displaced or moved by the interaction of the electron's charge and the energy of

World Radio History

the passing wave. Because of the "drag" resulting from the electron's mass and the influence of the earth's magnetic field, the reradiated energy is slightly out of phase with the transmitted wave. This has the effect of changing the velocity of propagation of the wave as it passes through the medium, and thus alters its direction. Ions have a similar effect, but because of their larger mass respond more sluggishly and thus influence the radio wave less.

As altitude increases the atmosphere becomes more and more rarified. At the same time, the effect of solar radiation is greater since there is less absorption and shielding by the atmosphere itself. Because of these two factors and the fact that the atmosphere has slightly different composition at various altitudes, ionization tends to occur in stratified layers which vary somewhat according to the season, the degree of solar activity, and whether it is day or night. The "lifetime" of the free electrons is determined by the rarefaction of the atmosphere; random collisions between free electrons and positive ions cause them to recombine and become neutral. Where the air is dense these collisions are very frequent, whereas at higher altitudes recombination is much less likely. For this reason, the lower layers illustrated in Figure 1 may vanish at night. The *F* layer (or layers) is able to persist through the night, although the number of free electrons decreases from about 10^6 electrons per cubic centimeter to about 10^5. In the lower *E* region, electron density is typically about 10^5 electrons per cc, largely disappearing at night. These *E* and *F* layers make possible virtually all radio propagation beyond the horizon.

Although the return of radio energy to the earth by these layers is commonly

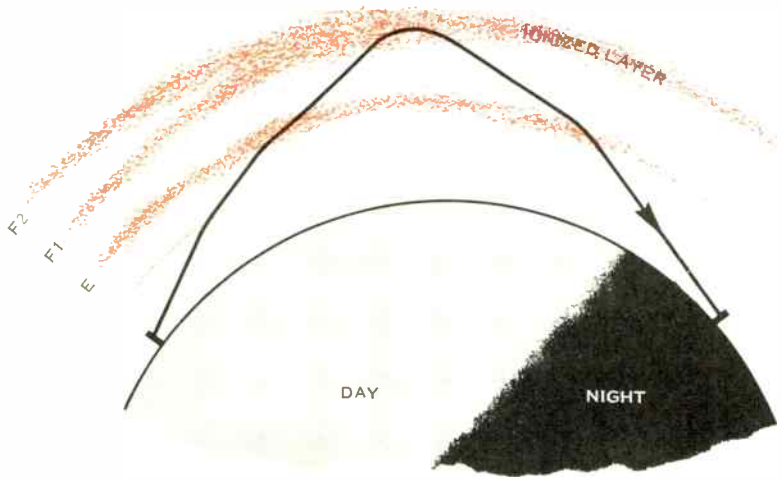


Figure 1. Intense solar radiation disassociates some gas molecules and atoms into ions and free electrons. The free electrons bend HF radio waves back to earth, permitting radio communication around the curvature of the earth. Stratification into layers results from varying composition of atmosphere at high altitudes. At night, recombination lowers electron density, causes lower layers to disappear.

World Radio History

called "reflection", it is actually a form of refraction which is controlled by both the frequency of the signal and the density of the electrons. At high frequencies the radio wave is less strongly affected than at lower frequencies. For a given frequency, the greater the electron density the greater the refraction or bending of the wave. In other words, the *index of refraction* of the ionized medium, which determines the amount of deflection of the ray, is a function of both the electron density and the operating frequency.

If the ray enters the ionized medium vertically, it will penetrate the medium only if it is above a certain so-called *critical frequency*. Frequencies higher than this frequency will penetrate the layer, while all lower frequencies will be returned to earth, as if by reflection. Frequencies higher than the critical frequency may return to earth if they enter the ionized layer at such a shallow angle that only a little deflection is necessary to direct them back toward the earth. The critical frequency can be calculated from the relationship $F^2 = 81N$, where N is the number of electrons per cubic centimeter and F is the frequency in kilocycles. Thus, the higher the electron density, the higher the critical frequency.

Figure 4 illustrates how a radio signal may be propagated from one point to another using ionospheric refraction. Radio energy enters the ionosphere at several angles. Note that as the angle of incidence becomes more nearly vertical, the radio waves are not returned to earth but are merely deflected as they penetrate the ionized layers. As the angle of incidence becomes greater some rays are deflected just parallel to the ionized layer. At a slightly lower angle radio energy is returned to earth. Thus, this minimum angle at which reflection can occur determines the "skip distance" which must be exceeded before reflected signals can be received. Beyond the skip distance, radio waves may be able to reach the receiver by several different paths.

One of the natural results of this is *multipath fading*, in which signals which have traveled slightly different routes arrive slightly out of phase with each other, thus causing cancellation. Complete cancellation of the signal occurs when two components of equal strength arrive exactly 180° out of phase. At a frequency of four megacycles this will occur if the two transmission paths differ by some multiple of about 125 feet; at thirty megacycles a path length difference of about 40 feet will produce the same cancellation.

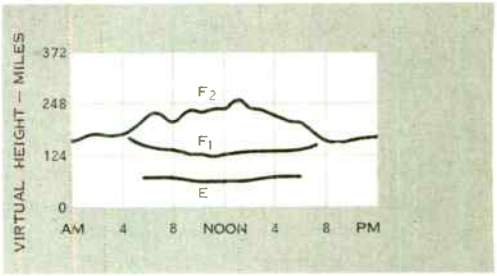


Figure 2. The E layer is lowest "full-time" ion belt. During peak sunlight hours F layer divides into two well-separated regions called F1 and F2. Not shown are D and "sporadic E" layers, both of which occur randomly.

As the length of one or more paths change, due to turbulence or changes in electron density, the frequency at which complete cancellation occurs shifts. A channel will often suffer severe fading at discrete frequencies within the channel, but not at others. As the multipath characteristics change, a fade may

World Radio History

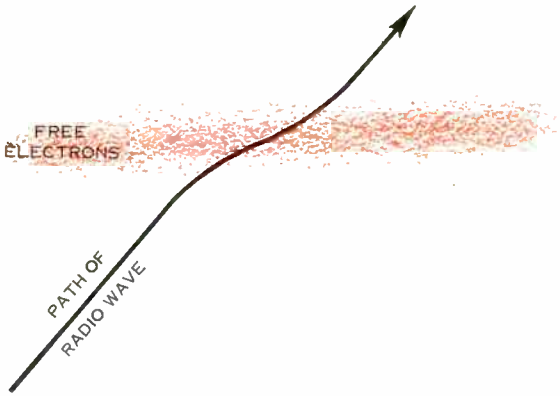


Figure 3. Free electrons alter direction of radio waves by a refraction-like process. Amount of bending is proportional to electron density and inversely proportional to radio frequency. At some lower frequency, the wave would not penetrate, but would be "reflected" to earth.

"drift" from one end of the channel passband to the other, perhaps attenuating low frequencies at one moment and higher frequencies the next.

This is a major reason why double-sideband amplitude modulation is no longer used for HF transmission. When both sidebands are transmitted, a carrier frequency having an explicit phase relationship with both of the two sidebands is required. Although the carrier may be suppressed in order to save power, it must be accurately reinserted at the receiver. If the reinserted carrier is as much as 90° (¼ cycle) out of phase with the original carrier frequency the transmission is completely garbled. In the case of double sideband transmission over HF circuits, selective fading may remove the carrier. The two sidebands then attempt to demodulate each other and the signal is destroyed.

This can be avoided by eliminating one of the two sidebands. Now, the phase relationship between the two sidebands is no longer a factor and the reinserted carrier need only be held to within a few cycles of the transmitter carrier frequency. Selective fading within the single transmitted sideband may be virtually unnoticeable. In addition to reducing the effects of selective fading, single-sideband transmission requires only about one-eighth the power of a double-sideband signal for comparable quality, and results in much greater transmission reliability.

Propagation Reliability

It is important to note that the ionosphere is not homogeneous or "smooth." The atmosphere is subject to winds, turbulence and irregularities. Solar activity has several effects: solar storms and eruptions cause upper atmosphere ionization to increase very greatly. At such times, the atmosphere literally "swells", reaching farther into space than during periods of lower solar activity times. On a smaller scale, hour to hour variations occur so that the ionosphere behaves as though it consisted of numerous separate reflecting surfaces, all shifting and changing independently. The received signal seems to "twinkle," suffering continuous phase irregularities and disturbances. Although this has little effect on voice signals, certain types of data transmission are strongly disturbed.

Since the ability to send a signal around the curvature of the earth depends on a so-called reflection from the ionosphere, the natural changes in the atmosphere may rapidly alter the suitability of a given frequency for trans-

World Radio History

mission between two points. For instance, it may be possible, because of strong solar activity, to transmit 30-mc signals across an ocean. As solar activity abates or as darkness nears, electron concentration may drop to a value which no longer reflects the 30-mc signal back to earth. Alternatively, the signal may not be deflected back to the particular spot where the desired receiver is located, but may instead reach the earth at a greater distance. In either case, transmission fails at 30 mc, but may be possible at some lower frequency.

Improving HF Reliability

It is difficult to overstate the value of reliable communications in either war or commerce. A superior method of communicating was worth untold millions to Lord Rothschild in his stock market dealings during the battle of Waterloo. Over a century later, an urgent message from General George C. Marshall to the commanding officer at Hawaii warning of the possibility of impending attack was blocked by failure of the military HF circuits. The message, which would have been received one-half hour before the Japanese raid on Pearl Harbor, was re-routed over a commercial submarine cable and was delivered about eight hours *after* the attack.

An effective way of minimizing the vagaries of HF propagation is the use of *ionospheric sounding*. In one version an HF transmitter sends a series of pulses which sweep across the entire HF band At the far end, a synchronized receiver records the reception of each of the transmitted pulses which get through. This technique reveals the presence of multipath distortion at each frequency and shows which frequen-

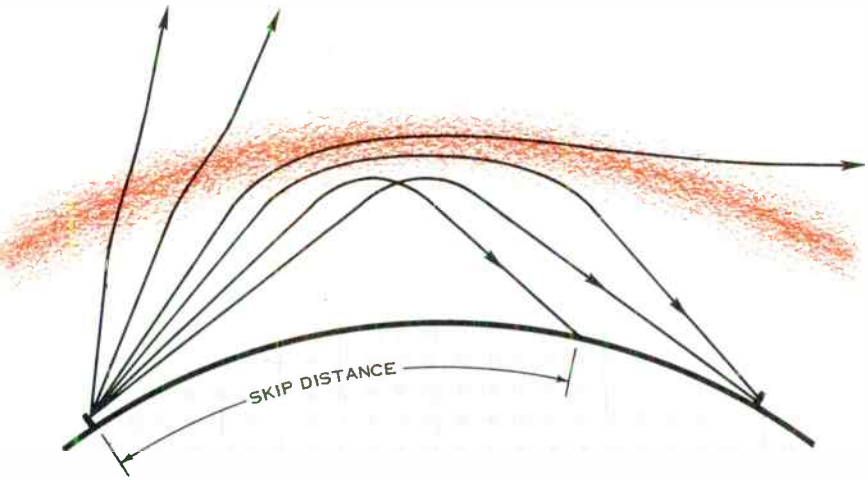
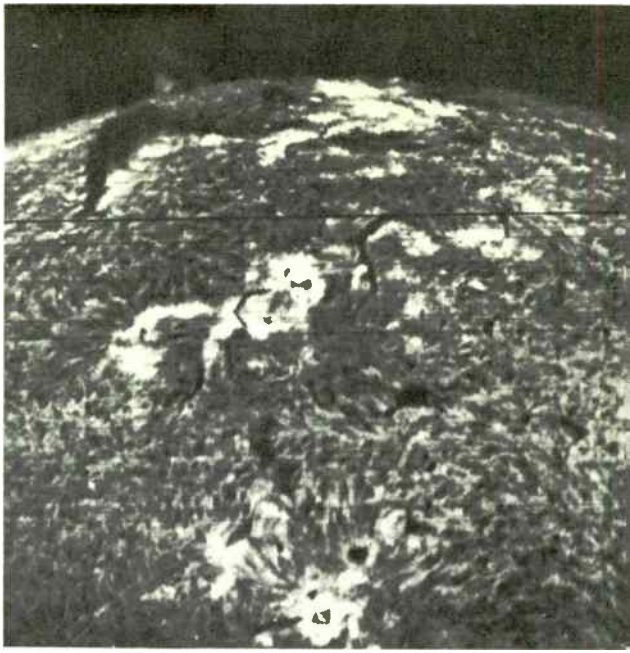


Figure 4. Typical HF propagation. Note that rays which enter ionosphere nearly vertically are only slightly deflected and penetrate the layer. As angle of incidence increases, rays are deflected into layer or refracted back to earth. "Skip distance" is the nearest distance at which radio waves return to earth. At greater distances, signal may arrive by several routes and thus suffer multipath distortion.

World Radio History



RADIO ASTRONOMY LAB, UNIVERSITY OF CALIFORNIA

Figure 5. Vast solar flares flood space with high-energy particles and intense radiation, causing earthly auroras and magnetic storms. These have an important effect on almost all terrestrial long-distance communications. Solar activity follows an 11-year cycle, reaching a minimum next year (1964).

cies are most suitable for transmission. Reception characteristics can be displayed on a tube or printed on a strip of paper.

Figure 6 shows a series of ionographs made in this fashion. The horizontal marks in each of the three transmissions represent reception at Palo Alto, California of a signal transmitted from Hawaii. Note that in the first trace the maximum usable frequency on this particular path was approximately 10 mc. At another hour (shown in the second trace) maximum usable frequency had risen to 20 mc. In the last trace adequate signals were received at 28 mc. These ionograms also reveal multipath reception in the lower frequency portions of the band. Obviously, such a technique may occasionally reveal that *no* assigned frequencies are suitable. However, trends in propagation characteristics can be revealed so that communications may be resumed as soon as possible when a usable frequency appears.

A similar technique called "backscatter sounding" may be thought of as a sort of HF "radar." In this technique only a single station is required. As in synchronized sounding, pulses are transmitted at many frequencies across the band. Where they come to earth some of the energy is scattered or reflected back over the transmission path. By recording the time required for each frequency to return, it is possible to calculate the distance traveled and therefore the location of distant points at which reception is possible. Where multiple reflections are involved each will tend to return a signal, thus revealing the number of skips.

Figure 7 shows a typical recording of a backscatter sounding. Between 16 and 32 mc it is possible to transmit more than 4500 kilometers using two

World Radio History

low-angle hops. Note that although each frequency appears to be limited to certain specific distances, this is for *optimum* communication. Successful transmission is usually possible at frequencies lower than those indicated in the display. By using backscatter sounding it becomes possible to select the optimum frequency for a particular destination. Had such a technique been available on "Pearl Harbor Day," the loss of eight battleships, hundreds of aircraft and thousands of lives might have been avoided.

Transmission of Digital Signals

More and more communications are transmitted in the form of digital sig-

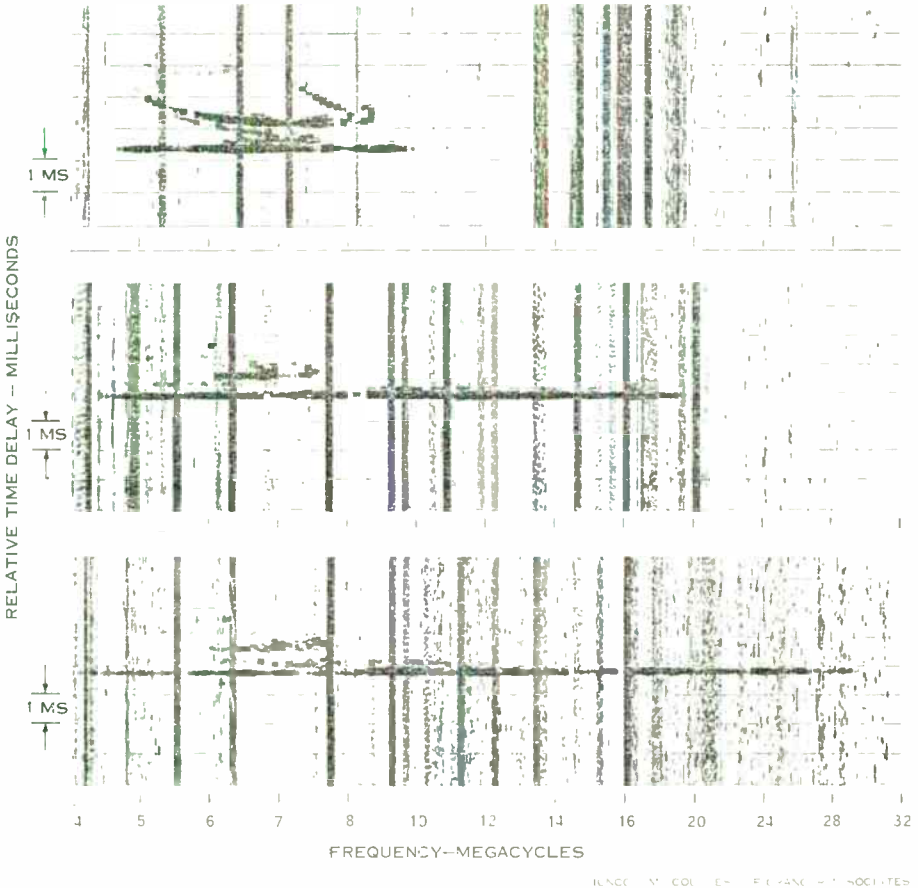


Figure 6. Typical ionograms obtained by oblique ionospheric sounding between Hawaii and California. The vertical streaks are radio transmissions. Horizontal lines represent 1 millisecond relative time. Note multipath reception between 4 and 7 mc. Maximum usable frequency increased from about 10 mc (top) to about 30 mc in the bottom record. Periodic soundings reveal propagation trends, permit optimum use of transmitting time.

World Radio History

nals. The digital signal form of information can be handled and transmitted more efficiently than such other forms as speech. It can be recorded in the form of punched tape or cards for transmission at any rate which the handling equipment and transmission medium will allow. For one reason or another, it is often desirable to transmit even speech in digital form. In the telephone network, pulse code modulation affords the possibility of improving signal quality when transmitting over great distances, and provides compatibility with electronic switching techniques which will probably dominate future telephone networks.

Another approach for transmitting digital speech has the advantage of reducing bandwidth considerably. So called "Vocoders" analyze speech sounds and generate digital code signals which represent the average energy level in various portions of the sound spectrum. At the receiving end a synthesizer interprets the codes and regenerates the speech. In this way it has been possible to reduce the bandwidth required for intelligible speech from about 3000 cycles to about 300 cycles while maintaining a surprisingly high degree of intelligibility and quality.

This technique lends itself admirably to the transmission of "secure" speech, that is, voice transmissions which cannot be understood if intercepted. This is useful in military and government operations where privacy of communication is essential. This type of security can be obtained by taking the digital signals from a Vocoder and scrambling them into an unpredictable code pattern

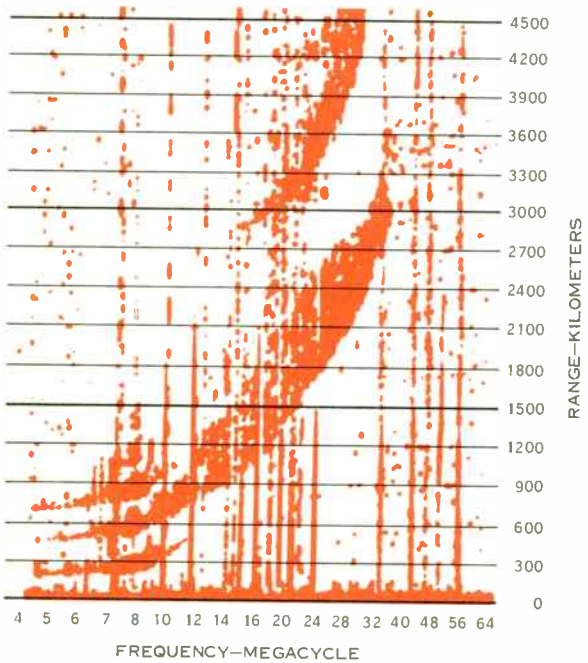


Figure 7. Backscatter sounding is made from a single location, reveals distances at which reception is optimum. Gain of receiver is set to reveal best transmission, usually at or near maximum usable frequency. Good reception is also usually possible at frequencies lower than those shown in ionogram.

World Radio History

which only the intended receiver can interpret correctly.

Although digital transmission is efficient and useful, it encounters a number of problems when going through HF circuits. HF circuits are crowded, noisy and erratic. Although these characteristics have little effect on voice or other analog signals, they tend to cause transmission errors in digital signals. This can be very serious because of the inherent lack of redundant information

GRAHGEF ASSOCIATES

Figure 8. Typical receiver for ionospheric sounder. Heart of device is a precision time standard to sweep receiving frequency across band in precise step with transmitter.

in most digital signals. Noise bursts which produce only "static" in a voice transmission could destroy or alter digital characters or the larger blocks of information contained on a punched card. Although the error rate can be reduced by using more transmission bandwidth, this is generally impracticable in HF radio due to spectrum crowding.

Most approaches for obtaining better transmission have been based on improving the modulation technique. The earliest systems used on-off keying of the radio carrier. This was disadvantageous on two counts: average transmitted power was only half the average capability of the transmitter, since energy was radiated only during "marks." Thus, at the receiver a space was indicated only by background noise. If the circuit were marginal or if the carrier frequency suffered a momentary fade, communication would almost certainly be lost.

To overcome this, *frequency-shift-keying* (FSK) was introduced. Two adjacent alternate frequencies represent either mark or space. Since the transmitter is sending either the mark or the space frequency at all times, the full average power rating of the transmitter is attained. Furthermore, FSK provides greater protection against fading, since the receiver detector can respond to signals that are almost lost in the background noise. By contrast, the envelope detector used in the on-off keyed transmission required that the signal clearly exceed a higher arbitrary power threshold.

Although FSK transmission greatly improves the reliability of digital transmission over HF channels, it is considered to be wasteful of bandwidth; the two sidebands seem to cost too much transmitter power and frequency spectrum. In a further effort to reduce the bandwidth required for digital transmission over HF channels, various methods of *phase*-shift-keying (PSK) have been introduced within the last decade. In this method, a single carrier frequency is transmitted. At appropriate instants, the carrier is shifted 180° in phase to signify mark or space for binary transmission. Higher information rates can be transmitted by shifting the carrier between 0°, +90°, −90°, and 180°, which is the equivalent of a quaternary or four-level system. PSK

World Radio History

reduces the bandwidth requirements for transmission and thus provides a theoretical improvement in the vulnerability to noise, since the receiver bandwidth can also be reduced.

Since all the information transmitted in this fashion is conveyed by the phase shift of the carrier, it is important that the carrier be extremely stable. Successful operation requires that the transmitter and receiver not depart from perfect synchronization more than a very few degrees of phase. Because of the stringency of this requirement, PSK transmission may be very difficult over HF radio.

The main difficulty stems from the inherent phase instability of HF radio. Random phase variations of the signal occur in transmission. The receiver can only interpret phase shifts as information bearing signals. It would appear, therefore, that such transmission channels would be unsuitable for PSK modulation. This is not quite true, since the rate of the phase change is usually not very great. One system takes advantage of this to establish a phase reference at the receiver, based on comparison of each received signal pulse with the preceding pulse. In effect, phase shifts introduced by the transmission medium are "averaged out." For such a "differentially coherent" PSK system to be successful, the bit rate must be substantially faster than the random phase variations introduced in the channel. This approach is, in fact, a compromise between an ideal phase coher-

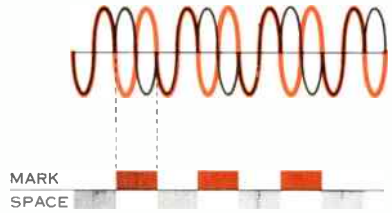
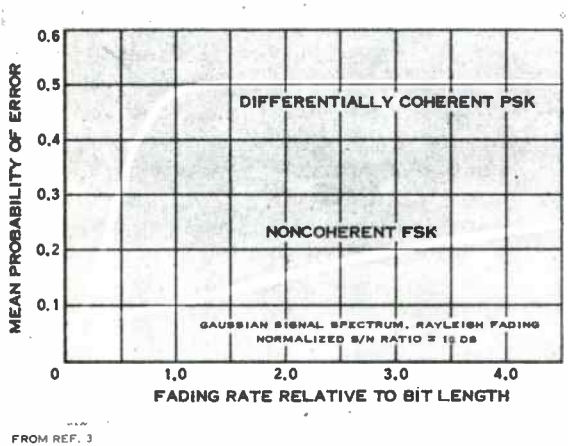


Figure 9. Phase shift keying (PSK) alters the phase of a stable carrier ± 180° to achieve binary transmission. Use of four phases doubles information rate but also increases errors caused by phase changes in HF medium.



FROM REF. 3

Figure 10. PSK signals are very vulnerable to fading and phase shifts in transmission medium. PSK systems have lower probability of error than FSK systems only when fading rate is very slow. As the fading rate increases, PSK error probability becomes many times that of FSK signals.

World Radio History

ent system and FSK systems, which are essentially phase insensitive but require nearly twice the bandwidth of PSK for a given information capacity. Even under the most favorable conditions, the averaging technique degrades performance enough that twice as much signal power is needed than would be the case with an *ideal* PSK system. Averaging of pulses over longer periods (in an attempt to obtain a more stable phase reference) tends to degrade system performance even more. Actually, the primary advantage of PSK systems has been in the conservation of bandwidth. Although this leads to improved performance under very noisy conditions, PSK systems are particularly vulnerable to fading, whereas FSK systems are the least vulnerable to fading and to phase discontinuities, as shown in Figure 9. Furthermore, PSK transmission shows less improvement in error rate when transmission is very good and the signal-to-noise ratio becomes large.

Coding to the Rescue

The bandwidth conservation that PSK provides can be likened to single-sideband transmission, and this is its principal advantage. FSK suffers the disadvantage of inherently requiring more bandwidth than PSK, but is less vulnerable to transmission disturbances.

Lenkurt's recently discovered Duobinary Coding technique (described in the February 1963 DEMODULATOR) has tipped the balance definitely in favor of FSK transmission by doubling the information capacity of the channel for a given signal bandwidth. In effect, this is the same as reducing the bandwidth required by FSK systems without giving up any of the advantages.

This new performance capability has been implemented in Lenkurt's Type 27A Data Transmission System which permits 2400 bits per second to be sent over a 3-kc voice channel. Future development will permit the operation of thirty-two 100-words-per-minute channels over the 16 basic 150-baud channels. Because of the inherent simplicity of the Duobinary Coding technique, the 27A system is far less complex than a differentially coherent PSK system of the same capacity, while providing far better performance over HF radio channels.

Although new types of submarine cable are being laid which have much greater channel capacity than the old, and communications satellite experiments are very successful, HF radio will have to continue bearing the major burden of global communications for many years. By effectively doubling the capacity of these circuits, the Duobinary Coding technique has, in effect, doubled the availabe frequency spectrum. •

BIBLIOGRAPHY

1. A. T. Brennan, B. Goldberg, and A. Eckstein, "Comparison of Multi-Channel Radio Teletype Systems Over a 5000 Mile Ionospheric Path," *IRE National Convention Record, Volume 6, Part 8 (1958)*, pp. 254-260.
2. H. B. Voelcker, "Phase-Shift Keying in Fading Channels," *Proceedings of The Institution of Electrical Engineers* (British) Part B, No. 31; January 1960, pp. 31-38.
3. A. B. Glenn and G. Lieberman, "Performance of Digital Communications Systems in an Arbitrary Fading Rate and Jamming Environments," *Paper 19.2, 1962 WESCON*.
4. Ya. L. Al'pert, *Radio Wave Propagation and the Ionosphere* (translated from the Russian) Consultants Bureau Enterprises, Inc.; New York, 1963

World Radio History

World Radio History

the Lenkurt Demodulator

VOL. II, NO. 5 MAY, 1962

EARTH SATELLITE COMMUNICATIONS

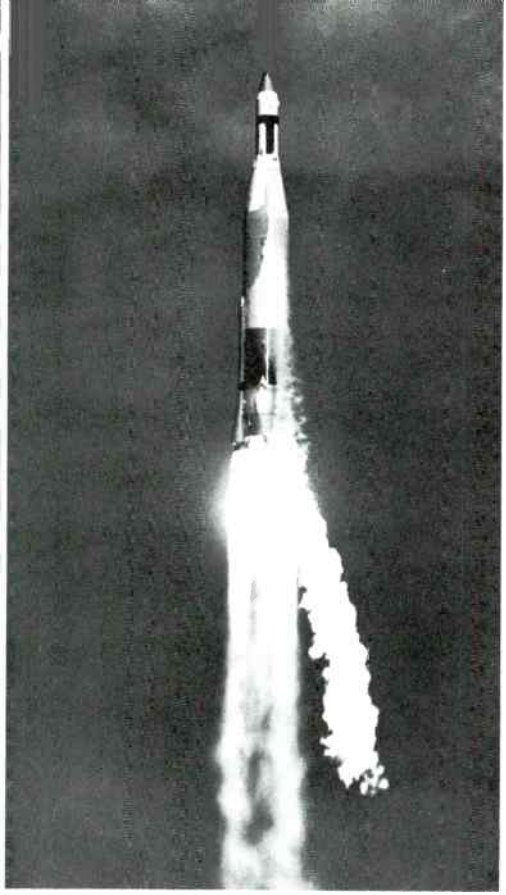
A giant step forward in the art of telecommunication is now in the making. Unlike such advances in the past, this one results not so much from new discoveries in electronics or communications technique, but from man's rapidly increasing ability to overcome the raw forces of nature.

Artificial earth satellites are the most recent benefit from this growing power, and now enable man to place television "weather eyes," radio repeater stations—and even communications switching centers—high above the earth where they can serve very large areas. This article surveys some of the fundamentals of earth satellites and how they can be used for telecommunications.

In the year 1267, Roger Bacon suggested that distant communication by magnetic means might be possible. By 1746, the deed had been accomplished over two miles of wire. Shortly after Samuel F. B. Morse's successful demonstration of his code in 1844, telegraphic communication spread swiftly throughout the world. The subsequent invention of telephony and radio, then microwave and television, helped push our global civilization to new heights.

Despite the tremendous improvements made possible by new techniques, communication has never been really cheap. The physical plant required—the miles of wire or cable, the radio repeaters, the complex terminal equipment—all have placed a price on communication that has sharply limited it to areas where demand is sufficient to justify the cost. Thus, while an underseas cable able to carry a few dozen channels of voice and telegraph might be justifiable, nobody has been able to afford a cable that could carry a television transmission across the oceans. Similarly, where traffic is lighter, perhaps no cable at all is economically feasible. In such areas, the burden of communications has fallen on fairly dependable "high-frequency" radio communications. Even this is fairly expensive in terms of communications

World Radio History



LOCKHEED MISSILES AND SPACE CO.

Figure 1. Atlas rocket carrying Agena satellite roars into sky. Hundreds of tons of fuel are required to place even small objects in orbit. Once difficult to achieve, the control and reliability necessary for successful orbiting is now routine.

capacity, since only single channels are usually practicable; available bandwidth and the propagation characteristics may not permit more.

Communication satellites appear able to provide a very efficient solution to this problem. Circling the earth beyond its atmosphere, satellites can provide a direct link between many distant points below.

Why Satellites?

Actually, there is very little novel or revolutionary about satellite communications except the ability to place the required equipment in orbit. Certainly, the satellite-borne equipment must be specially designed to fit the needs of the application, but no new or novel basic principles are required.

Many technical factors favor satellites for communication. Microwave frequencies, with their tremendous bandwidth and information capacity, are the natural choice for this service. Signals between one and ten gigacycles (thousand megacycles) pass through the atmosphere with relatively little attenuation. Below this range, radio energy is scattered by the ionosphere, and atmospheric noise smothers the signal; at higher frequencies, atmospheric oxygen and water vapor rapidly absorb the transmission.

This microwave "window in the atmosphere" makes it possible to transmit television signals or hundreds of voice channels up to a satellite for relay to points thousands of miles away. This is in contrast to conventional methods which may require hundreds of repeaters to transmit the signal between two points. The satellite's biggest advantage is that it has a direct radio "view" of very large areas of the earth, and can provide a common link between all the distant points in sight. This single fact makes it feasible to establish truly global communications on a scale never before possible. In addition, the ability to span thousands of miles using only a single repeater promises communications of a quality not possible with conventional techniques.

How to Orbit

Artificial satellites are now possible because we have learned how to produce and control the huge amounts of energy that are necessary to place even small

World Radio History

objects in orbit. All objects on or near the earth are drawn toward it by a relentless gravitational force that is proportional to their combined mass, but inversely proportional to the square of their distance from the earth's center of gravity. Thus, the farther away an object is from the earth, the less it "weighs," and the weaker the pull of gravity upon it.

In order for an artificial satellite to achieve orbit, it must be lifted above the atmosphere and started moving around the earth at the exact speed required to produce a centrifugal force just equal and opposite to gravitational force at that altitude. In effect, the satellite is falling freely toward the earth, but because of its tangential or sideward velocity (which would carry it away from the earth in the absence of gravity), it remains at the same altitude, as diagrammed in Figure 3. If the tangential velocity is greater than required to just balance the pull of gravity, the object tends to fly off into space. If the satellite is slowed down — by atmospheric drag, for instance — it is pulled back to earth by gravity.

Since the earth's gravitational attraction becomes weaker at greater altitudes, high-altitude satellites do not need to circle the earth as rapidly as low satellites to overcome gravity. Thus, the period required for each orbit becomes a function of the satellite's altitude. Actually, the period should vary with the mass of the satellite, since gravitational force is proportional to the product of the masses of the earth and the satellite. However, since the satellite's mass is so very tiny compared to that

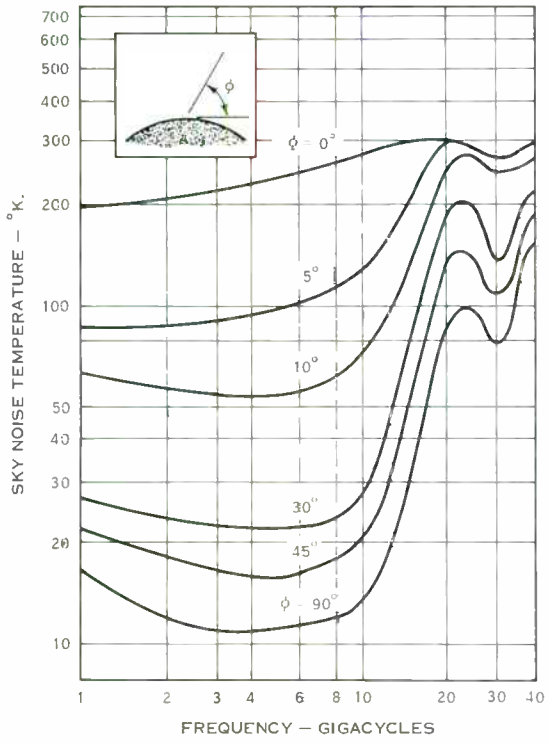


Figure 2. Atmospheric noise increases sharply above ten gigacycles (kmc), particularly when antenna angle with horizon is small. Increase in noise corresponds to attenuation by oxygen and water in atmosphere. Galactic noise from stars and interstellar gases increases rapidly at frequencies lower than one gigacycle.

World Radio History

of the earth, its effect is negligible. Only when the satellite approaches the size of the moon—which has a mass about 1% that of the earth—does its mass have a noticeable effect on orbital period.

The period of an artificial satellite in circular orbit can be easily calculated when altitude is known, by using the relationship

$$t^2 = \frac{a^3}{k}$$

where a is altitude or distance *from the earth's center of gravity* (not the surface), and k is a constant. For convenience, assume that the radius of the earth is 4000 miles. Then $k = 8.9 \times 10^6$, when altitude a is given in statute miles, and the period or time t is in minutes. Due to the rounding of actual values, the period determined in this way is only approximate, but is still accurate to within a few seconds of the true value.

We find that a satellite 100 miles above the surface circles the earth in about 87.5 minutes, while a satellite 600 miles up requires 104 minutes. At an altitude of a thousand miles, a satellite completes its orbit in 118 minutes. As the altitude becomes still greater, the orbital period becomes longer, until at an altitude of 22,270 miles, a satellite orbits the earth in exactly the same period of time as the earth's rotation—just under 24 hours. Such a satellite is called a *synchronous* satellite because its orbit is synchronized with the rotation of the earth. When a synchronous satellite travels eastward directly over the equator, it appears from the surface to be stationary in the sky (if it could be seen), and is sometimes referred to as a "stationary" satellite.

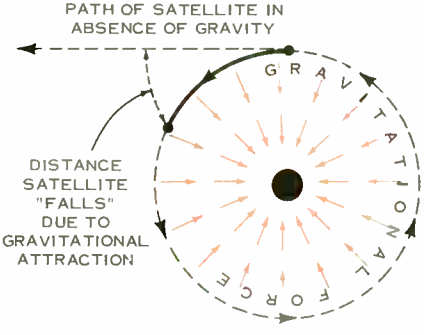


Figure 3. Tangential velocity of satellite keeps it clear of earth, although it is constantly "falling" freely. When velocity, direction and gravitational field just balance, satellite "falls" in circular orbit. Because of small errors of speed or direction, most orbits are eliptical rather than circular.

Passive versus Active Satellites

Any communication satellite system will require compromises between many design factors which tend to conflict with each other. Some of these factors are quality (a function of the signal-to-noise ratio), cost, reliability, and coverage. One basic decision is whether to use "active" or "passive" satellites.

A passive satellite is one used merely as a reflector of signals transmitted from the earth. An active satellite is one which carries radio equipment for receiving and re-transmitting the signal.

The passive satellite has the tremendous advantage of economy, simplicity, and reliability. With no functions to perform except "to be there," and with nothing to fail or get out of adjustment, the passive satellite may be able to last indefinitely. Such a communications satellite has already been tried successfully. *Echo I*, the first experimental passive communications satellite, was placed in a 1000-mile orbit in August, 1960. *Echo*, shown in Figure 5, is a metal-covered balloon 100 feet in diameter. Since its launch, *Echo* has been

World Radio History

used to relay voice, music, and even television transmissions over great distances. Although it is still in orbit, it has now become wrinkled and distorted so that reflected signals now scintillate or "twinkle"—varying in strength by a factor of about ten to one.

The main disadvantage of passive satellites is the extremely large amounts of transmitter power required to obtain a useful signal at the receiver.

Disregarding antenna sizes and operating frequency, the transmitter power required to obtain an adequate signal is proportional to the fourth power of the distance to the satellite, and inversely proportional to the square of its diameter:

$$\frac{\textit{Transmitting power}}{\textit{Received power}} = \frac{\textit{altitude}^4}{\textit{diameter}^2}.$$

This indicates that if, for instance, 10,000 watts of transmitter power are required (as in the case of *Echo I*) to return a barely adequate, narrow-band signal from a 100-foot satellite orbiting 1000 miles above the earth, effective transmitter power would have to be increased to 810,000 watts if the satellite altitude were increased to 3000 miles. Alternatively, the same results would be obtained if the diameter of the satellite were increased to 900 feet. If the satellite were located 5000 miles above the surface, six million watts of transmitter power would be required or the satellite would have to be nearly 2½ miles in diameter.

Since the satellite itself is a passive reflector, it imposes no important restrictions on bandwidth, but increased bandwidth requires a proportionate increase in transmitter power, thus further restricting the type of transmission that could be handled economically.

Even if extremely large amounts of transmitter power should prove to be economical, a very difficult problem of

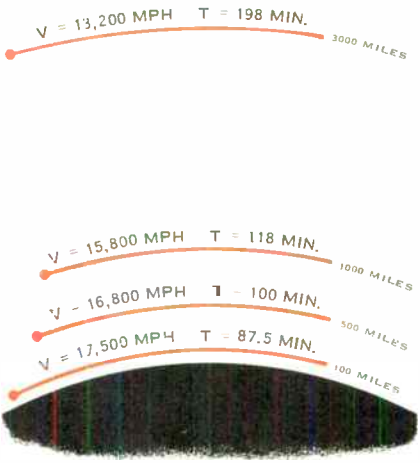


Figure 4. Gravitational force decreases with increased altitude, requiring lower velocity to maintain orbit. Some typical velocities and periods for earth satellites of different altitudes are shown. Although less velocity is required at high altitudes, much more energy must be expended in reaching it.

interference with other communications would very likely result. Even when extremely good antennas are used, antenna side lobes, and energy dispersed and reflected in the atmosphere could interfere with other services in the same frequency band over great distances.

If we seek to reduce power or increase bandwidth by *lowering* satellite altitude, we encounter other problems. Below about 250 miles, the life of the satellite will be shortened by drag from the outer fringes of the atmosphere. More important, however, the area that can be covered by a low-altitude satellite is greatly reduced. The low-altitude satellite crosses the sky very rapidly, making it difficult to locate and track, and leaving little time during which it can be shared or viewed by both terminals.

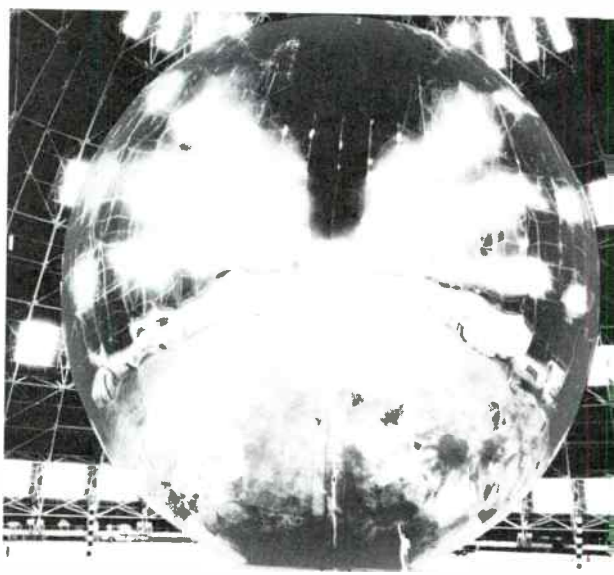
World Radio History

By contrast, the active communications satellite is freed of the altitude limitations of the passive satellite. It has many unique problems of its own, however, the most pressing being that of reliability. An active satellite consists of one or more radio transmitters and receivers, and a suitable source of power. This equipment must be rocketed into orbit and then operate properly for years without any possibility of receiving routine maintenance and occasional adjustment or repair.

The environment beyond the atmosphere is incredibly harsh. Floods of radiation and ultra-high energy particles from the sun erode the proper function of components, and sap the ability of silicon solar cells to continue providing vital electric power. Microscopic meteorites travelling many times the speed of a rifle bullet riddle the satellite, giving it a sort of cosmic sandblasting, occasionally hitting some vital element and putting the satellite permanently out of service.

However, the tremendous signal power advantage of an active satellite makes this approach worth a serious effort. A very small satellite transmitter returns a much greater signal to earth than would be reflected even from large, low passive satellites. It has been calculated that an output power of only two watts will suffice for transmitting broadband signals from a satellite closer than three thousand miles, even if the transmitting antenna radiates equally in all directions. Of course, this requires rather large receiving antennas at the ground stations. No matter how far out the satellite is located, however, the two-watt signal will still be sufficient if the transmitted energy is directed into a beam which just covers the disk of the earth. Thus, the farther the satellite is from earth, the narrower the transmitted beam must be. At the synchronous satellite altitude of 22,300 miles, antenna beamwidth should be about 17.5°.

The need for directive antennas introduces a new problem: that of keeping



NASA PHO

Figure 5. First passive communications satellite, Echo I, during trial inflation before launch. Although 100 feet in diameter, Echo weighs only 130 pounds. Satellite is a balloon of strong Mylar plastic covered with thin film of aluminum. Launched into nearly perfect circular orbit on August 12, 1960, Echo still orbits. Pressure of sunlight has caused shape of orbit to change with time.

World Radio History

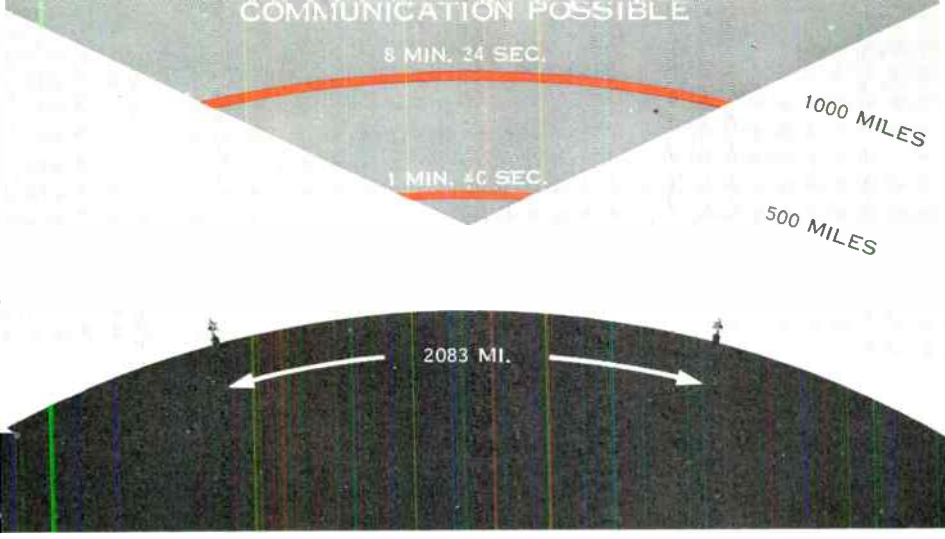


Figure 6. A major difficulty of low-altitude satellites is their rapid transit time and limited coverage. For communication, satellite must be within view of two stations simultaneously. Experimental low-level satellite soon to be launched will be placed in highly eliptical orbit designed to increase its time over northern hemisphere at expense of southern hemisphere.

the antenna pointed toward the earth. For a satellite in orbit there is no "up" or "down". Since gravity and centrifugal force exactly offset each other, the satellite is weightless, and gravity cannot practicably be used as a reference in aiming the antenna. Some satellites have been oriented by sighting the earth's horizon in several directions, then altering the attitude of the satellite by small jets of compressed gas, or by rotating small flywheels in the appropriate direction, thus moving the satellite in the opposite direction by reaction.

If some sort of stored propellant such as compressed gas or hydrogen peroxide is carried aloft with the satellite to keep it oriented, the useful life of the satellite will end shortly after the propellant is exhausted. A system using electrically-driven flywheels might last much longer. In this method, a flywheel (which might be the motor itself) is spun in the opposite direction to that desired of the satellite. Since nothing impedes the movement of the satellite, it would rotate until stopped by a brief reversal of the flywheel. However, this technique appears to be suitable only for overcoming small oscillatory or irregular forces acting on the satellite, and must be supplemented by some other method of attitude control when some constant force or torque acting on the satellite must be overcome.

Another possible orientation method which has been suggested is to use suitably placed coils aboard the satellite, which would be energized as required to work against the earth's magnetic field, in much the same fashion as an electric motor operates. The magnetic field is weak at high altitudes, and the resulting torque would be small. However, only a little force is required in the friction-free vacuum of space. Regardless of the

World Radio History

method or combination of methods used for directing satellite antennas toward the earth, and keeping the panels of solar cells pointed toward the sun, it is imperative that they continue to work reliably—or the satellite dies.

High or Low?

Although communications satellites are still highly experimental, intense planning is going into the configuration of the future systems. One of the most important considerations is the orbital altitude of the satellites, for altitude has a profound effect on the cost and function of the system. Many space scientists believe that the ultimate system will use synchronous satellites located 22,300 miles above the equator. Because a signal relayed through a satellite requires about 0.3 second to travel from one terminal to the other, some have feared that the resulting 0.6-second delay before a reply could return would be too objectionable in two-party conversations. However, tests indicate that talkers adjust rapidly to this delay.

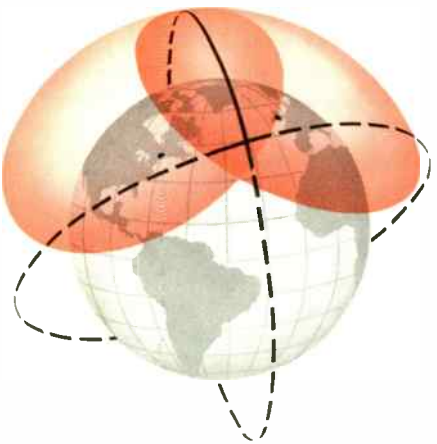


Figure 7. Time during which communication is possibe between stations will vary with location of orbit relative to stations. This will change with each pass. Tinted area shows coverage of each station and where they overlap. Communication between the two is possible only in overlap area.

All so-called two-wire talking circuits experience "echoes" due to unavoidable characteristics of the telephone equipment, but these echoes usually return to the speaker so fast that they are masked by some of the original sounds themselves. The long delay imposed by a high-altitude satellite relay allows the echo to be heard, however. The effect is extremely distracting and may even be intolerable to the talkers. It was once feared that this problem was so formidable that synchronous satellites would necessarily be limited to such one-way services as relaying television or business data. However, a new type of echo suppressor developed by the General Telephone Laboratories appears to have overcome this problem quite satisfactorily.

A more important consideration is the effect of satellite altitude on system *cost*. A synchronous satellite appears to hang nearly motionless in the sky. At somewhat lower altitudes, the satellite travels slowly across the sky, while at low altitudes, it races from horizon to horizon very swiftly. For instance, a satellite 250 miles high and passing directly overhead is in view less than 8½ minutes. At best, only about 6½ minutes of this would be suitable for communications, since signal quality is poor when the tracking antennas are within 10° of the horizon. Below 10°, noise from the earth and its atmosphere degrade and mask the signal. The problems of anticipating, locating, and tracking a satellite at this altitude are formidable and their solution very costly.

At greater altitudes, the problem becomes progressively easier; a thousand-

World Radio History

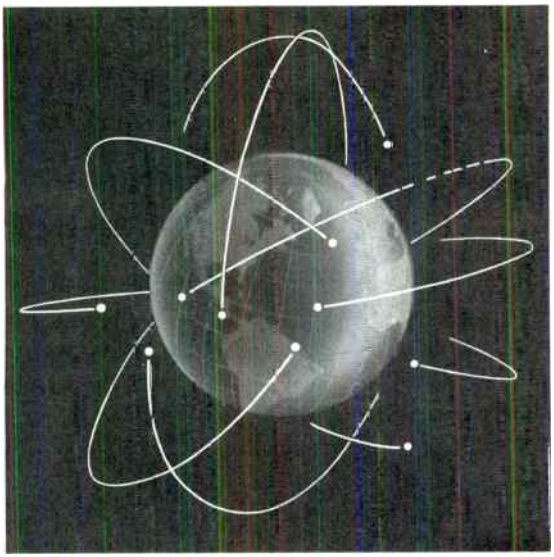


Figure 8. Fifty or more low- or medium-altitude satellites in random orbits will be required to assure continuous communication between any two stations. Number must be even larger to allow communication with other stations. Although numerous, these satellites can be relatively cheap. Terminal costs are high, due to finding and tracking difficulties.

mile satellite (which passes overhead) remains in view about 20 minutes.* However, only about half of this time is useful for communications, since the satellite must be tracked by *two* stations located far apart. Thus, in the simplest case, a 1000-mile satellite could be tracked simultaneously by two stations 1950 miles apart for only about ten minutes before it would be lost by one. A 3000-mile satellite could be tracked for 24 minutes by stations located 3100 miles apart, providing the satellite passes directly over both stations.

In a system using relatively low altitude satellites—that is, lower than synchronous altitude, many satellites will be required to assure that at least one will be in view of a pair of terminal stations most of the time. Multiple access to the satellite system—i.e., use of the system by several pairs of terminals in the same part of the world—is possible only by increasing the number of satellites visible at one time.

Each terminal will require at least two tracking antennas; while one tracks, the other searches for the succeeding satellite. A computer will be required to store information about the orbits of all the satellites, and to point the antennas at the proper point in space at the right instant. In such a system, the necessary ability to follow a satellite across the sky limits the size of the antenna and therefore its performance. However, large, expensive tracking facilities might be easily justified for linking points which share very heavy traffic. If the satellite orbits are low enough to avoid the need for aiming satellite antennas downward, the satellites themselves can be relatively cheap.

Synchronous Satellites

The synchronous satellite system has an entirely different set of conditions. The satellites themselves are relatively expensive, while the ground stations are

*The earth's rotation will increase these figures somewhat, depending on the direction travelled by the satellite, and its period. These examples disregard this factor.

World Radio History

Figure 9. Single "synchronous" satellite above equator at about longitude 25° W. can provide communication between any of more than 100 nations which have nearly 92% of world's telephones. Television broadcasts could be relayed to all stations simultaneously without tracking interruptions.

cheap. The greatest expense is getting the satellites into orbit. In addition to the much more expensive rocket boosters needed to achieve the desired altitude, a very "sophisticated" procedure is required to alter and adjust the direction of the satellite in order to orbit it in the equatorial plane at the correct distance from the earth.

The synchronous satellite provides multiple access by many terminals. Coverage is so great that three satellites can cover the entire earth—excluding only the polar regions. Even a single satellite located over the equator at longitude 25° W will be able to provide service to more than 100 nations, as shown in Figure 9. This includes all of Central and South America, Africa, western Europe, and eastern Mexico and North America.

Simultaneous multiple access would require that each terminal maintain extremely high precision and stability of transmitting frequencies—on the order of 1 part in 10^{10}. (See DEMODULATOR, *January*, 1962 for a discussion of how this may be done.)

However, terminal requirements are still greatly simplified. Since the satellite always occupies the same region of the sky, there is no need for a second tracking antenna and a computer with which to "acquire" successive satellites. Although there are a number of forces which may "perturb" the orbit of a synchronous satellite, causing it to drift from its original position, these are

World Radio History

quite small, and may permit large antennas to be constructed which have only a limited tracking capability, thus reducing their cost very greatly. Already several experimental antennas of this type have been built, and which are able to track objects in the sky by moving the "feed" or focal point of the antenna, rather than the large reflector itself. Other antennas have been designed which are able to track moving objects electronically by changing the phase relationships between elements of large antenna arrays.

Thus, the most expensive elements of a terminal station are eliminated in a synchronous system. Since economics largely determines the availability of communications to areas of light traffic, the synchronous satellite with its much lower terminal costs promises to extend wide-band transmission and communication to many areas of the world that otherwise could not support the more elaborate facilities required for low-altitude satellites. Of course, such a system will have to be compatible with existing world-wide communications networks.

Despite the desirability of synchronous satellites for certain types of service, many problems must still be solved before they can become more than an experiment. Although we have now achieved a fairly high degree of reliable capability in placing various objects in low orbits, this has not yet been achieved for synchronous altitudes. In addition to requiring much more powerful rockets than we now have available, these launching systems will have to be able to make major course deviations, and well-controlled fine adjustments of orbital speed in order to achieve the desired ultimate orbit. Even after achieving a perfect orbit, minor corrections will have to be made from time to time in order to minimize the tracking requirement at the terminals. This will require operational reliability of a sort still rare.

"Orbiter Dictum"

Accomplishment of a truly global communications system capable of exchanging television and other cultural information between the people of many lands will have profound benefits for all. We are now on the threshold of this achievement. Although there were many bitter disappointments in early attempts to place objects in orbit, this has now become quite routine. Similarly, early failures may be expected in the more advanced communications satellite experiments now impending. However, even if these failures occur, they will be stepping stones leading to eventual success. •

BIBLIOGRAPHY

1. David C. Hogg and W. W. Mumford, "The Effective Noise Temperature of the Sky," *The Microwave Journal:* March, 1960.
2. William Meckling, "Economic Potential of Communication Satellites," *Science;* June 16, 1961.
3. W. C. Jakes, Jr. et al, "Project Echo" issue, *The Bell System Technical Journal,* July, 1961.
4. John R. Pierce, "Communication Satellites," *Scientific American;* October, 1961
5. Leonard Jaffe, "The NASA Communications Satellite Programme," *Telecommunication Journal* (ITU, Zurich); March, 1962.
6. Samuel G. Lutz, "A Survey of Satellite Communication," *Telecommunication Journal* (ITU, Zurich); April, 1962.

World Radio History

World Radio History

the Lenkurt Demodulator

VOL. 12, NO. 6 JUNE, 1963

PROTECTIVE RELAYING

Vital Communications for Power Transmission

The transmission of electric power, the "lifeblood" of modern civilization, demands reliability of a magnitude almost unmatched elsewhere in our complicated society. Like so many things, power distribution is growing more complex, and modern life is becoming more dependent on it. Very effective techniques have been developd by power transmission engineers to assure continuous day-and-night service. Certain special types of communication play a vital role in maintaining this reliability. One of the more important is called "protective relaying" — the subject of this article.

The unique nature of power transmission introduces communication problems not found in telephony and similar industrial communications. "Faults" or outages in a telephone network interrupt valuable and important communications, but neither the users nor the communications network are otherwise affected; the loss of communication and information is the greatest harm that results.

In power transmission, however, a different type of commodity must be moved — raw power, electrical energy which may achieve an incredible magnitude When power fails, almost everything around is affected. A breakdown in the power distribution system can literally paralyze any part of the country that is deprived of electricity.

No less important is the effect of a transmission breakdown on the generation and distribution network itself. Modern power transmission is a finely balanced operation in which great amounts of energy are transformed by generators into electricity and moved efficiently through distribution networks to the consumers. The power that is generated and transmitted is carefully matched to consumer demand and the capacity of the transmission network Faults upset this balance, possibly releasing the load from a generator with-

World Radio History

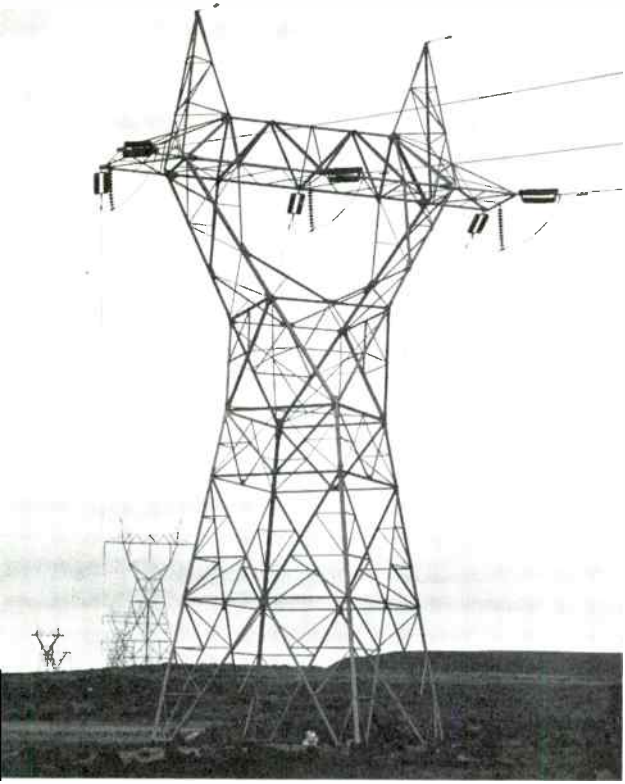


Figure 1. High-voltage transmission lines sweep across country, bringing cheaper power for industry and home use than could be possible in most locations if power had to be produced near the consumer. Like most long lines, these carry power at a potential of 230,000 volts in order to lower transmission losses due to high current flow.

out warning, or maybe doubling it, in the case of short-circuited lines. Such faults can suddenly release millions of watts of power which, if unchecked, could severely damage or destroy generating and transmitting equipment worth hundreds of millions of dollars.

Because of the extreme importance of protecting equipment and maintaining service despite the almost inevitable occurrence of faults, many techniques have been developed for minimizing the effects of such occurrences. The most important of these is the organization of power "grids" or multi-terminal transmission networks which link many power sources and load centers. Successful operation of the power grid requires a rather elaborate combination of sensing relays on each transmission line for detecting the presence of a fault on the line and swiftly triggering circuit breakers to isolate the fault before serious damage can occur.

The Power Grid

In the earliest days of electric power, transmission was necessarily local. Generating stations were located close to the consumer, direct-current power was used, and voltage was low. Because of the low available voltage, current had to be high for a given amount of power. This limited the distance over which power could be transmitted, since transmission loss is proportional to I^2R, (where I is current and R is line resistance).

With power distribution restricted to short distances, standby generators capable of meeting peak load were required to permit routine maintenance

World Radio History

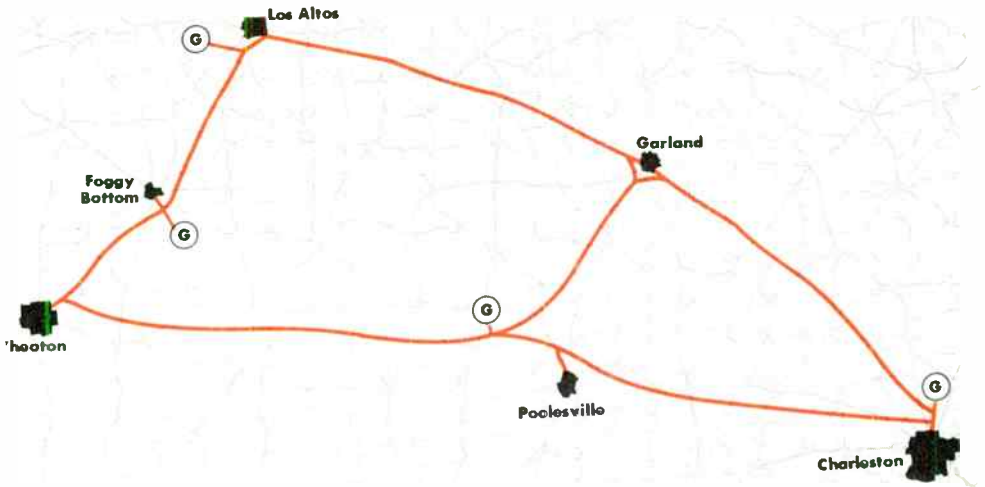


Figure 2. By linking distant load centers and generating sites together by long transmission lines, it becomes possible to equalize demand and generating capacity. Reduction of standby generating capacity at each load center reduces cost of power.

and to protect against failures. Naturally, this was very expensive.

Since then, the practice has changed. Alternating current permits transformers to be used to step the voltage up to very high values for transmission. By raising the voltage and lowering current, transmission loss is reduced, thus permitting power to be transmitted efficiently over great distances.

It soon became evident that widely spaced generators, each serving its own area, could be connected together to provide "mutual aid." If one generator or transmission line should fail, another could take on its load, thus maintaining service. As long-distance transmission became more efficient, it became possible to connect a greater number of widely separated power plants together in the network.

Obviously, power grids are more efficient and beneficial to all parties as they are increased in size and area. It becomes possible to adjust generating capacity to load changes more smoothly, since periods of peak demand in various parts of the grid may no longer coincide. For instance, daily peak load often comes at nightfall. This will obviously occur at different times in cities which are widely separated in longitude. Similarly, there may be a considerable seasonal load difference between cities of the far north where winter daylight hours are few, and in the far south where electricity is used in large quantities for summer air conditioning. The ability of each area to help the other meet its peak loads permits more modest investment in generating plant.

A catastrophic fault in large power grids, however, can affect far more consumers and more actual power than in simpler systems. For instance, without some means of rapidly isolating a serious fault, literally hundreds of millions of watts from the grid could be released,

World Radio History

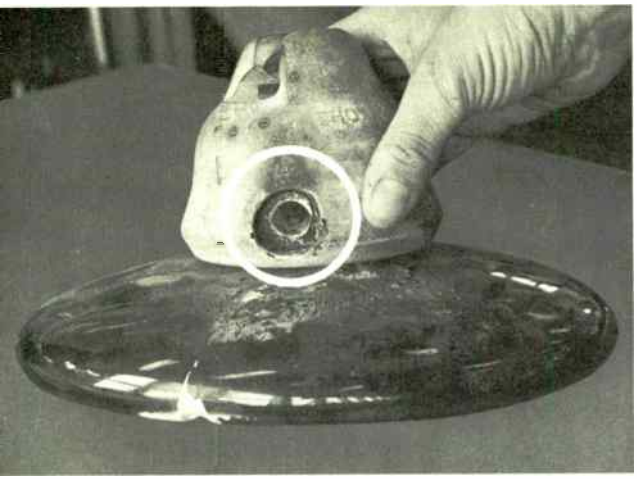


Figure 3. High transmission voltage may subject equipment to severe strain. Fault may cause substantial damage. Shown here is one section of a multi-section insulator which succumbed. Powerful arc chipped surface and edge, then burned its way through metal and thick porcelain (in circle.)

causing serious damage to the transmitting and generating equipment. Even individual generating sites may have a tremendous power capability. For instance, some large generating plants are able to produce currents as high as 40,000 amperes at 230,000 volts if the main bus is short-circuited to ground. In this case, more than nine *billion* watts would be dissipated in the fault and equipment. At this rate, enough power would be liberated *in one second* to supply all the power requirements for 500 average American homes for a year (assuming 5000 kilowatt-hours (kwh) per home per year)! A sudden shock of this much power may seriously damage tranformers, generators, or transmission lines, if unchecked.

To guard against such damage, fast-acting circuit breakers must be used to disconnect a "faulted" circuit as rapidly as possible. Speed of operation is very important in minimizing or preventing damage. Typically, modern high-powered circuit breakers are built to break the flow of current within three cycles (1/20th second) of the occurrence of the fault. Even at this speed, a major fault of the size mentioned above would still dissipate 130,000 kwh before the circuit could be broken.

Even where less power is involved, speed is very important in isolating a fault. This presents a problem, since accurate discrimination between true faults and such natural occurrences as switching transients requires time. For instance, switching transients or momentary heavy loads which are within the capacity of the system should not cause circuit breakers to trip as if a fault were present. However, actual faults must be promptly recognized and isolated swiftly.

Protective Relaying

To reduce uncertainty as much as possible, many special fault-sensing arrangements have been developed, usually employing specialized sensitive relays which are able to distinguish between "normal" conditions and faults. One of the most basic types of protective relay is called an "overcurrent" relay, which protects against loads which are beyond the ability of the equipment to accommodate safely. Such loads may occur because of unusual peak demand, faults, or a combination of these. The overcurrent type of relay

World Radio History

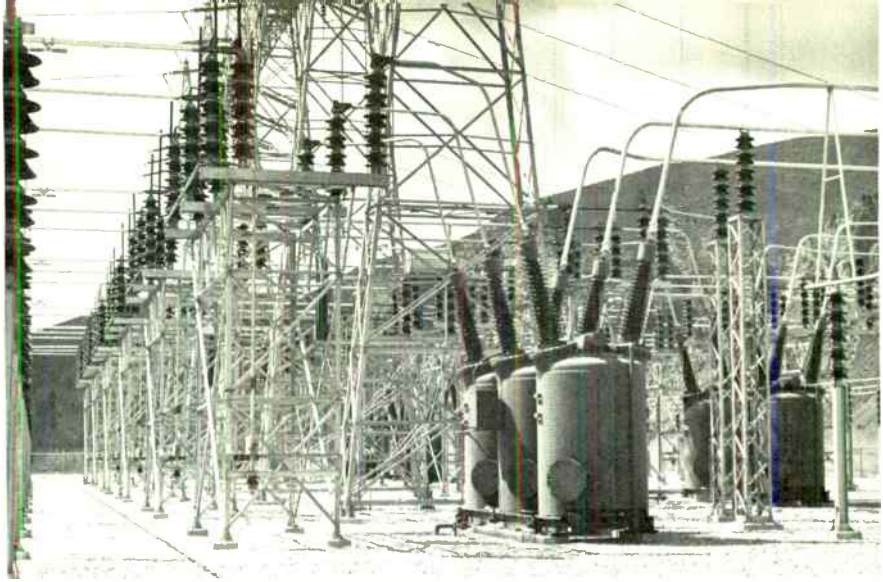


Figure 4. Typical switch yard, showing large circuit breakers for high-voltage line. Massive contacts are actuated by hydraulic force, and are usually immersed in oil bath to quench arc. The tremendous forces which must be arrested usually limit the life of high-voltage circuit breakers.

does not discriminate between faults and heavy demand.

One of the basic types of protective relay which can distinguish between faults and normal overload is the so-called "distance" relay. These devices independently monitor both the voltage and the current on the line. Under normal conditions, increased voltage will result in increased current. In some fault conditions, however, (such as a ground fault or transmission line short-circuited to ground), line voltage will drop and current increase. The relay is sensitive to the relative values of current and voltage, and to the resulting impedance that is established. When a fault occurs, line impedance will change in a characteristic way and the relay will trip if not prevented from doing so by other relays which guard against false trips. Because the transmission line is reactive, the impedance change varies with the distance from the fault. It thus becomes possible to adjust the relay to respond to faults which occur within a certain distance, but ignore those beyond.

Another way of sensing line faults is to compare the phase of the current at one end of the line with the phase at the other. Under normal conditions, the two ends will be essentially in phase. If the line is short-circuited or grounded, however, the phase at one end will reverse with respect to the other. The phase reversal will be detected and cause circuit breakers at both ends of the line to trip.

In order to achieve reliable tripping, yet avoid false trips, conventional relay arrangements may be quite elaborate. Often two or three back-up relay arrangements are employed to provide high speed and supplementary protection and to prevent false tripping for faults in adjacent transmission sections. Many different relaying methods

World Radio History

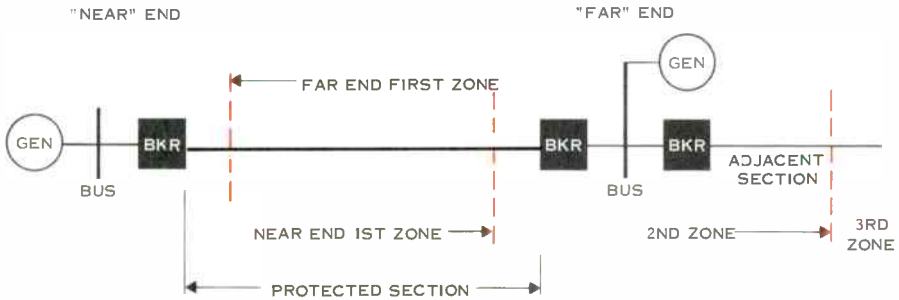


Figure 5. Schematic representation of a typical transmission line. Distance relays must be adjusted to respond only to faults in "first zone," thus avoiding false trips due to ordinary switching transients in far-end switch yard. Different type of sensing arrangement is used to detect faults in second zone. Second zone includes portion of adjacent section, third zone includes everything beyond.

may be used, according to circumstances and line characteristics. A typical basic arrangement for a two-terminal transmission section is diagrammed in Figure 5. High speed distance relays are used at each end of the line to provide fast "first zone" protection. The relays are adjusted to respond to faults occurring within 90% of the distance to the far end of the transmission line. It is necessary to adjust the relay for less than the full length of the line to avoid responding to momentary transients and surges caused by normal switching at the distant switchyard. At the distant end, a similar arrangement is used. Thus, faults occurring within the center 80% of the line will be detected at each end and quickly isolated by tripping the breakers at both ends.

If a fault occurs within the far 10% of the line, the distant breaker will immediately trip, thus disconnecting its end of the line. However, because the fault is beyond the reach of the near-end relay, it cannot respond, thus leaving the generator at the near end still feeding power to the fault. To prevent this from continuing, an "overreach" relay is used. This is a sensitive distance relay that is adjusted to respond to faults occurring not only on the protected transmission section, but also in the first 20% of the adjacent section.

In order to prevent this relay from tripping the near-end circuit breakers for faults in the next section, a blocking signal is transmitted, which, in effect, identifies the fault as lying in the next section and prevents breakers in the local section from operating. In order to allow the blocking signals "first priority," the overreach relay is normally made slightly slower acting than the first zone relays and the blocking signals. This time delay also may tend to prevent tripping in response to routine switching transients occurring at the far end.

Obviously, some sort of independent communications channel or so-called "pilot" circuit is required to link each end of a transmission line in order to trip the breakers at both ends of the line in case of a fault. Traditionally, these pilot channels have taken the form of wire or cable physical circuits, "power line carrier" channels, or channels transmitted by microwave. Each of these methods has its own advan-

World Radio History

tages and disadvantages. Physical circuits are generally limited to short distances, usually 10 or 15 miles, mostly because the shunt capacitance and series resistance of the line alter the currents which are put on the line to detect faults, and this effect becomes excessive as distance increases.

Power line carrier is widely used for protective relaying, but is gradually giving way to microwave because of its limited information capacity, relatively high cost, and dubious reliability at the moment it is needed most — during a fault. Power line carrier systems transmit tones in the frequency range 30 to 200 kc directly over the power lines themselves. Normally, the carrier transmitter and receiver are coupled to one phase wire of a three-phase transmission line. If the fault occurs on the particular line carrying the signal, there is some chance that the signal may still be able to get through the fault by inductive and capacitive coupling to the adjacent phase wires.

During a fault (which may consist of a short circuit between phases or between one or more phases and ground), noise is extremely high, and this may obscure communication even if the phase wire carrying the signal is not involved. Because of these possible hazards to communication, many protective relaying arrangements which use power line carrier are arranged to prevent or *block* the tripping of a circuit breaker. Thus, it is not necessary to transmit through a fault. If a blocking signal continues to be received, it tends to confirm that the fault lies in another transmission section. Of course, many other blocking arrangements may be used to prevent a circuit breaker from tripping in error. These blocking arrangements are effective and dependable but are generally most suitable for simple two-terminal transmission lines.

As power grids grow more complex, there is greater use of multi-terminal transmission sections, that is, sections which have one or more branches. A fault in such a section is much harder to detect accurately than in a two terminal network since it may occur in a branch carrying some fraction of the energy appearing at the other terminals. Such fault detecting techniques as phase reversal are particularly difficult due to a substantial loss of sensitivity to the fault condition.

Multi-terminal sections also greatly increase the communication problem because it is necessary that each terminal be able to signal directly to each of the others. In the case of power line carrier, this uses up the limited signal bandwidth very rapidly. For example, a two-

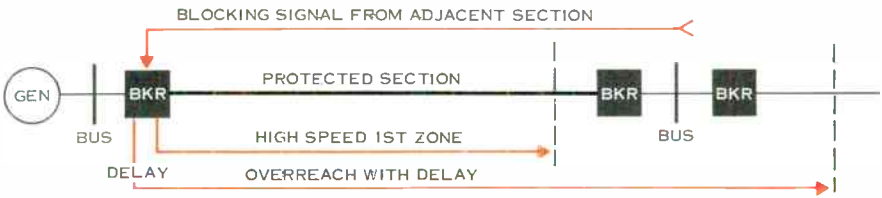


Figure 6. Typical "overreach" arrangement with third zone blocking. High speed distance relays at both ends detect first zone faults. Sensitive overreach relay detects second zone faults, but is blocked by signal from far end which indicates when the fault comes from the adjacent section. A slight delay is built into the overreach circuit to allow the blocking signal to have "first priority" in acting.

World Radio History

terminal line would require only two frequencies, one for each direction. A three terminal line requires six frequencies, while a four-terminal line must employ 12 frequencies. Once used, these frequencies should not be used again in nearby sections of the grid because of the difficulty of removing them from the power lines. Although frequency traps such as shown in Figure 8 are customarily used, these are necessarily simple devices (since they must operate at hundreds of thousands of volts), and are thus not completely effective in blocking these carrier frequencies. Since most carrier tones are used as blocking signals, undesired tones from a distant transmission section, even though attenuated, might prevent a breaker from tripping during a fault. As a result, limited bandwidth available on power lines limits the use of power line carrier in large or complicated power grids.

Transferred Trip

One solution to these problems is the use of remote tripping or *transferred trip*, as it is often called. The principle of transferred trip is directly opposite to that of conventional blocking schemes in which a transmission prevents a circuit breaker from tripping. With transfer trip, distant breakers are tripped on command of a signal from a terminal where a fault has been identified. Thus, all breakers in a section — even where there are several branches — may be tripped rapidly to isolate the fault. It is necessary to arrange the fault detecting relays to overlap their areas of sensitivity so that a fault anywhere in the line will cause at least one terminal to trip and transmit a signal to the other terminals. Naturally, each terminal is still free to trip at high speed if the fault is detected by that terminal's high speed relays.

Some principal objections to transferred trip operation are that it lacks security — that is, may produce false trips as a result of noise or other interference — and that it is not "fail safe." In the event that the communications channel over which the trip signal is sent should fail, the protection of the grid would be reduced.

Offsetting these arguments are the inherent simplicity and adaptability of the transferred trip method, and the fact that it provides 100% backup for the relaying methods on the transmission line itself, thus improving overall reliability.

Security, Speed, and Reliability

Most faults consist of a short circuit between phases of the transmission line or between one or more phases and ground, and these will almost inevitably result in an arc. At such a time, electrical noise may be tremendous. This, of course, is the basic reason that many conventional protection methods use the

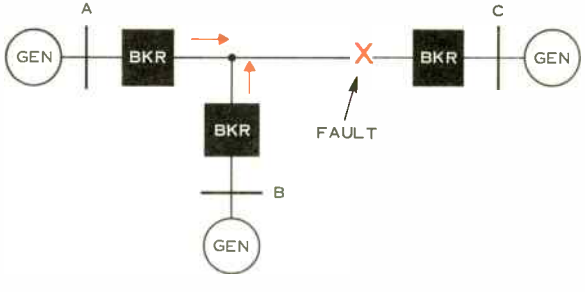
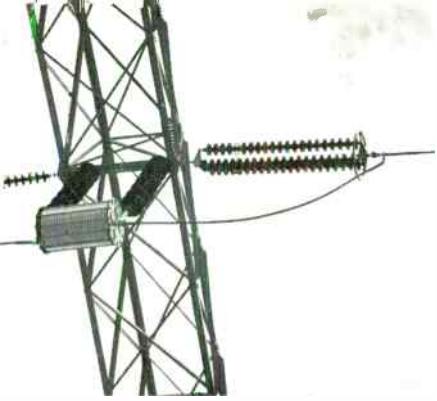


Figure 7. Three-terminal lines seriously complicate fault detection. Phase comparison methods are particularly handicapped because flow of current from other legs of line reduce ability of relay to identify faults.

World Radio History



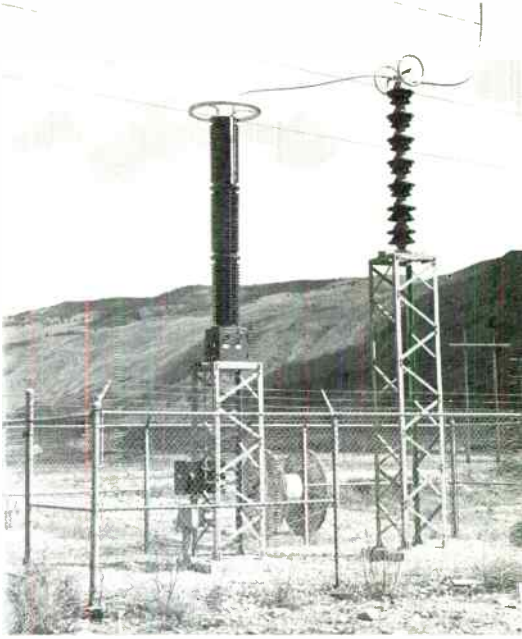


Figure 8. Very elaborate coupling capacitor (right) and frequency trap (above) is required for power lines. Cost may exceed that of a high capacity microwave terminal, particularly on very high voltage lines.

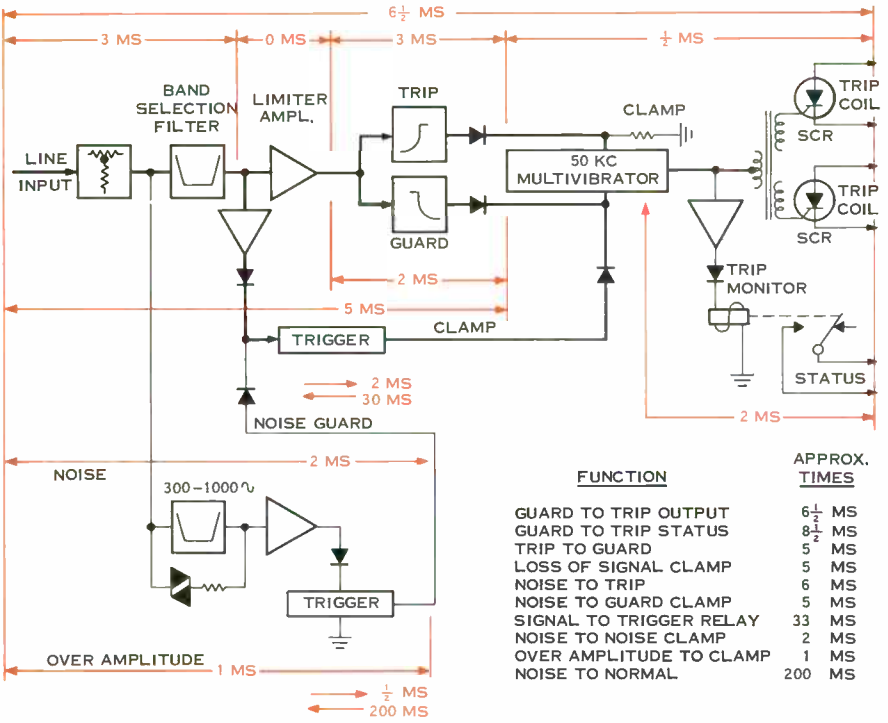
transmitted signals to *block* the tripping of breakers, since a power line carrier channel must be assumed to be unreliable at the time of a fault.

Modern microwave systems provide relief from this problem by establishing reliable communication channels that are not associated with the line, and which are therefore free from the interference caused by line faults. In the past, microwave has seemed too costly and relatively unreliable. Two factors are changing this, however. Modern solid-state microwave systems are now able to operate directly from batteries, thus eliminating outages due to power failures. In addition, substantial improvements in equipment reliability have been achieved by increased use of transistors and by effective techniques for switching to standby equipment very rapidly in case of equipment failure. Furthermore, the cost of the coupling and frequency trapping equipment required for power line carrier has tended to increase greatly as higher and higher transmission voltages are used, and this increased cost has in no way relieved the shortage of carrier frequencies or increased information capacity.

Although microwave provides communication channels that can be used for any of the conventional relaying methods which normally use a pilot wire or power line carrier, it also provides the major advantage of being suitable for remote tripping. Accordingly, its use in power transmission is growing very rapidly.

With higher transmission voltages, a transmission line carries much more power than formerly. This makes it especially important to react to a fault with utmost speed. Modern relaying equipment and circuit breakers are able to isolate a line within about three cycles after a fault occurs, but additional time delay in responding to a fault may be introduced by the communications equipment and additional relays which may be used to trip the circuit breaker. Relays are often used in tone equipment of older design to control the fairly large amount of current required to trip the circuit breaker. However, with almost no exception,

World Radio History



FUNCTION	APPROX. TIMES
GUARD TO TRIP OUTPUT	6½ MS
GUARD TO TRIP STATUS	8½ MS
TRIP TO GUARD	5 MS
LOSS OF SIGNAL CLAMP	5 MS
NOISE TO TRIP	6 MS
NOISE TO GUARD CLAMP	5 MS
SIGNAL TO TRIGGER RELAY	33 MS
NOISE TO NOISE CLAMP	2 MS
OVER AMPLITUDE TO CLAMP	1 MS
NOISE TO NORMAL	200 MS

Figure 9. Lenkurt Type 937A achieves very high security against false trips due to noise by using overlapping and interlocking circuits which distinguish between noise and desired signals. Special care is devoted to time constant of filters used to separate noise and signals. Interposing relays are eliminated by SCR's.

they contribute more delay and more unreliability than any other component in the systems. In protective relaying the equipment may not be actuated even once a year, with the result that contacts may oxidize or become dusty, and fail to function when energized. In addition, relays add as much as 18 milliseconds further delay to the tripping of the breaker.

Both objections can be overcome by using modern solid state components such as silicon controlled rectifiers (SCR's). These heavy-duty devices have no contacts or moving parts, and switch within a microsecond after being keyed, thus eliminating a major source of needless delay. Furthermore, the SCR is not subject to chatter or dropout. Once triggered, it remains conducting, thus assuring that tripping will be completed.

One of the biggest problems encountered in transferred trip protection is to provide positive tripping on command, but avoid false trips caused by noise, hum, or other sources of interference which may occur in a power switching yard. This is essentially a communications problem which yields readily to techniques developed for multichannel carrier equipment.

The highly successful approach taken in the Lenkurt Type 937A protective relaying equipment is diagrammed in

World Radio History

simplified form in Figure 9. One of the unique features of this equipment is the careful attention given to the absolute time delay required for electrical energy to get through the various electrical filters used. The time delay imposed on a signal by a filter is inversely proportional to its bandwidth; the narrower the bandwidth of the filter, the slower the operation. In the 937A equipment, noise and signal tones follow separate paths having different time delays. The noise is given a "fast" path to a clamping circuit which overrides or blocks a multivibrator trigger circuit and prevents a false trip. Similarly, the guard tone filter is designed for a faster response time than the trip filter. This gives the guard signal "priority" and thus helps prevent spurious tripping. The circuit is arranged so that a trip signal can actuate the SCR output devices only if the trip signal is applied simultaneously with the removal of the guard tone. Several additional protective circuits are also incorporated to eliminate improper operation under various other conditions which could occur in service.

Total elapsed time between keying and the rise of current in the circuit breaker trip coil is 8 milliseconds in the 937A, almost all of it contributed by the filters. Although it would be possible to reduce this delay by increasing the bandwidth of filters, this would impair the ability of the equipment to discriminate between noise and authentic tripping signals. In actual service, this design approach has proved more than adequate, transmitting tripping signals reliably, yet consistently failing to cause false trips even during a major fault in an adjacent section which caused intense noise and heavy current surges in the switch yard.

Future Trends

Higher and higher transmission voltages are being undertaken. Today, 230,000 volt lines are fairly common, but 345, 500, and even 775 kilovolt lines are being undertaken. Although basic protective relaying practices in general will be little different, the very high investments in transmission equipment and the much heavier flow of power will demand unusual care and diligence in obtaining higher tripping speeds and greater reliability.

Already there is a pronounced trend toward the greater use of microwave for all sorts of power transmission communications. In some cases, microwave is being used to free the power line carrier circuits for exclusive use in relaying. In other cases, microwave channels themselves serve as reliable pilot channels for protective relaying. This will undoubtedly increase as microwave becomes more widely accepted in the power industry. •

BIBLIOGRAPHY

1. C. Russell Mason, *The Art and Science of Protective Relaying;* John Wiley and Sons, New York, 1956.
2. R. E. Dietrich, H. S. Lorentson, and T. W. Stringfield, "Bonneville Power Administration Experience with Transferred Trip over Microwave Radio," *AIEE Transactions, Part III, Power Apparatus and Systems;* August, 1961, pp 405-411.
3. A. M. DiCioccio and G. R. Brandenberger, "Problems in the Operation and Development of Audio Tone Protection on the Duquesne Light Company System," *Paper presented at the Spring Meeting of the Relay Committee, Pennsylvania Electric Association;* State College, Pennsylvania, May 23, 1963.
4. J. L. Blackburn, "Protective Relaying for EHV Systems," *Paper presented at the Spring Meeting of the Relay Committee, Pennsylvania Electric Association;* State College, Pennsylvania, May 23, 1963.
5. C J. Holloman and A. E Olson, "Protective Relaying Using Carrier and Microwave Channels," *CP 63-691, Institute of Electrical and Electronics Engineers Region 3 Conference;* Richmond, Virginia, April 25, 1963.

World Radio History

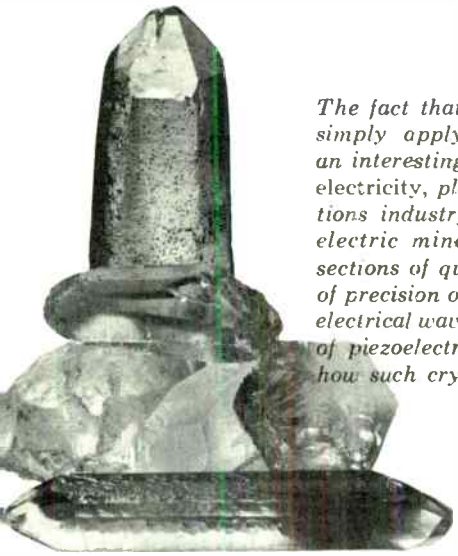
World Radio History

the Lenkurt Demodulator

VOL. 13, No. 7 JULY, 1964

The Properties and Uses of

PIEZOELECTRIC QUARTZ CRYSTALS



The fact that electrical charges can be developed by simply applying pressure to a mineral is indeed an interesting phenomenon. This effect, called piezoelectricity, *plays an essential role in the communications industry. Probably the most important piezoelectric mineral is quartz. Properly shaped, thin sections of quartz are used to stabilize the frequency of precision oscillators and to produce highly selective electrical wave filters. This article discusses the theory of piezoelectricity, the properties of quartz crystals, how such crystals are made, and some of their uses.*

The demand for precise frequency control and frequency discrimination is inherent in the field of carrier and radio communications. Accordingly, a great deal of effort has been spent in developing highly stable oscillators and extremely selective electrical wave filters

In the early days of radio broadcasting, transmitters contained a plate-modulated oscillator whose frequency tended to vary slightly during each modulation cycle. This instability in the transmitter frequency would, at times, produce a rather unintelligible signal in the home

World Radio History

receivers. Many radio listeners objected to such a condition, and their complaints led to the use of piezoelectric quartz crystals to control the frequency of these oscillators. These crystals, because of their highly sensitive frequency characteristics, provided a remarkable improvement in the stability of broadcast signals.

The advancement of military radio communications greatly increased the demand for crystal-controlled oscillators. The armed services required that their radio receivers be almost instantly adjustable to several frequencies, thus permitting immediate communications between battle groups. This vital requirement led to the use of crystal-controlled local oscillators in radio receivers.

Piezoelectric crystals are also used to construct excellent electrical wave filters. Such filters are used in high-frequency transmission systems to separate the simultaneous messages that may be transmitted over a single wire, cable, or radio circuit. In addition, these crystals are used as transducers to convert mechanical or sound energy into electrical energy in such things as microphones, phonographs, and in sound and vibration detection systems.

Piezoelectricity was first observed in 1880 when Pierre and Jacques Curie put a weight on a quartz crystal and detected a proportional electric charge on its surface. A year later the converse effect was demonstrated — that is when a voltage is applied to a crystal, a displacement occurs which is proportional to the voltage. Reversing the polarity of the voltage reverses the direction of displacement. The term piezoelectricity is derived from the Greek word *piezein* meaning *to press*. Hence, a piezoelectric crystal is one capable of producing electricity when subjected to pressure.

Crystal Structure

Since all solid matter consists of electrical particles, the piezoelectric phenomenon was not unexpected. In electrically *uncharged* crystals, the positive and negative charges are balanced, and no piezoelectric properties are observed. It is necessary, then, to *unbalance* these charges in order to produce piezoelectricity. Only crystals possessing certain types of atomic symmetry can become electrically unbalanced.

Crystals form when a gas or liquid solidifies into a definite atomic pattern, called a lattice. This atomic lattice may be symmetrical with respect to a point, a line, a plane, or any combination of these. Crystal lattices that are symmetrical with respect to a point cannot be electrically unbalanced. The atomic lattices of these crystals are said to have a *center of symmetry*. To better understand the nature of crystal symmetry it is helpful to consider the crystal classification system.

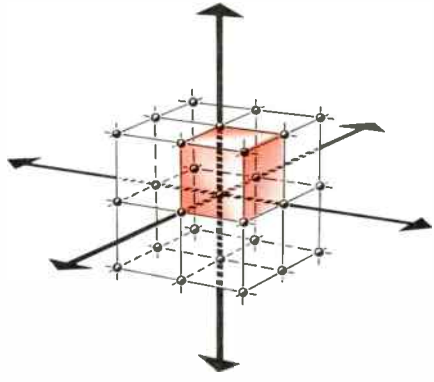


Figure 1. Atomic lattice of simple cubic crystal, showing unit cell (shaded) and the crystallographic axes.

World Radio History

Crystals are divided into seven crystal *systems* and 32 crystal *classes*. The lattice of each crystal is composed of a series of discrete three-dimensional atomic patterns, each forming a six-sided prism called a *unit cell*. The unit cells (Figure 1) can be considered as the *building blocks* which form the crystal. The edges of the unit cells are parallel to a set of imaginary reference lines that intersect at the ideal center of the crystal. These reference lines are called the crystallographic axes and, in piezoelectric crystals, are commonly identified as X, Y, and Z. The seven crystal systems are determined by the directions of these axes, with respect to each other, and by the length of the unit cell measured along each axis.

The symmetry exhibited by the surface of a crystal is merely an expression of the *arrangement* of the atoms within each unit cell. A total of 32 such arrangements is possible within the seven crystal systems. The 32 arrangements, or types of symmetry, constitute the 32 crystal classes. Only 20 of the 32 crystal classes exhibit piezoelectric properties.

Theory of Piezoelectricity

As previously stated, a crystal possessing a center of symmetry cannot be piezoelectric. When this type of crystal is subjected to pressure the same displacement of positive and negative charges occurs in any direction. Hence, there is no separation of the centers of opposite charges relative to each other. A distribution of charges having a center of symmetry is shown in Figure 2.

An example of the distribution of charges for a crystal *without* a center of symmetry is shown in Figure 3A. Note that the lines connecting like charges form an equilateral triangle and that the geometric centers of the two triangles coincide. As long as the centers coincide the charges remain neutral. If, however, a so-called longitudinal stress is applied to this crystal in the direction of the Y axis, a displacement of the charges occurs as shown in Figure 3B. In this example, the center of each set of like charges has shifted in opposite directions along the X axis, thus creating a dipole.

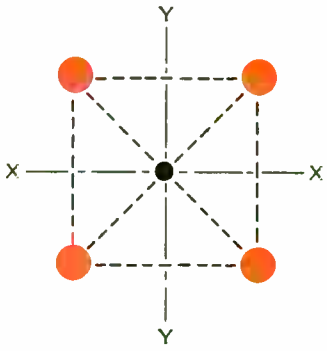


Figure 2. A molecular structure that has no charge separation when deformed. A crystal with this type of molecular symmetry has no piezoelectric property.

If the stress is applied in the direction of the X axis, as shown in figure 3C, the charge separation still occurs along the X axis, but is of opposite polarity. For this reason the X axis is called the *electrical* axis and the Y axis is called the *mechanical* axis. (These names are slightly misleading because under certain types of stress a displacement of charge centers will occur along the Y axis.) Perpendicular to these two axes is the Z axis. Because of the optical techniques used to locate this axis in a raw crystal, it is called the optical axis.

World Radio History

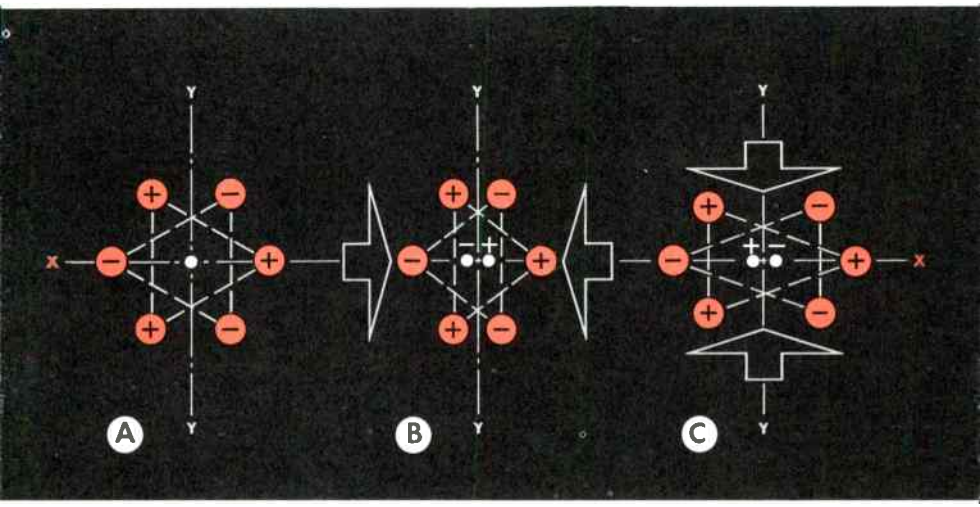


Figure 3. A model of one molecule of a quartz crystal showing the distribution of charges. When a longitudinal force is applied a charge separation occurs along the X axis. The charge appears along the Y axis if the crystal is subjected to a shearing force.

No piezoelectric effect is associated with the optical axis.

Quartz Crystals

The most commonly used piezoelectric crystal is quartz. Quartz is silicon dioxide (SiO_2) and crystallizes in the trigonal trapezohedral *class* of the trigonal *system*. This system includes all crystals which can be referred to 4 axes as shown in Figure 4. The 3 lateral axes are always equal and intersect each other at 60° angles. The fourth vertical axis may be shorter or longer than the lateral axes. A drawing of an ideal quartz crystal is shown in Figure 5. In nature, crystals of such perfect symmetry are seldom found and usually only the top formations and parts of the prism faces are visible. The different faces associated with quartz crystals are customarily designated by the lower-case letters m, r, s, x, and z.

In its original form, a raw quartz crystal is not usable in an electronic circuit. The quartz crystals found in communications equipment have been formed into various sizes and shapes, called plates, to give them particular piezoelectric properties. By slicing a raw quartz crystal at various angles with respect to its axes it is possible to obtain a variety of plates with different frequency and temperature characteristics. Certain plates have become standard and are classified into two groups; the X-Group and the Y-Group. The thickness dimension of X-Group plates is parallel to the X axis of the raw crystal from which it was cut. In Y-Group plates the thickness dimension is parallel to the Y axis. These standard plates are iden-

World Radio History

tified by symbols such as AT, BT, CT, or 5°X. Figure 6 shows the orientations of several commonly used plates. Listed in the accompanying table are the principal quartz plates of the two groups including the frequency ranges in which they are ordinarily used.

X-Group

Name	*Frequency Range (kilocycles)*
X	*40 to 20,000*
5°X	*0.9 to 500*
−18°X	*60 to 350*
MT	*50 to 100*
NT	*4 to 50*
V	*60 to 20,00*

Y-Group

Name	*Frequency Range (kilocycles)*
Y	*1000 to 20,000*
AT	*500 to 100,000*
BT	*1000 to 75,000*
CT	*300 to 1100*
DT	*60 to 500*
ET	*600 to 1800*
FT	*150 to 1500*
GT	*100 to 550*

The resonant frequency of a quartz crystal is determined generally by the size of the plate combined with the *mode* in which is is vibrated. Resonant frequencies achieved in standard quartz plates range from about 400 cycles per second to about 125 megacycles. The lower frequency limit is set by the dimensions of the largest usable plates obtainable from the raw crystal. The upper frequency limit of a quartz plate is reached when its size becomes so small that it is difficult to handle and is apt to shatter when put into use.

The three basic *modes of vibration* associated with quartz crystals are the flexure mode, the extensional (or longitudinal) mode, and the shear mode. Flexure motion is a bending or bowing motion, while extensional motion consists of a displacement along the length of a plate and away from the center. The more complicated shear motion involves sliding two parallel planes of a quartz plate in opposite directions with both planes remaining parallel. Each type of vibration can occur in a fundamental mode or in an overtone (harmonic) mode. The fundamental vibration of each mode is illustrated in Figure 7. The practical frequency ranges achieved in vibrating quartz plates in the three modes, including overtones, are:

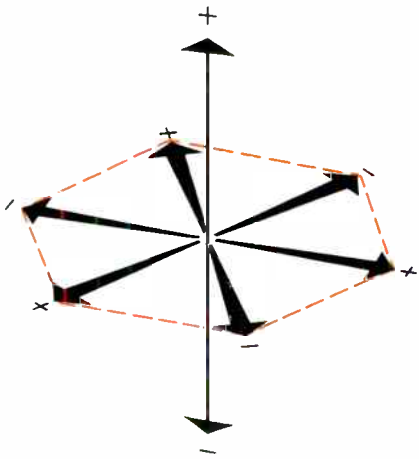


Figure 4. Crystallographic axes of the trigonal crystal system.

Flexure Mode . . . . 0.4 to 100 kc
Extensional Mode . 40 to 15,000 kc
Shear Mode . . . 100 to 125,000 kc

The excellent electromechanical coupling of quartz plates makes it possible

World Radio History

to use them in electronic circuits. In addition, quartz plates exhibit an extremely high *quality factor,* or Q, which makes them highly stable and efficient. In general, the Q of a circuit can be stated as the ratio of reactance to resistance. At the resonant frequency of a circuit, the capacitive and inductive reactances are equal and opposite and therefore neutralize each other. This leaves only the resistance of the circuit to oppose the flow of current. When this resistance is high, the Q will be low, and more power must be supplied to sustain oscillation. This added power contributes to instability and drift. Thus, the higher the Q, the more stable the oscillations and the less the resistive losses.

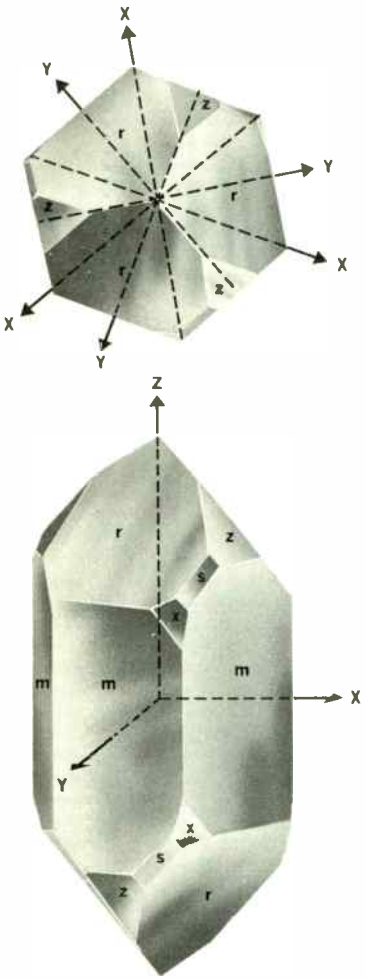


Figure 5. Ideal quartz crystal, showing the directions of the crystallographic axes. The faces of the crystal are identified by the lower-case letters.

Most of the losses in electrical resonant circuits are caused by the high resistance of coils. As a result, the Q of these circuits is comparatively low, ranging from about 10 to 400. The losses of a crystal are in its internal dissipation, mechanical mounting, and the damping of its motion by the surrounding air. The sum of these losses is very small when compared to an electrical circuit. Because of this, the Q of a crystal is comparatively high and may range from *ten thousand* to over *one million.*

Crystals possess two resonant frequencies; a series-resonant frequency and a parallel-resonant (or anti-resonant) frequency. The series-resonant frequency is determined by the distributed inductance, L_1, and the distributed capacitance. C_1, as shown in Figure 8A. The parallel-resonant frequency is determined by these same factors plus the parallel capacitance, C_0. Figure 8B shows the impedance versus frequency characteristic of a typical quartz plate. Note that the impedance of a crystal is lowest at its series-resonant frequency and highest at its parallel-resonant frequency.

Either the series-resonant or the parallel-resonant characteristics of a crystal may be used in an oscillator circuit. The mode of operation is deter-

World Radio History

mined primarily by the impedance of the circuit in which the crystal is connected. Figure 9 shows two transistor oscillator circuits, each employing a quartz crystal plate to control its frequency. The first circuit (Figure 9A) employs a crystal in the series-resonant mode. In this circuit the feedback loop from collector to base is established through transformer T1 and crystal Y1, which causes the circuit to oscillate at the series-resonant frequency of the crystal. Because of the high Q of the crystal the oscillator frequencies will be extremely stable over a long period of time.

The common-base Pierce oscillator, shown in figure 9B, employs a crystal in the parallel-resonant mode. Feedback is established from collector to emitter through capacitor C1. The frequency of this oscillator is controlled by crystal Y1 and the parallel capacitances of C1 and C2. The crystal operates just below its parallel-resonant frequency providing an inductance that resonates with the capacitances of C1 and C2. Hence, this circuit oscillates at a frequency

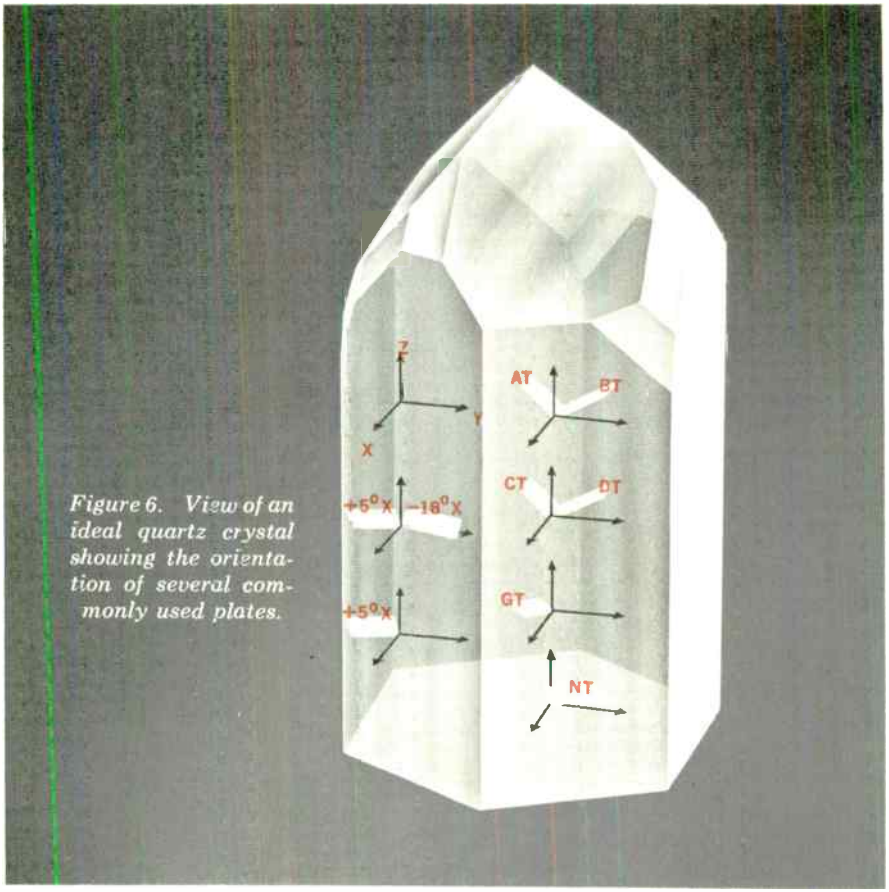


Figure 6. View of an ideal quartz crystal showing the orientation of several commonly used plates.

World Radio History

slightly below the crystal's parallel-resonant frequency.

Under certain conditions it may be desirable to adjust the resonant frequency of a crystal-controlled oscillator. Slight adjustments can be made by adding a variable capacitor or inductor in series with the crystal in the series-resonant mode and in parallel with the crystal in the parallel-resonant mode. The amount of adjustment can only be very small, but it does provide a means of offsetting frequency drift.

Because of their extremely high Q and stability, quartz crystals are also

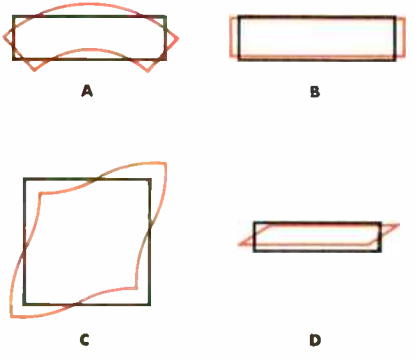


Figure 7. Fundamental vibrational modes associated with quartz crystal plates. (A) Flexure, (B) extensional (or longitudinal), (C) face shear, and (D) thickness shear.

used to produce very selective filters. Crystal filters are widely used in narrow-band applications, such as intercepting the pilot frequency in a carrier system, or separating the sidebands in a single-sideband radio system. Excellent wider band filters have been constructed using crystals with inductors and capacitors, usually in lattice-type

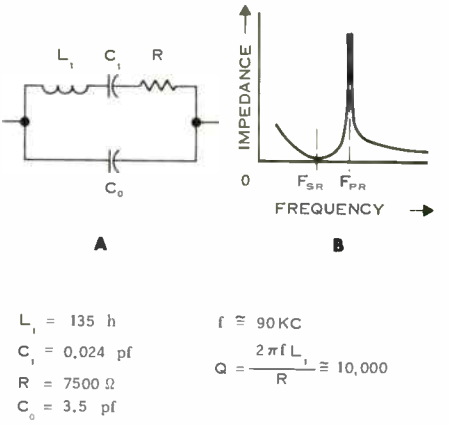


Figure 8. (A) The equivalent circuit and typical values for a 90-kc quartz crystal. (B) The impedance-frequency characteristic of a quartz crystal.

networks. Such networks exhibit superior frequency cut-off characteristics and, therefore, are used advantageously as carrier channel filters to multiplex voice frequency signals. The sharp cut-off feature permits closer spacing of carrier channels, thereby providing more channels within a given frequency range.

Manufacturing a Quartz Crystal

There are two types of raw quartz crystals, natural and cultured. Natural quartz occurs in veins and geodes (stones) and is mined primarily in Brazil. Cultured (or synthetic) quartz is grown in a process that is analogous to growing the familiar cultured pearls. Cultured quartz crystals are usually better formed than natural quartz, resulting in a higher yield of quality plates per crystal.

Making a quartz crystal unit begins by carefully inspecting the natural or cultured crystal for imperfections and then by locating its optical (Z) axis. A

World Radio History

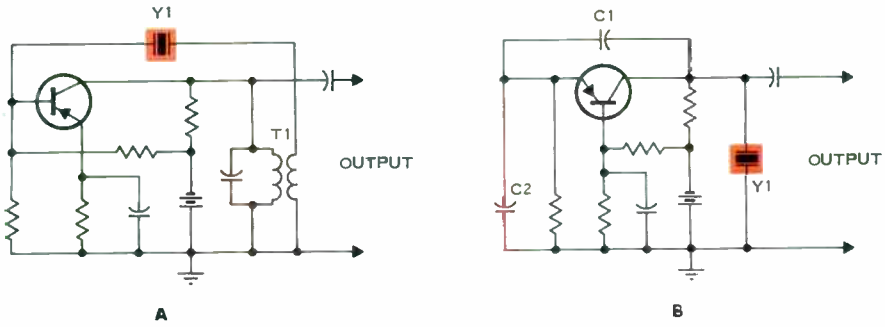


Figure 9. (A) Transistor oscillator employing a crystal in the series-resonant mode. (B) Transistor oscillator employing a crystal in the parallel-resonant mode.

face is ground on the raw crystal parallel to the optical axis. Next, the electrical (X) axis is located by an X-ray process. The mechanical (Y) axis is then known to be perpendicular to the X and Z axes.

After locating the axes, the raw crystal is sawed into sections and then into thin wafers along the Z axis so that the length of the plate is parallel to either the X or Y axis. Each wafer is then diced into small sections called *blanks*. These blanks are ground to the desired dimensions using a precision grinding process, known as lapping, and then inspected, cleaned, and etched in chemical solutions. After cleaning and etching, a spot of silver is applied to two opposite sides of each blank where the wire leads are to be attached. These same sides are then gold plated to form the electrodes required to electrically connect the crystal to the leads. Next, the leads are soldered to the silver spots. The crystal can now be frequency calibrated and placed into a holder — usually a vacuum-sealed glass tube or a hermetically-sealed metal can. After the crystal is mounted in its holder, a final frequency check is made. The crystal unit is then allowed to *age* for a short period of time, after which it is ready for use in an electronic circuit.

Aging

It is impossible to construct a crystal unit that is completely free of imperfections. Consequently, the resonant frequency of a crystal may vary slightly over a period of time. The degree to which the frequency varies depends primarily on the *quality* achieved in the manufacturing process. Poorly constructed crystals *age* faster than those made to more exacting standards. Many factors contribute to aging, including such things as leakage through the container, corrosion of the electrodes, material fatigue, frictional wear, presence of foreign matter, various thermal effects, and surface erosion of the crystal.

If, however, a crystal is very carefully

World Radio History

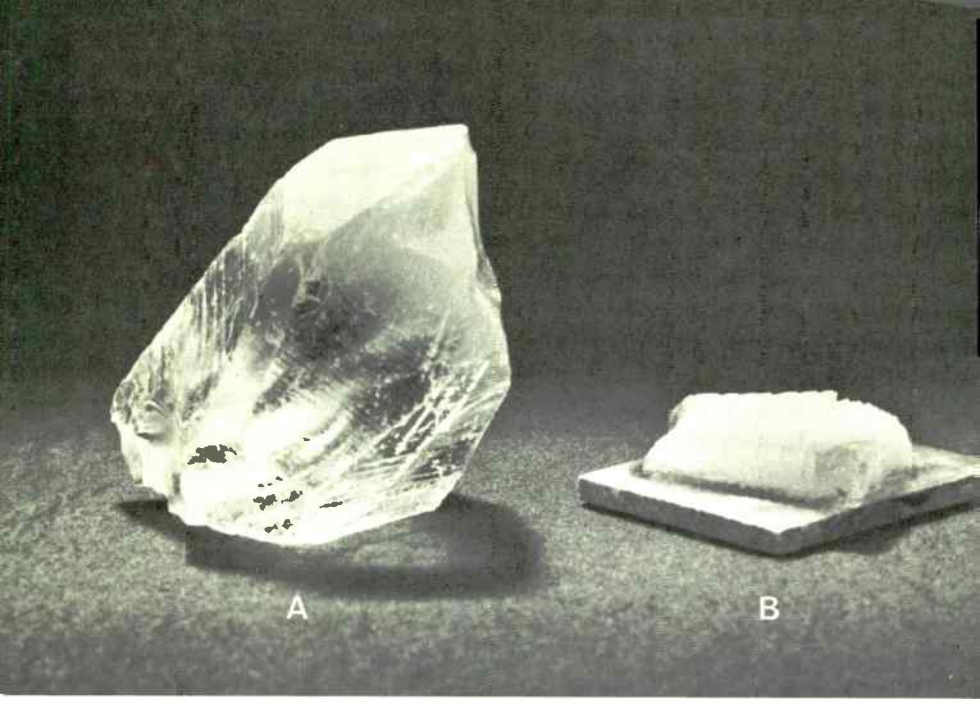


Figure 10. The various stages in the manufacture of a crystal plate installed in a vacuum-sealed glass tube. (A) Raw quartz, (B) section sliced into blanks, (C) blanks ground to the desired dimensions, (D) blanks gold-plated with leads attached, (E) crystal plate mounted in holder, and (F) finished quartz crystal unit.

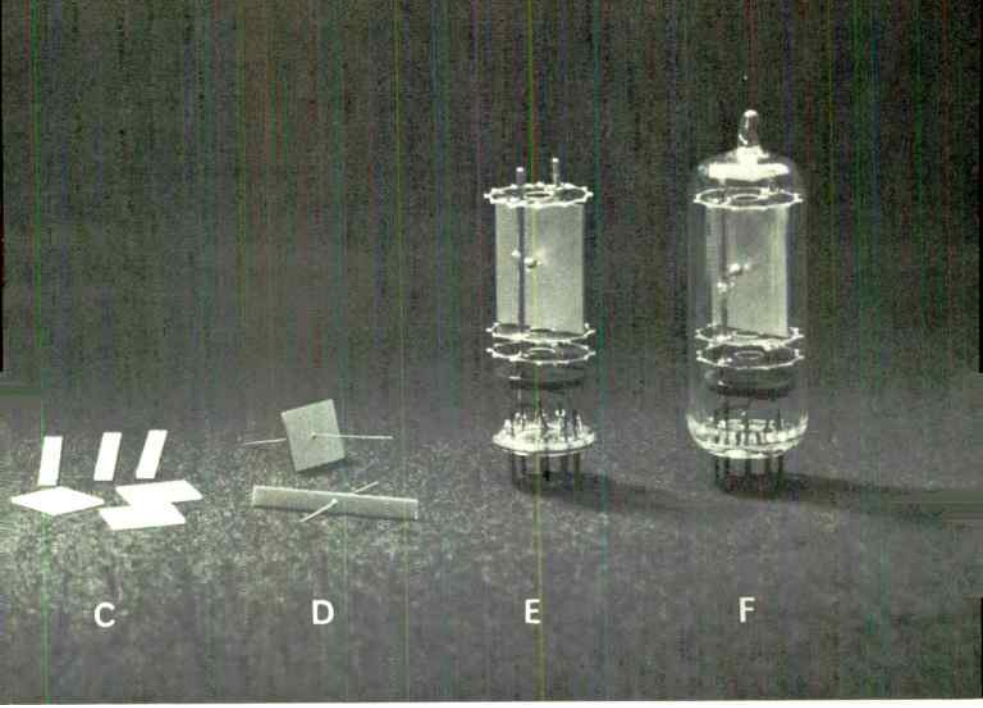
constructed it will last a long time before deteriorating. To produce a high-quality crystal unit, particular care is required in the finishing processes of lapping, cleaning, and mounting. For these reasons, a few electronic equipment manufacturers, such as Lenkurt, develop and build their own crystal units to ensure that the crystals used in their equipment are of the highest standards.

Conclusion

Quartz crystals enjoy a very useful position in the communications industry and, from all observations, their usefulness will continue for some time. In the past, several possible substitutes for quartz crystals were developed, but none were too successful. Prominent among these were ammonium dehydrogen phosphate (ADP), ethylene diamine tartrate (EDT), and dipotassium tartrate (DKT). These so-called synthetic crystals were developed because of a shortage of the large size raw quartz crystals required for making filter crystal plates. When the process of growing cultured quartz crystals developed, these substitutes were no longer needed.

In recent times, a disc-shaped ceramic device with piezoelectric properties has been designed and used in filter circuits.

World Radio History



One advantage of the ceramic disc is its characteristic bandwidth, which is about 10 times larger than the bandwidth of quartz crystals. However, its Q is limited to below 2000 and its frequency stability to about 0.1 percent variation (as opposed to about 0.01 percent for quartz crystals). It is doubtful, therefore, that these devices will compete seriously with quartz crystals in filter applications, except in certain special cases.

Quartz crystals possess a fundamental quality which has not been practical to achieve in electrical resonant circuits. Consequently, they have become an integral part of communications equipment and have played an essential role in advancing the development of transmission systems. •

BIBLIOGRAPHY

1. R. A. Heising, *Quartz Crystals for Electrical Circuits — Their Design and Manufacture*, D. Van Nostrand Co., Inc., New York, 1946.
2. *Information Bulletin on Quartz Crystal Units*, Armed Services Electro Standards Agency, Fort Monmouth, New Jersey, August, 1952.
3. "The Theory and Applications of Piezoelectric Crystals," *The Lenkurt Demodulator*, December, 1955.
4. "Oscillators for Carrier Systems," *The Lenkurt Demodulator*, February, 1956.
5. J. P. Buchanan, *Handbook of Piezoelectric Crystals for Radio Equipment Designers*, Wright Air Development Center, Air Research and Development Command, United States Air Force, Wright-Patterson Air Force Base, Ohio (WADC Technical Report 56-156), October, 1956.
6. D. R. Curran and F. Kulcsar, "Piezoelectric Ceramic Filters and Transducers," Final Report, June 1, 1958 through February 29, 1960, United States Army Signal Supply Agency, Contract No. DA-36-039-SC-78039, March 15, 1960.
7. "Precise Frequency Control," *The Lenkurt Demodulator*, January, 1962.

World Radio History

World Radio History

the Lenkurt Demodulator

VOL. 13, NO. 1 JANUARY, 1964

HYBRIDS

Hybrids, used so widely in modern telephone systems, are really a 'necessary evil.' They provide the means for accomplishing a vital compromise between the superior performance of four-wire circuits and the lower cost of two-wire circuits, but each hybrid introduces some transmission impairments. Since hybrids make the transitions from four-wire to two-wire circuits, a single long-distance call may go through several such junctions — and transmission performance suffers by the sum of their imperfections. This article discusses the operation and limitations of hybrids, how their performance is measured, and the effect of Direct Distance Dialing on hybrid performance requirements.

Although they find many other uses now, hybrids were developed primarily to allow repeaters to be placed in two-wire lines. Since most amplifiers can operate in only one direction, two amplifiers are commonly used at a single repeater point — one for each direction of transmission. Therefore, when a repeater (other than a negative-impedance device) is placed in a two-wire line, a hybrid must be used on each side of the repeater point to provide a short section of four-wire circuitry. Today most long circuits operate over four-wire carrier facilities for their entire length. Such lines do not need hybrids at intermediate amplification points, but they do need hybrids for connection to two-wire drops and switching equipment.

Any device which provides impedance matching between certain circuits

World Radio History

and isolation between other circuits may be referred to as a "hybrid junction," or more commonly as a "hybrid." It may be a three-winding transformer, a resistance bridge, or a waveguide device for microwave frequencies. But in common telephone usage, the term refers to a junction between a balanced four-wire circuit and a balanced two-wire circuit.

For illustrative purposes, a hybrid can be considered as simply a network having four arms, or ports, as shown in Figure 1. The function of the hybrid is to permit signals to pass freely between *adjacent* arms of the network, but to block signal passage between opposite arms. Various types of hybrids attain this in different ways, but the principle is the same. An incoming signal is "split" so that part of the power is applied to each of the adjacent arms. Portions of the signal power are then recombined in such a way that they "cancel" each other; thus, no net power is delivered to the opposite arm.

In a typical telephone arrangement, the transmit and receive branches of the four-wire circuit connect to opposite sides of the hybrid. One of the other connections goes to the two-wire line, and the remaining one goes to a balancing network required in the "cancelling," or "balancing" process.

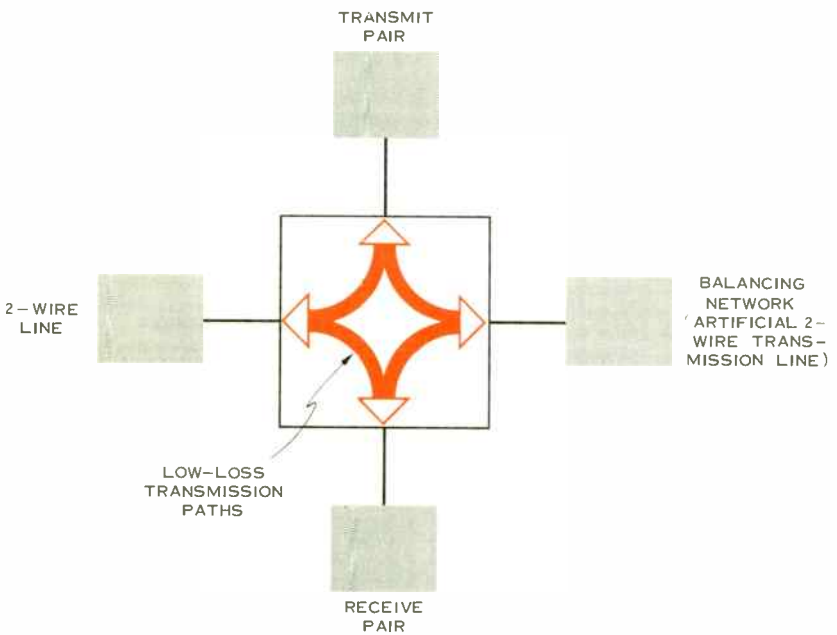


Figure 1. A hybrid may be any network which offers a low-impedance path between adjacent connections, but a high degree of isolation between opposite connections. In telephone practice, it provides the junction between four-wire and two-wire circuits.

World Radio History

Ideally, a hybrid would have infinite loss between opposite sides, thus providing complete isolation between the two branches of the four-wire circuit. At the same time, there would be no loss between adjacent arms — a signal could go unhindered from the two-wire line to the four-wire transmit branch, or from the four-wire receive branch to the two-wire line.

Since such performance is never attained in practice, actual hybrids are judged by how closely they *approach* the ideal. The isolation between the transmit and receive branches of the four-wire line is often called *transhybrid loss*. Since high transhybrid loss is directly related to the balance achieved between opposite legs, transhybrid loss is also known as *transhybrid balance*. The undesired loss between the two-wire line and the four-wire line is usually called *insertion loss*.

If the transhybrid balance is too low, enough of the power from the receive branch of the four-wire line "leaks" across the hybrid to go out on the transmit branch. This power appears at the other end of the four-wire line as an echo. Figure 2 shows how this might occur at a repeater point in a two-wire line. If the transhybrid balance at the east hybrid is too low, a portion of the eastbound signal returns to the west. If the same thing happens at the west hybrid, the echo goes back again to the east. If the loss around this echo loop is greater than the gain of the amplifiers, the echo will die out. But if the gain is almost as great as the loss, the echo may go around the loop several times before vanishing. This produces a "ringing" effect which may give a talker the impression that he is speaking into a rainbarrel, with its hollow, cavernous sound.

If the gain in the loop exceeds the loss, the echo does not die, but builds up and becomes self-perpetuating. In effect, the loop becomes an oscillator in the same way that an amplifier can be made to oscillate by using positive feedback. This condition is known as "singing." It creates a howl in the subscriber's earpiece, usually making the connection unusable.

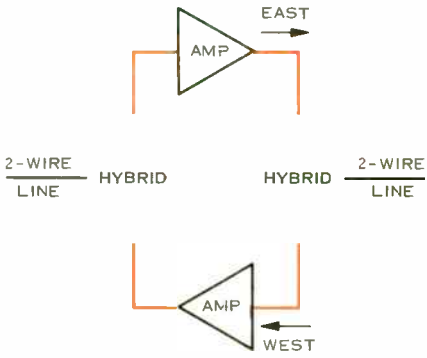


Figure 2. One of the earliest hybrid functions was to provide a short section of four-wire circuitry at a repeater point, permitting one-way amplifiers to be used in a two-wire line. Red line indicates possible "sing path" when some power "leaks" across hybrid.

A hybrid may also produce echoes by reflecting power back down the two-wire line. Such reflections occur at any impedance irregularity; that is, if the input impedance of the hybrid fails to match the characteristic impedance of the two-wire line, some power will be reflected rather than transferred through

World Radio History

the connection. How much power is reflected depends primarily on how closely the hybrid's balancing network matches the impedance of the two-wire line. The quality of this match can be expressed as *return loss*, the ratio (in db) between the transmitted power and the reflected power. (The reflected power may be the sum of reflections from *several* mismatches plus the power which "leaks" across the hybrid despite the transhybrid loss.)

Performance Measurements

From the telephone subscriber's viewpoint, singing and echo are among the more annoying telephone transmission defects. Furthermore, a singing circuit can overload amplifiers or other devices common to more than one circuit, thus degrading the performance of several channels. Singing can also cause crosstalk in adjacent channels.

Since hybrids are intimately involved in the problems of echo and singing, their transmission performance is evaluated in terms of these factors. *Echo return loss* (ERL) refers to an average of return loss measurements made every 500 cps between 500 and 2500 cps. Echo is most noticeable to the subscriber in this frequency range because his telephone receiver is more sensitive to these "middle" voice frequencies. Studies of people's tolerance to echo indicate that louder echoes can be tolerated if they closely follow the original signal. However, an echo of a given magnitude becomes more noticeable as the delay between a signal and its echo increases. Thus, the echo problem is compounded on long lines where propagation time is greater. There are likely to be more impedance irregularities to cause echoes, and because of the distance involved the additional time delay makes these echoes more noticeable. Present minimum ERL objectives for Direct Distance Dialing vary with the type of connection, reaching 27 db for intertoll applications.

Figure 4 shows a typical test set-up for measuring ERL. In this arrangement an oscillator and a voltmeter are connected to opposite sides of a test hybrid (Hybrid A) at the four-wire

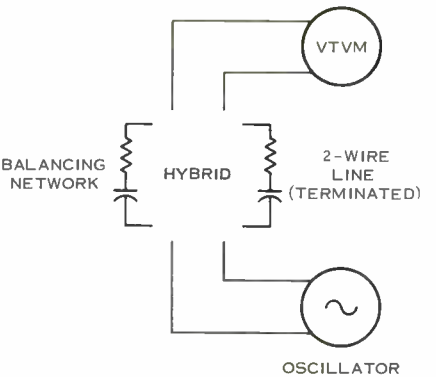


Figure 3. Typical test arrangement for measuring how much isolation the hybrid provides between the branches of a four-wire line. Ideally, no oscillator power would reach VTVM.

connections. The distant end of the two-wire line is terminated by the hybrid under test (Hybrid B). The four-wire connections to Hybrid B are terminated in 600 ohms, while each balancing network consists of 600 ohms in series with 2.1 microfarads. (This is a typical impedance used to simulate the impedance of a two-wire

World Radio History

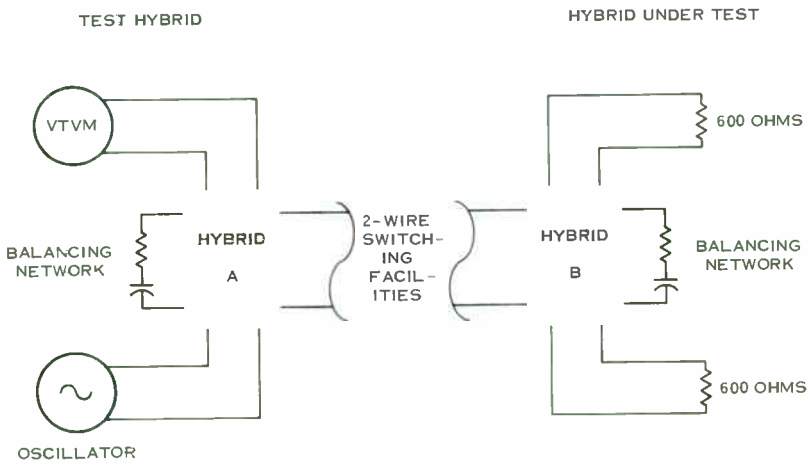


Figure 4. Typical test arrangement for measuring echo return loss. ERL measurement here depends primarily on the quality of the termination which Hybrid B provides for the two-wire line.

line in an office which uses a 600-ohm standard impedance.) Under ideal conditions, none of the oscillator output would reach the voltmeter. However, in any practical situation, some of the power *does* reach the meter. The amount of power depends primarily on whether the impedance of the Hybrid A balancing network is the same as the impedance of the two-wire line when terminated by Hybrid B. Any impedance mismatch between Hybrid A and the terminated line causes power to be reflected directly back into the hybrid.

The tendency of a circuit to sing is measured in terms of the *singing margin*, the net loss around the "sing path." In other words, the singing margin is the gain which must be added to a particular circuit to make it sing. The *critical frequency* is the frequency of lowest singing margin — the frequency at which singing first occurs when gain is added. The critical frequency usually occurs between 250 and 500 cps or between 2500 and 3400 cps — within the passband but outside the echo range. The critical frequency is usually near the "edges" of the voice band because the return loss is lower there.

Transformer Hybrids

The transformer hybrid is one of the oldest and most widely used types. Its operation can be visualized from the diagram of Figure 5. A signal coming into the hybrid at the four-wire receive terminals produces a current flow through the primary winding of transformer 1. The current through the two halves of this winding induces equal

World Radio History

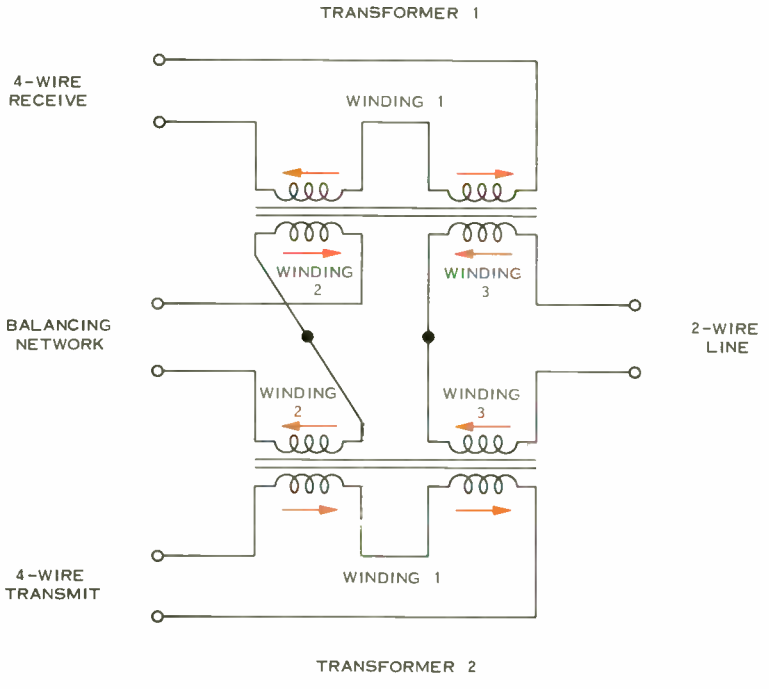


Figure 5. Two transformers may be connected to form a high-quality hybrid. Red arrows show how current flow in four-wire receive branch induces secondary current flow in the two-wire line and the balancing network. No current flows in four-wire transmit branch because opposite potentials are induced in the two halves of the winding, cancelling each other.

currents in secondary windings 2 and 3, which are connected to the balancing network and the two-wire line, respectively. Thus, half the applied power goes to the two-wire line and the other half is dissipated in the balancing network. Since half the power is wasted, the minimum theoretical loss in going through the hybrid is 3 db.

Windings 2 and 3 are connected in series to two corresponding secondary windings of transformer 2. Thus, the current to the balancing network flows also through winding 2 of transformer 2, and the two-wire line current flows through winding 3. Both these currents cause induced potentials in the halves of the transformer 2 primary — but these induced voltages are equal and opposite, effectively cancelling each other. Thus, no current flows in the transmit branch.

This, of course, is the ideal case, which provides complete isolation be-

World Radio History

tween the two sides of the four-wire circuit. In a practical case, several factors make the ideal unattainable. Both transformers must be carefully constructed for the voltages involved in the final "cancelling operation" to actually be equal and opposite. A slight difference in the windings, for example, would result in a net difference between the voltages, producing a current through the transmitter.

But such a current could result even if the transformer windings were perfectly matched — because the impedance of the balancing network must also match that of the two-wire line. If these two impedances are not the same, different currents will flow and different voltages will be induced in winding 1 of transformer 2, thus preventing full cancellation. The result is the same as though the transformer windings were not matched. Through careful construction of the transformers and matching of the balancing network to the two-wire line, however, satisfactory isolation can be obtained between the sides of the four-wire circuit.

For transmission in the other direction, the signal enters the hybrid from the two-wire line and is split equally between the transmit and receive branches. This current flow in the primary windings of both transformers induces voltages in the number 2 secondary windings of both transformers.

Figure 6. Two transformers in foreground, connected as shown in Figure 5, are the heart of this four-wire terminating unit. Pencil points to balanced attenuator pad (microcircuit package) which permits loss in both transmit and receive directions to be adjusted by front-panel strapping.

World Radio History

But these induced voltages are equal and opposite, so no power goes to the balancing network. Again, the minimum loss is 3 db because half the power is wasted in the receive branch. In practice, the loss in either direction of transmission is likely to be about 3.5 db because of transformer core losses and winding resistance.

Resistance Hybrids

A hybrid can also be made of a resistance network in the form of a Wheatstone bridge. Figure 7 (A) shows a resistance hybrid as usually drawn, while in Figure 7 (B) the same hybrid is redrawn to illustrate the bridge configuration more clearly.

Consider a signal applied to the receive terminals and assume that both R_1 and R_2 are equal to the resistances of the two-wire line and the balancing network — (typically these would all be 600 ohms). Signal power is then split, with half going to R_1 and the balancing network and the other half going to the two-wire line and R_2. Points A and B are at the same potential so no power flows to the transmit branch. Furthermore, since the power dissipated in the resistors and the balancing network is wasted, the two-wire line receives only 1/4 of the power. In other words, the resistance hybrid has a minimum loss of 6 db when all arms of the "bridge" are equal.

For transmission in the other direction, consider a signal applied at the terminals of the two-wire line. This signal is divided equally among the transmit and receive terminals and their associated resistors, R_1 and R_2. No power goes to the balancing network because points B and C are at the same potential. Since that portion of the signal applied to the resistors and to the transmit branch is wasted, the receive branch gets only 1/4 of the power and again the minimum possible loss is 6 db.

The foregoing discussion has assumed that all arms of the bridge have equal impedance, but this is not a necessary condition for hybrid balance. The requirement for balance is only that the product of resistances R_1 and R_2 must equal the square of the nomi-

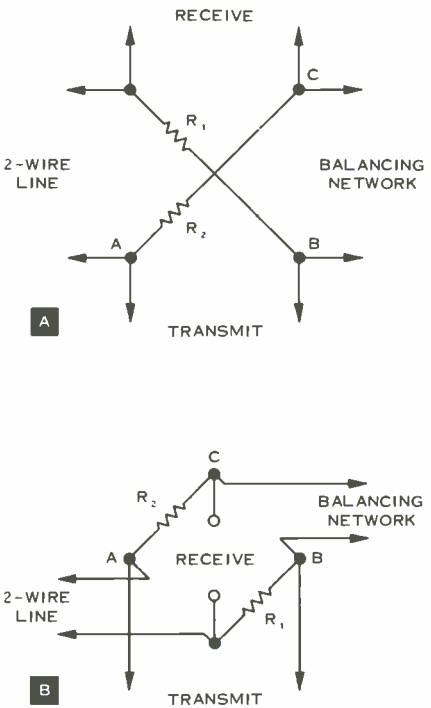


Figure 7. Resistance hybrid as usually drawn (A), and redrawn to illustrate bridge configuration (B). Although less expensive than transformer type, resistance hybrids introduce more transmission loss.

World Radio History

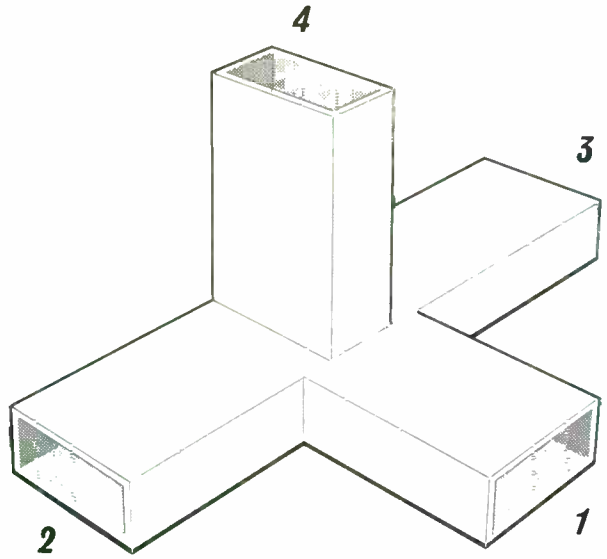


Figure 8. "Hybrid tee," used for microwave frequencies, provides isolation between opposite arms, but maintains low-loss path between adjacent arms. A signal in arm 4 produces equal and opposite signals in arms 2 and 3, nothing in arm 1. With input to arm 1, equal in-phase signals exist in arms 2 and 3, with no coupling to arm 4.

nal hybrid impedance. Stated mathematically,

$$R_1 R_2 = Z_0^2.$$

This means that if R_1 is made small and R_2 is made large, loss in one direction of transmission can be made smaller than 6 db. But this small loss must be paid for by a correspondingly larger loss in the other direction. Lower loss in one direction can also be obtained in a transformer hybrid, but the price is the same — higher loss in the other direction.

Comparison of Hybrid Types

The decision whether to use a transformer or a resistance hybrid may depend on several factors, but in some cases there may be literally no choice. For example, if the circuit is to be used at frequencies much higher than 1 Mc, a transformer hybrid is not likely to be seriously considered. Such factors as iron loss in the transformer core and interwinding capacitance severely limit the high-frequency performance of transformer hybrids, while resistance hybrids are nearly independent of frequency — at least until the microwave region is approached. (For microwave applications, a waveguide device such as the "hybrid tee" shown in Figure 8 would be used).

On the other hand, if the proposed application is, say, a voice-frequency repeater connection, the choice will probably be a transformer hybrid, simply because its loss is so much lower than that of the resistance hybrid. With a hybrid on each side of a repeater in a two-wire line, the two resistance hybrids would increase the transmission loss by 6 db — the equivalent of perhaps 100 additional miles of open-wire line.

If physical size and weight enter the problem, the resistance hybrid has an

World Radio History

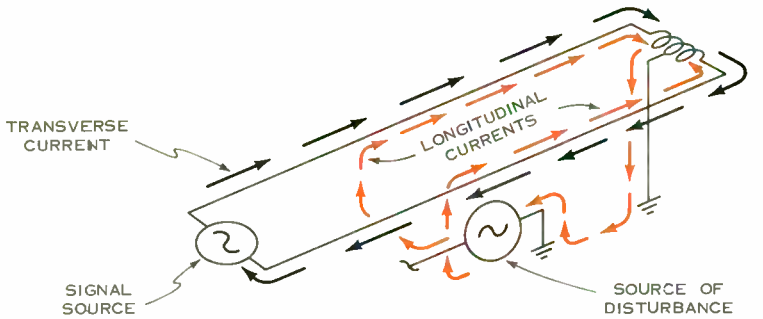


Figure 9. In a perfectly balanced line, longitudinal currents (red) cancel out in transverse circuit, but seek a ground return. Any imbalance converts some longitudinal current to transverse, producing interference.

advantage over the transformer type. Iron-cored transformers are heavy and considerably bulkier than resistors.

Cost is nearly always a factor in the choice of a hybrid. A resistance hybrid is less expensive than one which uses transformers, but the resistance type also has inherently more loss. Hence, in many cases, the cost of the hybrid itself must be balanced against the cost of additional gain.

Longitudinal Balance

Another factor which should be considered in choosing a hybrid is "longitudinal balance." In a conventional telephone system, speech or carrier signals are carried on a "balanced" transmission line consisting of two conductors at the same electrical potential above ground. Normal signal current flows in opposite directions in the two conductors, but interference often produces *longitudinal* currents which flow in the *same direction* in both conductors (as shown in Figure 9). These longitudinal currents tend to cancel each other, but they seek a path to ground. If any imbalance exists, these currents do not cancel, and interference with the desired signal results. If the ground path is through the primary winding of a hybrid transformer, imbalance can be introduced by any difference between the halves of the winding (for example, a different number of turns in the two halves). Not only does this unbalance the line on the primary side, but the interfering currents induced in the secondary windings do not cancel either. Thus, "longitudinal balance" applied to a transformer hybrid is a measure of how well the transformer resists interference.

Longitudinal balance is often measured in db. When, for example, a figure of 60 db is specified, it means that the net interference, (which is not cancelled by the transformer balance) is 60 db below the level of the desired signal. The longitudinal balance can also be measured in ohms (called Z_u, for unbalanced impedance). Here it is a direct measure of the impedance imbalance between the halves of the winding. A Z_u figure of 0.5 ohm is usually

World Radio History

considered good for a 600-ohm impedance.

If a resistance hybrid is used, any unbalance of the two-wire line is presented to both sides of the four-wire circuit. In many cases isolation transformers would be required to correct this situation. Therefore, a transformer hybrid is often the easier solution.

Impact of Nationwide Dialing

The increased incidence of long-distance calls, particularly with the advent of Direct Distance Dialing, has changed the performance requirements for much of the telephone plant. With the subscriber dialing his own connection thousands of miles away, no operator is present to check the quality of the circuit and perhaps use another if the first one is bad. This means that all possible connections must be good. Thus, almost without exception, performance requirements have become more stringent. Certainly this is true of the requirements for hybrids.

The system of random interconnection used in the United States and Canada permits as many as seven toll links (eight on calls between the two countries) to be connected in tandem — in addition to two terminating links (tool-connecting trunks). The significance of this, in terms of hybrid requirements, is that there are many more opportunities for impedance irregularities to produce echoes, and a multitude of "sing paths." Therefore, it becomes particularly important to maintain the minimum values of echo return loss and singing margin for the random-interconnection system to consistently meet the objectives of nationwide dialing.

The increasing use of four-wire transmission facilities means that hybrids are no longer used in some of their "traditional" applications. Hybrids used with repeaters in two-wire lines are seldom used now because few new long circuits are two-wire, and those that are often use negative-impedance repeaters. Most switching facilities, however, are still two-wire — simply because of the expense of four-wire switching. So hybrids must be used to connect the four-wire circuits to the switching equipment.

A far-off ideal would be to use four-wire circuitry exclusively from drop to drop. This would eliminate many of the transmission problems now encountered. But such a system is not in the foreseeable future. In the meantime, hybrids form a vital and integral part of the modern telephone plant, and their performance requirements are becoming more stringent. •

BIBLIOGRAPHY

1. "Operation and Application of Hybrid Junctions," *The Lenkurt Demodulator;* September, 1953.
2. *Transmission Systems for Communications,* Bell Telephone Laboratories; New York, 1959.
3. *Notes on Distance Dialing,* American Telephone and Telegraph Company, 1961.
4. J. C. Mau, "Terminal Balancing of Independent Telephone Offices," *Automatic Electric Technical Journal;* April, 1963.
5. R. W. Ruth and J. S. Stuehler, "Four-Wire Terminating Unit Uses 'Thin-Film Pads,'" *Automatic Electric Technical Journal;* January, 1964.

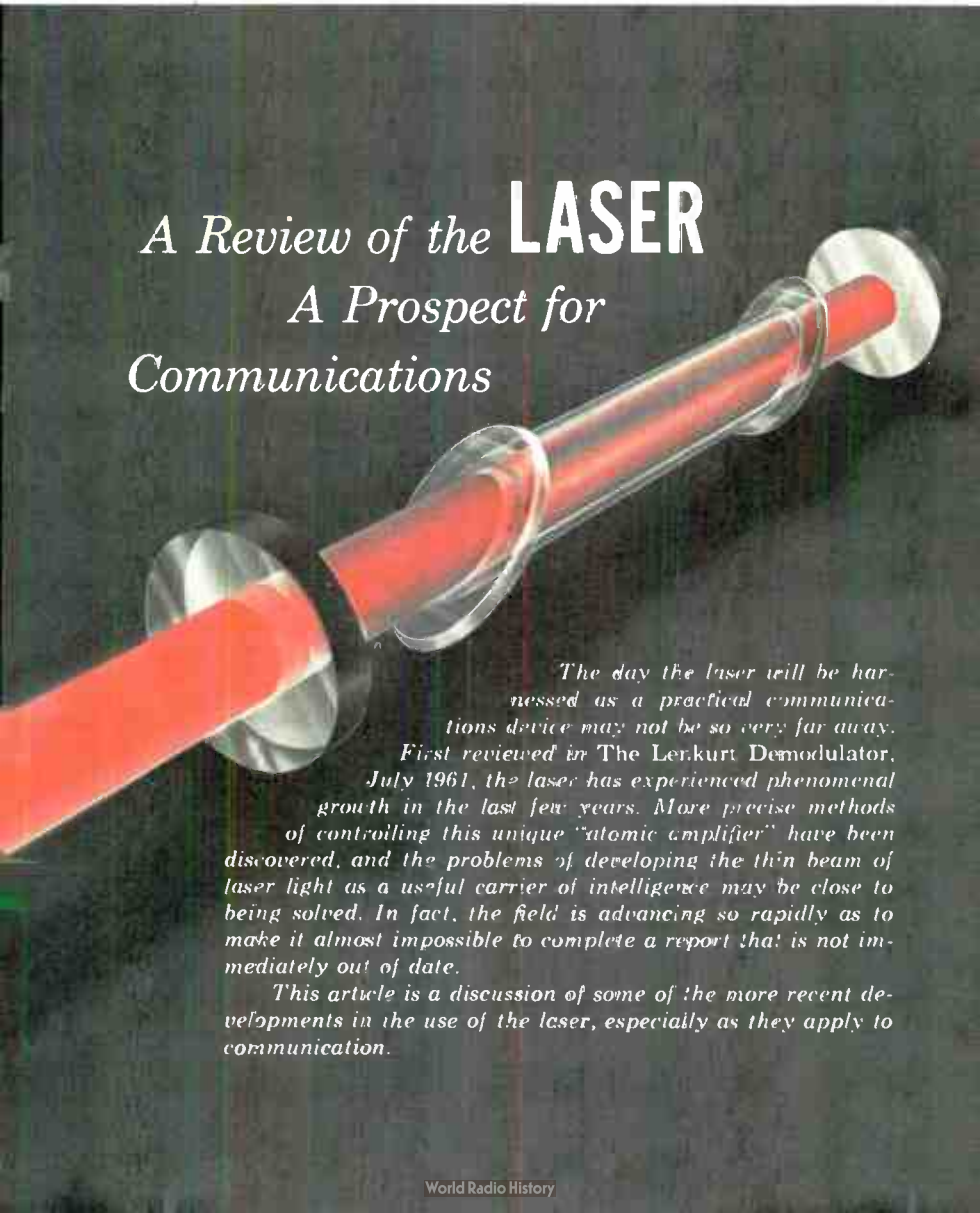
World Radio History

World Radio History

the Lenkurt

Demodulator

VOL. 14, NO. 11 NOVEMBER, 1965



A Review of the LASER
A Prospect for Communications

The day the laser will be harnessed as a practical communications device may not be so very far away. First reviewed in The Lenkurt Demodulator, *July 1961, the laser has experienced phenomenal growth in the last few years. More precise methods of controlling this unique "atomic amplifier" have been discovered, and the problems of developing the thin beam of laser light as a useful carrier of intelligence may be close to being solved. In fact, the field is advancing so rapidly as to make it almost impossible to complete a report that is not immediately out of date.*

This article is a discussion of some of the more recent developments in the use of the laser, especially as they apply to communication.

World Radio History

Courtesy of Spectro-Physics

Figure 1. Commercial helium-neon laser receives final inspection from optical technician. Sensitive mirror adjustments tune the laser, while output is monitored with a photocell power meter.

The laser is basically an oscillator operating at light frequencies. But instead of tapping the energy from a stream of electrons as in the common vacuum tube, the laser stimulates the emission of stored energy from the atom itself. An acronym for light amplification by stimulated emission of radiation, the term *laser* is accepted throughout most of the world to describe the action. Practitioners have even coined a new verb, *to lase.*

Common laser types may be catagorized as crystal, gas, liquid and semiconductor, each having properties identifying it with certain uses. A crystal laser, such as the ruby, can be pulse-operated at very high power outputs, measured in megawatts. Gas and the more recent liquid lasers provide less power, 10 to 20 watts, but are more suited for continuous operation. They also offer the advantage of single frequency operation. Semiconductor lasers are very small, operate at low power, but boast very high efficiency. Crystal and gas lasers may currently be purchased as off-the-shelf items, and are being pushed into service in a variety of fields.

Referred to by some as "an answer looking for a problem," the laser exhibits a number of qualities just now finding application in the fields of medicine, industry, physics, chemistry, optics, and electronics. One of the most

World Radio History

pronounced of these qualities is the laser's ability to produce coherent light —of particular interest to the communicator wishing to find a new carrier for increasingly huge volumes of information.

Communicating with Light

The concept of communicating with light is certainly not a new one. For hundreds, if not thousands of years, men have met the need to "talk" over long distances by using light. In early history a torch served as a very effective semaphore. Flashing heliograph mirrors reflect the sun's brilliance for many miles, and the lighthouse certainly communicates a most urgent message. (Paul Revere's "one if by land, two if by sea' lantern code, incidentally, was not only an early use of light for communication, but illustrative of the binary code in primitive form.)

As early as 1880 Alexander Graham Bell transmitted voice by light, but only for a short distance. A variation of Bell's experiment is often used as a demonstration for budding high school scientists. Similarly, the sound tracks on many motion pictures today are produced on film by controlling a narrow source of light. But one major handicap prevents ordinary light from becoming a practical message carrier. It is *incoherent* light!

Light from common sources is very unsystematic, like the waves from a handful of pebbles thrown on a pond. An ordinary light bulb emits a literal jumble of light waves, completely at random and at different frequencies (or colors). Only the most gross pieces of information can be transported by these waves. The high school science demonstration, for example, does not attempt to tamper with the frequency of light, but merely varies the intensity of the beam proportional to the amplitude of an audio signal such as voice or music.

Coherent Light

Laser-produced coherent light is extremely stable and precise in frequency, with waves exactly in phase with each other. Like the waves from a single pebble dropped in the water—or more accurately like ranks of marching soldiers all in step—these coherent light waves are now potential information carriers. Just as one soldier out of step is immediately obvious among disciplined marchers, a laser light wave may be disturbed (modulated) and detected at another point (demodulated) in a very orderly manner.

In addition to excellent phase coherence, the beam leaves the laser as a "wall" or flat plane of light exactly parallel to the laser's face. The beam will diverge so very slightly that even in the earth's thick atmosphere it theoretically spreads out only about an inch per mile. And in space, where there is no dust or water vapor to deflect the laser beam, it conceivably would reach the surface of the moon, over 250,000 miles away, as a circle only a half-mile in diameter.

Of equal if not more interest to the communicator is that laser light contains only one frequency—and a very high frequency at that. The center of the visual spectrum is about 10^9 megacycles, or about 160,000 times higher than the 6 gc microwave band. Since available bandwidth increases with frequency, a tremendous bandwidth is available at light frequencies. Currently, laser modulators are being developed with bandwidth capabilities of 1 gc and higher. How much more bandwidth is possible remains for the experimenter to find out. But compare this with a common microwave system

World Radio History

carrying 960 voice channels on a bandwidth of 4 mc. Even with current techniques, the laser's bandwidth is 250 times this size.

Stimulated Atoms

The inherently unique properties of laser light have their origin in basic atomic theories governing electrons, photons, and their energy levels. The same facts apply to the *maser,* or microwave version of the laser, but for clarity only the visible light spectrum will be used in this explanation.

Atoms in nature are usually in a relatively undisturbed or *ground* state. The energy of orbiting electrons is balanced with the energy in the nucleus of the atom. These electrons occupy specific orbits determined by their own energy, but when "excited" by an outside source of energy, may jump to a higher second level raising the total energy of the atom (Figure 2). In lasers, adding this outside energy is known as *pumping.*

The excited state for the atom is unnatural and it will tend to relax to its ground state. As this happens, the stored energy is dissipated by emitting a photon of radiant energy. The energy of the photon is exactly proportional to its frequency—the higher the energy, the higher the frequency.

The common neon tube is an example of this action. Molecules of gas are excited to upper energy states by high voltage. As the atoms drop to their ground state, they emit light of a characteristic color or frequency—various gases produce various colors.

If left uncontrolled, the atom's spontaneous relaxation occurs in a random manner and results in incoherent light. But during the period when the atom is still excited, it is possible to stimulate the drop to ground level by striking the atom with an outside photon of the same energy it would have otherwise emitted spontaneously. Relaxation is no longer random, and the emitted photons leave the system as coherent light.

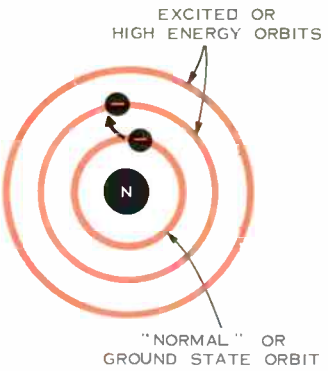


Figure 2. Atomic electrons normally occupy an orbit around the nucleus, representing a certain fixed energy level. A stimulated atom acquires energy as one of its electrons jumps to a higher level.

Another remarkable feature of laser action results when an emitted photon strikes another excited atom within the laser. As that atom returns to its ground state, another photon is added to the stream exactly in phase with the first, producing amplification.

Various methods have been discovered to improve the efficiency of the lasing action. The *three-level* method pumps atoms not to the second energy level, but to a still higher third level (Figure 3). The atoms are very unstable at this level, and quickly fall back to the intermediate, or second level. Here, by the nature of the laser material, the atoms tend to accumulate and are available in greater numbers for outside stimulation. Cooling the material, say to liquid nitrogen tempera-

World Radio History

ture (70°K), increases the effect. These techniques are typical of crystal lasers such as ruby. Gas lasers of the helium-neon type rely on the difference in energy levels between the two gases to provide for more effective pumping.

The Wave Grows

As emitted photons of light travel along the laser tube bumping into more excited atoms, the light wave continues to grow (Figure 4). These waves are reflected back and forth inside the tube by mirrors, one of which is slightly transparent. Forming a resonant cavity at light frequencies, the laser now builds up standing waves which continue to multiply on each pass through the cavity. When the gain is strong enough to overcome the loss in the mirrors, an intense beam is emitted from the partially transparent mirror. As the light bounces back and forth, any waves moving at angles to the axis between the mirrors will soon leave the system through the walls of the tube. Therefore, the output beam of the laser will be extremely parallel.

Because of the coherence of laser light, it is possible to construct extremely efficient optical lenses to further focus the beam. A laser can be focused into a spot no wider than a wavelength of light, or about 0.0001 cm. The result is intense heat at the focal point, useful as a precision cutting tool; for micro-welding; in delicate surgery, such as eye operations; and in chemistry, where individual molecules may be subjected to the unique radiant energy. The laser should also provide an invaluable laboratory tool for research in optical physics.

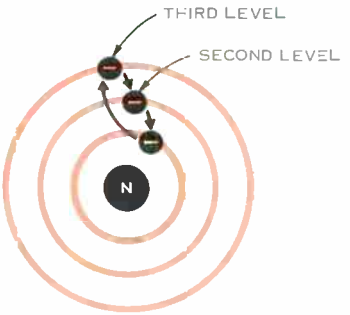


Figure 3. Atoms can be pumped to higher third level, allowed to fall back and accumulate at second level, then controlled in their return to ground state. More atoms available at second level improves laser action.

Communications Problems

The communications industry, faced with an already overflowing radio spectrum and the ever increasing demand for more and more communications service, quickly became interested in the laser's ultra-high information carrying potential. But engineers found that before the laser could compete with microwave radio for point-to-point

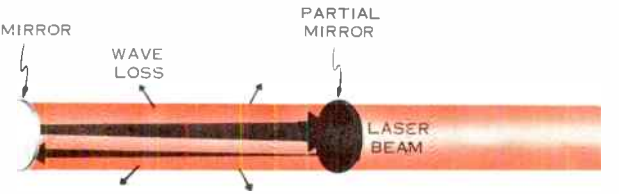


Figure 4. Growing photon stream bounces back and forth, emerging as brilliant, coherent light. Extraneous waves are lost through laser walls.

World Radio History

communications on earth, many problems remained to be solved.

The most serious of these problems is finding a suitable transmission path. The earth's atmosphere presents many natural barriers to light—haze, rain, snow—which do not seriously affect microwaves. Several studies have been undertaken to establish path reliability over relatively short ranges. To date, television pictures have been transmitted over a few miles without serious losses, but any system matching microwave reliability and quality over 25-mile links seems improbable with current techniques.

A number of *covered* transmission paths have been proposed, including the use of fiber optics, shiny hollow tubes, and tubes with regularly spaced lenses to re-direct the beam around mild corners. Among these is a unique suggestion for a continuous gas-lens tube. A simplified arrangement (Figure 5) would pump a steady stream of cool gas into a warm tube. The heated gas near the surface of the tube would be less dense than the cooler gas at the center. This difference, represented by a varying refractive index, would be enough to produce a positive lens for the laser beam.

The slight divergence of the laser's beam—advantageous for keeping energy concentrated theoretically over vast distances—does become a problem in transmitter to receiver visual-path alignment. Experimenters have plotted the expansions and contractions of a building caused by heat from the sun by monitoring changes in signal strength between a laser mounted in the building and a receiver nearby. In the communications field this could be a definite problem. But to others this sensitivity to angular change makes the laser an extremely good device for measuring minor physical displacements. For instance, a system using lasers has been proposed to record instantaneously the land shift around California's infamous San Andreas fault. The same capability is being used to detect micromovement in laboratory experiments.

The laser's prime contribution to communications may come in space, where interference from a *dirty* atmosphere ceases to be a problem. Optical links from known positions, such as earth-orbiting space stations and the moon, could carry tremendous quantities of information on a single laser channel. Tracking of vehicles moving freely in space could be more of a problem for the thin-lined laser beam. But techniques developed here could also apply to laser radar, producing a highly sensitive system with greater resolution than ever before possible. Military applications now being tested include a laser fire control radar system to permit low-flying aircraft to see

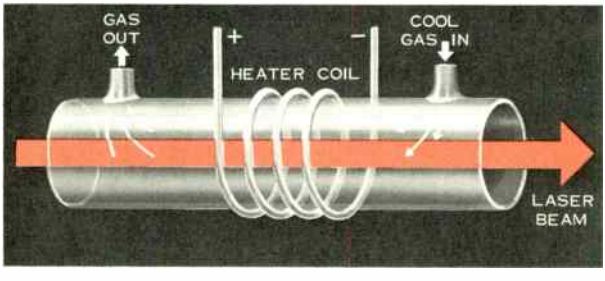


Figure 5. Cool gas forced through a hot tube can form a continuous positive lens, suggested as a means of "piping" laser beams between cities.

World Radio History

targets normally obscured by ground clutter on microwave radar.

The basic component needs of a laser communications system are the same—in name at least—as any similar radio device. The information to be transmitted must be amplified, modulated, demodulated, and recreated in its original form. The techniques needed here are not all new. For example, earlier developed masers operating at microwave frequencies have for some time been employed as amplifiers in radio astronomy because of their ability to provide high gain and very low noise.

Modulation

Apart from the extensive pure research being done with lasers, considerable effort is being put into finding suitable modulators and demodulators. Modulation of the relatively new semiconductor or injection-type laser is easily accomplished by simply controlling the pumping current. However, semiconductor lasers, such as gallium arsenide (GaAs), have a low output power and are not as coherent as other sources. Also, the output is a less desirable flat sheet of light as opposed to the solid round beam of other lasers. For the present, communicators are putting more faith in gas and crystal lasers for communications purposes, with further study being given liquid lasers.

One of the first successful devices for amplitude modulating a laser beam uses the polarization properties of the clear crystal potassium dihydrogen phosphate (KDP). As illustrated in Figure 6, the KDP device amplitude modulates a laser beam projected through it. The first polarizer blocks all light wave polarizations except, for example, the vertical.

The resulting beam may be thought of graphically as a number of ribbons of light, all parallel to each other and at right angles to the direction of propagation. For simplicity of illustration, this example treats the polarized light as only one *ribbon*. It is characteristic of the KDP crystal to shift polarization

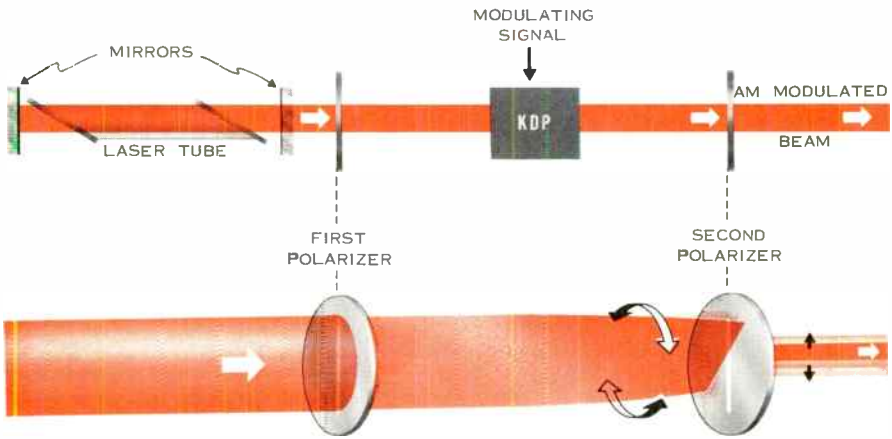


Figure 6. As beam passes through an AM laser modulator, it may be visualized as a ribbon of light twisting proportional to a signal applied to the KDP crystal.

World Radio History

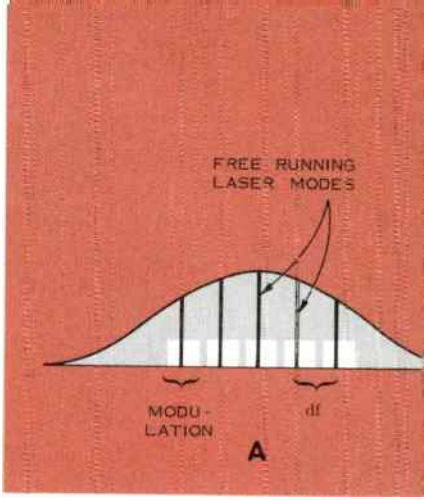


Figure 7. Laser modes representing the (A) free-running laser, (B) laser with FM modulator inside the cavity, and (C) super-mode laser, where power in the FM sidebands is combined into one output frequency

in a circular direction proportionate to a stimulating voltage. For a higher voltage, there will be more circular change in polarization. If a signal is applied to the KDP crystal and the now vertically polarized laser beam shone through it, the beam will be what might be called *polar modulated*. In the diagram, the ribbon will be twisted in accordance with the modulating signal. The second polarizer, known as the analyzer, will sense this twist as a decrease in amplitude. The output intensity will then vary in relation to the signal, hence, amplitude modulation.

Bandwidth in the order of 200 mc can be achieved with amplitude modulation, but this by no means takes full advantage of the capabilities of coherent laser light. On the other hand, frequency and phase modulation methods are producing bandwidths of over 1 gc. One such device, known as a wideband traveling wave phase modulator, consists of two parallel brass rods about one meter in length, with an electro-optical material (such as KDP) sandwiched between. A microwave signal voltage applied to the device varies the velocity of light in the crystal, resulting in phase modulation. The length of the modulator allows longer interaction time between signal and light beam, and hence greater depths of modulation.

Laser Modes

It should be noted that since the laser cavity is thousands of times longer than any wavelength at light frequencies, a number of frequencies will resonate in the tube at the same time. This results in the laser having in its output a number of separate and distinct frequencies or modes (Figure 7). The separation of these modes is determined by the mirror placement in the laser, and may be calculated by the formula:

$$df = \frac{c}{2L}$$

where:

df = *the difference frequency between modes*

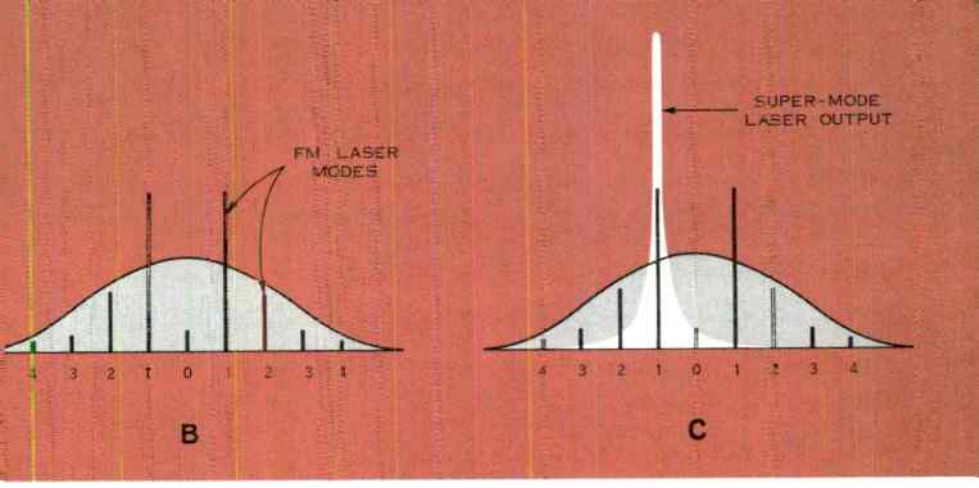
c = *speed of light, and*

L = *length between the mirrors.*

Since the obvious desire is to transmit only one frequency, power distributed in modes other than the one to be used is wasted. Likewise, each mode acts as a carrier frequency for any modulation. As sidebands are added to each mode (Figure 7A), it can be seen that the bandwidth of modulation on any one mode is limited by the difference in frequency between the modes.

One solution is the super-mode laser (Figure 8). By inserting a phase modu-

World Radio History



lator inside the laser tube, driven at a frequency nearly equal to the difference frequency between modes *(df)*, the output is converted to a typical FM configuration with sidebands occupying the positions formerly held by the various modes (Figure 7B). By using this method, the bandwidth limitation in the mode structure has been eliminated. Another phase modulator outside the tube will additionally affect the beam by compressing all the modes together into a single frequency. The final output contains most of the power formerly held in the many modes, plus the highly desirable single frequency (Figure 7C). This "super-mode" beam can be successfully modulated by any chosen method with much superior performance.

A similar means of arriving at the same end is found by placing a device known as a *Fabry-Perot etalon* (not shown) inside the tube, along with the FM modulator. The output frequency is determined by the spacing of two highly reflective mirrors forming the etalon, creating a resonance at the desired frequency. The beam leaving the tube will be of single frequency if the etalon is

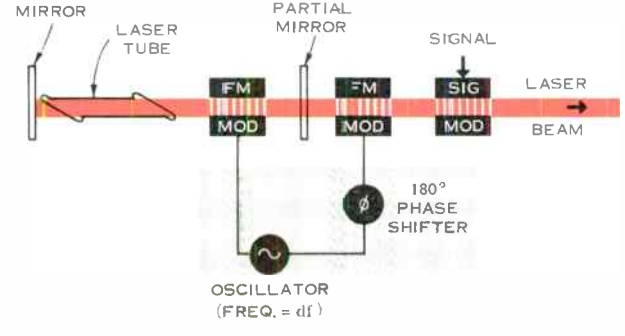


Figure 8. Supermode laser has single frequency output of comparatively high energy. Oscillator frequency is nearly equal to the difference frequency between free-running laser modes.

World Radio History

tuned to one of the FM sidebands, and will be at the laser's full power. The device combines the power of the other sidebands into the output. This can be done because the FM laser modes are sidebands of a single carrier, rather than a set of independent oscillations.

Demodulation

Demodulation of optical radiation is typically accomplished with either a photomultiplier tube or a microwave phototube, each relying on the secondary emission of electrons from a cathode when struck by light photons. The photomultiplier technique redirects emitted electrons onto other secondary-emitting surfaces, producing considerable amplification. The current is eventually collected on an output electrode. The photomultiplier tube has a range from d-c to many megacycles, thereby detecting signals directly to baseband frequencies. The microwave phototube (Figure 9) is designed with a traveling wave tube helix output, and is effective at the higher microwave frequencies. A modification of the microwave phototube, known as the crossed-field electron multiplier, amplifies the signal before the electrons reach the helix. In both cases, since light frequencies are outside the bandwidth capabilities of the phototubes, the electron stream represents only the original modulation placed on the laser beam.

Optical heterodyning is also possible using the photomultiplier tube, as seen in Figure 10. A laser local oscillator beam beats with the incoming laser signal in the phototube, resulting in an IF frequency equal to the difference between the two light frequencies. This IF signal is typically in the microwave region and may be amplified and demodulated by conventional methods. A discriminator supplying a control signal to the laser local oscillator maintains frequency stability.

Other Applications

Interest has been shown in using lasers, possibly of the semiconductor type, for inter-component communication in high speed computers. The technique would eliminate the delay asso-

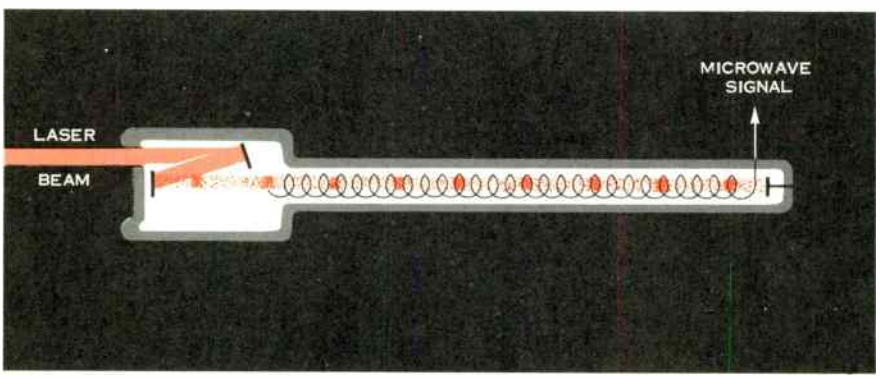


Figure 9. The microwave phototube converts laser light to electrons, bunched proportionate to the original modulation. The current induced in the helix is then processed by usual microwave techniques.

World Radio History

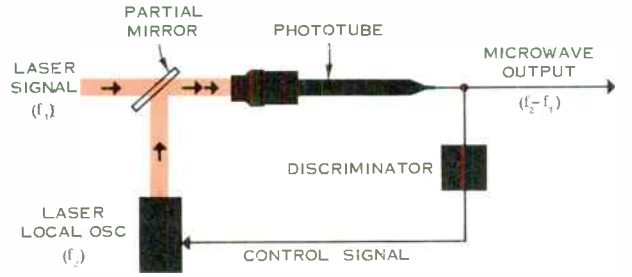


Figure 10. Optical heterodyning combines laser signal with that of laser local oscillator at phototube. Microwave output at IF frequencies is then amplified and demodulated.

ciated with placement of components and connections, and would offer much faster switching speeds than now possible with ordinary electronic computer circuits.

Closer to the communications field, coherent laser light has opened a "magic window" to successful wavefront reconstruction photography, known as *holography*. Holograms record both amplitude and phase variations of laser light reflected from a subject, allowing near-perfect three-dimensional reproduction. Interference patterns between the reflected light and a reference beam are recorded on film without the use of lenses. Applications of the future may include three-dimension color television, and medical or industrial X-ray holograms for studying the interior of an object.

Conclusion

While the laser is becoming a valuable tool in many pure-science areas and may have useful military and industrial applications, its high information-carrying potential for communications seems barely tapped. Techniques for the use of the laser in communications are gaining in sophistication, but a great number of obstacles must be overcome before a practical system becomes feasible. Only a continuing refinement of the art will take the laser out of the laboratory and into the field.

BIBLIOGRAPHY

1. Brotherton, M., *Masers and Lasers,* McGraw-Hill, 1964.
2. Gordon, J. P., "Optical Communication," *International Science and Technology,* No. 44 (August, 1965), pp. 60-64.
3. Leith, Emmett N., and Juris Upatnieks, "Photography by Laser," *Scientific American,* Vol. 212, No. 6 (June, 1965), pp. 24-35.
4. Reed, John S., "The Wonderful World of the Laser," *Telephone Engineer and Management,* Vol. 68, Nos. 6, 8, 9 (March 15, 1964 pp. 43-47, April 15 pp. 64-67, May 1 pp. 40-43).
5. Schawlow, Arthur L., "Optical Masers," *Scientific American,* Vol. 204, No. 6 (June, 1961), pp. 52-61.

World Radio History

World Radio History

the Lenkurt Demodulator

VOL. 11 NO. 1 JANUARY, 1962

Precise Frequency Control

Modern communications depend as never before on frequency and time standards of remarkable accuracy and stability. Precise time-keeping and extremely accurate frequency control are actually the same problem, since the accurate measurement of either time or frequency implies exact knowledge of the other. Although the basic principle of accurate timekeeping has been known for centuries, modern techniques have improved accuracy and stability many thousands of times. This article discusses some of these methods and how they are used in modern communications.

The precision demanded by a civilization or technology in measuring time and frequency is a good indicator of its state of advancement. Not very long ago, when communication as well as travel depended on ocean winds or beasts of burden, errors of a few minutes or even a few hours in a day mattered very little.

As the world and its affairs have grown more complicated, our need for better timekeeping has become much greater. Today our best time and frequency standards appear to vary only about one part in 10^{11} per month; if this error were to add up steadily in one direction (rather than occasionally cancelling), a clock controlled by such a standard would show less than one second error after more than two and a half centuries! More workaday quartz crystal oscillators have been built which drift less than one part in 10^9 per month, or one second in two and a half years.

Although such stability seems extraordinary, it is not just precision for the sake of precision, or mere technical "showmanship." Many new communications methods now proposed will require time standards at least this good

World Radio History

in order to function adequately. For example, the telephone network of the not-too-distant future is envisioned as transmitting many thousands of telephone and data channels through waveguide and other transmission means, in the form of millions—even billions—of pulses per second, and using time-division multiplex to keep track of each channel and its information. As the number of channels is increased (which is economically very desirable), the tiny intervals separating the pulses necessarily become shorter and are squeezed closer and closer together. Only if both the transmitting and receiving terminals operate at exactly the same frequency, and exhibit essentially no frequency drift can the information be recovered accurately. For large numbers of channels or long distances, it is very likely that this objective can be realized more easily by very stable free-running oscillators than by using the transmitted signal for synchronization.

Another application in which extremely precise frequency control will be required is a new type of international communications service using earth satellites. Frequency stability on the order of two or three parts error in 10^9 will be necessary in order to prevent distortion of the single-sideband, suppressed microwave carrier transmissions from the ground stations to the satellite, and this tolerance would have to be met by all the stations communicating through the satellite. This results from the fact that the 6000 mc radio carrier would not be transmitted, but would be reinserted at the satellite, and would have to match the carrier frequency of the ground stations within a very few cycles in order to avoid objectionable distortion.

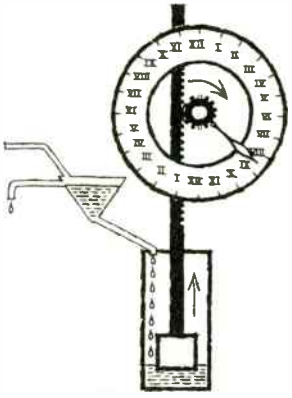


Figure 1. Early "clocks" used steady flow of liquid to mark time intervals. Device shown is refinement of simple graduated vessels from which water slowly escaped.

This is the counterpart of the problem in conventional frequency-division multiplexing systems which suppress the carriers of individual channels and groups of channels, but restore them at the distant terminals. As in the case above, serious distortion may result if the carriers provided at transmitter and receiver are not within a very few cycles of each other. Although this may be overcome by synchronizing the system with some sort of reference signal or pilot, it does not, however, eliminate the noise and distortion which results if the transmitted channels have drifted somewhat and no longer quite match the passband of the receiver filters. Frequency drift is more likely to be a problem in synchronous systems because oscillators of relatively low stability can be used, since they are easier to "pull" into synchronism and because they are cheaper.

World Radio History

Earlier Methods

From the earliest times, man has apparently been concerned with measuring time and dividing it into intervals to help him regulate his activities. Primitive methods of keeping time were based on the rising and setting of the sun and moon, and on noting the position of shadows of fixed objects. The sundial, however, was of very limited value, since it required sunshine and was therefore useless at night and in bad weather.

Early attempts to bypass this limitation were always based on the continuous flow of some medium—water or sand as a rule. Water clocks, or clepsydra as they were called, and the similar sand clocks marked the passage of time by passing the medium from one vessel to another in a regular way that could be used to define specific intervals. Similarly, the rate at which special candles or tapers burned also provided a rough indication of time intervals.

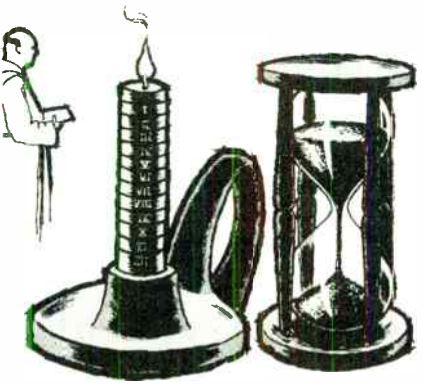


Figure 2. Other ancient methods of keeping time include calibrated candles and the hour glass, an offshoot of the clepsydra or water clock. Neither was noted for convenience or accuracy.

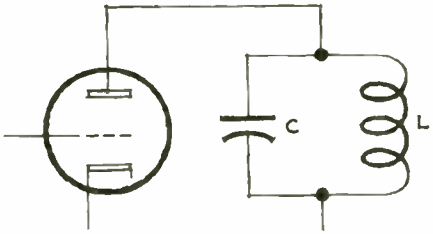
Around the year 1360 A.D. a major improvement in the basic design of timepieces appeared; the escapement, a means of changing a steadily-applied force into a reciprocating, periodic motion, soon became the basis for almost all timekeeping devices. These clocks were inherently faulty, however, since they had no basic time "consciousness." The rate at which they operated was entirely dependent on how much force was applied, the inherent loss or friction in the system, and the moment of inertia of the parts. Only the last item could be depended on to remain constant. Increased force would speed up the escapement rate, while bad lubrication would slow it down. Despite these shortcomings, however, the escapement was a great step forward, laying the groundwork for further improvements while providing better packaging and convenience than any of the various forms of clepsydra which preceded it.

The Discovery of Resonant Control

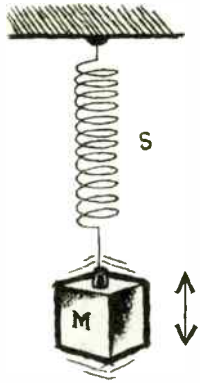
Galileo is said to have noticed that a lamp hanging in a church appeared to swing at a constant rate, regardless of the width of its swing, and suggested that this phenomenon might be used to measure time. Nearly a century later, Christian Huygens produced the first pendulum clock, immediately revolutionizing the art of timekeeping. At once the error in timekeeping was reduced from about 20 minutes a day to less than two.

The secret behind the new timekeeping technique was *resonance,* which still forms the basis of all frequency control. Resonance occurs in many forms—mechanical, electrical, or a combination

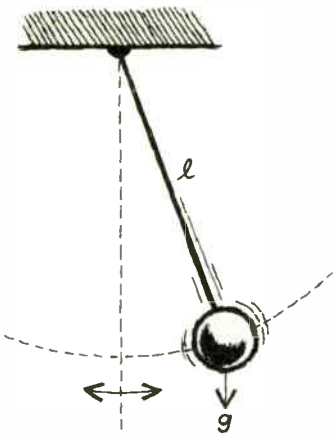
World Radio History



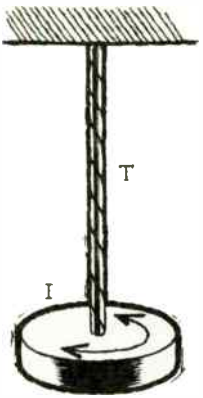
$$\text{PERIOD} = 2\pi\sqrt{\frac{L}{\frac{1}{C}}} = 2\pi\sqrt{LC}$$



$$\text{PERIOD} = 2\pi\sqrt{\frac{M}{S}}$$



$$\text{PERIOD} = 2\pi\sqrt{\frac{\ell}{g}}$$



$$\text{PERIOD} = 2\pi\sqrt{\frac{I}{T}}$$

Figure 3. Several typical examples of resonance, and the basic relationships which determine frequency of oscillation. Note that all equations are essentially the same, save for the terms used. As long as these basic physical properties remain unchanged, frequency of oscillation remains constant. External forces and conditions, however, usually cause change, and must be overcome in stable oscillators.

World Radio History

of these. In all of them, the resonant element must have two essential properties—*mass* and *elasticity*, or their electrical equivalents. When the resonant element is deflected from its resting condition and then released, the elasticity tends to restore it to its former position. But when it reaches its resting position, inertia carries it by until the elasticity arrests it and again turns it back toward the resting position. This continues until various losses in the system finally use up all the energy supplied by the original deflecting force.

This type of oscillation is very obvious in the pendulum, the tuning fork, and various mass-and-spring combinations, but is also present in tuned electrical circuits, piezoelectric materials, and even in molecular and atomic particles. Figure 3 diagrams some familiar examples of resonance and the basic relationships which control their resonant frequency. In all of these resonant systems, it is vital that the "circuit constants" do not change if the frequency of oscillation is to remain constant. For this reason, a pendulum will change its period if either its length or the local force of gravity changes. Thus, if a pendulum rod is made of a temperature-sensitive material such as steel or copper, the lengthening of the rod with higher temperature will slow the pendulum about ½ second per day for each °C change. Similarly, a pendulum clock that keeps accurate time at ground level will lose more than a second a day at the top of a tall building, due to the slight reduction in the force of gravity. In the case of electrical resonance, slight changes in either capacitance or inductance will change the resonant frequency. These changes might be caused by the effect of temperature on the capacitor's dielectric constant, or on the magnetic permeability of the inductance.

Effect of Q on Stability

A very important factor in determining the accuracy and stability of a resonator is its Q or quality factor. In general, the Q of a resonator is directly proportional to its freedom from loss. Thus, the higher the Q, the less power that must be supplied to sustain oscillation. When the Q of a device is low, more power must be added. This unavoidably "smears" the frequency of oscillation and contributes to instability and drift. Thus, the higher the Q, the more stable the resonant frequency. A resonator of infinite Q would oscillate indefinitely and would exhibit perfect frequency stability.

Due to resistive losses, Q's on the order of only 150-200 are typical of electrical resonant circuits. Resonant cavities and transmission lines normally achieve a Q of a few thousand, while the Q of a good pendulum will range from 10,000 to 100,000.

Quartz Crystal Resonators

Among the best man-made resonators yet developed are plates, bars, or other shapes cut from quartz crystals. The quartz crystal resonator, developed about 1922, has proven to be the most versatile and dependable resonating device yet produced. Carefully prepared quartz resonators have demonstrated a Q of 5,000,000, although this class of performance is available only from certain special cuts which operate at relatively high frequencies. Only recently has the accuracy and stability of the best quartz resonators been substantially

World Radio History

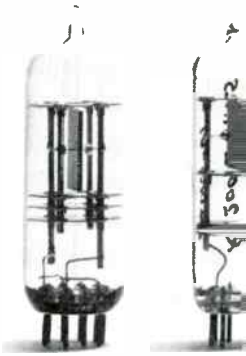


Figure 4. Modern quartz crystal resonators for multi-channel carrier systems. Size, shape, and proportion determine resonant frequency and other properties. Angle of cut relative to crystal structure, is also very important in determining frequency and temperature characteristics.

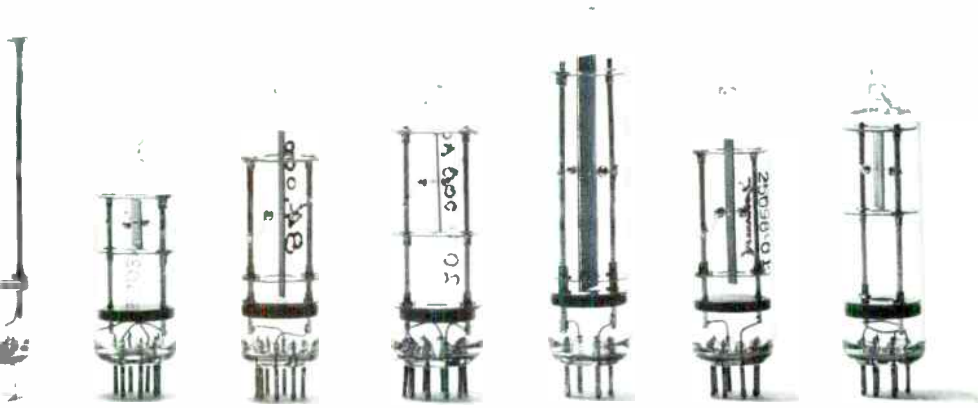
exceeded, and then only by so-called "atomic clocks", rather complicated devices which use atomic resonance for their control element.

Quartz appears to be very nearly ideal for use in resonators. Although almost all materials suffer internal losses from internal friction or elastic hysteresis, most if not all of the loss in good quartz crystals appears to be caused by minor surface irregularities and the effect of mounting. Very tiny amounts of surface contamination have been shown to reduce quartz crystal Q by one-half! A good comparison of the effect of Q on resonator performance can be made by noting how long the device will vibrate when no additional power is supplied. Although a good tuning fork vibrating in a vacuum can sustain about 2000 cycles before the amplitude of vibration is reduced to half, and a high-Q electrical circuit about 100 cycles, well-mounted quartz crystals have been able to "coast" more than a million vibrations before reaching half amplitude.

Quartz is unusually stable, mechani cally and chemically, so that it is generally unaffected by almost any conditions to which it may be subjected. In addition, quartz has a very low temperature coefficient of expansion, thus reducing temperature effects. Strangely quartz slabs cut from the crystal at a certain angle will exhibit a *negative* temperature coefficient, while slabs cut at another angle may have a positive coefficient. By carefully choosing the angle of cut, it has been possible to produce crystals having essentially a zero temperature coefficient over the range 0-100° C. Unfortunately, this cut (the "GT") is limited in the frequency range available and is expensive. Accordingly, other cuts are more frequently used, according to the mechanical and electrical properties of the finished crystal most desired. Figure 4 shows some of the quartz crystals manufactured at Lenkurt for use in Lenkurt carrier equipment.

Despite the high Q available from

World Radio History



quartz units, resonant frequency is sensitive to crystal current (a function of driving power) as well as temperature. Temperature and driving power also affect crystal *aging*, a change in operating frequency which is relatively great when the crystal is new, but which declines with time. Aging is apparently caused by local surface imperfections introduced during fabrication, migration of material between electrodes and leads, and the "shaking out" of microscopic contaminating substances. Once the initial crystal aging period is passed, it is important to restrict crystal activity, and all precision oscillators use some form of current limiting to stabilize crystal operation.

Temperature Control

Despite the inherently low temperature coefficient of quartz, the frequency stability required may be so great that crystal operating temperature must be carefully restricted. As shown in Figure 5, the temperature-frequency charac-

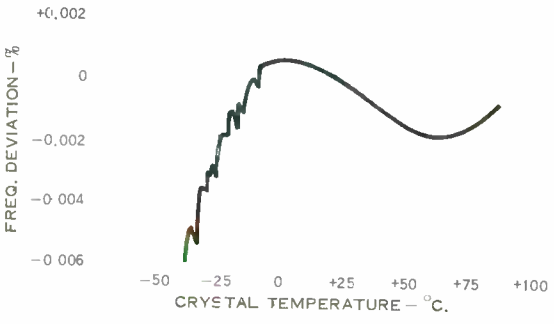


Figure 5. Temperature-versus-frequency characteristic of crystal designed for overtone or harmonic oscillation. Irregularities in left portion of curve are spurious modes of oscillation. Note "turning point" at about 60°C.

World Radio History

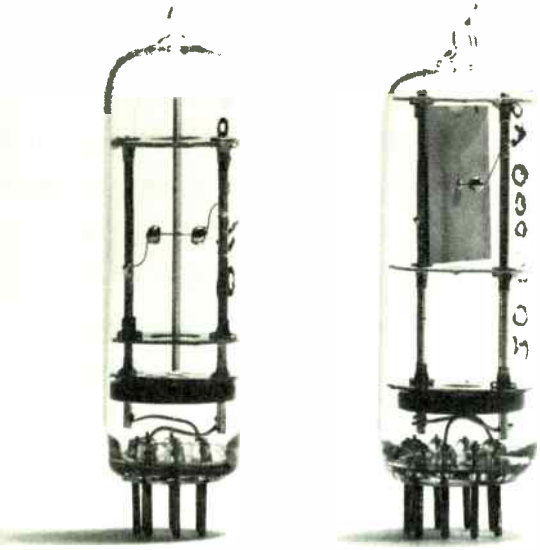


Figure 6. Extreme care is required in mounting crystals. Leads are soldered to gold coating at exact nodal point of vibration. Solder beads damp vibration in leads, should be located an odd number of quarter-wavelengths (at resonant frequency) from crystal.

teristic of crystals is not uniform, but changes according to temperature. Note that the temperature characteristic reverses direction at about 60° C. Over a very narrow range of temperature in this region, temperature has a minimum effect on frequency. This point of "zero" temperature coefficient is sometimes called the "turning point" of the crystal, and can be varied by slight changes in the angle at which the crystal is cut.

In order to take advantage of a crystal's zero temperature coefficient, it is usually operated within an insulated, temperature-controlled "oven." This will usually consist of an insulated enclosure having a high specific heat (requiring a large number of calories to change the temperature a small amount) and a temperature-controlled source of heat.

Where stability requirements are not very strict, an on-off thermostat may be used. Ovens of this type can be designed to maintain the temperature constant within a few tenths of a degree. This type of temperature control has the merit of simplicity, but permits relatively large excursions of temperature, since a definite temperature change must be detected before heater power is turned on or off. Nevertheless, a high degree of frequency stability may be achieved, depending on the type of crystal used.

Proportional Control

A more refined technique is the use of *proportional* temperature control. Proportional control has been known for many years but used very little except where utmost temperature stability was essential. Its use is now growing because reliable transistor circuits make it more convenient, and because the need for improved frequency stability is becoming much greater.

In this method, instead of turning a heater on and off (so that oven temperature swings above and below the desired temperature), heat is introduced continuously, but is continuously varied to

World Radio History

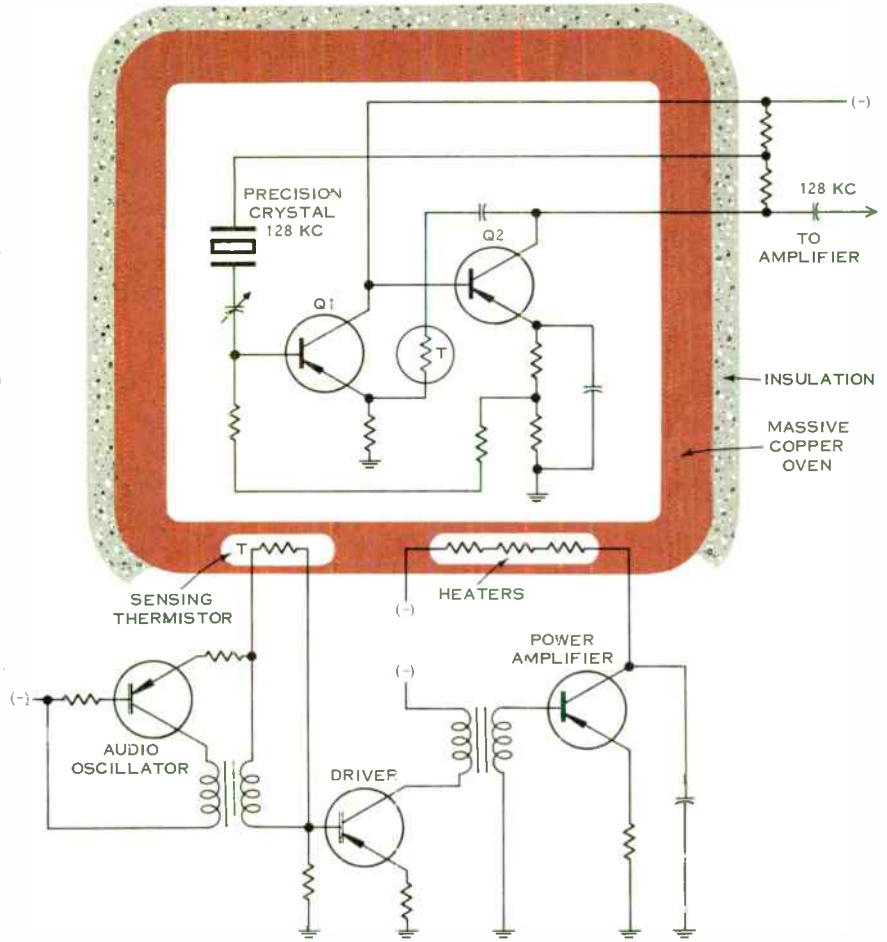


Figure 7. Simplified schematic diagram of Lenkurt 46A Master Oscillator and proportional temperature control oven. Crystal oscillator is thermistor-limited to maintain stable output. Temperature change in copper oven wall is sensed by thermistor which continuously varies audio oscillation and the amount of current supplied to oven heaters.

match the heat lost by the oven.

Figure 7 shows a simplified schematic diagram of the master oscillator and its proportional temperature control circuit used in the Lenkurt Type 46A Carrier system. The 128-kc output from this oscillator is used to derive all channel, group, and supergroup carriers for the 600-channel system. In order to avoid effects of temperature variations on transistor performance, the entire oscillator circuit is enclosed within the oven.

World Radio History

The thermistor (temperature-sensitive resistor) shown in the feedback path from Q2 to Q1 acts as a "governor", holding oscillator activity and crystal current to a very constant value.

The oven itself is a massive copper block into which holes have been drilled to hold the oscillator, heating resistors and sensing thermistor. The sensing thermistor is connected into an audio oscillator circuit in such a way that if oven temperature drops slightly, oscillator feedback increases, thus driving the power amplifier harder and drawing more power through the heating resistors which comprise its load. In this particular design, crystal temperature is maintained within .01° C over a 60-degree range of ambient temperatures; the 128-kc output drifts less than 1/50 cycle in three months if error adds up in the worst way.

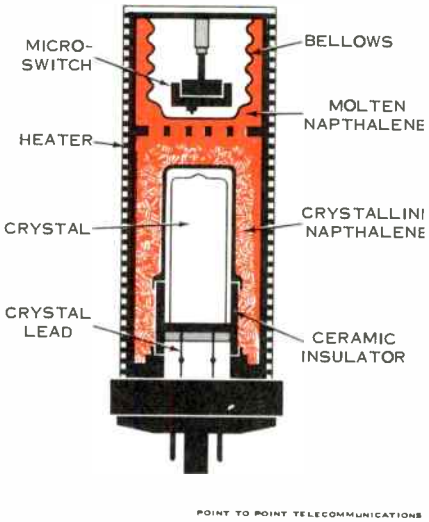


Figure 8. Napthalene in change-of-state crystal oven is kept in half-molten, half-solid state to maintain near-perfect temperature stability. Bellows and microswitch control heat by sensing expansion and contraction of napthalene.

Change-of-State Oven

A recently-developed crystal oven provides what may be the ultimate in temperature control — essentially zero temperature change. Even excellent proportional control ovens must exhibit some temperature differential, since an "error signal", no matter how tiny, must be present before heater power is altered. In the change-of-state oven, however, no temperature differential at all need exist except for the heat lost through the electrical leads to the crystal.

This device takes advantage of the fact that the melting and freezing temperatures of crystalline substances are identical, and that the change from the solid to the molten state is dependent on the *heat* stored in the material, rather than its temperature. Water, for instance, requires about 80 calories per gram to melt. When both physical states are present, the temperature remains exactly at the melting point until the material is either melted or solidified. If a substance is chosen which expands or shrinks when going from one state to the other, it becomes possible to detect changes in the heat content of the system before there has been any change of temperature. As the substance cools and contracts (expands, in the case of water) a bellows and switch activates the heater, thus restoring some of the material to the molten state, still with-

World Radio History

out a change of temperature.

In one such device, napthalene (also used in "mothballs") was the substance chosen. Napthalene melts at 79.5° C and requires about 36 calories per gram for fusion. In this particular oven design, temperature variation within the oven was found to be ± 0.0014° C.

Atomic Timekeeping

Until a few years ago, most national frequency standards consisted of extremely stable quartz crystals located in temperature-controlled vaults. Today, these have been replaced by cesium beam "atomic clocks" which maintain stability on the order of one part in 10^{11}, or about two parts in 10^{12} over a few hours. These standards use the resonance of cesium 133 atoms in making quantum transitions between energy states (see DEMODULATOR, *July*, 1961 for a discussion of quantum resonance). The cesium resonance frequency in these standards is 9,192,631,770.0 cycles per second.

Other ultra-stable atomic resonators have been developed which are both simpler and possibly more stable than the cesium beam standard. One such development is the rubidium vapor cell, which has a Q of 170,000,000, more than twice that of the cesium resonance. The rubidium cell is not really as suitable for an absolute standard of frequency as the cesium beam because the resonant frequency can be altered slightly by the way the device is constructed. Once built, however, it is fantastically stable. Several rubidium gas cells built for use in satellites have shown less than one part drift in 10^{11} per month over a period of a year, about 1/30 that of a cesium standard with which they had been compared.

Oscillators capable of accuracies in this range are certain to become relatively simple and cheap, thus permitting their widespread use in many old and new communication applications, particularly ultra-high-speed data transmission and satellite communications. The stability and cost of conventional communications equipment will benefit and its information-handling ability can be expected to increase. •

BIBLIOGRAPHY

1. Warren A. Marrison, "The Evolution of the Quartz Crystal Clock," *The Bell System Technical Journal;* July, 1948.
2. T. C. Anderson and F. J. Merrill, "Crystal-Controlled Primary Frequency Standards; Latest Advances for Long-Term Stability," *IRE Transactions on Instrumentation;* September, 1960.
3. R. C. Mockler, R. E. Beehler, and C. F. Snider, "Atomic Beam Frequency Standards," *IRE Transactions on Instrumentation;* September, 1960.
4. R. J. Carpenter et al, "A Prototype Rubidium Vapor Frequency Standard," *IRE Transactions on Instrumentation;* September, 1960.
5. F. G. Merrill, "Frequency and Time Standards — A Status Report," *IRE Transactions on Instrumentation;* September, 1960.
6. John P. Buchanan, *Handbook of Piezoelectric Crystals for Radio Equipment Designers,* WADC Technical Report — 56-156; October, 1956.
7. "On Means To Improve Frequency Stability," *Point To Point Telecommunications,* Marconi's Wireless Telegraph Co., Lt., October, 1961.

World Radio History

World Radio History

the Lenkurt Demodulator

VOL. 13, NO. 3 MARCH, 1964

the universal VOICE CHANNEL

Most terminal equipment for modern telecommunications service is designed to operate over a voice channel, a fraction of a voice channel, or several such channels put together. The transmission medium may be cable, open wire, or radio, but nevertheless the basic unit for defining the facility is usually the voice channel. But the question naturally arises, what is a voice channel? This article attempts an answer, which necessarily includes discussion of services other than speech transmission.

If several types of communications users, such as a telephone man, an industrial user, and a military man, were asked to define the term "voice channel" there would be as many different definitions as there were types of users. This is to be expected, since each one puts his definition in terms of his own particular needs and applications.

One man may be concerned only with the essential information content in speech transmission. Another may be concerned with the reactions of paying customers to how well the communications equipment reproduces the quality of their friends' voices. Still another man may not be concerned with speech at all, but rather with the transmission of data or some other service over the "voice channel."

But regardless of these diverse needs, and of the various qualities of the facilities required to fill them, all voice channels have a single common denominator. They are designed primarily around the characteristics of the human voice and the human ear.

Speech Characteristics

From the listener's point of view the quality of a voice channel can be measured in terms of two parameters, intelligibility and intensity, which together determine the quality of reception of sounds transmitted over the channel. Interestingly enough, intelligibility and intensity are virtually independent of each other over quite a broad frequency range. Most of the speech energy, and hence the intensity,

World Radio History

is concentrated in the lower frequencies, while the high frequencies contribute most to the intelligibility. If *no* frequencies below 1000 cps were transmitted, articulation would be about 86 per cent perfect, but the received energy content would be only about 17 per cent of the original energy. On the other hand, if *only* the frequencies below 1000 cps were transmitted, articulation would be reduced to about 42 per cent, while 83 per cent of the total energy would be transmitted (Figure 1 illustrates this graphically). This means that any voice channel must include both the low frequencies and the high frequencies. Furthermore, some compromise is usually necessary because available bandwidth is limited.

Perhaps the most important factor in determining bandwidth, however, is the reaction of the people using the facilities. Many experiments have been performed to test subjective reaction to various transmission impairments, including restricted bandwidth. As a result of such tests, the "standard" voice channel bandwidth has come to be accepted as about 3 kc. Typically, the range of transmitted frequencies is from about 200 cps to about 3200 cps.

Another significant characteristic of speech is its redundancy. As much as 75 per cent of the information content in normal speech is redundant. If a syllable is lost (or often even a word), the listener automatically fills in the gap from the context. The result is that the requirements for speech transmission are often much less stringent than those for other types of transmission.

Services Other Than Speech Transmission

Manufacturers of equipment for the transmission of telegraph, data, and facsimile often find it convenient to design their equipment for operation over voice channels. The reason is simply that voice transmission facilities are almost universally available. However, difficulty may arise because the voice channel is designed around the peculiarities of speech transmission, while these other services may have quite different requirements. It then becomes necessary to evaluate the quality of the voice channel in terms of the technical factors which are important for the other types of transmission.

One of the first characteristics which comes to mind in defining any type of communications channel is bandwidth. The bandwidth required for voice transmission is, of course, determined by speech characteristics, as previously mentioned. The bandwidth required for digital transmission, however, is primarily determined by the speed of transmission. The higher the speed the more bandwidth required. Thus, tele graph service with its relatively slow speed requires a small fraction of a voice channel — typically, 170 cps for 100 words per minute. Conversely, high-speed data transmission, measured in kilobits and even megabits per second, may require enormous bandwidth, equivalent to many voice channels. Where a data system operates over a voice channel with a nominal 3-kc bandwidth, the transmission speed is often 1200 or 2400 bits per second.

One of the items of major concern in evaluating any communication channel is noise. But there are various types of noise, and one kind may affect a particular type of communications more than another kind. For example, white noise (background noise) measurements are usually used in evaluating a channel for speech purposes because the human ear hears this kind of noise — white noise tends to mask the desired message. Data, however, is relatively insensitive to white noise. Here it is impulse noise which causes the most trouble. Impulse noise consists of

World Radio History

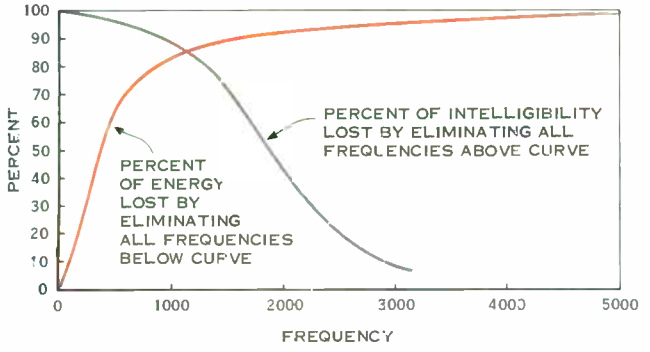


Figure 1. Curves indicate that speech energy is concentrated in the lower frequencies, while the higher frequencies contribute most to intelligibility.

short "spikes" which may reach quite a high amplitude, but which have a relatively short duration. The length is often measured in fractions of a millisecond. Even though they have a high amplitude, these spikes are often too short for the human ear to hear. Therefore, they cause very little trouble in voice transmission. In data transmission, however, they can easily reach the level of the data signal, thereby causing an error, as shown in Figure 2.

White noise has little effect on data transmission until the noise reaches a very high level because data is not affected until the noise peaks reach the detection level of the signal. Thus, even though the ear may hear a significant amount of white noise, the chances are slight that it will have much effect on data.

Attenuation and Delay Distortion

Services other than voice transmission may be transmitted in three forms. They may use a serial digital transmission arrangement with one information bit transmitted after the other; they may use a parallel digital transmission arrangement where more than one bit is transmitted simultaneously; or transmission can be in the form of analog signals, without the "quantizing" of digital techniques. But regardless of the way the information is transmitted, there is one thing in common: the various frequency components which make up the complex signal bear a very definite relationship to each other both in magnitude and in time. This means that all frequencies within the passband should suffer the same loss if *attenuation distortion* is to be avoided. It also means that all frequencies within the passband should propagate through the transmission medium at the same speed to avoid *delay distortion* — an altering of the phase relationship between the frequency components.

Attenuation distortion is perhaps equally important in either speech or data transmission. Typically, the higher frequencies in the passband are attenuated more than the lower ones. For example, on nonloaded wire pairs the attenuation is usually proportional to the square root of frequency within the voice band. Inductance loading is used to reduce both attenuation distortion and overall loss on most cable pairs longer than about three miles. This is fine for speech transmission, but it makes the line resemble a lowpass filter with a cutoff frequency. As this cutoff frequency is approached, phase or delay distortion increases rapidly.

Phase distortion is relatively unimportant in voice transmission because the components of speech need not

World Radio History

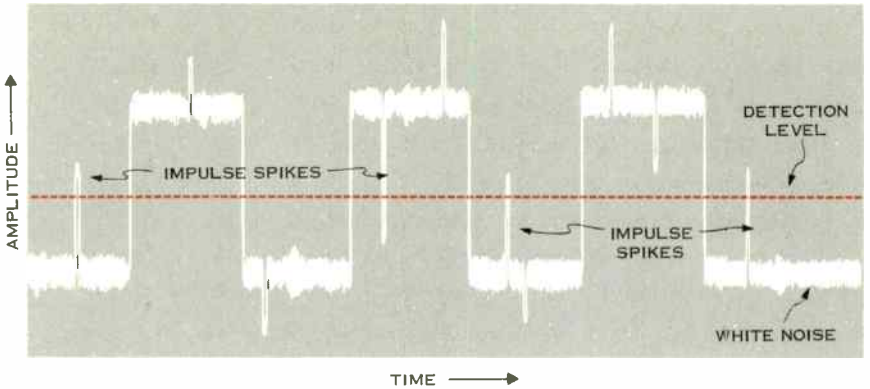


Figure 2. Unlike speech transmission, data is relatively unaffected by noise below the signal detection level. Thus, errors are quite likely to be caused by impulse "spikes," while white noise has little effect on error rate.

bear an exact time relationship to each other for intelligibility. Furthermore, the large amount of redundancy in normal speech tends to mask the effect of any phase distortion. For these reasons phase distortion is rarely considered (unless it is extreme) in evaluating a voice channel for speech transmission.

In a typical channel the frequency of minimum delay is approximately 1600 to 1700 cps. A curve of relative delay plotted as a function of frequency usually reaches a minimum in this region and forms a more or less symmetrical "U" about the midpoint. The two ends of a typical curve reach a relative delay of perhaps a millisecond or more at about 400-800 cps on the low end and at about 2700-2900 cps at the high end. Figure 3 shows the characteristics of a typical carrier system voice channel. The channel shown has a "one-millisecond bandwidth" of approximately 3000 cps and it has a half-millisecond bandwidth of 2500 cps. A relative delay of one millisecond will have little affect on speech transmission, but it may well render a voice channel completely unusable for data, particularly at higher transmission speeds. For this reason data transmission systems seldom use the entire voice frequency bandwidth, but are placed near the center of the frequency band where distortion is lowest. (Data is often transmitted in the 1000-2600 cps band).

Equalization

A common technique to increase the percentage of bandwidth usable for data transmission is known as *equalization*. A channel is equalized by introducing a network which produces either attenuation or phase shift characteristics (or both) opposite to those already inherent in the channel. Figure 3 illustrates how the equalizer characteristics add to the channel characteristics to produce a relatively smooth curve.

Of course it is seldom possible to obtain a truly flat curve in this way. Equalization is used simply to get the attenuation and phase delay within the acceptable tolerance for data transmission. Ripples in the attenuation and

World Radio History

delay characteristics frequently represent echoes caused by impedance mismatches. The edges of the band normally exhibit more pronounced ripples because more discontinuities, and hence more reflections, occur here. This is one more reason why often only the center of the bandwidth is used for data transmission. Echoes from "close-in" discontinuities usually result in ripples which can be equalized, whereas echoes from remote transmission discontinuities often cannot be equalized. The reason is that an individual transmitted pulse is usually affected primarily by its own echo if the source of the echo is close in. Echoes from remote discontinuities tend to affect later transmitted pulses. Thus, the effect is random for a train of pulses, and equalization is not generally effective.

Compandors and Echo Suppressors

The interfering effects of noise, crosstalk, and echo on speech are often reduced by the use of compandors and echo suppressors. These devices take advantage of certain peculiarities of human conversation. Unless they are removed from the transmission path, however, they often render a circuit unfit for many types of data transmission.

Compandors give an apparent reduction in noise on voice circuits by increasing the circuit loss during speech pauses — between syllables and words. If amplitude-modulated (on-off) type data signals are transmitted through a compandor, some of the pulses may be badly distorted by the varying loss characteristic. However, frequency-shift data signals are not appreciably distorted by a compandor since their power level remains essentially constant.

Echo suppressors are frequently used on long circuits to prevent the echo of reflected speech from annoying the talker. They are simply devices which inhibit the return path when a person is talking. Most echo suppressors in use on commercial telephone circuits today permit transmission in only one direction at a time. During data transmission they must be removed from the transmission path or they will not permit a data receiver to ask for the repeat of a message in which an error has been detected. This "disabling" of the echo suppressor is accomplished by a device which is sensitive to a tone of 2000-2250 cps generated by the data transmitting equipment. When the disabler receives this tone for some specified length of time, it holds both directions of transmission open while data is being transmitted.

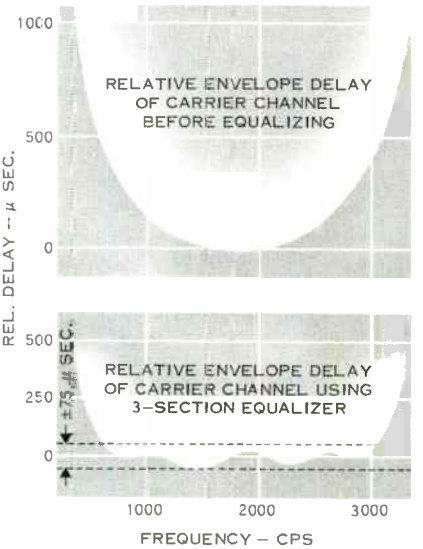


Figure 3. Top panel shows the envelope delay characteristics of a typical carrier channel before equalization. Addition of a 3-section equalizer reduces relative envelope delay to ±75 microseconds over a large portion of the bandwidth.

World Radio History

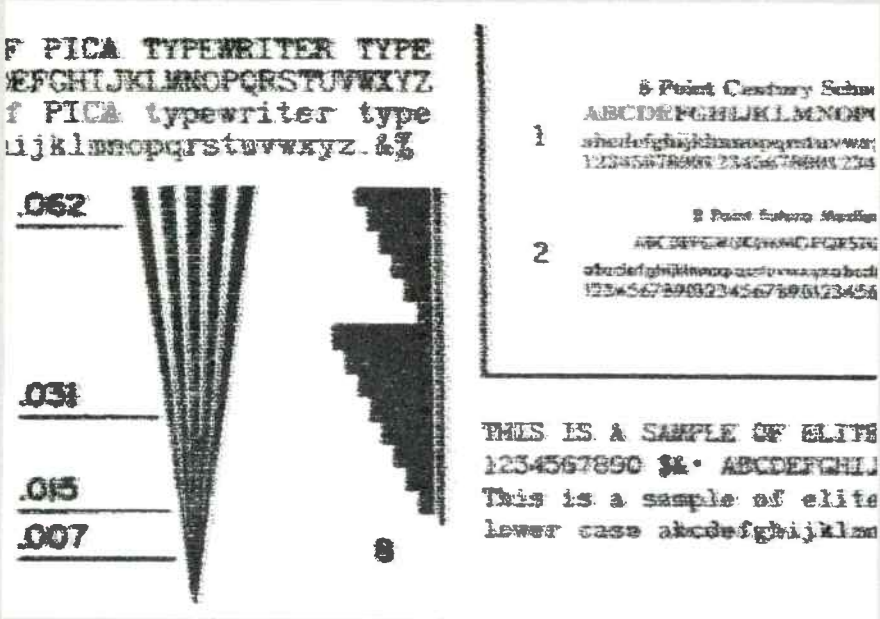


Figure 4. Actual sample of facsimile transmission over a circuit not corrected for delay distortion. Note the fine echoes following each transition from black to white or white to black. This same effect occurs in high-speed data transmission, causing errors to increase.

Future Trends

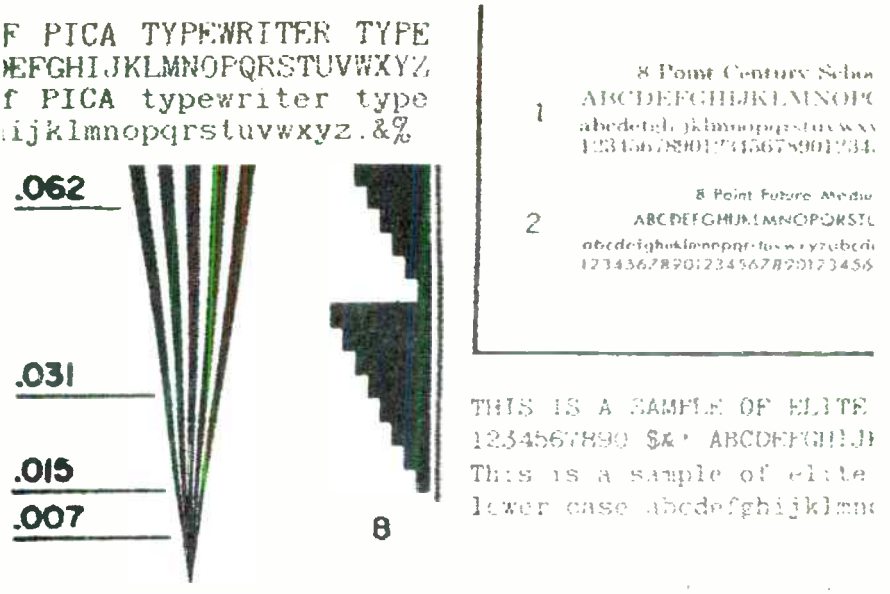
The history of the communications industry is one of increasingly stringent performance requirements, with no end to this trend in sight. This comes about because communications users are demanding that their transmitted information be reproduced at the receiving end with ever-greater fidelity.

But there is another important factor influencing the shift toward "tighter" specifications for voice channels. This is the changing pattern of communication. It includes not only the ever-greater quantities of information being transmitted, but also the trend toward the transmission of many kinds of information in digital form. Equipment manufacturers today must consider factors which, in the not-too-far-distant past, were only academically related to voice channel performance.

Consider delay distortion for example. Serial data at 2400 bits per second has a bit length of only 0.417 millisecond. If the delay distortion is several milliseconds, the data will be hopelessly garbled because parts of a pulse may be delayed enough to overlap with the "faster" portions of the following pulse. Newer equipment designs recognize this problem and allow for it. For example, Lenkurt's long-haul 46A carrier equipment holds delay distortion to 0.165 millisecond (without equalization) for the usual data band, 1000 to 2600 cps.

The 46A is also unique in its data-handling capabilities. Up to 65 per cent of its channels can be loaded with

World Radio History



Transmission is greatly improved when the delay distortion present in the transmission medium is corrected by delay equalizer. Lines and other fine detail are reproduced much more sharply. The ability to transmit finer detail by equalizing delay is similar to higher speed data transmission.

data at the same time — more than any other commercial multiplex equipment. (Older equipment may falter under as little as 25 per cent data loading.) In recognition of the fact that data-handling capacity is becoming more important, the extra power-handling capability necessary to accommodate such a load is an integral part of the design.

Thus, it is becoming increasingly apparent that the performance standards of the past do not apply to the present, much less the future. The direction of the trend is clear. A system barely adequate for today's needs may be hopelessly obsolete in a few years; whereas a system which is "over-designed" by present standards will have a working life far longer. •

Bibliography

1. A. L. Albert, *Electrical Communication*, Third Edition; John Wiley & Sons, Inc., New York, 1950.
2. *Transmission Systems for Communications*, Bell Telephone Laboratories; New York, 1959.
3. *Telephone Transmission Quality, Local Lines and Telephone Sets*, C.C.I.T.T. Red Book, Vol. 5; IInd Plenary Assembly, New Delhi, 1960.
4. A. A. Alexander, R. M. Gryb, and D. W. Nast, "Capabilities of the Telephone Network for Data Transmission," *The Bell System Technical Journal*: May, 1960.

World Radio History

World Radio History

the Lenkurt Demodulator

VOL. 14, NO. 9 SEPTEMBER, 1965

Requirements of a HUMAN COMMUNICATIONS CHANNEL

The objective of a human communications channel is to extend our conversational range by permitting direct oral communication between two or more people over some geographic distance. Because the success of such a system is largely subjective, it is necessary not only to make the service easy to use, but to enable the user at each end of the line to communicate in a manner approximating as nearly as possible face-to-face discussion.

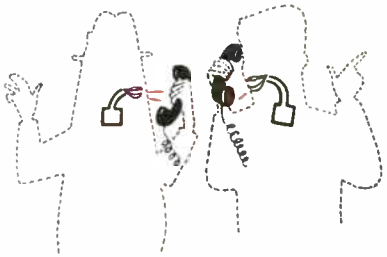
With this human factor in mind, it is mandatory that the effects of various interfering factors be judged on a personal basis, not easily correlated in terms of electronic meter readings.

In this article, some of the factors which affect communication are discussed, and an attempt is made to relate these interfering effects to the requirements of a human communications channel.

About this article

This article is a revision of the February, 1959 issue of The Lenkurt Demodulator. It provides a basic introduction to some of the problems of human telephone communications, and is being reissued for the benefit of our new readers.

World Radio History



What we say in face-to-face conversation is emphasized by facial expressions and gestures. In a telephone conversation, we depend upon the voice alone.

A human communication channel (or system) must offer the user more than just a means of transferring information from one point to another. To enjoy wide acceptance as a means of communication, it must present to the user a good "second best" to direct face-to-face conversation, along with being convenient and sufficiently economical. Otherwise, the use of the facility will be restricted to messages which are urgent or absolutely necessary and cannot wait to be mailed or sent by telegraph.

In voice communication, we are not only interested in transmitting information—that is, sounds which can be interpreted as words—but we are also interested in conveying shades of meaning through variations in voice amplitude and inflection. In effect, we wish to capture the feeling of presence that is obtained in face-to-face conversation. The rather extensive use of the telephone as a means of communication indicates to a high degree that this objective has been achieved.

In direct conversation, communication is emphasized in a number of ways: variations in speech amplitude, inflection, facial expressions, and gestures. In a telephone conversation the visual communication is, of course, missing. The listener is forced to rely on word context and voice inflection for the transfer of meaning.

Admittedly, the intelligence alone can be transmitted through a communications system with a minimum of variation in amplitude and in the presence of high noise levels. But while it is possible to get oral information through such a system, this type of low quality communications circuit would certainly not find the wide acceptance that today's telephone industry enjoys.

What then is necessary for a successful human communications channel? To answer this question, let's consider some of the things affecting the transmission of voice over a telephone.

In direct face-to-face conversation, the sound intensity (amplitude) and the frequency (pitch) of the sound

World Radio History

waves may vary over wide limits. Sound intensities may range from the threshoid of hearing (a whisper) to very high levels. However, in a normal conversation the variation is between 30 to 40 db for the usual talker.

But since many different people use the telephone channel, the relationship between the sound intensity ranges of different talkers must be considered. Studies have shown that the intensity range between the weakest syllable of a soft talker and the loudest syllable of a loud talker (that is, the two extremes) is in the order of 70 db. Under the same conversational conditions, the frequency range of sound can vary from about 50 cps to about 10 kc depending on the individual.

Within this scope, it would be difficult to construct a multichannel transmission system that could include all ranges of sound intensity and frequency without unduly restricting the ultimate channel capacity.

From the listener's point of view the quality of a voice channel can be measured in terms of two parameters, intelligibility and intensity, which together determine the quality of reception of sounds transmitted over the channel. Interestingly enough, intelligibility and intensity are virtually independent of each other over quite a broad frequency range. Most of the speech energy, and hence the intensity, is concentrated in the lower frequencies, while the high frequencies contribute most of the intelligibility. If *no* frequencies below 1000 cps were transmitted, articulation would be about 86 percent perfect, but the received energy content would be only about 17 percent of the original energy. On the other hand, if *only* the frequencies below 1000 cps were transmitted, articulation would be reduced

Restricting the conversational frequency range does not seriously impair telephone communication if normal levels are maintained. Where wide level variations are encountered, communication is seriously affected.

World Radio History

to about 42 percent, while 83 percent of the total energy would be transmitted. This means that any voice channel must include both the low frequencies and the high frequencies. In practice, some compromise is usually necessary because available bandwidth is limited.

For toll circuit engineering, the normal range in talker volume can be considered to be between 0 dbm0 and −31 dbm0, and the *standard* voice channel bandwidth is about 3 kc. Typically, the range of transmitted frequencies is from about 300 cps to about 3400 cps.

The discussion thus far has been concerned with restrictions which may be imposed on speech communication without incurring serious degradation, but has not included any effects which may be encountered during the transmission of speech. When transmission is considered, other factors are introduced that affect the intelligibility and identification of the message signal. These include: (1) level variation, (2) changes in frequency, (3) distortion caused by non-linearities in the circuit, and (4) interference such as noise and crosstalk.

Level Variations

In any transmission medium, the loss in signal strength is not constant, but varies from instant to instant. These variations are a result of changes in circuit loss caused by such things as varying temperature and other weather conditions. This means that unless some type of level control is used, the signal level at the receiver will also vary. The manner in which level control is accomplished depends on the transmission system.

It is desirable to keep level variations which may occur over a short time interval to about 0.25 db. Systems may be expected to drift more than this over a long period of time, but proper routine maintenance usually prevents any serious impairment in transmission quality.

Frequency Stability

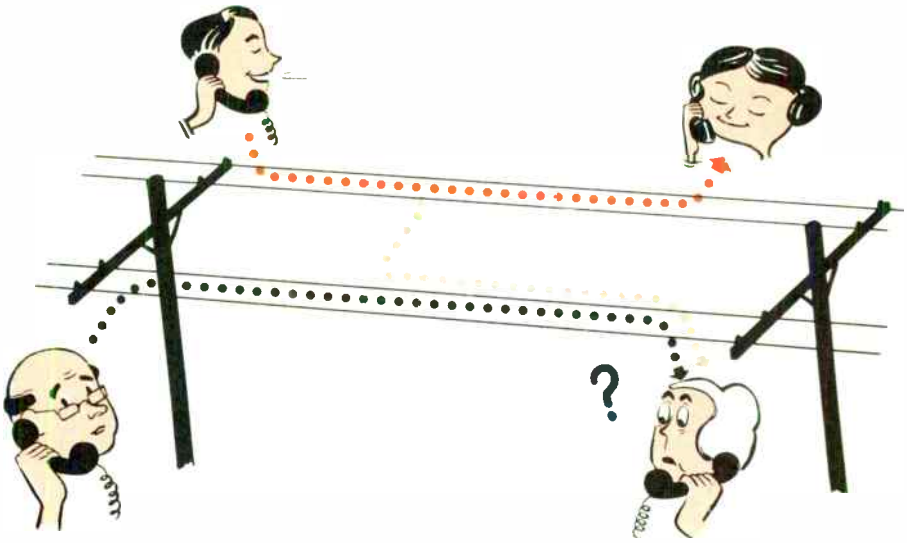
The use of oscillators in carrier telephone systems introduces the problem of frequency stability. This problem is commonly associated with the change in frequency of an oscillator over a period of time. But in a telephone circuit, the frequency stability that is of concern is the net change in frequency that occurs between the transmitting and receiving ends of the circuit. The amount of frequency change tolerable is directly related to the amount of change discernable to the ear. Analyses of hearing acuity tests suggest that a certain amount of frequency change, even over a short period of time, can be made without the ear detecting the shift as a different sound.

For voice circuits, a frequency stability (end-to-end) in the order of ± 3 to ± 5 cps provides a very good circuit. In actual operation, frequency shifts approaching ± 15 cps may occur over a long time period without seriously impairing the quality of the voice circuit. However, the increased use of switched networks and dial-up connections for data transmission has necessitated much greater frequency control, measured usually in fractions of a cycle.

Crosstalk

Wherever telephone circuits follow adjacent paths, they are susceptible to crosstalk interference. Crosstalk may be produced by inductive or capacitive coupling in parallel lines, or at junctions, producing unwanted signals in the disturbed circuit. Generally, three

World Radio History



Excessive crosstalk interferes with the desired conversation, and reduces the privacy of this means of communication.

types of crosstalk are considered to exist: (1) intelligible crosstalk, which is in the same frequency range, but lower in amplitude than the original or desired signal, (2) unintelligible crosstalk, which is translated in frequency, or appears in the disturbed circuit in an inverted order, and (3) babble, which is crosstalk from a number of sources, either intelligible or unintelligible. With babble, the resulting sound has an apparent syllabic rate, but because of the number of interfering signals, does not appear as intelligible crosstalk. Babble is normally evident during the busy-hour periods and is similar to noise.

Crosstalk performance could be readily calculated if it were only necessary to measure coupling between two circuits. However, the magnitude of crosstalk in the disturbed circuit will vary depending on talker volumes in the disturbing circuit, or circuits. Other factors which change the interfering effect of crosstalk are variations in subscriber loop losses, and the masking effect of circuit and room noise. In addition, the reactions of different people to a given crosstalk volume will vary widely. The total range of tolerance is about 30 db. These are among the factors considered in determining the crosstalk index, often used in establishing grades of performance for telephone circuits.

Noise

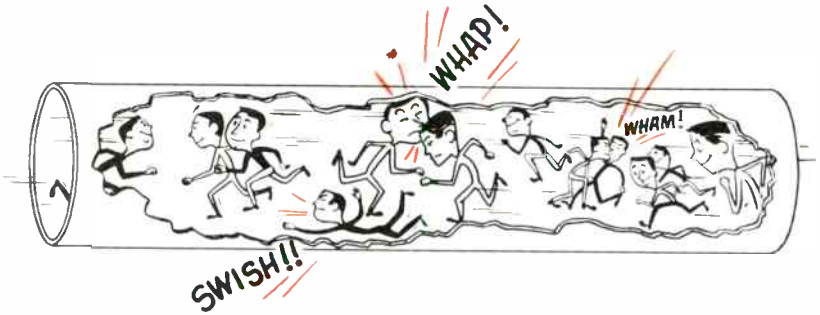
Another type of interference that is important in telephone communication is noise. Any type of interfering signal may be classified as noise. However,

World Radio History

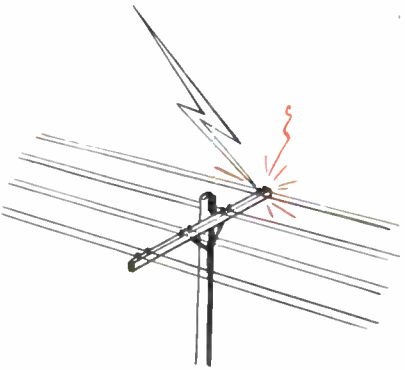
crosstalk, which is included in this broad definition, is generally treated as a separate problem.

Types of interference which may affect voice communication include: (1) thermal noise, (2) impulse noise, and (3) extraneous tones. The disturbing effect of these types of noise is generally less than that of crosstalk because in most instances no recognizable syllabic pattern is discernible. However, the disturbing effect on any one circuit will depend upon the type of noise and its frequency distribution. Many types of noise are man made, and can be eliminated, or at least reduced in magnitude.

Since noise cannot be entirely eliminated, objectives for noise amplitude



One source of random noise is from thermal motion of the conduction electrons in a resistor.



Lightning is another source of noise which can affect telephonic communication.

Room noise is present even during a normal conversation. While a high level of room noise is always distracting, in a telephone conversation room noise helps to mask noise and crosstalk in the telephone circuit.

World Radio History

and frequency distribution have been established. For example, in many toll applications, the overall objective for end-to-end connections is to maintain total noise (including crosstalk) at 20 dbrnc0, as measured at the subscriber terminal under normal busy-hour conditions. This total value is then divided into separate objectives for subscriber loops, terminal equipment noise, various types of line noises, and crosstalk.

Distortion

Distortion is the general term used to describe any change in waveform of a signal. Even where noise and crosstalk requirements are met, distortion may reduce the intelligibility and identification of the speaker. The three basic types of distortion are: (1) amplitude distortion, (2) frequency distortion, and (3) delay distortion.

Amplitude distortion is caused by non-linearities in the circuit and is often called non-linear distortion. This type of distortion is characterized by the generation of harmonics which are multiples of the speech frequencies being transmitted. Depending upon the point in the circuit at which the non-linearity exists, these harmonics may appear as crosstalk in other carrier telephone channels. Because of the possible interfering effects, amplitude distortion is kept to a minimum in carrier equipment design.

Frequency distortion, unlike amplitude distortion, does not generate harmonics, but simply appears as selective attenuation of some frequencies with respect to the overall frequency spectrum. If frequency distortion appears within a voice channel and is excessive, the effect is readily apparent. Where low frequencies are greatly attenuated, the resulting speech will sound "tinny"; if the high frequencies are greatly attenuated, the resulting speech will have a "booming" sound.

Delay distortion is the result of differences in the propagation velocities between the various frequencies in a complex wave. For speech circuit, delay distortion is not a problem because the ear is relatively insensitive to phase variation. However, on voice-frequency circuits used for high-speed data transmission, delay distortion is an important factor.

Conclusion

While many of the factors affecting the transmission performance of a telephone circuit may be readily measured, it is difficult to correlate these figures meaningfully to the satisfaction of the user. The telephone user remains subjective in his evaluation of the human communications channel. As advances are made in techniques and equipment, so will the user raise his standards in demanding better and more complete service. Only through continual review can the telephone industry insure that it is in agreement with the user on what he expects of his telephone.

World Radio History

World Radio History

the Lenkurt Demodulator

VOL. 13, No. 10 OCTOBER, 1964

The Theory and Use of

COMPANDORS

in Voice Transmission Systems

Progress in the field of electrical communications records a relentless fight against interference caused by noise and crosstalk. Through the proper use of a specialized device, known as a compandor, *toll quality voice transmission can often be achieved over telephone circuits otherwise unsuitable because of excessive noise or crosstalk. This article is a general introduction to compandors and includes a brief discussion of their application and of the characteristics of speech energy that required their development.*

Sound energy that is converted to electrical energy in ordinary telephones consists of a complicated wave made up of tones of different frequencies and intensities (or magnitudes). These speech frequencies and intensities are the fundamental signal characteristics which must be dealt with when designing a voice transmission system.

Most of the energy of speech signals is concentrated in a frequency band that ranges from about 200 cycles per second to about 3200 cycles per second. The intensity range is determined by two factors—the talker and the words or syllables spoken. The difference between the loudest syllable of a loud talker and the weakest syllable of a soft talker may be as much as 70 db (equivalent to a power ratio of *10 million to one*). The average talker produces an intensity range of about 30 to 40 db.

Such a wide range of speech powers presents a significant problem to the designer of transmission systems. Weak signals must be transmitted at a higher level than the noise and crosstalk encountered in the circuit, while strong signals must *not* overload the amplifiers. These factors essentially set an

World Radio History

upper limit and a lower limit to the power range that a typical communications system can handle effectively. Building communications systems with greater load capacity and with greater noise and crosstalk protection is not always practical or economical. Coping with the noise performance of communications circuits, therefore, is usually an economic problem.

Reducing noise and crosstalk in communications systems is not an easy task. Some types of induced line noises are unavoidable because telephone lines are often necessarily placed near power transmission lines or other sources of electrical noise.

The maximum amount of power that can be transmitted over wire lines or cable carrier systems is generally limited by established transmission standards. Increasing the power would, of course, provide some increase in the ratio of wanted signals to line noise. However, there would not be any crosstalk improvement since the amount of crosstalk is independent of the line levels (provided all parallel systems operate at the same relative levels).

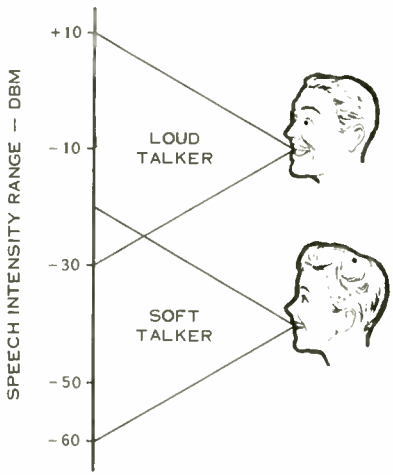


Figure 1. The dynamic speech range of an individual is about 40 db; however, the difference between the strongest sounds of a loud talker and the weakest sounds of a soft talker may be as much as 70 db.

Reducing crosstalk between carrier systems on different pairs of a pole line sometimes requires extensive line transposition. Most lines used for carrier systems are constructed so that both noise and crosstalk are as low as practicable.

When the first New York-to-London radiotelephone circuit was installed back in the late 1920's, it was recognized that noise would be a major problem on such a long-distance radio circuit. Noise in radio systems consists mostly of stray electrical energy commonly known as static which usually causes more interference than the noise encountered in physical circuits.

To help overcome this noise problem, a volume indicator and a manual volume control were placed in the wire circuit between the overseas toll office and the radio. The volume indicator displayed the intensity of the speech signals transmitted over the circuit. A *technical operator* monitored the volume indicator and adjusted all signals to a level that fully loaded the transmitter. Manually adjusting the speech signals was effective in reducing the range of signal intensities applied to the circuit to about 30 db. However, it was rather difficult for the operator to follow the rapidly varying signal amplitudes, and static or noise was still annoying to the listener, especially during speech pauses and other silent periods.

Efforts to improve the signal-to-noise performance on the transatlantic radio-

World Radio History

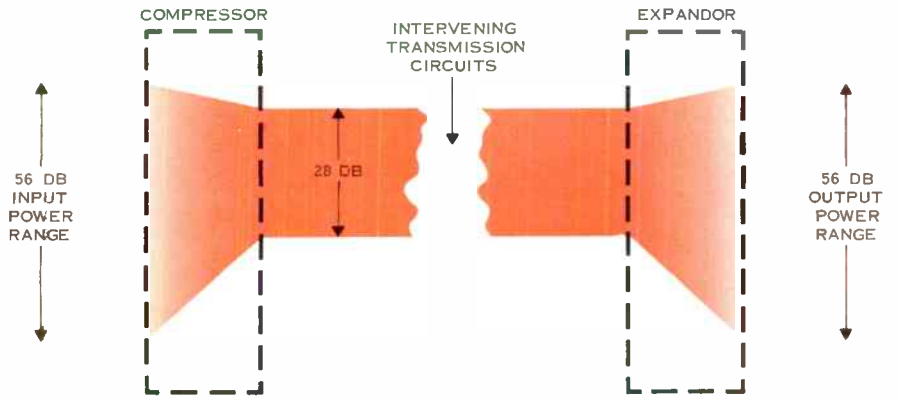


Figure 2. The wide power range of speech signals is reduced by the compressor and restored to its natural power range by the expandor.

telephone circuit led to the development of a voice-operated device called a *compandor*. The compandor, first used in the New York-to-London radiotelephone circuit in 1932, provided a great improvement in overcoming the problems of static, thus making the radio circuit usable for a greater portion of the time.

Cost and space requirements of the early compandors prevented their general use on wire circuits until the late 1930's. The Bell System first employed compandors on a wire line in 1941 on the Charlotte, North Carolina-Miami, Florida and the Charlotte, North Carolina-West Palm Beach, Florida routes.

Technically, the type of compandor described in this article is a *syllabic* compandor. There is another type of compandor, known as an *instantaneous* compandor, that is used to reduce quantum noise in pulse-code modulation (PCM) transmission systems. Since the instantaneous compandor has such a distinct function, it has not been included in this general discussion of the voice-operated compandor.

What Is A Compandor?

A compandor is a combination of an intensity range COMpressor and an intensity range exPANDOR, from which it derives its name. The compressor is employed in the voice *input* circuit of a communications channel to compress the intensity range of speech signals by imparting more gain to weak signals than to strong signals. At the receiving end, or voice *output* circuit, of the same channel the expandor performs the opposite function of restoring the intensity back to its original range.

The compressor automatically raises the level of weak speech signals so that they can be transmitted above the noise and crosstalk encountered in the circuit *between* the compressor and the expandor. Thus, the signal-to-noise improvement for weak signals is produced by the compressor. In addition, the compressor attenuates strong signals, there-

World Radio History

by preventing them from overloading the transmission equipment.

The expandor, at the receiving end of the circuit, presents more attenuation to weak signals than to strong signals and adjusts to a condition of maximum loss between speech syllables. The ordinarily weak crosstalk and noise signals that are very disturbing to the listener during silent periods or between syllables are greatly attenuated, thus quieting the circuit.

How A Typical Compandor Works

As stated previously, the compandor consists of two devices: the compressor at the transmitting end of the circuit, and the expandor at the receiving end of the same circuit. The compressor and expandor each contain a variable loss device (variolosser), an amplifier, and a rectifier control circuit, as shown in Figure 3.

Speech signals entering the compressor first pass through the variolosser and then the amplifier. A portion of the speech energy leaving the amplifier is routed into the control circuit where it is rectified. The resulting direct current is fed into the variolosser circuit where it is used to control the amount of signal attenuation. The level of this direct-current signal is *directly proportional* to the strength of the speech energy, which is constantly varying. As the level of the direct current increases, the attenuation of the variolosser also increases. If a weak speech signal is present in the compressor, the control current is small and the attenuation of the variolosser is low. When the input speech energy increases, the attenuation of the variolosser increases in direct proportion. Thus, strong signals are attenuated more than weak signals, resulting in a compression of the speech energy range. The amplifier sets the proper level of the speech energy leaving the compressor.

At the receiving end of the circuit, the expandor restores the energy of the compressed speech signal to its original intensity range. This is done by introducing a loss that is inversely proportional to the level of the input speech energy and equal to the gain introduced by the compressor. Operation of the expandor is complementary to that of the compressor. A portion of the input speech energy (rather than the output energy as in the compressor) is routed to the rectifier control circuit to form the direct-current control signal. The attenuation of the expandor variolosser is *inversely proportional* to the level of the direct-current control signal. Thus, weak signals are attenuated more than strong signals, thereby restoring the speech sounds to their natural power range.

The performance of a compandor is controlled by three characteristics: 1) the compression-expansion ratio; 2) the companding range; and 3) the attack and recovery times. Each of these characteristics is discussed separately in the following paragraphs.

Compression-Expansion Ratio

The degree to which speech energy is compressed and expanded is expressed by the ratio of *input* to *output* power (in db) in the compressor and the expandor, respectively. The compression ratio is always greater than 1. The expansion ratio is the inverse of the compression ratio and, therefore, is

World Radio History

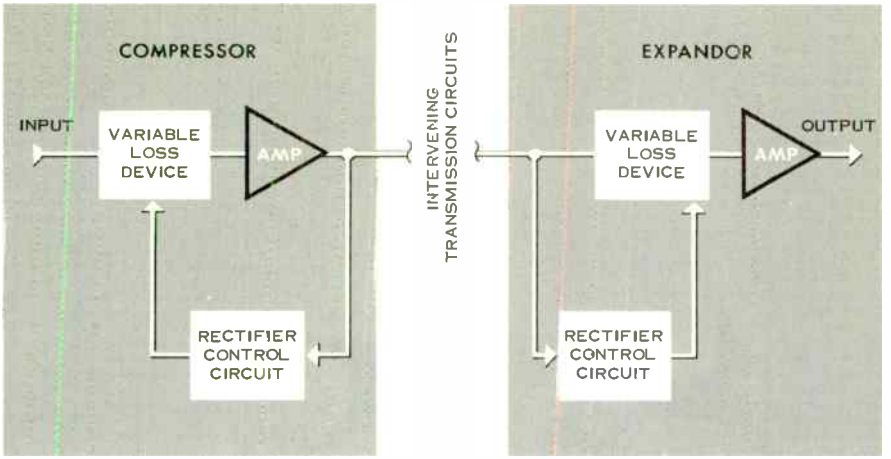


Figure 3. Block diagram of a typical compandor. Both the compressor and the expandor contain a variable loss device, an amplifier, and a control rectifier.

always less than 1. (The degree of expansion is sometimes expressed as the ratio of *output* to *input* power so that the compression and expansion ratios are equal.) If the compression (or expansion) ratio were 1, companding action would not occur and the compandor would behave like an ordinary linear amplifier.

Selection of the proper compression-expansion ratio usually involves a compromise. If the compression ratio is too high, minor irregularities in performance are likely to be magnified by the companding action, thus causing excessive distortion. On the other hand, if the compression ratio is too small, (approaching 1), the range of speech energy will not be compressed enough to realize a sufficient improvement in the signal-to-noise ratio.

A compression ratio of 2 (or 2 to 1) with a corresponding expansion ratio of ½ (or 1 to 2) provides satisfactory performance for compandors used in most telephone circuits. This means that the speech energy traveling in the circuit between the compressor and the expandor will have an intensity range of one-half its original value.

Companding Range

To be effective, companding action must occur over the wide intensity range of speech energy. Distortion may occur if the companding range is less than the usual range of speech signals. A companding range of 50 db to 60 db is usually sufficient to avoid distortion and to provide the optimum signal-to-noise improvement. The few high or low intensity signals that appear outside of this range can be attenuated or limited without seriously affecting intelligibility.

Compression and expansion of speech energy in the compandor occurs around a focal point known as the un-

World Radio History

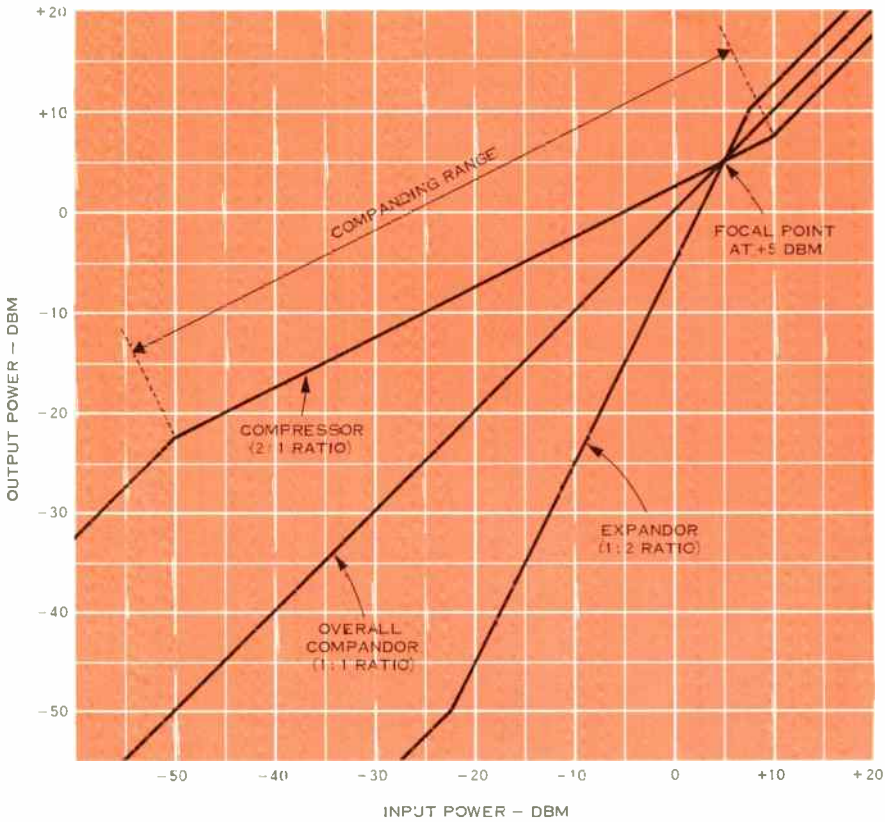


Figure 4. Ideal load characteristics of a typical compandor with a 2:1 compression ratio, a 60 db companding range, and a +5 dbm focal point. Note that the 2:1 compression ratio applies to an intensity range of −50 dbm to +10 dbm and that the expansion ratio of 1:2 applies to the compressed intensity range of −22.5 dbm to +7.5 dbm.

affected level. The focal point refers to that energy level, within the companding range, that is not affected by compandor action. Energy at the focal point level passes through the compressor and the expandor with zero loss or gain. As an example, the focal point of the companding action shown in Figure 5, occurs at +5 dbm.

Maximum noise advantage is achieved when the focal point coincides with the highest level of the companding range. In such an arrangement, all speech powers, except those at the focal point, are amplified in the compressor and attenuated in the expandor. However, to make allowances for the increase in mean power introduced by the com-

World Radio History

pressor, and to avoid the risk of increasing the intermodulation noise and the overloading that might result, the focal point is sometimes reduced by as much as 10 db to 15 db below the top of the companding range. Selection of the actual focal point depends on the desired noise advantage and the power level capability of the particular system in which the compandor is used.

Attack and Recovery Times

If the gain or loss of a compandor changed instantaneously with a change of input signal, the output signals would be badly distorted. Consequently, the operating time constants of the compandor are set so that the gain or loss varies as a function of the speech signal *envelope* and not the instantaneous amplitudes. Gain or loss, therefore, is controlled by syllabic variations of the input signal, rather than by individual speech peaks.

The time constants, which are designed into the compandor control circuits, are referred to as the attack and recovery times. Both the attack and recovery times are established in respect to a voice-frequency test signal. In accordance with CCITT Recommendations, the attack time is that interval

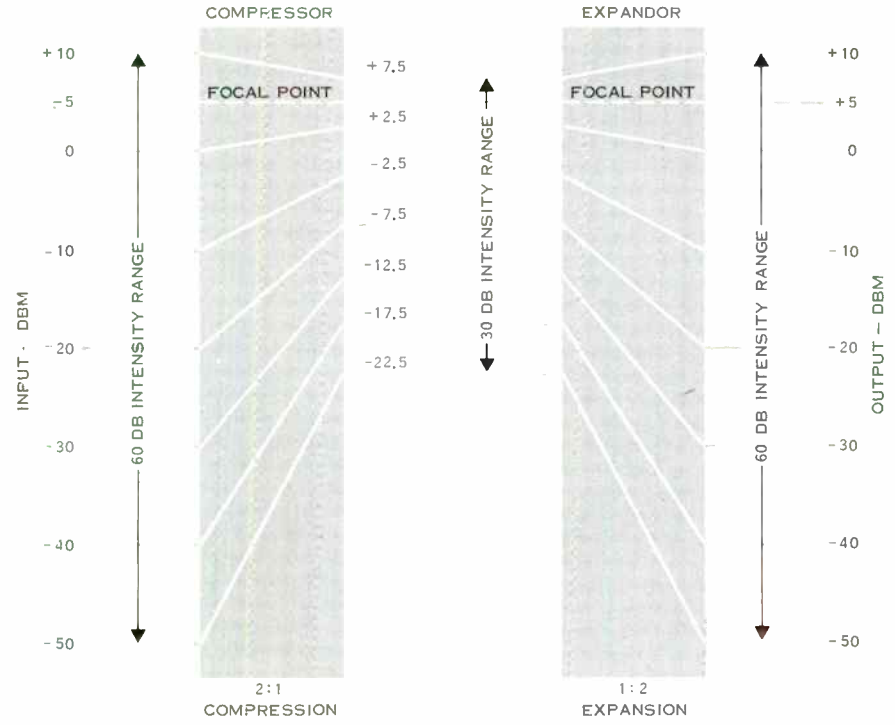


Figure 5. The effect of a compandor with a 2:1 compression ratio on various power levels. In this particular example, compression and expansion occur around a +5 dbm focal point known as the unaffected level.

World Radio History

between the instant when the power of a test signal, applied to the compressor input, is increased from −16 dbm0 to −4 dbm0, and the instant when the output voltage envelope of the compressor reaches 1.5 times its final steady-state value. The recovery time is that interval between the instant when the power of the test signal is reduced from −4 dbm0 to −16 dbm0, and the instant when the compressor output voltage envelope reaches 0.75 times its final steady-state value. The attack and recovery times for the expandor are determined in an equivalent manner. Normal values for these time constants are about 3 milliseconds for the attack time and about 13.5 milliseconds for the recovery time.

If the attack and recovery times are too abrupt, modulation products may cause distortion and noise. If the attack time is too slow, the initial parts of syllables may be mutilated. If the recovery time is too slow, the full loss of the expandor will not be inserted between syllables. In addition to establishing proper time constants, it is essential that the attack and recovery times of the compressor and the expandor exactly coincide to avoid overshoots.

Noise Advantage

The effect of a compandor in a typical carrier channel is shown graphically in Figure 6. This illustration compares the operation of two carrier channels, one with a compandor and the other without. A line noise intensity of −51 dbm was assumed to exist at the input to the receiving carrier terminal. Gains and losses are shown for a high intensity signal of 0 dbm and a low intensity signal of −31 dbm, both at the zero reference level.

In Figure 6A, where compandors are not used, the low intensity signal reaches the input of the receiving carrier terminal at −54 dbm. This is 3 db below the assumed noise power. Since the noise will also be amplified in the carrier terminal, it will reach the listener at a level 3 db higher than the low intensity speech signal. For intelligible transmission, noise power should be more than 20 db *below* the weakest speech signal.

Now consider how the same range of signals would be transmitted over the carrier channel now equipped with a compandor, as shown in Figure 6B. The −31 dbm signal is fed into a compressor where it is amplified. The amount of amplification depends upon the signal power, and in this case, is 18 db. Therefore, the signal enters the carrier terminal with an intensity of −13 dbm and reaches the receiving carrier terminal at an intensity of −36 dbm instead of −54 dbm, as occurred in the example without a compandor. The line noise power is, of course, still −51 dbm. Both the signal and the line noise are amplified 23 db in the carrier terminal, but in this case they both enter the expandor instead of going directly to the toll switchboard. The weak speech signal enters the expandor with an intensity of −13 dbm, and the noise enters the expandor with an intensity of −28 dbm. Since the expandor signal is attenuated by an amount inversely proportional to its power (in this case, 18 db for the speech signal and 28 db for the noise), the margin between the signal and the noise has been increased. For the same signal which was 3 db

World Radio History

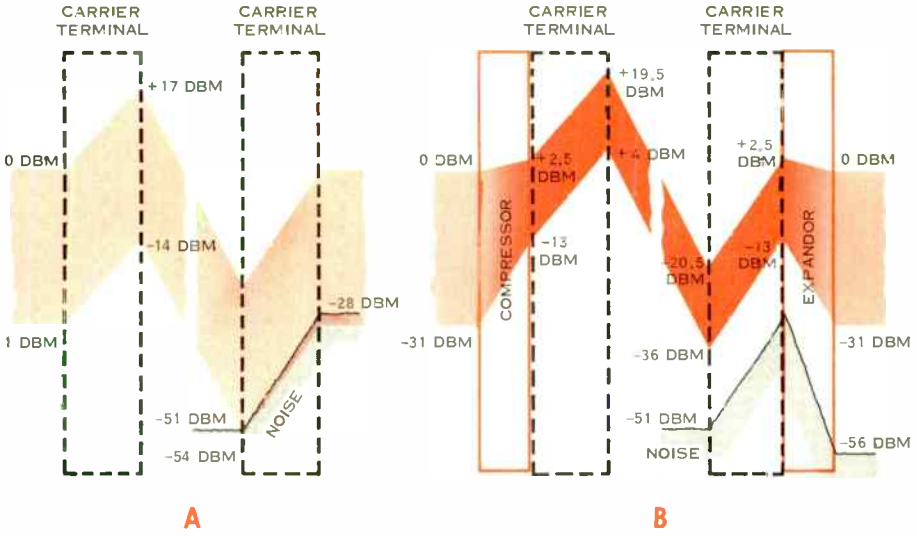


Figure 6. Using compandors, toll quality transmission can often be achieved over a circuit otherwise unsuitable because of noise. Diagram A shows a typical circuit, without a compandor, where the weakest speech signal at the receiving end is 3 db below *the noise level. Diagram B shows the same circuit with a compandor added. The weakest speech signal at the receiving end is now 25 db* above *the noise level, resulting in a 28 db noise improvement.*

below the noise when compandors were not used, the circuit now achieves a signal-to-noise ratio of 25 db. Compandor action is not apparent to the telephone listener, except for the desirable reduction in interference.

Crosstalk Advantage

Compandors are especially helpful in reducing the crosstalk problems encountered in multi-channel carrier systems. High speech signal peaks constitute the greatest source of crosstalk. At the compressor, the amplitude of such loud signal peaks is reduced, thus preventing crosstalk to adjacent channels due to circuit overloading.

A crosstalk advantage is also realized through the action of the expandor at the receiving end of a circuit. The expandor takes advantage of the fact that crosstalk, like noise, is not too noticeable during speech but becomes objectionable during silent periods.

The expandor, as explained previously, adjusts to a condition of maximum loss when speech signals are not present. As a result, the relatively weak crosstalk signals that enter the receive circuit are greatly attenuated during such silent periods and thus do not disturb the telephone listener.

Applications

Compandors are used in telephone voice channels to make noisy circuits satisfactory for toll transmission service. On physical and phantom voice-

World Radio History

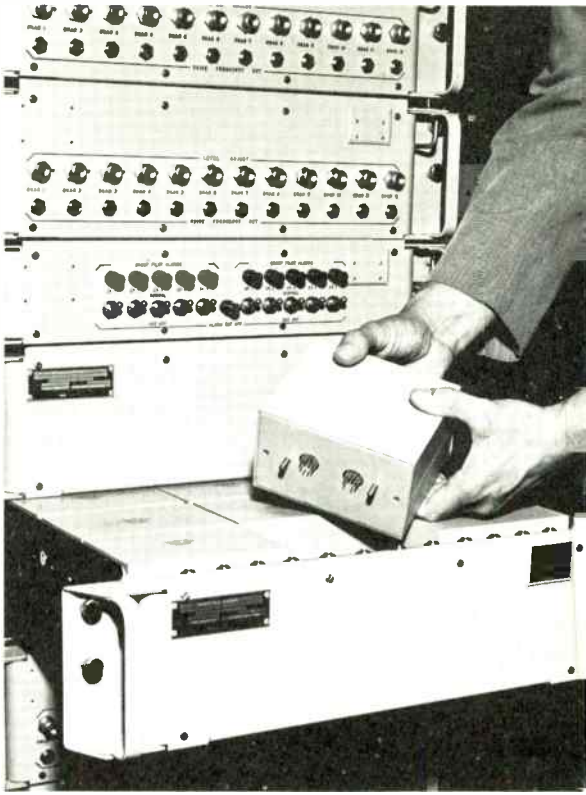


Figure 7. The transistorized compandor developed by Lenkurt for the AN/FCC-17 multiplexer is designed to process AM and FM data signals with little distortion. Compressors and expandors are packaged separately in hermetically-sealed containers to withstand rugged military use. Photograph shows a plug-in compressor unit withdrawn to exhibit its sturdy mechanical construction. The compandor is especially useful in military networks to assure privacy in cable carrier systems where crosstalk may occur or to counteract the effects of enemy jamming in radio systems.

frequency circuits the compandor is particularly valuable in compensating for the effects of power line induction and noise pickup from other random sources. Many older circuits which have fallen below transmission standards can be restored to toll quality by using compandors.

Compandors are employed effectively on amplitude-modulated carrier systems to reduce the effects of crosstalk as well as to improve the signal-to-noise ratio. Repeater spacing of carrier systems operating over wire lines or cable pairs is often limited by line noise conditions or excessive near-end crosstalk. With the noise advantage offered by compandors, repeater spacing may be limited only by the maximum system gain.

Multi-channel microwave radio carrier systems can achieve greater fading margins, longer transmission paths, and make use of more repeaters to extend the system because of the additional signal-to-noise advantage of the compandor. Such an advantage can also reduce the radio antenna gain requirements, thus permitting the use of smaller antenna systems. When compandors are employed on multi-channel radio-carrier systems, it is necessary to consider an average power increase of about 5 db for voice circuit loading. This additional power must either be

World Radio History

attenuated or added to the nominal figures used to calculate the loading effect on radio equipment.

Substantial savings can often be realized in designing and manufacturing carrier systems with compandors built into the channel equipment. Because of the noise and crosstalk improvement, design requirements of line filters, in addition to other carrier terminal and repeater equipment, can be lowered—resulting in lower equipment costs.

Data Transmission

In modern transmission systems, there is the ever present need to transmit data signals over already existing voice channels. However, it is not desirable to send data over voice circuits containing compandors since they offer little or no noise improvement for such signals. In addition, compandors tend to introduce intermodulation distortion, and can be especially detrimental to pulse-type data signals with changing power levels. When practical, therefore, compandors should be removed from voice circuits that are to be used for data. Nevertheless, it is possible to transmit data signals through compandors without too much signal degradation. It is currently being done over telephone networks that provide data transmission service on a dial-up basis, and for systems employing in-band signaling.

Conclusions

When measured under idle circuit conditions, compandors provide a noise advantage equal to the maximum gain of the compressor. For the compandor characteristics shown in Figure 6B, the noise advantage would be 28 db. In actual practice, however, the effective noise advantage is always less than the value achieved during idle conditions, and typically ranges from about 20 db to 25 db. The actual noise advantage depends on such things as the speech power level, the noise level, the compandor focal point, and the companding range.

The compandor offers relief, but not a cure, in the fight to overcome the disturbing effects of noise and crosstalk—the principal enemies of communications. It is, therefore, a remedial device that offers a practical method of improving the quality of voice transmission over otherwise marginal or unsatisfactory telephone circuits. •

BIBLIOGRAPHY

1. R. C. Mathes and S. B. Wright, "The Compandor — An Aid Against Radio Static," *AIEE Transactions*, Volume 53, pages 860 to 866, June section, 1934.
2. J. Lawton, "The Reduction of Crosstalk on Trunk Circuits, by the use of the Volume Range Compressor and Expandor," *Post Office Electrical Engineers Journal*, Volume 32, pages 32 to 38, London, England, April, 1939.
3. A. C. Dickieson, D. Mitchell, and C. W. Carter, Jr., "Application of Compandors to Telephone Circuits," *AIEE Transactions*, Volume 65, pages 1079 to 1086, August, 1946.
4. "Basic Principles of Compandors and Their Application to Carrier Systems." *The Lenkurt Demodulator*, March, 1953.
5. "Crosstalk," *The Lenkurt Demodulator*, November, 1960.
6. "The Universal Voice Channel," *The Lenkurt Demodulator*, March, 1964.

World Radio History

World Radio History

the Lenkurt® Demodulator

VOL. 10 NO. 11 NOVEMBER, 1961

dba

and Other Logarithmic Units

Logarithmic units are widely used in the communications industry as the most practical and convenient means of expressing the extremely wide range of power ratios encountered. The ratio of very strong signals and noise may be 100 million billion to one, or more. When expressed in decibels, the most common logarithmic unit, this tremendous range becomes, simply, 180 db. Other, more specialized units, such as dba, dbrn, vu, and dbx are related to the decibel but require more complex definitions.

In this article dba, dbrn and vu are explained in detail. Shorter explanations of other units are also included.

The decibel is commonly used in the communications industry to express a ratio between two quantities of power. Neither quantity needs to be defined to express the ratio in decibels or *db*, and for many purposes, a knowledge of the ratio alone is sufficient. For example, the gain of a linear amplifier or the attenuation of a pad can be expressed in decibels without knowing either the input or output power of the device. Frequently, however, there is a need to know the ratio between signal (or noise) power at some point in a circuit and some fixed, known quantity of power. In this case, it is customary to express the ratio as so many db above or below the reference power.

The most common reference power used in the telephone industry is *one milliwatt*. Because signal power is almost always undergoing attenuation (a division process) or amplification

About this article

This article is revised from the January, 1954 article by A. M. Seymour, then Editor of the DEMODULATOR. *The original article has been reprinted many times and has been adopted without change as a Bell System Practice. Because of extreme continued interest in this subject, the article has been brought up to date by the addition of material on measurement units that have come into general use since the original article was written, and is reprinted here again.*

World Radio History

(a multiplication process) the expression of power directly in watts or milliwatts would often require lengthy and cumbersome calculations. A more convenient method of indicating an amount of power is to express it as being so many db above or below a reference power of one milliwatt, because adding and subtracting decibels provides the same result as multiplying and dividing power. Because of its widespread usage, "decibels above or below one milliwatt" is usually abbreviated $\pm dbm$.

In addition to dbm, there are several other logarithmic units in use in the telephone industry which are expressed as db above or below some reference power. The most important of these are *dba, dbrn, dbx,* and *vu*.

Dba and dbrn (formerly written db-RN) are used to indicate the actual interfering effect of noise in a communications channel, by relating it to the interfering effect of a fixed amount of noise power or reference noise. *Dbx* is a unit for expressing crosstalk coupling on transmission lines. *Vu* is used to designate the ratio between the "volume" (or loudness) of spoken or musical sounds and a reference volume.

Dba and dbrn

Dba and dbrn are closely related units; in fact, the abbreviation *dba* means, effectively, *dbrn adjusted*. Both terms originated from research conducted by the Bell Telephone Laboratories and the Edison Electrical Institute to determine the transmission impairment caused by noise interfering with speech. Since noise may consist of random frequencies with widely varying amplitudes, it was necessary to evaluate the interfering effects of single frequencies or relatively narrow bands of frequencies, to obtain usable data.

A large number of listening tests were made with different tones introduced as interference. The degree of interference was determined by comparing the power of each tone with the power of a 1000-cycle tone that created the same degree of interference. For example, if the interfering effect of a particular tone was to be determined, that tone was superimposed on a specially selected conversation at a reference power level. When the interfering tone was removed, a 1000-cycle tone was superimposed on the same conversation and its power level adjusted until the listener judged that it had the same interfering effect. The difference noted between the power levels of the selected tone and the 1000-cycle tone was then considered to be the difference in interfering effect.

When this same test was performed for a number of different tones in the voice frequency spectrum and for a number of different listeners using the same apparatus, it was possible to plot a graph such as Figure 1 showing the relative interfering effects of different frequencies in the voice frequency spectrum compared to 1000 cycles. Such curves are called *weighting* curves. With this information available it was possible to construct equalizing networks such that each component frequency of the voice frequency spectrum was attenuated in the same manner as it appeared to be attenuated by the average ear with the listening test apparatus. By using these equalizers in conjunction with a suitable amplifier, rectifier and d-c meter it was further possible to measure electrically the interfering effect of any frequency or combination of frequencies.

Since any noise or tone superimposed on a conversation has an interfering effect, it was desirable to express all quantities of interfering effect in positive numbers. To accomplish this, a power of 10^{-12} watt or -90 dbm at 1000 cycles was selected as the reference power because a 1000 cycle tone having

World Radio History

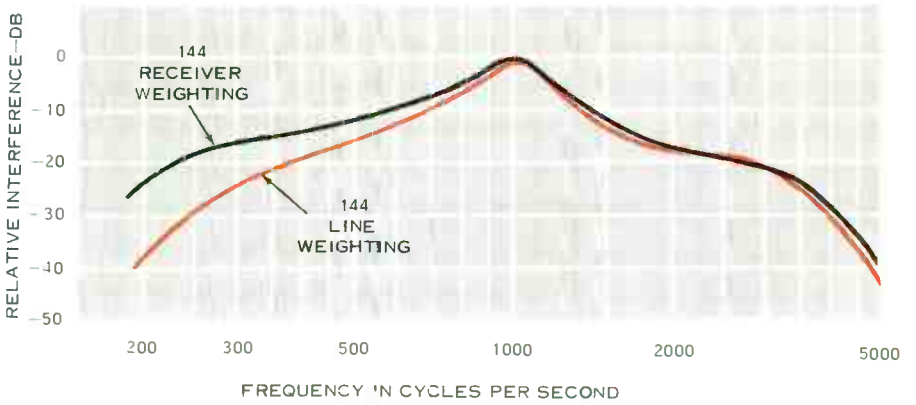


Figure 1. Relative interfering effects of different frequencies when using 144-type telephone sets. Red curve includes attenuating effect of complete loop, including exchange circuit. Black curve shows only attenuation effects of 144 set itself.

a power level of -90 dbm appeared to have negligible interfering effect Therefore, all other noise powers likely to be encountered would have a positive interfering effect that could be expressed in *db above reference noise* of -90 dbm at 1000 cycles or, in abbreviated form, as dbrn.

These early experiments and tests to evaluate the interfering effect of noise were made with Western Electric Type 144 handsets and resulted in the weighting curves in Figure 1. The line weighting curve (red) of Figure 1 includes the attenuating effect of a typical exchange circuit, subscriber loop, and telephone set. The receiver weighting curve (black) of Figure 1 includes only the attenuation effects of a telephone set itself.

Later, an improved type of handset (Western Electric Type F1A) came into general use with a type F1 transmitter and HA1 receiver. This equipment was less sensitive at the 1000-cycle reference frequency, but had a more uniform frequency response. When tests similar to those used for 144 telephone sets were conducted with this handset, the weighting curve shown in Figure 2 was obtained and designated F1A weighting. The tests indicated that the new handset gave approximately a 5-db improvement over the electrical and acoustical performance of the 144 handset when using line weighting. Because of the differences in sensitivity and frequency response, noise measurements with the F1A weighting provided indications about 5 db higher than with the 144 weighting. Rather than change existing standards, a new reference noise power of -85 dbm ($10^{-11.5}$ watt) was introduced which produced identical noise measuring set readings for equal transmission impairments. The change in reference noise power necessitated a change in the units used to express interfering effect and resulted in the adoption of a new unit called *dba* — which means "decibels adjusted."

Recently, a newer line weighting was introduced in the Bell System, reflecting the performance of more recent equipment. As in the change from 144 weighting to F1A weighting, the new

World Radio History

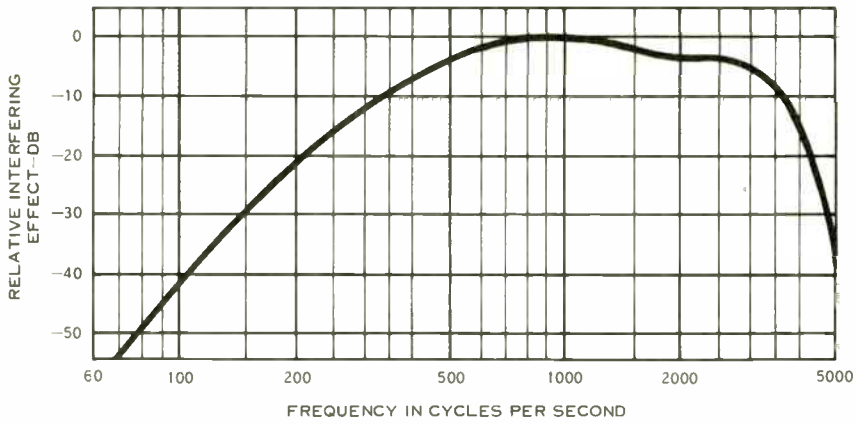


Figure 2. F1A line weighting characteristic. This curve closely approximates line weighting characteristics of most of the world's telephones, and has been adopted by CCITT.

characteristic — which is called C-message weighting — has required a change in the noise power to which it is referred. Since the new equipment improves on the older equipment, an even higher reference power than the −85 dbm F1A reference would be required to express equal interfering effects with equal numbers. Since this might have resulted in some measurements having to be made in terms of "negative" values of noise, the original −90 dbm (10^{-12} watt) reference power was reinstated. Interference measured with this weighting is expressed in "dbrn C-message" units to distinguish it from the earlier "dbrn 144-line."

Heretofore, all noise measurements made with either 144 weighting or F1A weighting have been expressed in dba, since, when using 144 weighting, dba and dbrn are numerically equal, represent the same amount of noise power, and have the same interfering effect. This is not true of dbrn C-message, however. Although the same reference noise power is used as in 144 weighting, the broader frequency response of C-message weighting results in much less noise power being attenuated by the weighting characteristic. Thus, a 3-kc band of uniform noise is attenuated 8 db by the 144 weighting characteristic, 3 db by the F1A weighting, but only 1.5 db by the C-message weighting.

Where the noise is known to be evenly distributed across a 3000-cycle voice frequency band, measurement can be made with a suitable transmission measuring set or ac voltmeter. Because the weighting networks attenuate various frequencies differently, one milliwatt of evenly distributed noise (flat noise) produces only 82 dba of interfering effect with 144 and F1A weighting. Therefore, measurements of *flat noise* in dbm can readily be converted to dba by adding 82 to the reading (−50 dbm +82 = +32 dba). In the case of C-message weighting, however, noise power is calculated by adding 90 −1.5, or 88.5 to the flat noise power. If noise power is again −50 dbm,

Noise=

$-50 + (90 - 1.5) = 38.5$ dbrn $_{\text{C-message}}$

World Radio History

In case of measurements made at 1000 cps, however, it is only necessary to make a comparison with the reference power, since there is no weighting effect on a 1000 cps tone. Thus, a 1000 cps signal having a power of 0 dbm would yield 90 dbrn (144 line), 85 dba, and 90 dbrn (C-message).

Although these conversions are quite straightforward for 1000-cycle tones and flat noise, conversion of other types of noise power to meaningful noise terms may become quite complicated. Noise, as found in transmission lines and electronic apparatus, varies widely as to its component frequencies and their relative amplitudes. It seldom consists of a single frequency. The chief source of noise in open-wire voice frequency circuits is induction from power lines and apparatus.

In open-wire carrier and radio systems, noise comes from a variety of sources including power lines, atmospheric disturbance and interference from radio transmitters. While cable circuits are not subject to any great degree of noise from atmospherics or power lines, the low power levels used make them subject to noise caused by the thermal agitation in circuit components. Noise at carrier frequencies is often distributed evenly across a channel bandwidth. In open-wire voice frequency circuits, however, noise is more often concentrated at certain frequencies, usually odd harmonics of the local power frequency.

Single interfering frequencies other than 1000 cycles can also be measured in dbm and converted to dba by use of the weighting curves. For example, if a 300-cycle interfering tone is measured to be -40 dbm, the F1A line weighting curve shows that it is 15 db less interfering than a -40 dbm, 1000 cycle tone;

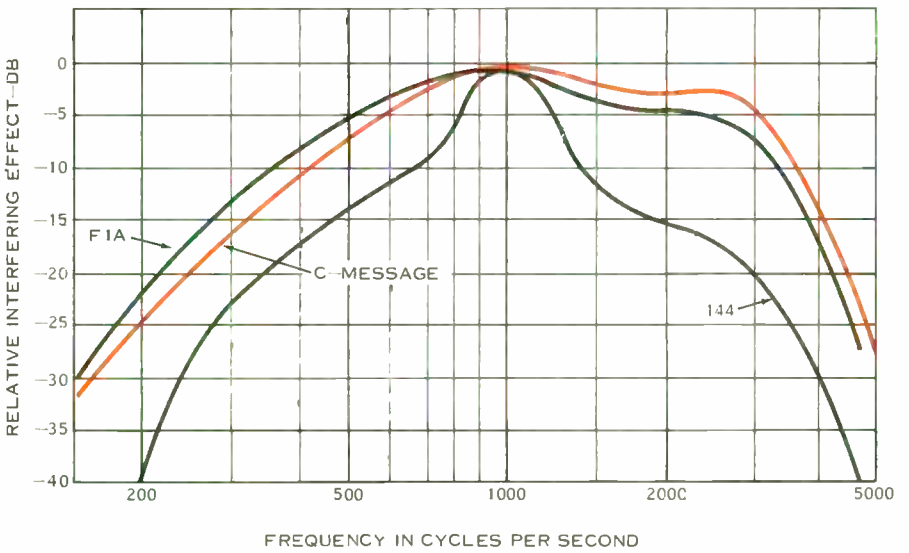


Figure 3. New C-message weighting characteristic, compared with 144-line weighting and F1A-line weighting. In all three curves, interfering effect of various frequencies is referred to 1000-cycle interference.

World Radio History

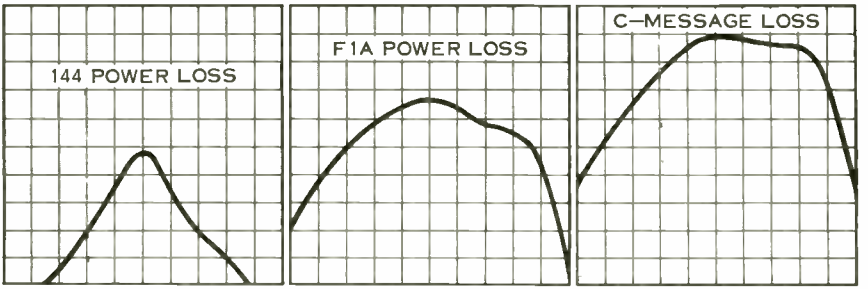


Figure 4. Comparison of loss characteristics of the three weighting curves. Attenuation of wide-band noise or signal is proportional to area outside the curves.

in other words, its interfering effect is the same as that of a 1000-cycle tone of -55 dbm. A -55 dbm, 1000-cycle tone is equivalent to 30 dba. Therefore, the -40 dbm, 300-cycle tone has an interfering effect of 30 dba.

Volume Units

Where programs and certain other types of speech or music are being transmitted, monitoring of the program volume level is necessary to maintain a constant average volume. Failure to maintain the program volume level constant may cause overmodulation of line amplifiers or radio transmitters and cause blasting and distortion from listeners' loudspeakers or handsets. If a simple db meter or voltmeter is bridged across the circuit to monitor the program volume level, the indicating needle will try to follow every fluctuation of program power and will be difficult to read and will have no real meaning. Also, different meters will probably read differently because of differences in their damping and ballistic characteristics.

To provide a standardized system of indicating volume, a special instrument and set of units were created. These instruments are called VU meters or volume indicators and the units of measurement are *volume units* or *vu*. Ordinarily, volume can be measured in vu only on these special instruments because the volume unit is based on the readings of these instruments under a specified set of conditions. One exception to this rule is the Western Electric 2B noise measuring set which can be calibrated to read in vu.

The indicating instrument used in vu meters is a d-c milliammeter having a slow response time and damping slightly less than critical. If a steady sine wave is suddenly applied to the meter, the pointer will move to within 99 percent of its steady state value in 0.3 seconds and overswing the steady state value by 1.0 to 1.5 percent.

A standard volume indicator (meter and associated attenuator) is calibrated to read 0 vu when connected across a 600-ohm circuit carrying 1 milliwatt of sine-wave power having a frequency between 35 and 10,000 cycles per second. For complex waves such as speech, a vu meter will read some value between the average and the peak values of the complex wave. There is no simple relation between the volume measured in vu and the power of such a complex wave. The actual reading will depend on the particular wave shape. For steady sine waves within the frequency range of the instrument, the reading in vu will be equal to the reading of a db meter

World Radio History

in dbm connected across the same circuit.

Dba0 and Dbm0

These terms are coming into widespread use in microwave and carrier systems to express noise and signal power in terms of the levels that should exist at a so-called zero transmission level point (0 TLP). As recently defined by the United States Electronics Industries Association:

The term dba0 is a measure of noise power with reference to zero dbm at the Reference (zero) Transmission Level Point. Noise powers measured at any Transmission Level Point can be expressed in dba0 by correcting the noise power measured for the difference in level between the point of measurement and the Reference Transmission Level Point. The relative noise power in dba is obtained from a power measurement of noise using F1A weighting.

Example: A noise measurement of +20 dba measured at a −4 db point is equivalent to +24 dba0.

The term dbm0 is a measure of power with reference to zero dbm at the reference transmission level point (RTLP). Powers measured at any transmission level point can be expressed in dbm0 by correcting the power measured for the difference in level between the point of measurement and the Reference Transmission Level Point.

Examples: (1) A tone of +36 dbm measured at a +19 db transmission level point is equivalent to +17 dbm0.

(2) A tone of +17 dbm0 is equivalent to +7 dbm measured at a −10 db transmission level point.

Other Units

Various other units are also used in the telephone industry and other sections of the communications field. They include *dbx, dbw, dbRAP, dbv,* and others. Of course, dbx is the most common in the telephone industry.

Dbx is used to indicate crosstalk coupling in telephone circuits (See DEMODULATOR, *November,* 1960). Like dba and dbrn, it may be measured with a noise measuring set. Dbx means decibels above reference coupling. Reference coupling is defined as the coupling necessary to cause a reading of 0 dba on the disturbed circuit when a test tone of 90 dba is impressed on the disturbing circuit. Both values of dba are for the same weighting.

The other units mentioned above are quite simple. Dbw are decibels referred to one watt. Dbk are decibels referred to one kilowatt. Both dbw and dbk are often used to indicate radiated power from radio transmitters. DbRAP means decibels above reference acoustical power which is defined as 10^{-16} watts. Dbv designates decibels referred to one volt. Other logarithmic units in use are similar to these mentioned. •

BIBLIOGRAPHY

1. H. E. Kent and R. G. McCurdy, "Relative Interfering Effects Of Different Single-Frequency Noises In Telephone Circuits," *Engineering Reports, Edison Electric Institute and Bell Telephone System,* 4; 1937, p. 163.
2. H. E. Kent and R. A. Shetzline, "Frequency Weighting For Message Circuit Noise," *Engineering Reports, Edison Electric Institute and Bell Telephone System,* 5; 1943, p. 183.
3. A. J. Aikens and D. A. Lewinski, "Evaluation Of Message Circuit Noise," *Bell System Technical Journal;* July, 1960.
4. W. T. Cochran and D. A. Lewinski, "New Measuring Set For Message Circuit Noise," *The Bell System Technical Journal;* July, 1960.
5. "Baseband Characteristics of the Microwave Radio and Multiplex Equipment," Electronic Industries Association, *Standards Proposal No. 714.*

World Radio History

World Radio History

the Lenkurt Demodulator

VOL. 10 No. 5 MAY, 1961

How to evaluate

Radio and Carrier NOISE PERFORMANCE

The tremendous growth in recent years of radio-carrier communication has been paralleled by an equally impressive, but vastly confusing, diversity of language and engineering practices. Manufacturers better known in other fields of electronics have appeared in the field of multichannel communications, and each new entrant appears to speak a different language inherited from a previous field of experience. Comparison or evaluation of systems — either existing or proposed — may be severely handicapped by the confusion of engineering practices and language used. This article reviews the noise performance of basic items of equipment, relates some of the different ways of expressing performance, and describes methods for measuring the noise performance of major items of equipment.

Of the various performance characteristics used to describe the performance of a communications systems, the amount of *noise* present in a communications channel provides one of the best immediate measures of system quality. Noise is the natural enemy of communication, working constantly to obscure the identity of a signal. Such technical characteristics as frequency and level stability, frequency response, and delay distortion are, more or less, under the control of the equipment designer and may be held to any desired value. *Noise performance*, on the other hand, is controlled not only by equipment design, but also by a combination of such other factors as increased traffic load, system layout, and operating practices.

The noise in a communications system will be found to consist of *intermodulation noise*, which increases with load, and *residual noise*. Residual noise is a useful "catch-all" term to designate all noise other than intermodulation noise. Basically, it is thermal noise from various sources such as electron tubes, transistors, modulators and the like and

World Radio History

is present even in the absence of a signal. For this reason, residual noise is often called *idle*, *background*, or *intrinsic* noise. Residual noise is not affected by the amount of traffic carried by the system. By contrast, intermodulation noise becomes greater with increased traffic, and after a certain "break point" in load-handling capability is exceeded, intermodulation noise becomes excessively high.

Carrier Noise Contribution

The amount of noise which the carrier equipment contributes to the system is largely dependent upon the design of the equipment itself. Although part of the carrier noise contribution is the result of intermodulation distortion, this can be minimized by good design. Speech limiting is often used in each channel modulator to restrict the very wide range of signal levels from individual talkers. Group amplifiers and modulators are designed to handle all but the most extreme loads that would be imposed if all channels were used simultaneously.

If a frequency plan is selected which allows each modulation step to handle less than one octave, second-order intermodulation products will fall outside the passband of the carrier filters and be suppressed. Because of these design features, there is usually more residual or idle noise than intermodulation noise in properly operated carrier equipment.

Radio Noise Contribution

The noise originating in the radio portion of a communications system comes from many sources and is more difficult to control than noise from the carrier system. Frequency modulation radio is able to overcome much of the residual noise appearing in the radio system by distributing the signal over a wide radio bandwidth. However, this exchange of bandwidth for lower noise is also proportional to the signal level appearing at the receiver. Thus, fading, poor system layout, or other factors which tend to reduce the input signal to the receiver, make an important contribution to system noise.

In a frequency-modulation radio system, the FM noise reduction is directly proportional, db for db, to the input signal level, once the "FM improvement threshold" is exceeded. The threshold improvement occurs approximately 10 db above the absolute noise threshold (or so-called "tangential" threshold) of the system, and is defined as the signal level at which the system begins to suppress background noise. Between the absolute threshold (carrier-to-noise ratio of unity) and the FM improvement threshold, noise is not reduced in the system by increased input signal levels.

In conventional communications practice, the noise which occurs at the FM improvement level is still excessive. Normally, a carrier-to-noise ratio of about 30 db is required to provide minimum-quality service. The more sensitive or noise-free the receiver, the lower the signal level at which adequate communications quality is obtained.

At the transmitting end of the system, the exchange of bandwidth for noise is improved by raising the level of the modulating signal, thus increasing the frequency deviation of the radio signal. Excessive deviation, however, increases intermodulation distortion tremendously, with the result that system noise increases despite the reduction of residual noise. For this reason, the best noise performance of a radio system results from a very careful balance between residual noise caused by low signal level, and intermodulation noise resulting from excessive signal level, as shown in Figure 1. The radio equipment cannot reduce either residual or

World Radio History

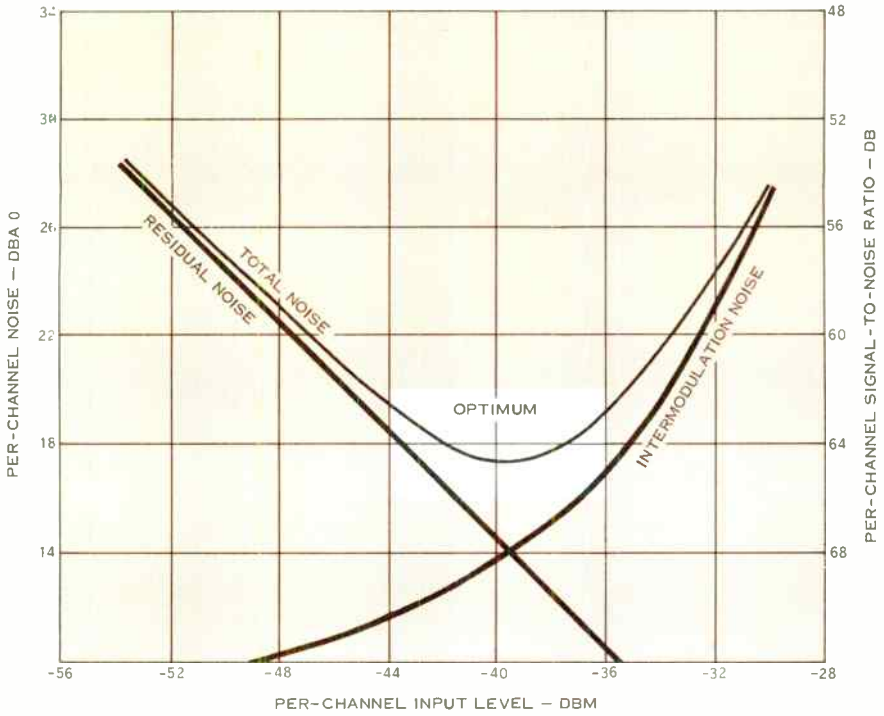


Figure 1. Increased signal levels reduce residual noise in FM radio system, but increase intermodulation. Correct operating levels provide balance between the two noise sources to achieve lowest total noise.

intermodulation noise that originates in the carrier system and accompanies the carrier signal.

Interpreting Noise Terminology

When planning a communications system, it may be more confusing than enlightening to compare the published noise performance of equipment from various manufacturers. One manufacturer may refer to "notch-to-no-notch ratio", another will specify intermodulation distortion in terms of so many db "signal-to-noise ratio" and yet another will specify noise in "per-channel dba."

The expression "notch-to-no-notch ratio" is jargon for *Noise Power Ratio*, one way of expressing the intermodulation distortion occurring in radio equipment when using one of the standard techniques for measuring intermodulation (DEMODULATOR, *December*, 1960). Noise Power Ratio (*NPR*) provides an excellent indication of intermodulation performance, when measured under standard conditions, but gives no direct indication of the *overall* performance of the system. The idle-noise performance of different types of equipment may vary widely. For example, the amount of residual noise

World Radio History

contributed by the receiver local oscillator klystron may be substantial, and intermodulation performance figures will not take this or other types of idle noise into account.

The term "signal-to-noise ratio" (S/N) originated in single-channel communications practice and generally took into consideration only the background or residual noise in a single radio channel. With the growth of multichannel communications, it is also used to express the total intermodulation and residual noise in a single radio channel, and is frequently referred to as "per-channel flat signal-to-noise ratio." Basically, it expresses the ratio, in decibels, of signal power to total noise power in a channel. It does not take into account the actual *interfering* effect of the noise on the signal in complete circuits.

Decibels adjusted or *dba*, originated in the telephone industry as an expression of overall system noise performance. Strictly speaking, the term dba implies that the frequency response or weighting of the voice frequency equipment used is "F1A" weighting. This method of specifying noise performance is especially practical. It takes into account not only special types of noise or noise in particular items of equipment, but also the effects of all system noise.

Converting NPR to S/N

Because all of these expressions are in common use, comparisons between equipment may require that one term be translated to relative values in another. The following method of converting Noise Power Ratio to signal-to-noise ratio has been proposed for adoption by the United States *Electronic Industries Association* (E.I.A.) in its standards.

Noise Power Ratio or so-called "notch-to-no-notch" ratio is normally measured in a narrow band roughly equivalent to the bandwidth of a typical communications channel. For this reason, the first step in converting *NPR* to *S/N* requires that the Bandwidth Ratio (*BWR*) of the system be calculated in order to obtain the proper relationship between the bandwidth of the entire baseband and the bandwidth of the slot or channel:

$$BWR = 10 \log_{10} \frac{\text{occupied bandwidth}}{\text{channel bandwidth}}.$$

Similarly, it is necessary to calculate (in decibels) the ratio of the noise power applied to the entire baseband, to the nominal signal power appearing in a single channel. If signal power values are expressed in watts, this *Noise Load Ratio (NLR)* is

$$NLR = 10 \log_{10} \frac{\text{Baseband Noise}}{\dfrac{\text{Test Signal}}{\text{Channel}}}.$$
$$\text{Test-tone power}$$

If the Baseband Noise Test Power is expressed in dbm0 (decibels referred to 1 milliwatt at the reference or zero transmission point), the Noise Load Ratio equals dbm0, and no calculation is necessary. Figure 3 shows the noise load power equivalent to various numbers of voice channels and tone signals, as recommended by the C. C. I. R. and the Electronics Industries Association.

When Bandwidth Ratio and Noise Load Ratio have been determined, the per-channel signal-to-noise ratio may be calculated from the Noise Power Ratio:

$$S/N = NPR + BWR - NLR.$$

As an example, the per-channel S/N of a 300-channel radio system in which intermodulation is quoted in terms of a Noise Power Ratio of 50 db. It was

World Radio History

specified that *NPR* was measured in a 3 kc "slot," and that the measurements were made with the baseband loaded with noise in the frequency range 60 to 1300 kc. Then,

$$S/N = NPR + BWR - NLR$$
$$= 50 + 10 \log \left(\frac{1300 - 60}{3} \right) - 9.8$$
$$= 50 + 26.2 - 9.8$$
$$S/N = 66.4 \text{ db.}$$

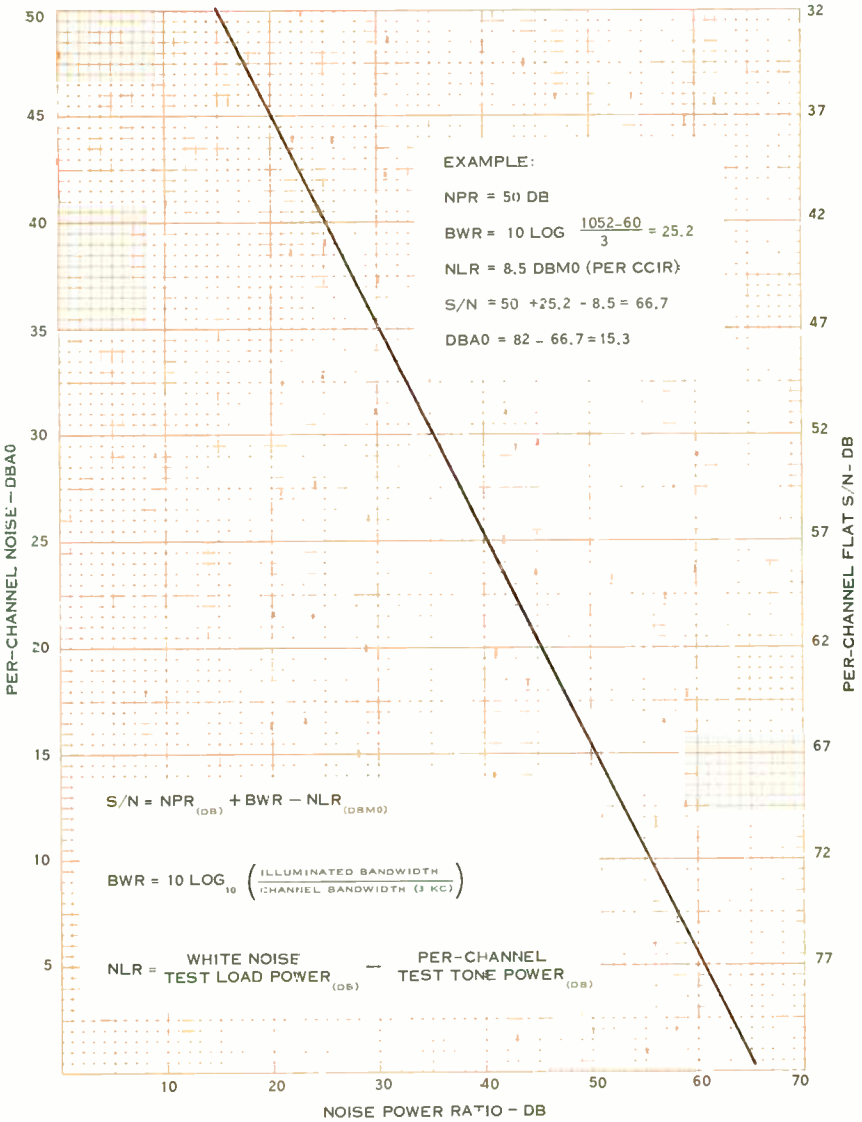


Figure 2. Relationship between Noise Power Ratio, per-channel Signal-to-Noise ratio, and dba is calculated for 240-channel system according to Electronic Industries Association formula.

World Radio History

The conversion of the *NPR* to *S/N* gives the signal-to-noise ratio of the channel without weighting or adjustment for the response of the human communicator and his handset. In systems where the channel response is essentially "flat" and the noise is being measured with an unweighted meter, the equivalent F1A-weighted signal-to-noise ratio may be obtained by adding 3 db to the unweighted meter reading. This addition of 3 db compensates for the noise power lost at the high and low portions of the channel passband.

Converting S/N to Dba

By definition, *dba* refers to decibels of noise power above a reference noise power, with an adjustment factor included to compensate for weighting.

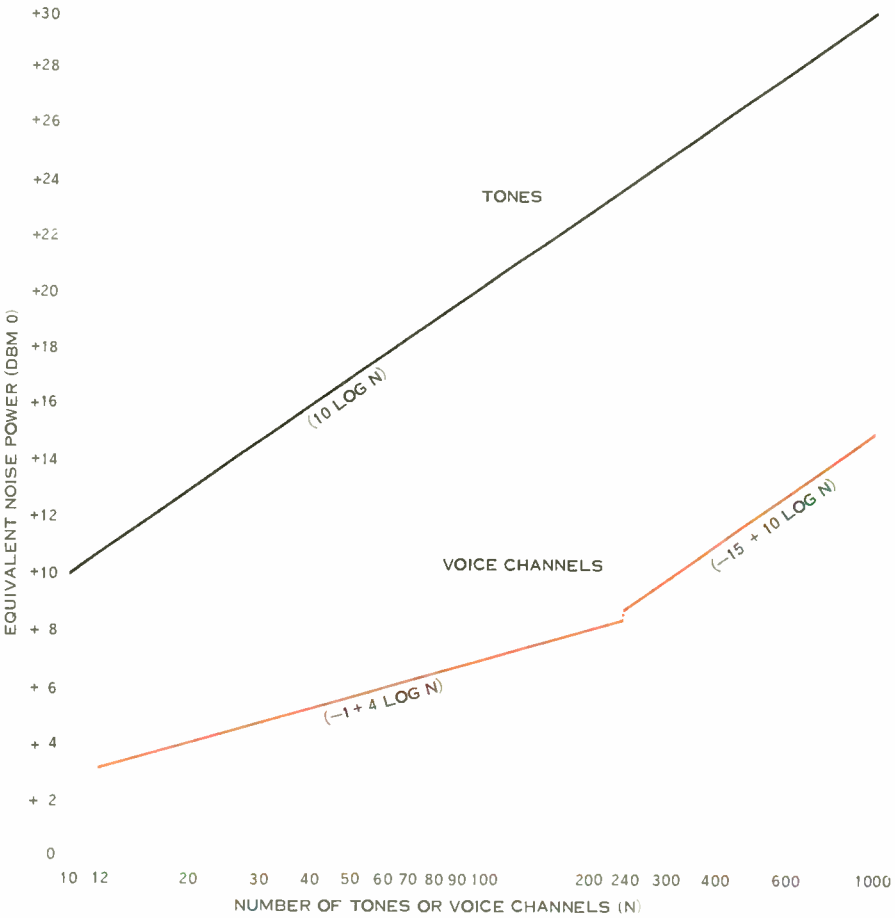


Figure 3. Equivalent white noise test signal for simulating load provided by various numbers of voice channels or tone signals. Lower values for simulating voice channels result from lower activity factor. Values shown may be used directly for Noise Load Ratio when converting NPR to S/N and dba.

World Radio History

Even though the equipment from which the F1A weighting was derived has been superseded by newer equipment having better performance, F1A weighting continues to be used almost universally because it provides a very close approximation to the performance of most of the world's telephone equipment.

The reference noise power to which dba is referred is −85 dbm. To obtain dba, it is only necessary to calculate how many db above this reference power the signal is. For flat voice channels, the corrected reference level is −85 +3 or −82 dbm. Therefore, in this case

$$dba = 82 - (S/N).$$

Thus, a channel having a signal-to-noise ratio of 60 db exhibits 22 dba noise.

To convert the per-channel S/N of the hypothetical 300 channel system to a value in dba:

$$S/N = 66.4 \text{ db}$$

$$dba = 82 - 66.4$$

$$Noise = 15.6 \text{ dba} \quad (F1A \text{ weighted}).$$

Measurements During Installation

At the time a new communications system is installed, or whenever additions are made to an existing system, careful measurements should be made of noise performance. When properly conducted, these measurements will provide reference standards with which the long-term noise performance of the system can be evaluated, particularly if the traffic load or size of the system is to be increased. Separate measurements of carrier and radio equipment are required in order to determine their individual noise contributions.

Carrier Noise Measurements

Generally, it is impractical to obtain accurate field measurements of the intermodulation noise contributed by the carrier equipment; this would require that all channels be loaded simultaneously by suitable test signals. However, the intermodulation noise contributed by the carrier system may be estimated at periods of peak traffic if the noise performance of the radio has been determined previously. Such estimates may not be very accurate, however, since the load imposed by a large number of voice channels varies from moment to moment and may never reach a value which can be identified as "peak" load. For this reason meaningful evaluations of carrier noise usually consist of "back-to-back" measurements of idle noise.

Unlike cable or open-wire carrier systems, carrier systems for use with radio transmit and receive on the same band of frequencies. This permits the output of the transmit terminal to be connected back into the receive terminal on a "back-to-back" basis, thus permitting measurements of the carrier residual noise at one terminal.

For back-to-back measurements of idle noise, the carrier equipment is disconnected from the radio equipment and the input of the channel in which the noise is to be measured ("MOD IN") is terminated in its characteristic impedance. If the normal output or transmit level of the carrier terminal is the same as the receive level, it is then only necessary to patch the transmitter output to the receiver input. If different transmit and receive levels are used in the system, it will be necessary to adjust the transmit level to match the required receive level, either by adding or removing attenuators, or by adjusting the transmit line amplifier gain. This may require that a test tone be applied to the channel input before it is terminated, then adjusting the line amplifier gain until the correct receive level is obtained. A

World Radio History

suitable terminated vacuum tube voltmeter connected across the demodulator output ("DEMOD OUT") will provide a direct indication of the channel idle noise power, provided that allowance is made for the difference in levels of the actual measurement point and the zero transmission point.

If the necessary jacks are available in the carrier equipment, noise measurements may be made at any modulation step in the system — channel, group, or super group. As in the case of voice channel measurements, the modulator input must be terminated in its characteristic impedance, and the transmit level adjusted, if necessary, to match the required received level.

Radio Measurements

The installation of radio equipment provides one of the best opportunities for obtaining detailed performance measurements, and provides a good check on path engineering, system performance, and equipment adjustment. In addition, installation measurements provide a reference standard of performance which is useful in maintaining the equipment and judging the effect of future growth on the system.

In an FM radio system, idle noise measurements are mainly useful for determining the sensitivity of the receiver, and then the level of the input signal to the receiver. Receiver sensitivity is determined by applying a signal to the receiver input, using a signal generator of the correct frequency, and monitoring the noise appearing at the receiver output. Input signal is increased until the output noise is reduced to the value specified by the manufacturer for minimum performance. At this point, the input signal is noted and recorded.

Even more useful than idle noise measurements are the intermodulation noise measurements of the radio. Unlike the case of a carrier, it is quite easy to measure the performance of a radio system under a simulated load. Two basic techniques are widely used, each using random noise loading to simulate the load encountered at periods of peak traffic. Both methods were described in DEMODULATOR, *December,* 1960. Figure 3 shows the noise power recommended by both the C. C. I. R. and E. I. A. for simulating various numbers of channels.

System Performance

The noise performance of a communications system is the sum of the noise contributed by the carrier equipment and the individual radio sections. In order to add these noise contributions directly, they must first be converted into watts (or rather, picowatts — 10^{-12} watt). It may be more convenient to use Figure 4 to make the addition in decibels.

The total noise introduced by a system having many repeaters may be easily calculated if the noise contribution of each section is known, and all contribute equally. Total noise = noise in one section $+10 \log N$, where N equals the number of sections. Accordingly, in a two-section system in which each section contributes 18 dba,

$$\begin{aligned} \text{Total noise} &= 18 + 10 \log 2 \\ &= 18 + 3 \\ &= 21 \text{ dba.} \end{aligned}$$

For a system having 16 sections, the per-channel radio noise would be

$$\begin{aligned} \text{Radio noise} &= 18 + 10 \log 16 \\ &= 18 + 12 \\ &= 30 \text{ dba.} \end{aligned}$$

Total per-channel noise for the system is found by adding the carrier and radio contributions. In our 16-hop system, this would be 23 dba +30 dba, or 30.8 dba.

World Radio History

Radio Channel Capacity

One of the important limitations on system performance is the intermodulation noise contributed by the radio equipment. As the number of radio repeaters is increased, this noise may become overwhelming if the system is operated marginally. A common misconception is that the channel capacity of a radio system is determined primarily by its bandwidth. Actually, the load-handling capacity is equally important. A radio system with bandwidth sufficient for 240 channels, but having the load capacity for only 120 channels cannot provide satisfactory noise performance when carrying 240 channels, even if only one or two sections are used. If additional sections are added, intermodulation noise may become overwhelming at times of peak traffic, and render the system virtually useless.

Radio Pre-emphasis

Idle noise is suppressed in an FM radio system in direct proportion to the modulation index of the signal. This is the ratio of the frequency deviation of the RF carrier to the frequency of the modulating signal. Since deviation is a function of the amplitude of the modulating signal, rather than its frequency, a high-frequency modulating signal and a low-frequency modulating signal of equal amplitudes will produce the same frequency deviation. The lower modu-

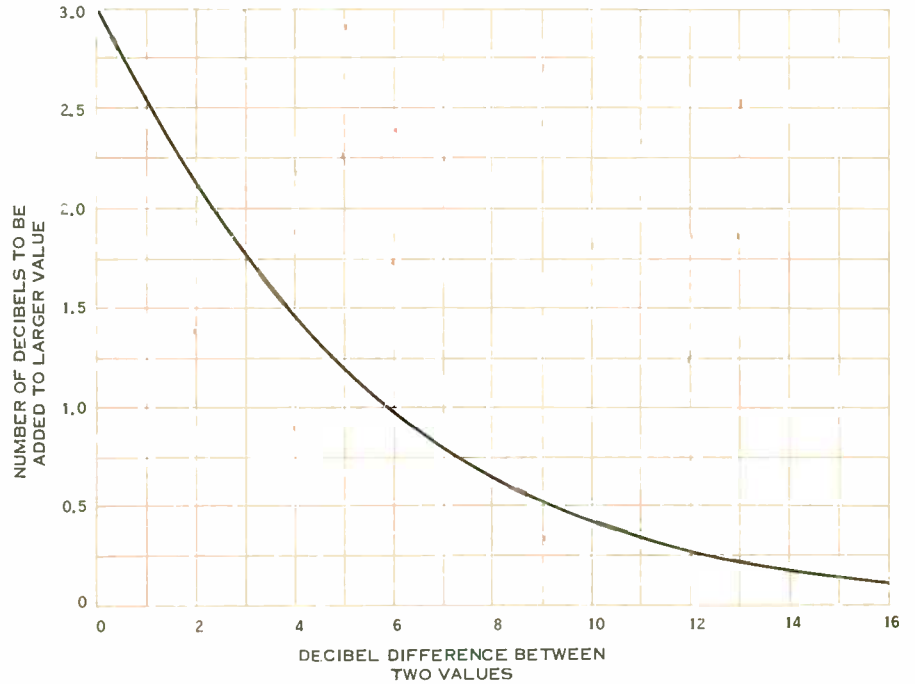


Figure 4. Graph for adding noise or signals expressed in decibels or dba. If signals differ by more than 16 db, smaller signal makes no significant contribution to total.

World Radio History

lating frequency, however, will have a higher modulation index than the high frequency signal and will suppress background noise more.

This is why high-frequency channels are noisier than low-frequency channels transmitted over FM radio, unless the modulation index of the higher channels is increased by increasing their amplitude. When this is done, it is called *pre-emphasis*, and serves to equalize the noise difference between high-frequency and low-frequency channels. In order that channels are restored to their correct level, a *de-emphasis* network must be employed at the receiver to compensate for the higher level of the upper channels, Figure 5 shows the effects of pre-emphasis and de-emphasis on noise across the baseband.

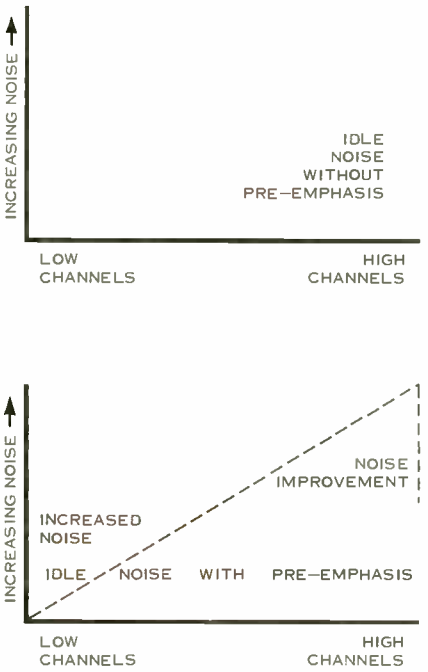


Figure 5. Idealized comparison of baseband noise distribution with and without pre-emphasis. Pre-emphasis improves high-channel noise at expense of lowest channels to achieve uniformity.

Since there is a limitation on the amount of load that can be handled by the radio system, the use of pre-emphasis requires that the level of the lower frequency channels be reduced so that the total power into the radio system remains the same. This causes noise in these lower channels to increase somewhat.

In systems intended for future expansion, the initial channels may occupy the lowest carrier frequencies, and pre-emphasis is sometimes omitted. Although this improves the noise performance of these channels during the period that the system is lightly loaded, the addition of more channels at a later time will require that some form of pre-emphasis be used if all channels are to have equal noise performance. If pre-emphasis is added later, the noise present in the original low-frequency channels will necessarily be increased. Figure 6 shows the pre-emphasis characteristic recommended by the C. C. I. R. for multi-channel communications.

For this reason it is important that a careful evaluation of system performance be made *at the time the system is installed* by loading the system to its *ultimate* capacity, and employing whatever degree of pre-emphasis will be used in the completed system. If no pre-emphasis is to be used on the system when it reaches full capacity, measurements of the noise characteristics of *all* channels, will simplify the future selection of traffic most suitable for the noisier channels.

Conclusions

Many types of communications may not seem to require the high noise performance standards that have become established in the telephone in-

World Radio History

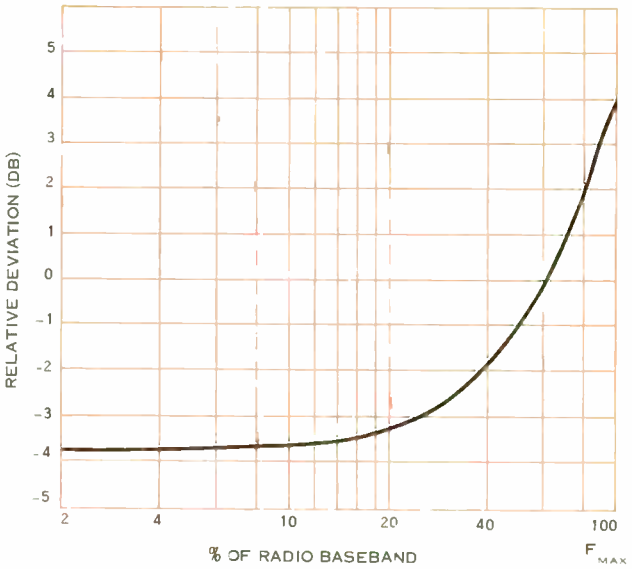


Figure 6. Pre-emphasis characteristic recommended by C. C. I. R. for multichannel voice communications. Different characteristic is required for other types of signal.

dustry and other long distance communications services. In some applications, a minimum "talking circuit" — that is, one that just permits intelligible speech, regardless of the hash of background noise also present — is considered adequate. However, this philosophy is proving more and more to be shortsighted. This type of performance is marginal and allows no room for getting more service out of the communications plant as new needs develop.

Although a poor circuit may permit inteiligible speech, other types of communication, such as data transmission, may be severely handicapped by noisy circuits. Data messages do not possess the same redundancy as speech, and which permits speech to be understood under adverse conditions. As a result, noise increases the data error rate or forces slower transmission rates. Conversely, the higher the circuit quality, the faster the transmission possible for any desired degree of freedom from error. Thus, even lightly loaded communications systems can benefit from the use of equipment and techniques designed for the highest freedom from noise. ●

BIBLIOGRAPHY

1. T. A. Combellick and M. E. Ferguson, "Noise Considerations on Toll Telephone Microwave Radio Systems," *Transactions of the A.I.E.E.*, Vol. 76, Part I.; March, 1957.
2. W. D. Baker and J. W. Joyner, "Channelization of Radio," *Proceedings of the A.I.E.E., Paper No. D.P. 59-519;* April, 1959.
3. C. A. Parry, "C.C.I.T.T. Recommendations for Multichannel Radio Relays and White Noise," *Communications and Electronics*, No. 42; May, 1959.
4. C.C.I.R., "Radio Relay Systems for Telephony Using Frequency Division Multiplex — Maintenance Measurements in Actual Traffic," *Drafting Committee Document, No. 602-E.* IX Plenary Assembly; 1959.

World Radio History

World Radio History

the Lenkurt Demodulator

VOL. 13, NO. 12 DECEMBER, 1964

Noise is the universal enemy of electrical communication. Therefore, the study of noise and its effects is immensely important to the operators and manufacturers of communications equipment. The widespread interest in the subject is indicated by the unusual popularity of this article, which originally appeared in the DEMODULATOR *in 1960. In response to a number of requests, the article has been brought up to date and is being reprinted here. It presents a basic review of the nature and sources of noise, and reflects the current trends in noise measurement.*

Across a crowded room one may have trouble being understood, even when speaking in a loud voice. When the room is empty, the same voice may seem too loud. What has obscured communication? In both cases the same amount of speech power was present to carry the message. The signal was there, but interference prevented it from being identified by the listener.

In communications, interference is called "noise," even though it may be electrical, rather than auditory in nature. A signal represents a certain degree of order or pattern. During transmission, disorganizing forces constantly

World Radio History

damage the signal. If this is allowed to go too far, the signal will eventually become lost in the background noise, thus destroying communication. The signal is always at a disadvantage because it constantly undergoes attenuation, whereas noise is generated afresh at almost every point in the transmission path.

Sources of Noise

The very nature of the universe gives rise to noise. Noise is generated in the flow of electricity through a conductor as electrons collide with some of the molecules of the conducting material. As the temperature of the conductor is increased, noise also increases as more of the electrons collide with the more agitated molecules of the conductor. The amount of noise generated is directly proportional to the temperature of the conductor.

In electron tubes and semiconductors noise is generated by the randomness of electrons or other current carriers. Electrons boil off a cathode irregularly. "Holes" or electrons slide through the lattice of a semiconductor randomly, taking different paths and varying amounts of time to travel from one electrode to another, thus adding to the noise in the circuit. This type of noise also increases as temperature rises. Since electricity consists of individual particles or charges rather than being a perfectly smooth homogeneous fluid, noise is bound to arise in connection with current flow.

Another fundamental source of noise is called "black body radiation," and is of interest primarily in radio transmission. All objects in the universe radiate energy over a broad spectrum. The most perfect radiator of energy would also be the most perfect absorber. Thus, a perfect black body—capable of neither transmission nor reflection—would be the ideal radiator. The hotter an object, the more energy it radiates, and the shorter the wavelength at which most of the energy is radiated. For instance, objects at room temperature radiate some energy at microwave frequencies, but most of the radiation from such objects is at very long infra-red wavelengths. Similarly, most of the sun's energy is radiated as visible light and ultra-violet rays, and most of the energy released during the first flash of a nuclear bomb consists of X rays and gamma rays.

Although the radio-frequency energy emitted by objects at moderate temperatures is very slight, some sensitive microwave receivers can detect a man as he crosses a field of snow by the extra microwave radiation that he emits. All radiation of this kind contributes to the background noise that must be overcome in radio communication.

The various types of noise based on thermal agitation or radiation are sometimes called resistance noise, thermal noise, Johnson noise, or white noise. The term "white noise" refers to the fact that white light has a uniform distribution of energy across the visible spectrum. Similarly, thermal or Johnson noise is uniformly distributed across the spectrum. This uniform distribution is caused by the great variety of noise

World Radio History

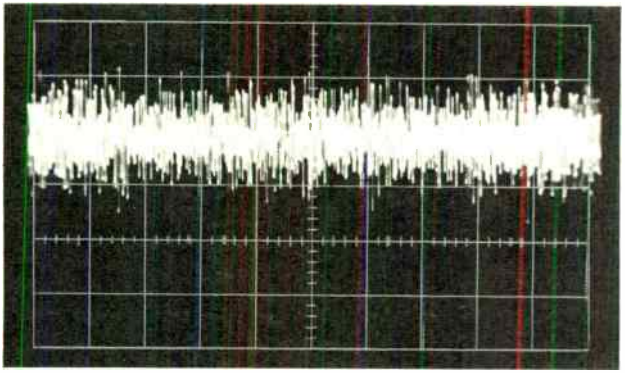


Figure 1. "White" noise as it appears in communication channel. Other names for white noise include "background noise," "random noise," and "fluctuation noise."

sources, and the extremely wide range of energy levels of the electrons and molecules that actually generate the noise.

Effect of Temperature

Johnson noise has been found to be directly proportional to bandwidth and temperature, regardless of the source of the noise. Actual calculations of the noise power to be found in an electrical circuit or detected by a radio receiver follow the relationship

$$\textit{noise power} = kTB \textit{ watts}$$

where

k is Boltzmann's constant (1.38 x 10^{-23} joule per degree),

T is absolute temperature in degrees Kelvin (0°C = 273°K), and

B is the bandwidth in cycles per second.

Since k does not change, the noise power for a *particular bandwidth* depends only on the temperature of the noise source. Thus a thermal noise source at room temperature of 290°K produces noise power of $(1.38 \times 10^{-23})(290) = 4.0 \times 10^{-21}$ watt per cycle of bandwidth. If heated to twice the temperature (580°K), the same source will produce twice the noise power because the noise is directly proportional to the temperature.

Thus, if a highly directional microwave antenna were connected to a suitable receiver having a bandwidth of 1 megacycle and pointed at an object the temperature of the sun (which has a surface temperature of about 6000°K), noise power of 8.28×10^{-14} watt, or —101 dbm would be received. The same microwave receiver, if pointed at a man (whose body temperature is about 310°K), would receive about 4.28×10^{-15} watt, or —114 dbm of noise power. Radio astronomers are now using this technique to uncover many new facts about other planets and other galaxies, by using very sensitive microwave equipment to measure *noise temperature.*

The concept of noise temperature, however, can be extended to cover many other noise sources. Any device

World Radio History

which produces random noise of 4.0 x 10^{-21} watt (—174 dbm) per cycle of bandwidth may be said to have a noise temperature of 290°K—even though that may not be its physical temperature. In other words, the noise temperature of a device is the temperature at which a thermal noise source would have to be operated to produce the same noise power as that produced by the device under consideration.

Assigning a noise temperature to each part of a system considerably simplifies the calculations. The total noise temperature is merely the arithmetical sum of the individual noise temperatures. Suppose the antenna noise temperature in a microwave system is 300°K and the receiver noise temperature is 2000°K. The noise temperature of the combination is then 2300°K.

Noise Figure

The noise contribution of a device can also be expressed in terms of its "noise figure." Noise figure is defined as the input signal-to-noise ratio divided by the output signal-to-noise ratio:

$$F = \frac{(S/N)_i}{(S/N)_o}$$

Both the signal-to-noise ratio and the noise figure can be specified in terms of pure numbers. The more common procedure, however, is to specify these quantities in db. For example, consider a signal-to-noise ratio of 60 db at the input to an amplifier, and an output signal-to-noise ratio of 50 db. Since both ratios are expressed in logarithmic terms, the noise figure in db is simply the difference between the two:

$$\begin{aligned} 10 \log F &= 10 \log (S/N)_i \\ &\quad - 10 \log (S/N)_o \\ &= 60 \text{ db} - 50 \text{ db} \\ &= 10 \text{ db} \end{aligned}$$

This 10 db may be considered a figure of merit for the amplifier—an indication of how much noise is introduced by the amplifier.

In microwave receivers, considerable noise is introduced by the first intermediate-frequency amplifier. Other major contributors of noise are diodes used in the mixer, and noise sidebands from the local oscillator. In addition, impedance mismatch between antenna, waveguide, or mixer can distort the signal, resulting in increased noise. Intermodulation or nonlinear distortion also produces noise which tends to obscure the signal.

Although there is little that can be done about noise originating outside the communication system, much progress has been made in reducing noise that originates within the system itself. New technological developments now permit the use of transmission performance standards which would have been beyond reach a few years ago.

Impulse Noise

A type of interference known as impulse noise is becoming increasingly important with the upsurge in digital communications. Unlike thermal noise, impulse noise is sporadic and may occur in bursts, rather than being uni-

World Radio History

formly distributed. Impulse noise consists of discrete impulses which occur on the circuit as the result of any of several causes. Some types of impulse noise are natural, being caused by lightning, aurora borealis, or other such electrical disturbances. Increasing amounts of impulse noise are man-made. Ignition noises, power lines and their associated switching are strong offenders. In telephone offices, impulse noise may be very great, due to dialing and switching impulses which are induced or otherwise coupled into transmission paths.

Figure 2 compares a speech signal and a typical noise impulse. Both traces are to the same scale; the speech signal level was approximately —20 dbm. Note that the amplitude of the noise impulse is greater than the maximum speech amplitude. These photographs also suggest why impulse noise hardly disturbs speech. A speech sound must be sustained to be understood, since the tone and other qualities of the speech are determined by a succession of waves over a continuing period of time. Noise impulses are too brief to produce a serious disturbance to speech. On the other hand, although noise impulses are usually of very short duration, they may have very great amplitude. Data pulses do not have the redundancy which permits speech to be understood in the presence of large amounts of interference. As a result, a burst of noise which might have little effect on even a single speech syllable could easily make a meaningless jumble of a block of information such as that contained on a punched card.

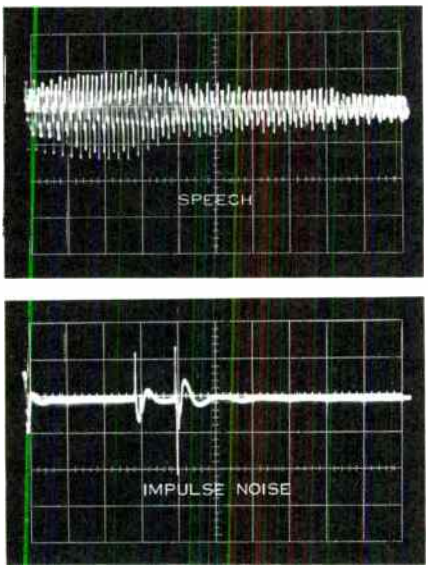


Figure 2. Comparison of speech and impulse noise recorded from actual telephone circuit. Speech sample is the word "two." Note slight ringing following each noise impulse.

The advent of dial-up data-transmission services makes the problem of impulse noise much greater. As data transmission rates increase (in a given bandwidth), the transmission becomes much more vulnerable to impulse noise, mostly because the data pulses are shorter and more nearly like the noise impulses. Even though noise impulses may be very brief, much shorter than the data pulses, they can cause serious interference by causing filters and other tuned elements in a communications channel to "ring." The resulting oscillations may interfere with the signal and cause errors.

World Radio History

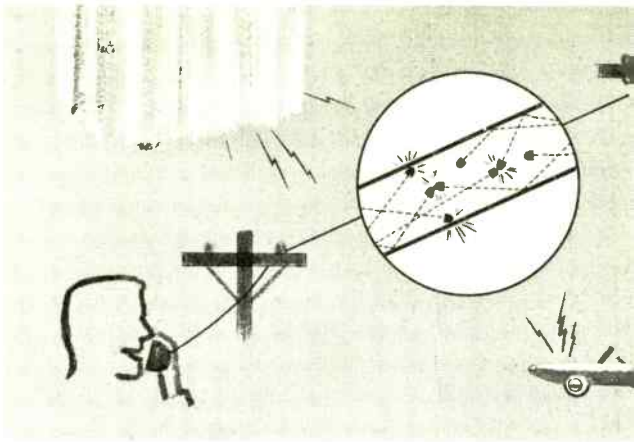


Figure 3. Noise originates outside a communication system as well as inside. Sources may include aurora borealis, molecular agitation in conductors, ignition systems, solar radiation, radio interference, atmospheric disturbances, and others.

Noise Measurements

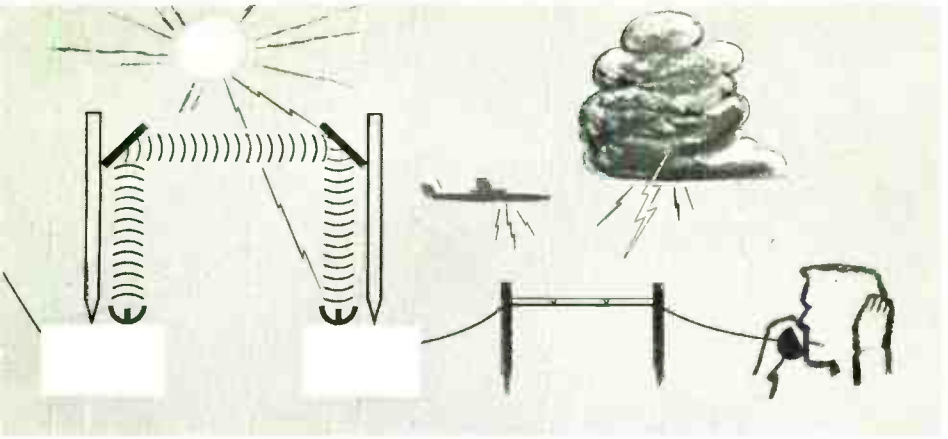
In order to establish performance standards for communications systems, it is necessary to be able to measure the interfering effect of noise. This is different from measuring the amplitude or power of the noise, since noise seems to create more interference at some frequencies than at others. Because the "waveshape" of noise is entirely different than that of speech or music, the ear reacts differently to noise than to speech. In telephone circuits, the type of receiver used has an effect on the amount of interference that a given amount of random noise will produce.

In establishing noise measurement standards, the interfering effect of noise was simulated by comparing the interference provided by a 1000-cycle tone at a reference level with other frequencies. The levels of the other frequencies were adjusted until they were estimated to have the same interfering effect as the 1000-cycle tone. In all tests, a carefully prepared recorded conversation was used to test the interference. Weighting curves such as those shown in Figure 4 were obtained. When a filter or weighting network having these characteristics is used in connection with a flat meter, the meter will give a direct measure of the actual interference produced on voice circuits by the noise.

Line weighting should not be used when measuring interference on data circuits, since the weighting is based on human response when using a certain type of instrument.

At the time of the first tests, the Western Electric type 144 handset was the most widely used handset in the United States. The 144 line weighting curve pertained to this instrument. With the advent of the type 302 set, new response curves were obtained by similar tests, resulting in F1A line weighting. With the development of the type 500 set, still another weighting curve, called C-message weighting, was introduced. However, it is important to

World Radio History



note that none of these weighting curves represent the frequency response of the telephone receiver alone. Instead, they combine the effect of the receiver and the listener's subjective reaction to the interference.

Noise Units

In the United States telephone industry, the reference power level for noise measurements was standardized at 10^{-12} watt or 90 db below 1 milliwatt at 1000 cycles per second. At the time this standard was established, the 144 handset was in general use and noise was measured in decibels above the reference power level (using the weighting network). The unit was called "dbrn" (formerly written dbRN) or decibels above reference noise. With the introduction of the 302 handset, which was about 5 db more sensitive than the 144 handset, the reference level was raised to —85 dbm so that established standards would still be meaningful. Measurements made with the F1A line weighting network were in terms of "dba" or decibels adjusted.

When C-message weighting was established in conjunction with the 500 set, the reference level was returned to —90 dbm. Thus, the noise unit again became "dbrn." At 1000 cps, this unit is the equivalent of the original dbrn, but it is not the same at other frequencies because of the different weighting characteristics. Therefore, noise measured with C-message weighting is specified in "dbrnc."

Rarely, if ever, is 144 line weighting used now. However, noise is often specified in dba, which normally implies F1A weighting unless some other weighting characteristic is specified. The trend in North America is toward C-message weighting with noise specified in dbrnc.

European Noise Units

In Europe and many other parts of the world, units established by the CCITT (International Telegraph and

World Radio History

Telephone Consultative Committee) are used to express circuit noise. The principal units of measurement, which are linear rather than logarithmic, are called "psophometric emf" and "psophometric voltage" (from the Greek *psophos,* meaning noise).

The psophometric emf is the electromotive force (or voltage) generated by a source having an internal resistance of 600 ohms and no internal reactance, which, when connected across a standard receiver having 600 ohms resistance and no reactance, produces the same sinusoidal current as an 800-cycle generator of the same impedance.

Psophometric voltage is defined as the voltage which would appear across a 600-ohm resistance connected between any two points in a telephone circuit. This value is one-half the psophometric emf since the latter is essentially the open-circuit potential necessary from a source to produce the psophometric voltage if the source has a 600-ohm internal resistance. Figure 4 illustrates the relationship between psophometric emf (E) and psophometric voltage (V).

Noise is measured by a psophometer —essentially a vacuum-tube voltmeter which includes a psophometric weight-

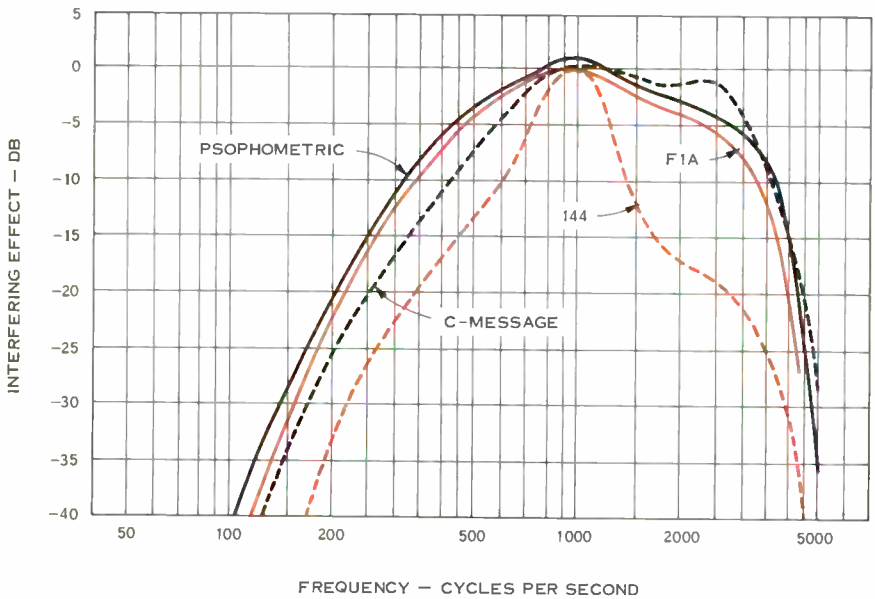


Figure 4. Weighting curves, based on listener response when using a particular type handset, show the relative interfering effect of noise on speech. All curves are referred to 1000 cps except psophometric, which is referred to 800 cps. Although no longer in wide use, 144 line weighting is included for historical interest.

World Radio History

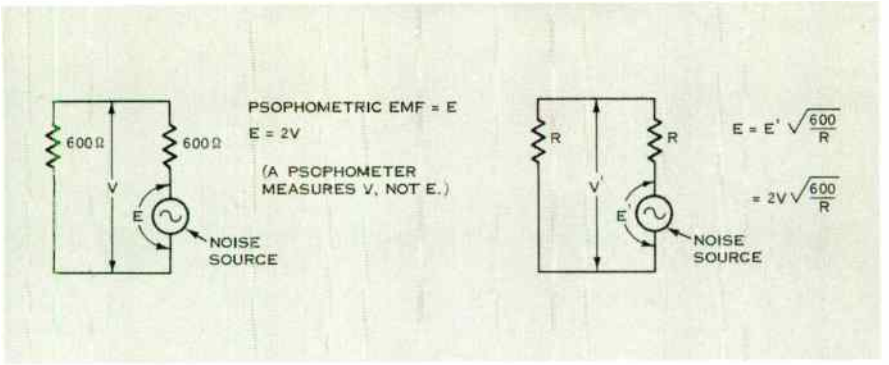


Figure 5. Psophometric noise is measured or calculated in volts (or millivolts) under conditions shown.

ing network. Psophometric weighting is nearly identical to F1A weighting. Minor differences include the fact that psophometric weighting is referenced to 800 cps instead of 1000 cps. The psophometer is calibrated so that an 800-cps tone at 0 dbm in a 600-ohm resistance will produce a meter reading of 0.775 volt, psophometrically weighted.

A common procedure is to specify noise in picowatts, psophometrically weighted (pwp), where

$$\textit{Psophometric power} = \frac{(\textit{psophometric voltage})^2}{600 \text{ ohms}}$$

The description of noise power in terms of picowatts, rather than logarithmic units, has the advantage of simplifying many calculations. Simple arithmetical addition gives the total noise power contributed by several sources. This is not true of logarithmic units, such as dba and dbrnc. However, many people feel that the use of db units gives a more accurate impression of the interfering effect of the noise. For example, a 3-db increase in noise power increases the interfering effect about the same amount whether the change is from 12 to 15 dbrnc or from 30 to 33 dbrnc. By contrast, a 50-pw change could have a considerable effect on interference at a low level, while it might go unnoticed at a high level.

While exact conversion from one noise unit to another is quite laborious, approximate conversions (accurate enough for most purposes) are as follows:

$$dba = 10 \log pwp - 6$$
$$dbrnc = dba + 6$$
$$dbrnc = 10 \log pwp.$$

Impulse Noise Measurements

Since impulse noise may be quite sporadic, it does not generate a "noise power" in the sense that thermal noise does. Although thermal noise may be measured with conventional instruments such as noise measuring sets and

World Radio History

voltmeters, these instruments are unsuitable for measuring impulse noise. One reason is that impulses last too short a time for the instrument to register. Another is that pulse amplitudes at one instant do not necessarily indicate the maximum pulse amplitudes that may be encountered. Because of this uncertainty, most impulse noise figures are estimates of the highest pulse amplitude that will occur. The estimates, in turn, are based on noting the highest amplitude which occurs in a given period of time—the longer the better.

One instrument used for such measurements is known as an impact meter. This device amplifies the impulse and charges a capacitor. The charge on the capacitor is directly proportional to the impulse amplitude. A vacuum tube voltmeter constantly indicates the charge. Thus, the meter reading will indicate the highest amplitude which occurred after the measurement began. Determining the impulse noise characteristics may take on the aspects of a statistical survey.

Trends

As communications users become more sophisticated, they demand "quieter" circuits, which in turn further refine the users' tastes. Thus, noise reduction appears to be a never-ending process. New standards for various types of service are nearly always more stringent than the ones they replace.

A major factor in providing consistently good noise performance is the acceptance of standardized noise measuring and specifying methods. Although F1A weighting is still used, the trend in the United States is toward C-message weighting, with noise specification in dbrnc. In other parts of the world, however, the more common unit is the pwp. Both units are widely recognized, and as international communication ties become stronger it is not uncommon to see noise specified both ways. •

BIBLIOGRAPHY

1. Aldert Van der Ziel, *Noise;* Prentice-Hall, Inc., 1954.
2. *Reference Data for Radio Engineers, Fourth Ed.,* International Telephone and Telegraph Corp.; New York, 1956.
3. Harold I Ewen, "A Thermodynamic Analysis of Maser Systems," *The Microwave Journal;* March, 1959.
4. "Noise," *The Lenkurt Demodulator;* April, 1960.
5. A. J. Aikens and D. A. Lewinski, "Evaluation of Message Circuit Noise," *The Bell System Technical Journal;* July, 1960.
6. Howard H. Smith, "Noise and Transmission Level Terms in American and International Practice," *Fifth National Symposium on Global Communications;* Chicago, May 22-24, 1961.
7. A. V. Balakrishnan (editor), *Space Communications,* McGraw-Hill Book Company, Inc.; New York, 1963.

World Radio History

the Lenkurt Demodulator

VOL. 9 NO. 10 OCTOBER, 1959

BASIC MEASUREMENTS

in communications

This is the first of a series of short articles which describe and discuss basic techniques and equipment used in measuring and testing the performance of carrier and microwave equipment. This introductory article discusses the philosophy behind the use of the decibel in communication measurements, and the growing practice of measuring all signal values in terms of power — a practice started by the telephone industry and now spreading into other branches of communications.

Communications engineers often prefer to measure signal levels in terms of power, and to express these power levels in terms of decibels. Power is preferred to voltage or current because it is an absolute quantity, rather than one that depends on other conditions or quantities to give it meaning. Voltage or current, for instance, can be traded off for one another without changing the actual power involved.

The decibel (db) is the preferred unit for expressing power because it is logarithmic. Two values can be multiplied or divided by adding or subtracting their logarithms. Since amplification and attenuation are multiplication and division processes, the decibel provides a handy means of expressing changes of power by simple addition and subtraction.

For example, if a signal is transmitted at a certain power and is received at *1/1,000* that power, it has suffered a 30 db loss. If this reduced signal is transmitted again and under-

World Radio History

goes similar attenuation, the final signal is *1/1,000,000* its original strength *(1/1,000 x 1/1,000)*. It is much simpler to add 30 db and 30 db to get 60 db as the total attenuation of the signal.

By definition, the decibel is 10 times the logarithm (to the base 10) of the *ratio* of two power levels. The resulting decibel value expresses the power difference between the two levels. If a standard or reference level is used for one of the two values forming the ratio, the resulting value expresses the actual power of the signal.

The reference power most widely used in communications is 1 milliwatt (.001 watt). When this reference is used, the resulting power level is usually abbreviated *dbm,* and means "decibels above or below a reference power of one milliwatt." Thus, 0 dbm is .001 watt, +10 db is .01 watt, and +30 dbm is 1 watt. Remember that *db* refers to a *comparison* of two powers and does not express a fixed value unless it refers to db above or below *some specific reference*. For this reason, an amplifier may have a gain of 30 db, but produce a maximum output of only +10 dbm.

Another reference power occasionally used is 1 watt, and the power levels expressed are abbreviated *dbw*. Power levels expressed in dbw may be converted to the more commonly used dbm by adding 30 db to the dbw value. Thus, -60 dbw = -30 dbm, and -10 dbw = +20 dbm.

The decibel was first used in communications work as an improvement on the "standard cable mile," a telephone unit expressing the loss occuring in one mile of a standard 19-gauge cable. As transmitted bandwidth increased, this unit was found to be unacceptable because loss was different at different frequencies. The decibel which replaced it as a means of expressing loss, is purely relative and always states a specific *percentage difference* between two powers, regardless of frequencies or other characteristics involved.

It is easy to estimate the difference between two signals, in decibels, without a slide rule or table of logarithms by remembering that when power is doubled there is a 3 db increase in level. Thus, raising the signal level four-fold increases the level 6 db. Changing power by a factor of ten changes the level exactly 10 db. One decibel is equal to a power increase of 1.26.

As an example, the attenuation in decibels of an attenuator which reduces a signal to *1/600* its original value may be estimated by dividing 600 by 10 and by 2 as many times as required to reduce the attenuation to the smallest convenient value, and adding the equivalent decibels. Thus, dividing 600 by a factor of 10 and then again by 10 reduces it to 6 and represents 20 db attenuation. Dividing 6 by 2 and then again by 2 represents 6 db attenuation, and leaves only 1.5, or slightly more than 1 db. The total attenuation in decibels is 10+10+3+3+1+, or 27+ db.

Occasionally, some confusion occurs when the decibel is used to express voltage or current ratios as though they were power ratios. Voltages and currents may be compared in terms of decibels if the decibels are computed on the basis of db = 20 log $voltage_1 / voltage_2$. This is equivalent to multiplying a normal "power" decibel by an additional factor of two. This is necessary because power is equal to $voltage^2/resistance$. If voltages are not squared, the power ratio is not correct. Since the expression involves the log of the voltage ratio, multiplying it by two is equivalent to squaring the voltages involved, and restores the correct relationship. Thus, doubling or halving voltage or current produces a 6 db change instead of the

World Radio History

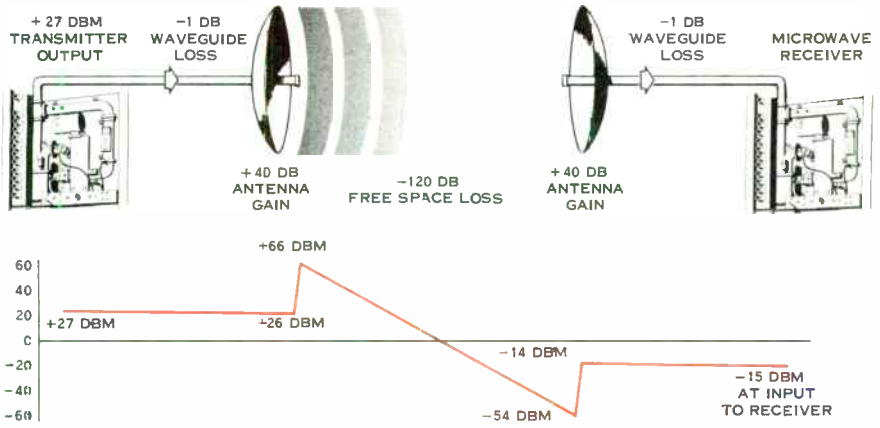


Figure 1. Rating each system element in decibels of gain or loss permits direct comparison of all elements on equal basis, and simplifies engineering calculations.

3 db change for doubling or halving power.

The advantage of measuring the signal levels in decibels becomes more evident when one calculates the performance requirements of a communications system. Carrier and radio equipment includes many elements which change the characteristics of a signal without adding to or taking away from its actual power. For instance, a transformer may raise signal voltage and reduce current without changing actual signal power. Other elements, such as antennas, amplifiers, and attenuators change signal level by a fixed amount. Thus, an antenna with a gain of 30 db always increases the effectiveness of a signal 1,000 times compared to a reference antenna. Such an antenna will provide this gain whether it is used for transmitting or receiving, and regardless of the signal level.

Figure 1 diagrams a simple radio system to show the ease with which system elements may be evaluated. If the transmission path causes too much loss, this may be compensated for by increasing the gain of other system elements such as antennas or transmitter power. Equivalent calculations using field strength, input impedance, peak-to-peak voltage, and other such values are much more complicated and provide no improvement in the results.

Measuring Power

Various methods are employed to measure power, depending on the conditions existing in the circuits where the power is to be measured. At high power levels or at radio frequencies, it is most convenient to measure the heat generated by the power. At the power levels and frequencies usually found in carrier systems, some form of voltmeter is used for measuring power.

As derived from Ohm's law,

$$power\ (watts) = \frac{(voltage)^2}{load\ resistance}.$$

In other words, a given power will cause a specific voltage to appear

World Radio History

across a known load according to the relationship stated in the formula. When the load resistance or impedance is known, power can be easily calculated from the observed voltage. For instance, assume that a 2-volt potential is measured across a 600-ohm load resistance. Then from the formula above,

$$power\ (watts) = \frac{(2)^2}{600} = \frac{4}{(600)}$$
$$= .0066\ watt,\ or\ 6.6\ milliwatts.$$

To convert this into dbm,

$$dbm = 10 \log \frac{6.6\ mw}{1\ mw\ (ref.)} = 10(.82)$$
$$= 8.2\ dbm.$$

In order to eliminate the need for calculating power levels, most voltmeters used in communications work are calibrated in terms of db or dbm. This

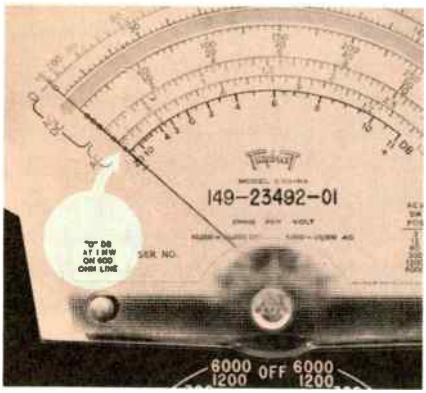


Figure 2. Typical meter face, showing basis for calibration of the meter decibel scale.

calibration is usually based on measuring the voltage appearing across a 600-ohm termination. This termination may be a resistance in the case of direct current circuits, or it may be an impedance in the case of alternating currents. Thus, when the circuit to be measured is terminated in the value specified on the meter (see Figure 2), the meter scale marked db or dbm provides a direct indication of the power level being measured.

If the circuit to be measured is terminated by a resistance value other than that for which the meter is calibrated, the indicated power level will be wrong. In such a case, the meter reading must be corrected by adding or subtracting a correction factor. In some alternating current circuits, an impedance-matching transformer may be employed to correct the termination and the meter indication.

When both the meter termination and the actual circuit termination are known, the correction factor may be calculated in exactly the same way that power levels in decibels are calculated. The correction factor in decibels is

$$db = 10 \log \frac{calibration\ termination}{circuit\ termination}.$$

As an example, when measuring a circuit terminated in 130 ohms, with a meter calibrated for 600 ohms termination,

$$db\ to\ be\ added = 10 \log \frac{600}{130}$$
$$= 10 \log 4.61 = 6.64\ db.$$

This means that the meter indication must be increased 6.64 db in order to correctly state the power in the measured circuit. (This particular value is often rounded off to 6.5 db for convenience.) If the circuit impedance (or terminating resistance) is higher than that for which the meter is calibrated, the meter reading will be too high. In this case it is customary to invert the ratio to avoid the nuisance of taking the logarithm of a value less than unity. When this is done, the sign of the correction factor changes, indicating that it must be subtracted from the reading rather than added to it.

World Radio History

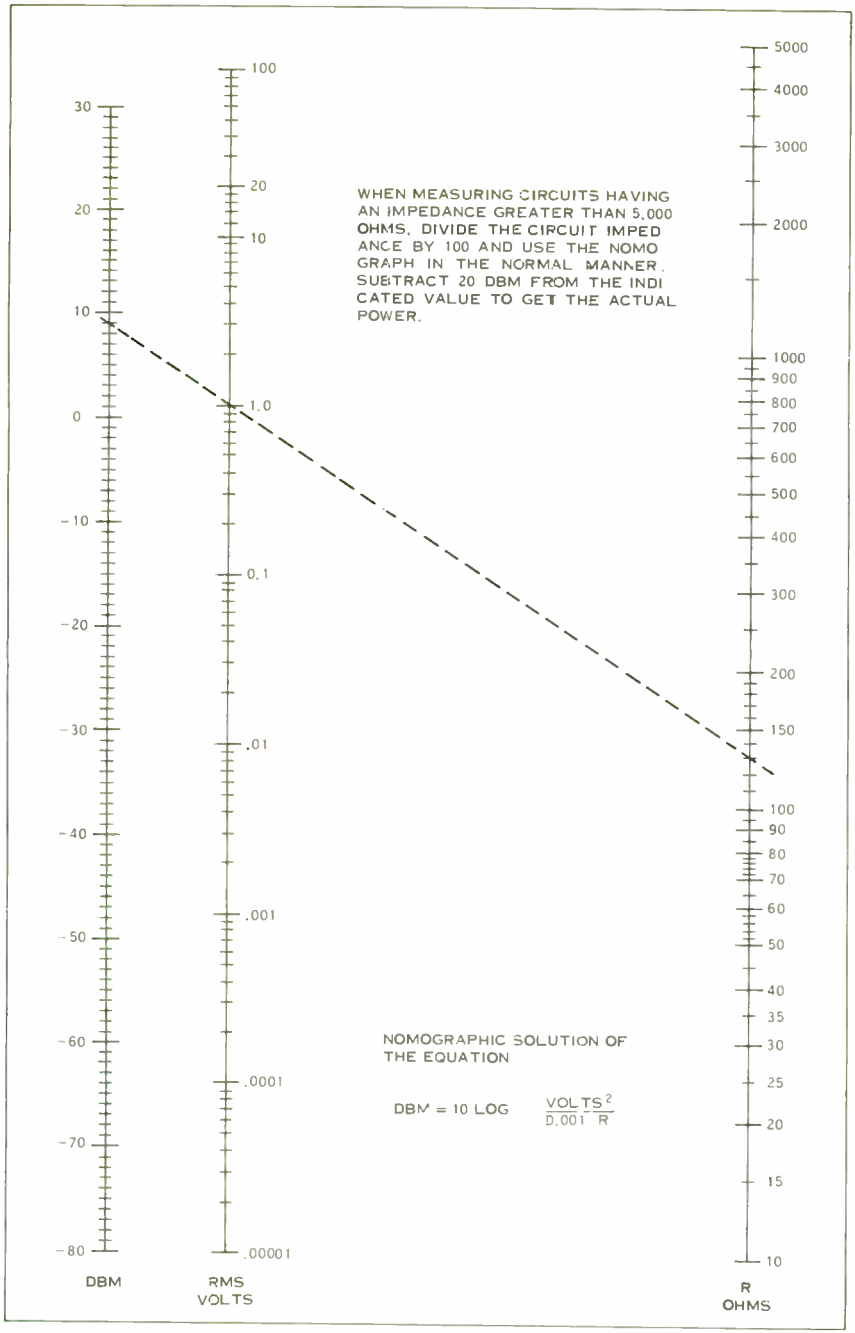


Figure 3. Nomograph for determining power in dbm from voltage measurements. Voltmeter should have high internal impedance.

World Radio History

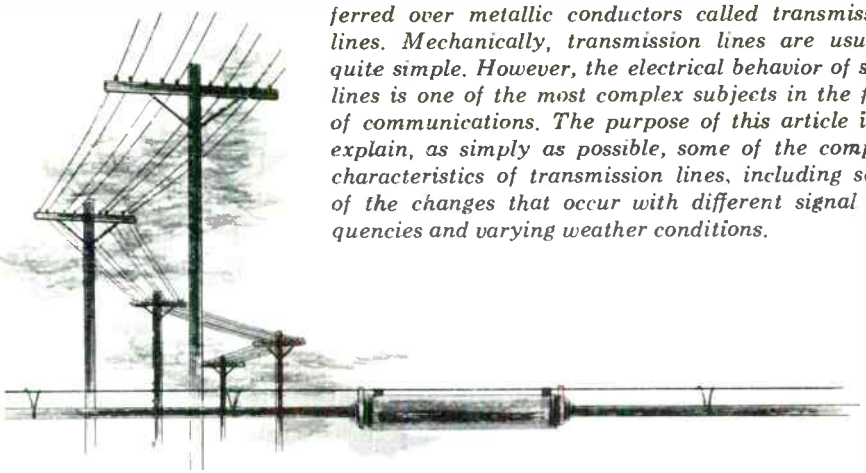
World Radio History

the Lenkurt® Demodulator

VOL. 13, NO. 11 NOVEMBER, 1964

Characteristics of

TRANSMISSION LINES



Signals in many communications systems are transferred over metallic conductors called transmission lines. Mechanically, transmission lines are usually quite simple. However, the electrical behavior of such lines is one of the most complex subjects in the field of communications. The purpose of this article is to explain, as simply as possible, some of the complex characteristics of transmission lines, including some of the changes that occur with different signal frequencies and varying weather conditions.

One of the most important elements in any communications system is the *transmission line* whose function is to transfer signals from one part of the system to another. The simplest transmission line consists of two parallel wire conductors with a power source at one end and a load at the opposite end. Since most transmission lines appear as relatively simple mechanical devices, their rather complex electrical behavior is not always fully appreciated. Because they are electrically complex, they can have a significant effect on the signals transmitted over them. For this reason, the effect of transmis-

World Radio History

sion lines must be determined and taken into account when evaluating the transmission performance of a communications system.

There are three general types of transmission lines used in communications systems: open wire, multipair cable, and coaxial cable. Open-wire lines consist of pairs of wire conductors suspended on poles. A multipair cable is an assembly of pairs of insulated wire conductors, wrapped in a protective sheath. A coaxial cable consists of two tubular-shaped conductors, one inside the other. The inner conductor is usually a small copper tube and is insulated from the outer conductor which is either a copper tube or copper braid. Multipair and coaxial cables are either suspended on poles, or buried underground or underwater. (The *waveguide,* a type of transmission line used to transmit signals at microwave frequencies, is not discussed in this article.)

Electrical Properties

Communications systems contain a number of different types of elements whose electrical properties (resistance, inductance, and capacitance) are considered to be *lumped.* The action of a so-called lumped element occurs at a concentrated point in the circuit—that is, at its location. The electrical properties of a transmission line, however, exist uniformly along its entire length and are considered to be *distributed* rather than lumped.

Each type of transmission line possesses four such distributed electrical properties which must be considered and properly adapted for operation with the different types of communications systems. These four properties are: (1) series resistance R, (2) series inductance L, (3) shunt capacitance C, and (4) shunt conductance G.

The particular values of these fundamental electrical properties depend primarily on the physical configuration of the transmission line and the material used in its construction. To a lesser degree, the values of these properties also depend on frequency, temperature, and weather conditions.

In analyzing the properties of a transmission line it is convenient to describe a line of infinite length consisting of sections of unit length (feet, yards, miles), each possessing the four fundamental electrical characteristics. The equivalent electrical circuit of such a line is shown in Figure 1. In this diagram, the distributed resistance, conductance, inductance, and capacitance are symbolized as *lumped* constants to show their effect more clearly.

Electromagnetic Waves

When power is first applied to a transmission line, energy from the power source does not appear all along the line simultaneously. Instead, it travels away from the source in the form of an electromagnetic wave, called the *incident wave,* and reaches the various sections of the line at different times. The time it takes to travel through each section depends upon the values of the four fundamental properties of the line. Since the series inductance and shunt capacitance of the line store energy for a period and then return it to the circuit, no loss of energy occurs as a result of these properties. (The losses in a line are caused mainly by its resistive properties which dissipate energy in the form of heat.)

Current that leaves the power source will start to charge the shunt capacitance of the first section. The charge is not instantaneous because the charging current is impeded by the series resistance and the series inductance of the section. When the shunt capacitance

World Radio History

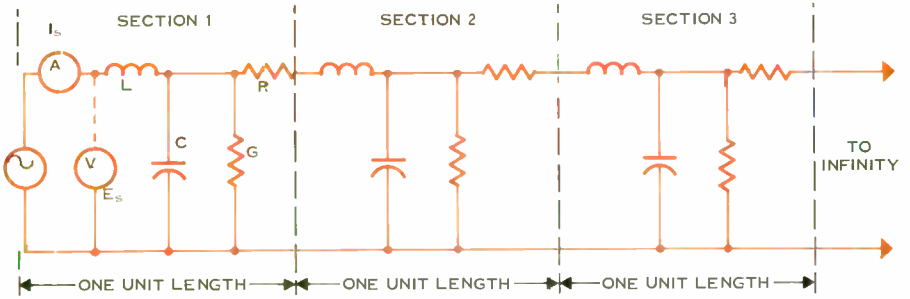


Figure 1. Equivalent circuit of transmission line of infinite length.

is fully charged, the current will begin to decrease. However, now the shunt capacitance in the second section begins to charge through the series inductance and resistance. The shunt capacitance in each succeeding section will add to the charging current as the current in the capacitance of the preceding section starts to decrease. The current, voltage, and the associated electromagnetic wave will progress down an infinite line in this manner until all the energy is diminished due to the resistance in the line. From an extension of Ohm's law, the amount of current that will flow in such a theoretical line is expressed mathematically as

$$I_s = \frac{E_s}{\sqrt{\frac{z}{y}}} \qquad (1)$$

where

I_s = sending end current
E_s = sending end voltage
z = series impedance per unit length
y = shunt admittance per unit length

Characteristic Impedance

It is important to note that in equation (1) the impedance, expressed as

$$\sqrt{\frac{z}{y}},$$

depends only on the four characteristic properties of the line, does *not* include any termination impedance, and is independent of length. An *infinite line* is used so that the effects of termination impedances can be ignored. Since the impedance includes reactance as well as resistance, it is also a function of frequency. This property of a transmission line is called its *characteristic impedance* Z_0, and is expressed mathematically* as

$$Z_0 = \sqrt{\frac{z}{y}} = \frac{E_s}{I_s} = \sqrt{\frac{R + j\omega L}{G + j\omega C}} \qquad (2)$$

where

R = series resistance
G = shunt conductance
$j\omega L$ = series reactance
$j\omega C$ = shunt susceptance

The resistance R and conductance G in transmission lines used in carrier systems are usually so small compared to the series reactance $j\omega L$ and the shunt susceptance $j\omega C$ that they are sometimes omitted when calculating the characteristic impedance. In such a case, equation (2) can be simplified to

$$Z_0 = \sqrt{\frac{L}{C}}$$

*Z_0 is a complex impedance and a detailed explanation of this expression is beyond the scope of this article.

World Radio History

If the values of the four fundamental characteristics of a particular transmission line are not known, the characteristic impedance can be obtained by measurement. This is done by first measuring the impedance of the line with the receiving end open circuited (Z_{oc}) and then measuring the impedance of the line with the receiving end short circuited (Z_{sc}). The characteristic impedance can then be obtained from the following equation:

$$Z_o = \sqrt{Z_{oc} Z_{sc}}$$

There is, of course, no such thing as a line of infinite length. All transmission lines contain a power source at the sending end and are terminated in a load of some impedance value at the receiving end. As mentioned earlier, energy placed on the transmission line at the sending end travels in the form of an electromagnetic wave (incident wave) toward the opposite end of the line. The value of the load impedance is very important. To transfer all of the energy that reaches the receiving end of the transmission line to the load, the impedance of the load *must* equal the characteristic impedance of the line. In such a case, the input impedance of the line also equals the characteristic impedance, and, as far as the power source is concerned, the line *appears* to be infinitely long. Therefore, for a given frequency, the input impedance of a transmission line is constant, *regardless of its length,* if the line is terminated in its characteristic impedance. If the load and characteristic impedances are not equal, an impedance mismatch exists and all of the energy will *not* be transferred to the load. Instead, some of the energy in the form of a *reflected* wave will travel back toward the power source and will interfere with the incident wave.

The amount of energy reflected because of a mismatch can be expressed by the *reflection coefficient* which is derived from the following formula:

$$\text{Reflection Coefficient} = \frac{Z_L - Z_o}{Z_L + Z_o}$$

where

Z_o = characteristic impedance
Z_L = load impedance

In communications systems, a more common method of expressing the degree of mismatch between the characteristic impedances of a transmission line and the load impedance is the *return loss,* expressed in decibels, which is defined as

$$\text{Return Loss} = 20 \log_{10} \frac{1}{\text{Reflection Coefficient}}$$

Since it is impossible to have a perfect match between the characteristic impedance of a transmission line and the impedance of the load, there is always some reflection. In communications systems, undesirable effects such as echos and singing may result if the return loss is too low.

Standing Waves

The incident wave and the reflected wave on a transmission line travel in opposite directions. At certain points along the line the voltages in the two waves will be in phase and will add, while at other points they will be out of phase and will subtract. The points along the line where the two voltages are in phase are points of maximum voltage and minimum current and are spaced one-half wavelength apart. The points along the line where the two voltages are 180° out of phase are points of minimum voltage and maximum current and are also spaced one-half wavelength apart. The distance be-

World Radio History

tween alternate points is one-quarter wavelength.

If the receiving end of the line is either a short circuit or an open circuit, all of the energy in the incident wave will be reflected back towards the power source (total reflection). In such a case, the voltage and current at minimum points are zero. When the receiving end is an open circuit, it is a point of maximum voltage and zero current. On the other hand, if the receiving end is a short circuit, it is a point of maximum current and zero voltage. Figure 2 is a plot of the rms voltage and current maximum and minimum points along transmission lines with various terminations.

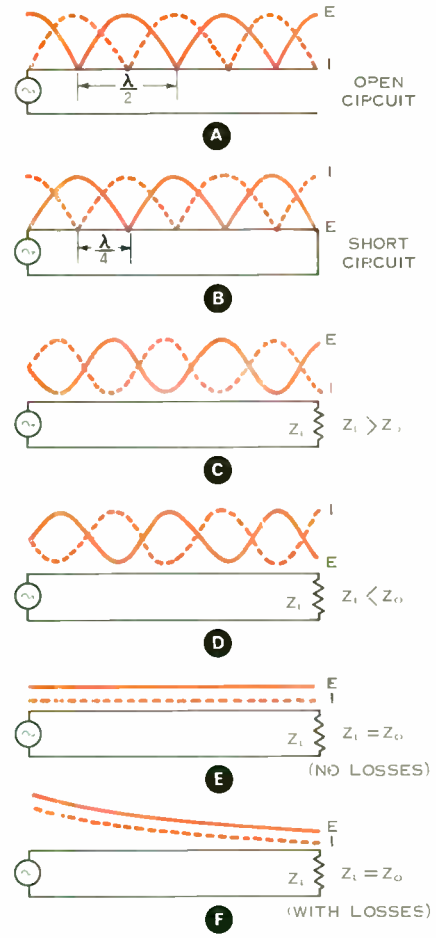


Figure 2. Diagrams A through D are plots of the rms voltage and current of lossless *transmission lines with various terminations. Diagrams E and F are plots of rms voltage and current along transmission lines terminated in their characteristic impedances.*

When the load impedance is greater than the characteristic impedance, the receiving end of the line will be somewhat like an open circuit, except that the rms current at minimum points will not reach zero. Conversely, when the load impedance is less than the characteristic impedance, the receiving end will be somewhat like a short circuit, except that the rms voltage at minimum points will not reach zero.

It can be seen in Figure 2 that when the line is not terminated in its characteristic impedance, the plots of the rms voltage and current appear as waves and, for this reason, are referred to as *standing waves*. Because they are motionless, however, they are not true waves in the same sense as the incident and reflected waves.

The ratio of the rms voltage or current at a maximum point to the rms voltage or current at a minimum point is referred to as the *standing wave ratio*. A standing wave ratio of 1:1 implies that there are no standing waves, and that the line is terminated in its characteristic impedance.

When a transmission line is terminated in its characteristic impedance, the current and voltage all along the line are in phase. If the line is considered to be lossless, a plot of the rms voltage and current is a straight line as shown in Figure 2E. However, there are always some losses which cause the energy to diminish as it travels down

World Radio History

a line. A plot of the rms voltage and current along a properly terminated line with losses is shown in Figure 2F.

Attenuation and Frequency Effects

The amplitude of the current that flows through a transmission line is continuously diminishing with distance from the power source because of the shunt conductance of the line. The voltage is also diminishing because of the series resistance of the line. Because the voltage and current are diminishing, the energy in the associated electromagnetic wave is also diminishing.

The series resistance of a transmission line increases with frequency as a result of the expanding and contracting magnetic field within the line which forces current to flow toward the outer surface of the wire. This action, known as the *skin effect,* reduces the total effective cross-sectional area of the wire, thereby increasing the series resistance.

In multipair cable where the individual wires of each pair are very close together, series resistance is further increased by *proximity effect.* The interlinking magnetic fields of the two wires force the current to flow in that portion of each wire that is closest to the other wire, causing a further reduction of the effective cross-sectional area of the conductors. Like skin effect, proximity effect in cable circuits increases with frequency. The external magnetic fields and current distribution in a cable pair are shown in Figure 3.

A slight reduction of the series inductance with increasing frequency is caused by the change in current distribution which reduces the net strength of the magnetic field within the wire.

Shunt capacitance changes so slightly with increasing frequency that the change is negligible at all frequencies below the microwave region.

Shunt conductance increases considerably with increasing frequency. The total shunt conductance consists of two parts. The first and most familiar part is the conductance of the line insulation. The second part is an apparent conductance caused by internal heating of the line insulation. When an alternating voltage is impressed across the line, the insulation is stressed first

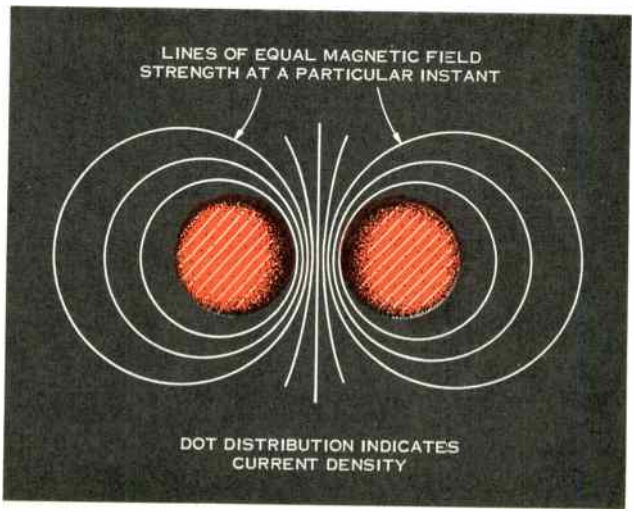


Figure 3. Increased line attenuation at high frequencies in cable pairs is partly caused by an increase in series resistance due to higher current density near the surface.

World Radio History

Figure 4. Current flow in a conductor covered with ice. Dot distribution indicates density of current flow.

in one direction and then the other. The heating that results is a power loss and appears to the line input voltage as an increase of the shunt conductance.

Influence of Weather

Transmission lines are often subject to a wide range of weather conditions. Cable lines are not usually affected by precipitation, although they are affected by temperature changes. Weather variations have significant effects on the properties of open-wire lines. The series resistance of an open-wire line is increased under high temperature conditions and also under wet weather conditions because of increased skin effect. During wet weather conditions, the film of moisture on the wires is slightly conductive. Since the magnetic field within the wires forces the current to flow more toward the surface, part of the current leaves the wire and flows in the film of water. The resistance of the water film is many times greater than the resistance of the copper wire. Therefore, the losses incurred by the current flowing in the water film are appreciable. During frosty or icy conditions, the coating of ice on the wires can become very thick with a considerable part of the current leaving the wire and flowing in the ice coating as shown in Figure 4.

Moisture and ice also affect the shunt conductance of the line. When the weather is dry, the shunt conductance is low and the loss is relatively low. During wet weather, dirt and dust collected on the insulators become much more conductive and the shunt conductance increases allowing more current to flow through the shunt leakage paths and thus increasing the attenuation. Under severe icing conditions the attenuation caused by skin effect and shunt conductance may be increased by as much as six or more times normal dry weather attenuation.

Conclusion

Transmission lines play an integral role in communications systems of all types. A good transmission line delivers as much of the signal energy as possible from the power source to its destination. The line should also be very stable and uniform. This means that each of the fundamental characteristics of the line must be the same throughout its entire length. It is also important that transmission lines be as free as possible from the effects of weather conditions such as temperature extremes, humidity, wind, rain, and snow.

Although transmission lines are being replaced by microwave radios in many medium- and long-haul communications systems, their usefulness has certainly not diminished. Transmission lines are essential in many types of communications systems and are capable of excellent service provided their characteristics are recognized and properly adapted for operation at the frequencies intended. •

World Radio History

World Radio History

the Lenkurt Demodulator

VOL. 5 NO. 9 SEPTEMBER, 1956

TRANSPOSITIONS

For Open-Wire Lines

Crosstalk caused by inductive coupling is one of the problems involved in engineering open-wire carrier facilities. An effective solution in general use is the practice of transposition.

This article describes the basic causes of inductive crosstalk and discusses some of the general considerations pertaining to the theory and application of transposition arrangements.

Crosstalk in open-wire carrier facilities results from the inductive couplings which exist between paralleling wire pairs. This crosstalk can be reduced to tolerable dimensions by a method of line treatment known as *transposing* in which the pin positions of the wires of each pair are interchanged (transposed) at systematic intervals along the length of the facility.

Because of the many possible interactions among various wire pairs, the technique of transposition design tends to be involved, particularly where carrier operation is concerned. However, the extra engineering effort and construction costs of good transposition practice represent a sound investment in performance.

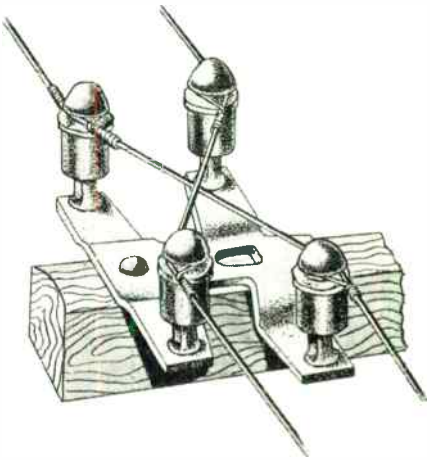


FIG. 1. Point-type transposition bracket. This is one of several methods used for interchanging pin positions of a pair of wires.

World Radio History

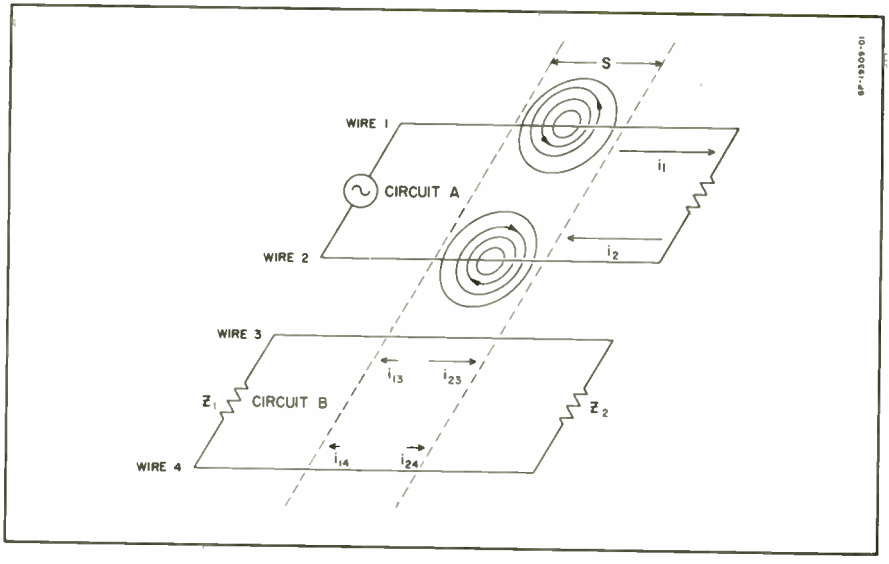


FIG. 2. Crosstalk by induction. Circles about wires 1 and 2 represent lines of magnetic induction and equipotential lines for a particular instant.

Source of Crosstalk

Inductive crosstalk between two wire pairs arises from the magnetic and electric fields that are set up when one or both of the pairs carry an alternating current. Although these fields differ somewhat in their actions in producing crosstalk, the over-all results of both are quite similar.

Figure 2 shows two pairs of telephone wires represented as two adjacent circuits. In a short segment, S, the alternating current in circuit A will set up an expanding and contracting magnetic field. The lines of this field will cut wires 3 and 4 of circuit B and induce voltages in them. These voltages, in turn, cause currents to flow in circuit B as shown in the figure. The induced currents are numbered to indicate that I_{13} is the current induced by wire 1 in wire 3, I_{14} is the current induced by wire 1 in wire 4, and so on.

Because of the differences in spacing between the wires and the directions of the inducing currents, the induced currents will differ in magnitude and phase. As a result, a small net current, the algebraic sum of all the induced currents, will circulate in circuit B. This resultant induced current, alternating at the frequency of the inducing current in circuit A, will set up near-end crosstalk voltages across Z_1 and far-end crosstalk voltages across Z_2.

The lines of magnetic induction of Fig. 2 may also be regarded as the equipotential lines of the electric field. This field will set up potentials on wires 3 and 4 which will not be equal. The resulting difference in potential will cause crosstalk currents to flow toward both ends of circuit B.

The short section S may be thought of as one of an infinite number of such sections contained along the length of

World Radio History

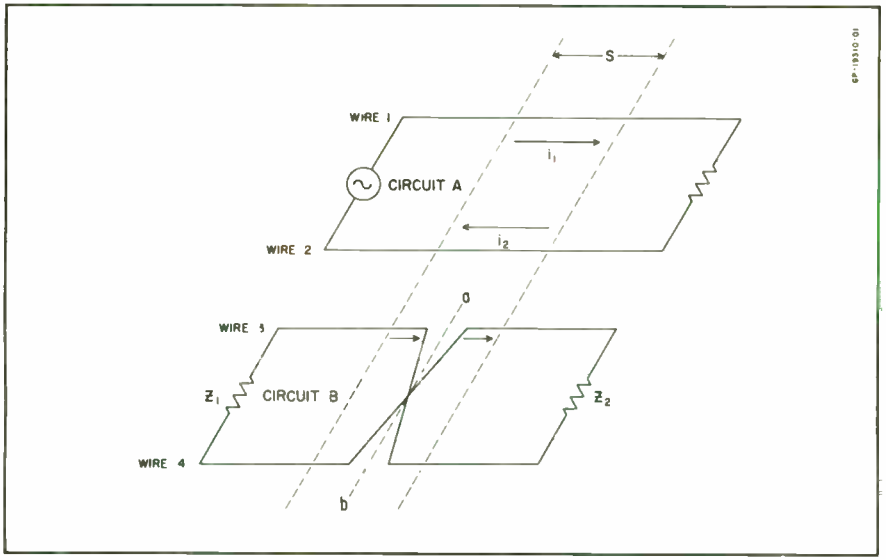


FIG. 3. Effect of transposing circuit B. Net induced current is made up of two components which oppose each other in circuit B.

the line, each contributing a share of the total crosstalk. In a similar manner, one circuit may crosstalk indirectly into another when the currents of the disturbing circuit induce currents in a third circuit (tertiary) which, in turn, induce currents in the disturbed circuit.

Principle of Transpositions

The basic principle of transpositions is shown in Fig. 3. Here the situation of Fig. 2 is modified by transposing the wires of circuit B at the center of segment S. The net induced current is shown as two components, one on each side of the line a-b. The transposition has now made these two components flow in opposite directions in circuit B so that they counteract each other.

The same effect would result from transposing circuit A instead of circuit B. If, however, both circuits were transposed at the same point, no relative transposition would exist between them and the original crosstalk situation of Fig. 2 would prevail.

Because of the propagation effects of line attenuation and phase change, the two components of induced currents in the transposed circuit B will not entirely cancel each other. Attenuation causes the inducing current to decrease in magnitude as it travels from the source toward the load. As a result, the induced currents to the left of line a-b will be greater than the induced currents to the right of a-b.

Phase change causes the instantaneous amplitude and direction of the inducing current to vary throughout each wavelength along the line, resulting in other inequalities. As a result of these combined propagation effects, a small residual induced current remains uncorrected by the transposition and the crosstalk which this current produces is

World Radio History

known as *type unbalance* crosstalk.

Another component of the total crosstalk in an open-wire facility is introduced by structural irregularities. Lack of uniformity in wire sag, pole spacing or wire spacing throughout the length of the circuits will cause random unbalances. The crosstalk resulting from such factors is random in nature and is known as *irregularity* crosstalk.

By transposing the line a number of times within each wavelength, type unbalance crosstalk can be considerably reduced. Careful construction will hold irregularity crosstalk to a minimum. The over-all effect of a properly transposed and well-constructed open-wire facility is the reduction of crosstalk to a point where it is not a controlling transmission limitation.

Transposition Arrangements

The basic building block of a transposition plan is the *transposition section*. A transposition section is a segment of line of arbitrary length in which individual pairs have been systematically transposed, each in accordance with a definite pattern. The patterns chosen are those which will limit the crosstalk between any two pairs to a value which is within the design objectives of the system. Transposition sections are then connected in tandem to make up the total length of the facility.

The actual length of a transposition section is chosen by taking into account the frequency of operation and the crosstalk tendency of the patterns used. Ideally, it is a length which meets the crosstalk objectives of the system for all pairs using the minimum number of transpositions. In practice, transposition sections are usually designed first for the longest lengths practicable. Shorter sections are then designed to provide for situations where these longer sections cannot be used.

Transposition patterns have been standardized and classified by the Bell Telephone System into 32 basic types. A few of the simpler patterns for typical transposition sections are shown in Fig. 4. The most complex of the 32 basic patterns (type a, not shown) has 31 transpositions within one section length.

When more than 31 transpositions per section are required, *extra* types of the 32 basic types may be formed by dividing the transposition section into 32 equal segments called *intervals* and inserting one or more transpositions within each interval.

Extra types are designated by the letter of the basic type and a subscript numeral indicating the number of transpositions within each of the 32 intervals. For example, Type M_1 is a basic type M pattern with its three between-interval transpositions per section plus one transposition within each of the 32 intervals to give a total of 35 transpositions per section.

Any two pairs are transposed in relation to each other only at points where one pair is transposed and the other is not. The relationship existing between any two transposed pairs is expressed as a *relative* type. Thus, for example, the relative of a type O pair and a type N pair is type M and the two pairs are said to have an M *exposure*.

The design of an actual transposition section for an open-wire carrier facility begins with a tabulation of the requirements to be met. These are factors such as over-all length, maximum length of

World Radio History

repeater sections, frequency of operation, wire configuration, wire gauge, and crosstalk objectives of the system.

The next step is the determination of the degree of crosstalk coupling existing between the various pairs on an untransposed basis. This is expressed as a *crosstalk coefficient* in units of crosstalk per mile per kilocycle and is a function of the configuration, gauge and material of the wires. Crosstalk coefficients may be determined from measurements, computations, or, for the more common configurations, from tables.

Equipped with this information, the transposition designer begins to lay out a transposition arrangement by choosing a pattern type for each pair of wires and a transposition section length. The pattern for each pair must be carefully selected so that the final arrangement will contain no two pairs which exceed the design objectives in crosstalk coupling, either directly or through a tertiary.

The procedure of choosing the pattern types is largely a trial and error method. For each tentative plan, the crosstalk (both type unbalance and irregularity) must be computed to ascertain that the relative types between all possible pair combinations are satisfactory for achieving design objectives. Whenever computations reveal that the crosstalk requirements are not met for any combination of pairs, the plan must be revised and the effects computed again. Fortunately, much of the data necessary for these computations is available in the form of tables and graphs.

Experience in transposition designing is a very valuable asset in this phase of the procedure. The seasoned designer often recognizes the pattern types which are not likely to meet his objectives. In general, type unbalance cross-

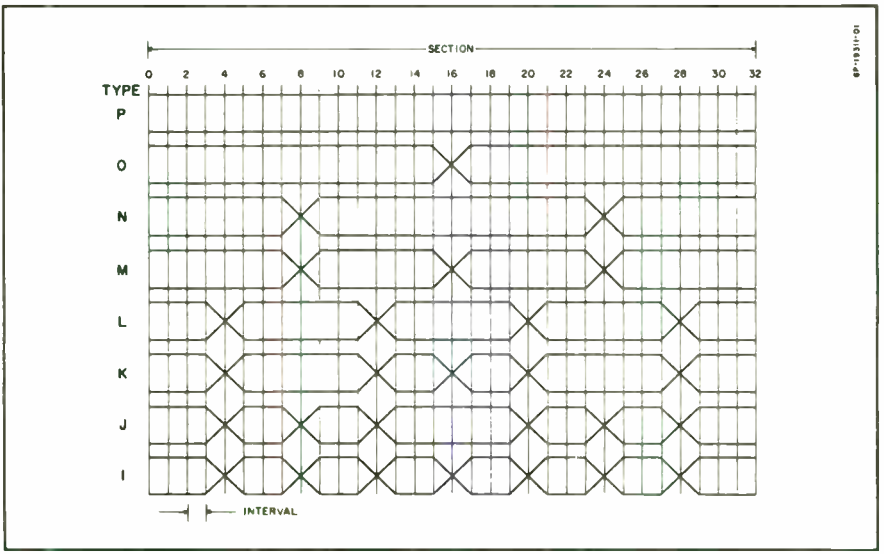


FIG. 4. Several of the 32 basic transposition pattern types.

World Radio History

talk will decrease with the number of transpositions in a section. Therefore, a good starting approach is to choose pattern types which give good crosstalk reduction (low type unbalance) for pairs with tighter coupling (higher crosstalk coefficient) and save the less effective types for pairs with looser couplings. In this manner, it is often possible to eliminate from consideration large blocks of types before the first tentative plan is chosen.

Throughout his analyses, the transposition designer must not lose sight of the relative importance of type unbalance versus irregularity crosstalk in the facility. Care must be taken not to overdesign from the standpoint of number of transpositions used. Transposition will not reduce the crosstalk due to random irregularities. Therefore, it is useless to transpose beyond a point where such crosstalk is controlling. In addition to increasing the cost, too many transpositions may even increase the crosstalk by introducing additional structural irregularities.

General

In applying transposition arrangements, it is important to keep in mind the possible future uses of the facility. The cost of transposing new wire is in the order of 30 per cent less than the cost of transposing existing wire on an out-of-service basis. Therefore, reasonable foresight in anticipating and transposing for future requirements may often result in substantial long-range savings.

The cost of transposing an open-wire carrier facility is frequently a large portion of the cost of the over-all project. However, this cost is outweighed by the advantages gained in the performance and economies of carrier. Therefore, transposing should be regarded not as an evil to be avoided but rather as a bargain in channel-miles of transmission performance.

World Radio History

the Lenkurt® Demodulator

VOL. 4 NO. 2 FEBRUARY 1955

CABLE TRANSMISSION CHARACTERISTICS

In the Carrier Frequency Range

In a previous issue of the Demodulator (October, 1953), some of the basic electrical properties common to all transmission lines were examined and discussed to show how they affect carrier transmission. In this issue, certain characteristics of multi-pair cable are described in greater detail to show how they influence the application of carrier transmission systems to such cable.

All two-wire transmission lines have four fundamental properties in common.

These are:

1. Series resistance.
2. Series inductance.
3. Shunt resistance (or conductance).
4. Shunt capacitance.

The values these properties assume for any particular transmission line depend primarily on the physical configuration (wire size, wire spacing, insulation, etc.) of the line. To some extent they also depend on other factors such as weather, temperature, and frequency. The performance of a transmission line, in turn, depends on how closely the relationships between these properties approach that ideal relationship which provides a low-loss distortionless line. This ideal relationship is realized when, for every frequency of interest, the ratio of the series inductance to the shunt capacitance is equal to the ratio of the series resistance to the conductance.

The theoretically ideal transmission line can be approached by a carefully constructed open-wire pair or coaxial cable. These two types of lines can be built with wire sizes and spacings designed for minimum attenuation and low distortion. The physical restrictions placed on multi-pair cable, however, are such that it is very difficult to manufacture such cable with transmission characteristics at carrier frequencies that even remotely approach those of open-wire or coaxial cable. Moreover, when many of the existing cables in the telephone plant were installed, their use at carrier frequencies was not contemplated. Because of these factors, the application of carrier to multi-pair cable is much different from simi-

World Radio History

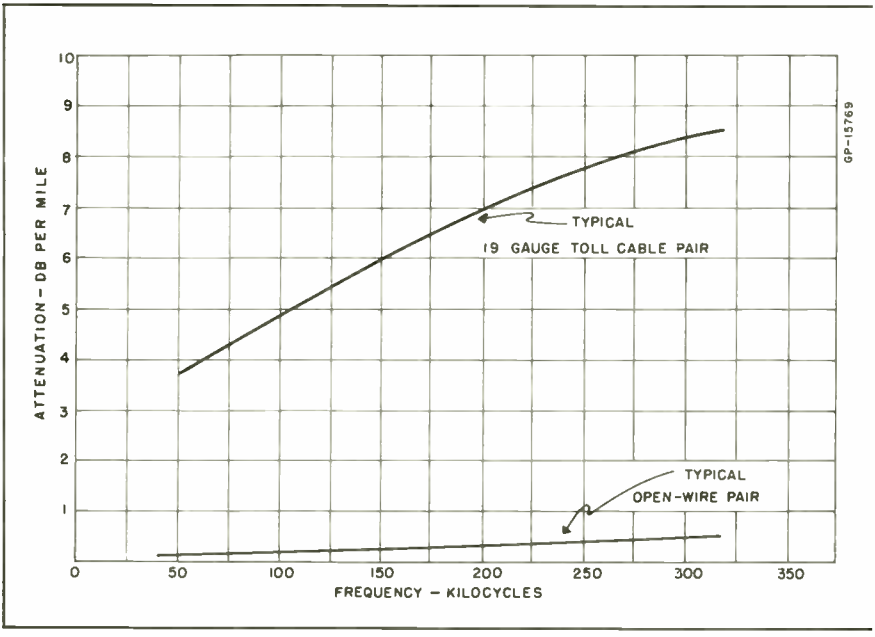


FIGURE 1. Attenuation at carrier frequencies is much greater in cable than in open wire.

lar applications to open-wire line, coaxial cable or radio.

Types of Cable

Although all types of multi-pair cable are subject to the basic restrictions of size and weight, there is a wide range of qualities and types in use. For the purposes of this article, all cable can be divided into two main classifications: exchange cable and toll cable. Exchange cable is often used for short inter-exchange office trunks where large numbers of circuits are required. Usually the lengths of exchange cables are so short that very small wire sizes can be used without too much degradation of transmission. Also, balance and uniform insulation of individual pairs are not critical.

Toll cable, on the other hand, is designed primarily for the long distance transmission of voice and carrier frequencies. Much greater care is taken in the manufacture of toll cable to insure uniform characteristics and minimum losses. Although all cable pairs are twisted to reduce noise and crosstalk, most toll cables are also quadded to further improve their electrical characteristics. A quad consists of two twisted pairs that are further twisted together to form a four-conductor group. The twisting and quadding of cable pairs has much the same effect on cable performance as transposition has on open-wire line. Attenuation, noise, and crosstalk are reduced because external fields are balanced out by the twisting process.

Cable Characteristics

The specific physical characteristics of both exchange and toll cable that affect the design and operation of cable carrier systems include:

1. Non-uniformity in spacing of wires and in distances between wires and sheath.

2. Small wire size.

World Radio History

3. Close spacing between individual wires and between pairs.

4. Relatively large changes in some electrical characteristics with normal changes in temperature.

These physical characteristics tend to cause high attenuation, wide and non-uniform variations of attenuation with frequency and temperature, and impedance variations. They also tend to increase the difficulty of noise and crosstalk control. A comparison between the attenuation characteristics of an open-wire pair and a 19-gauge pair of a quadded toll cable is shown in Figure 1.

The high attenuation of carrier frequencies in multi-pair cable is caused primarily by the small wire sizes and short leakage paths between conductors. In open wire lines and coaxial cable, conductors can usually be made large enough to keep the series resistance low. In toll or exchange cable, however, the necessity of keeping the total size and weight as low as possible requires a compromise between wire sizes that give low attenuation and wire sizes small enough for compact, lightweight cables.

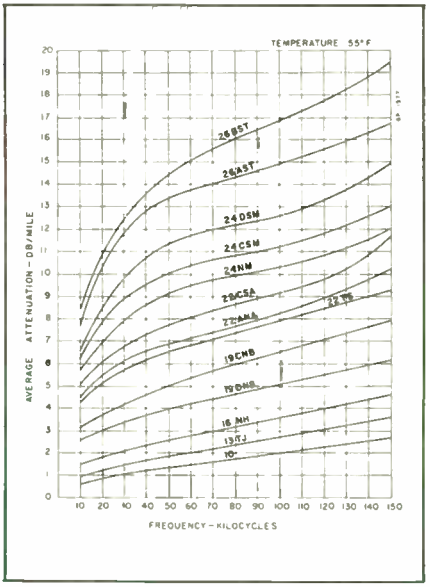
In general, wire sizes range from 10 gauge to 26 gauge depending on the type of cable. The most commonly used wire size in toll cable is 19 gauge; the smallest size usually considered practical for carrier transmission is 24-gauge. Figure 2 shows the effect of different conductor sizes on cable attenuation.

The type of line insulation has as much bearing on attenuation as does the size of the conductors. In open-wire lines and high quality coaxial cable, the dielectric (insulation) is usually air with solid insulators used only at regular intervals along the line for physical support of the conductors. Since air is one of the best insulating materials and has very low loss, these types of transmission lines have very low attenuation per unit of length. In multi-pair cable, however, the insulation is usually of paper or plastic which have relatively high losses compared to air, especially at carrier frequencies.

Several other important cable characteristics, besides high attenuation, that affect the design and operation of carrier systems are shown in the graphs of Figure 2. These include the marked increase in attenuation with frequency and the non-uniformity of attenuation increase with frequency. The general increase of attenuation with frequency is usually referred to as "slope" while the non-uniformity of the increase is commonly called "bulge".

Slope and bulge are characteristic of all transmission lines, open-wire as well as coaxial and multi-pair cable. They are more pronounced, however, in multi-

FIGURE 2. Cable carrier systems must be capable of operating over a wide range of cable characteristics



World Radio History

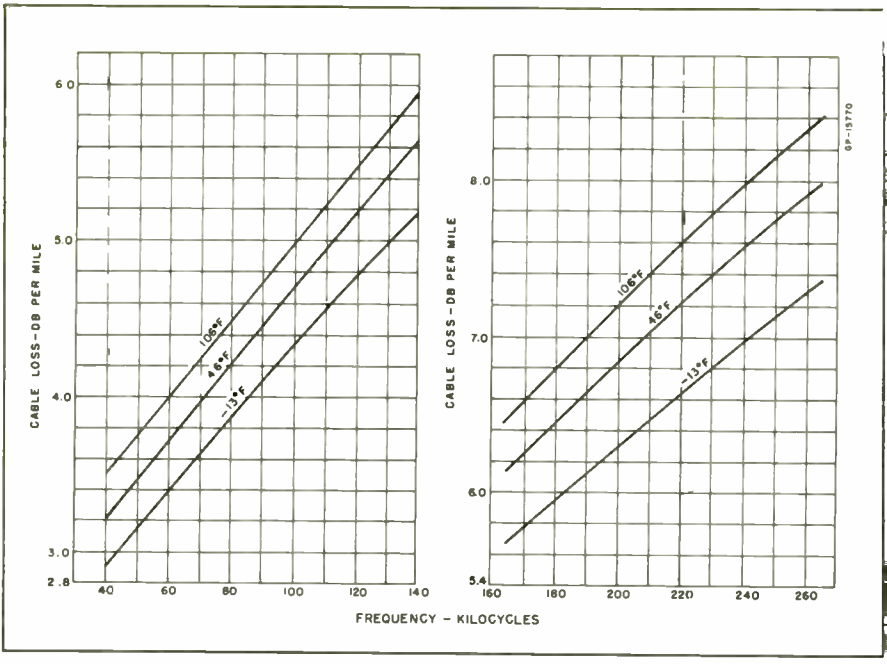


FIGURE 3. Typical variation of attenuation with temperature of a 19 gauge toll cable pair over the frequency ranges used in Lenkurt 45BN cable carrier.

pair cable. Slope compensation alone is possible by several different methods. Open-wire carrier systems usually make use of received regulating tones that adjust the gain-frequency characteristic of the terminal to compensate for slope caused by the transmission line. Another method involves the use of fixed equalizers that introduce a predetermined amount of negative slope across the transmission band. A third method, commonly used in modern cable carrier systems, is the inversion of frequency bands by modulation at each repeater point. In this method (usually referred to as frequency frogging), a channel occupying the lowest frequency space in one repeater section is inverted by the repeater to occupy the highest frequency space in the next repeater section. Thus, if two adjacent repeater sections are identical, the frequency frogging process will largely equalize the slope incurred in each section.

The frequency frogging process, though equalizing much of the slope in a cable characteristic, does not compensate for bulge. Except for very long systems, however, bulge has little effect on overall transmission. In carrier designed for long-haul use, special equalizers and regulating equipment are usually employed to eliminate bulge. In short haul systems, however, individual channel regulation corrects for most of the accumulated bulge.

Slope and bulge are variations of attenuation caused by variations of the fundamental properties of a transmission line with frequency. Another factor, temperature, also has a very pronounced effect on the attenuation of cable pairs. The amount of temperature variation that can be expected for a particular cable, depends on the climate and type of construction. If the cable is underground or under water,

World Radio History

variations will tend to be seasonal and of smaller range than if the cable is strung on poles. Aerial cable, under certain climatic conditions, may have internal temperatures that range from below zero degrees Fahrenheit at night to over one hundred degrees Fahrenheit in the afternoon sun. Figure 3 shows typical variations of attenuation with temperature for the frequency bands normally used for carrier transmission.

A secondary effect of the variation of attenuation with temperature is the change in slope of the attenuation characteristic with change in line temperature. This variation of slope (usually called "twist") is most noticeable at frequencies below 40 kc. Above 40 kc, variation of slope with temperature is slight. In long haul, low frequency cable carrier systems, cumulative twist requires special compensating amplifiers to reduce distortion from this source. However, in modern short haul types of cable carrier operating above 40 kc, twist compensation is unnecessary.

Noise in Cables

Although cable carrier circuits are shielded from direct external induction by a metallic sheath, they are nevertheless subject to certain types of carrier frequency noise. The sources of carrier frequency noises that occur in cables are:

1. Unsoldered or poorly soldered cable splices.
2. Telegraph, dial signaling, and relay transients as well as other office generated noise voltages.
3. Atmospheric disturbance and radio transmitters.
4. Interchannel modulation and stray tones from carrier systems.

Splices made in cable pairs used for voice frequencies are usually unsoldered. While such splices normally form a good joint for voice frequency currents, they often generate carrier frequency noise voltages that may make the pair unsuitable for the application of carrier. A steady transmission of direct-current over the pair is often used to improve contact at such joints and thus reduce the noise.

Carrier frequency plant noise such as caused by telegraph, dial signaling, relays, and switches is almost always present in cable carrier circuits. This noise enters the carrier circuit by first being induced into office wiring of non-carrier pairs and then being transferred again by induction to carrier pairs within the cable sheath. Reduction of this type of noise is possible through the use of special suppression coils in the non-carrier pairs; however, a more routine approach is the use of short repeater sections adjacent to the noisy office so that received levels will be high enough to override the noise.

Noise from atmospheric disturbances and low-frequency radio transmitters enters cable carrier circuits in much the same manner as office noise. Noncarrier pairs in a cable often are connected to open-wire lines or non-shielded subscriber drops which act like radio antennas in receiving noise from external sources. Coupling between carrier and non-carrier pairs within the cable, transfers the noise to the carrier circuit. The same suppression measures that are effective in reducing office noise are also effective in reducing atmospheric noise or interference from external sources.

Carrier systems generate a certain amount of internal noise. Interchannel modulation, signaling tone leaks, and tube noise are all

World Radio History

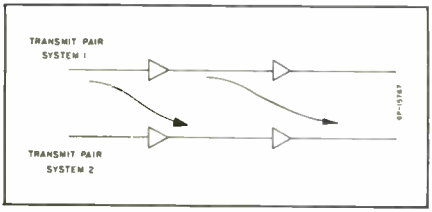


FIGURE 4. Far-end transverse crosstalk is caused by direct coupling between pairs of different systems transmitting in the same direction.

present in cable carrier systems.

The total noise that can be expected in a large number of exchange and toll cable pairs is large enough that special noise reduction techniques are a basic part of cable carrier engineering. Elaborate transposition and suppression schemes were used on early cable carrier systems to reduce noise to acceptable levels. In modern short-haul carrier systems, such expensive methods of reducing noise would make carrier uneconomical. Less expensive, yet perfectly adequate for short and medium haul carrier circuits, is a compandor used with the carrier channels which reduces the interfering effect of noise by as much as 22 db. This amount of reduction is sufficient to make the majority of exchange and toll cables suitable for carrier transmission over distances of 200 or more miles.

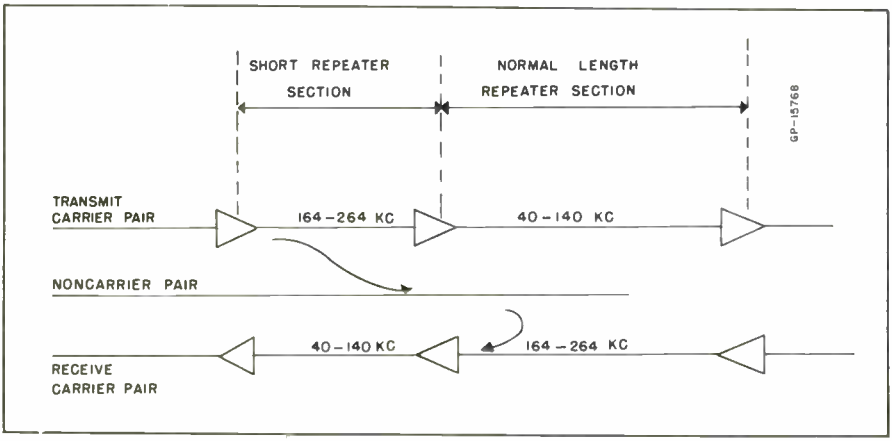
Crosstalk

Crosstalk problems in cable carrier systems are similar to crosstalk problems in open-wire carrier. For both types of systems, the close proximity of paralleling pairs creates the possibility of low-loss crosstalk paths. Slight unbalances in the electric and magnetic couplings between pairs and minor impedance variations are also contributing factors.

Two basic types of crosstalk occur between cable carrier systems operating within the same sheath: transverse crosstalk and near-end interaction crosstalk. Far-end transverse crosstalk is caused by direct coupling between paralleling systems. Near-end interaction crosstalk is caused by coupling between a carrier pair and a non-carrier pair which, in turn, is coupled to a second carrier pair in a different repeater section. This type of crosstalk is only of importance when adjacent repeater sections are of different lengths. Both types of crosstalk are illustrated in Figures 4 and 5.

Several methods are used to control the effect of crosstalk in

FIGURE 5. Near-end interaction crosstalk is caused by a noncarrier pair acting as a coupling link between a transmit pair in one repeater section and a receive pair in an adjacent section.



World Radio History

modern cable carrier systems. These methods include:

1. The use of compandors which reduce the effect of crosstalk by about 22 db.
2. Frequency frogging which eliminates far-end interaction crosstalk common in open-wire carrier repeater installations.
3. The use of separate pairs for opposite directions of transmission.
4. The use of different frequencies for opposite directions of transmission. This eliminates near-end transverse crosstalk as a problem.
5. Generation of masking noise to cover up low-level intelligible crosstalk.
6. Installation of improved cable types in new construction.

When all these measures are taken to reduce the effect of crosstalk, almost 100 percent of the pairs in a good quality toll cable can be used for cable carrier without excessive interference between systems.

Conclusions

Several recent advances in carrier techniques have made possible the economical application of short-haul carrier to multi-pair cable for distances as short as 15 miles. The use of compandors, frequency frogging repeaters, and low-cost channelizing equipment can multiply the message capacity of existing cables up to 12 times at less cost than the installation of additional cable.

World Radio History

World Radio History

the Lenkurt Demodulator

VOL. 9, NO. 11 NOVEMBER, 1960

CROSSTALK

Crosstalk, ancient enemy of electrical communications, may occur whenever more than one communications channel is sent over the same path. The causes of crosstalk are many, and its control is often difficult. This article reviews some of the more important aspects of crosstalk.

Privacy is one of the vital requirements of any communications system, whether the system be privately owned or commercial. Crosstalk tends to violate this privacy by "leaking" the signal from its alloted channel to other channels. Even where privacy is not particularly important, crosstalk has a very disturbing effect if it is intelligible. This disturbance is so great that special care is taken to make crosstalk unintelligible if it cannot be eliminated entirely. Because of this, crosstalk is generally classed as either *intelligible crosstalk* or *unintelligible crosstalk*.

One form of unintelligible crosstalk that may be almost as disturbing as intelligible crosstalk, is *babble*, which consists of "scraps" of sounds from several other communications channels. Although babble may approach noise in its randomness and its lack of intelligibility, it is usually syllabic in pattern, thus increasing its resemblance to speech and its disturbing effect.

There are many ways that signals may slip from one channel to another. One way is by simple leakage through an imperfect insulator. Good design and improved materials and manufacturing techniques have virtually eliminated this as an important source of crosstalk, however.

In radio and carrier, excessive or improper modulation may cause signal energy from one channel to appear in another. In frequency-division multiplex, channels are separated by filters which accept certain frequencies and attenuate others. If signal levels become

World Radio History

excessive, or if the filters don't have enough selectivity, some signals from outside the desired band may appear. Such crosstalk is relatively easy to control by good equipment design and by proper operating procedures. Much more of a problem is crosstalk that occurs between circuits consisting of open wire line and cable.

When an electrical signal passes through a conductor, it sets up electromagnetic and electrostatic fields in the space around the conductor. These fields vary in strength according to the strength of the signal itself. Where the fields encounter other conductors, they cause a current to flow in these conductors, due to inductive and capacitive coupling. Inductive coupling is caused by the electromagnetic field which surrounds the disturbing circuit, while capacitive coupling results from the electrostatic field.

The greater the coupling between circuits, the greater the strength of the crosstalk that will appear in the disturbed circuit. Coupling usually increases in proportion to how close together the two circuits are, how long they are, and the disturbing signal frequency. The crosstalk coupling transfers signal energy from one circuit to another in a fixed ratio which is independent of signal strength. If the signal level in the disturbing circuit is relatively high, the crosstalk will tend to be high. If the level of the disturbing signal is reduced, the crosstalk will be lower in the same proportion.

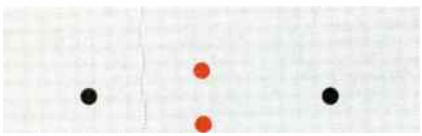


Figure 1. Two pairs arranged so that both conductors of each pair are equidistant from disturbing conductors. Such arrangements are impractical for many pairs.

Circuit Balance

The first telephone and telegraph circuits consisted of single wires between users, with the circuits completed through ground. This reduced the cost of wire, but made the circuits extremely vulnerable to interference, crosstalk, electrical storms, and even earth currents. Modern communications use balanced pairs or coaxial conductors to reduce these external influences. In theory, any disturbance which appears on one conductor of the pair will also appear on the other conductor, and the two will cancel each other out. While this is generally true for large-scale external influences such as electrical storms or ignition noise, it is not true for nearby disturbing influences such as adjacent pairs. The difficulty lies in the fact that it is impossible to achieve a *perfect* balance between the two conductors of a pair. Furthermore, it is impractical to arrange conductors so that both wires of each pair are equidistant from the others. As a result, there is more crosstalk coupling to a near conductor than to the more distant one. The two opposed crosstalk currents are not equal and cannot cancel completely. The excess crosstalk remains as a disturbing signal.

In cables, the problem is greater than in open wire lines. Many conductors are necessarily packed close together, some pairs spaced close together, others more separated. Without special techniques to neutralize crosstalk, it would be impossible to arrange conductors so that both wires of a pair are equidistant from all the nearby disturbing conductors.

World Radio History

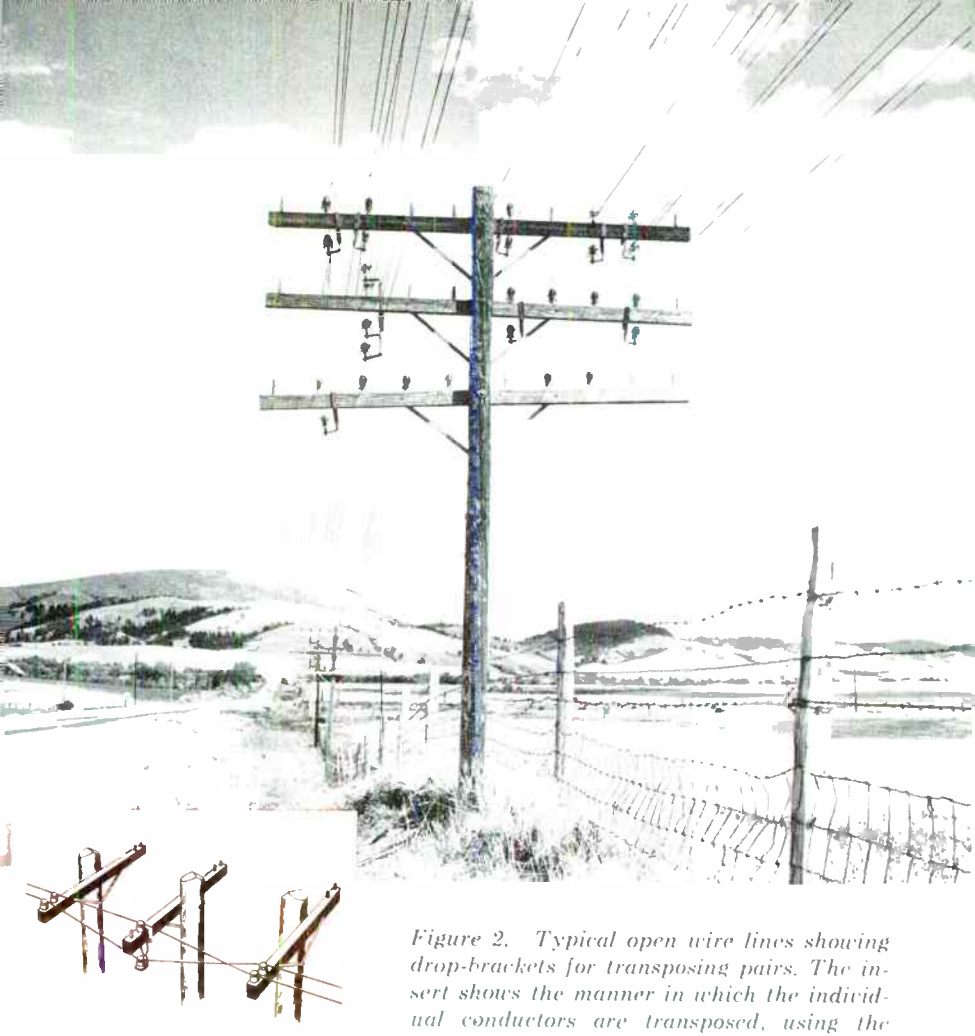


Figure 2. Typical open wire lines showing drop-brackets for transposing pairs. The insert shows the manner in which the individual conductors are transposed, using the brackets.

Even if both wires of a pair are equally spaced from a disturbing circuit, crosstalk will appear in the circuit if the pair is electrically unbalanced. One conductor might have greater resistance than the other, perhaps because of a poorly-made connection. One conductor or the other may have greater mutual inductance or capacitance with the disturbing circuit. Then, the crosstalk may be more strongly coupled to one wire than the other. Instead of being balanced out, the crosstalk in one conductor will predominate, and will appear at one end or the other of the circuit.

Transpositions

Since it is practically impossible to space each wire of a pair equally distant from all other disturbing conductors, the next best thing is to arrange the wires so that they "take turns" in sharing positions nearer and farther from disturbing conductors. This is done by transposing the wires systematically. Transpositions must be designed to can-

World Radio History

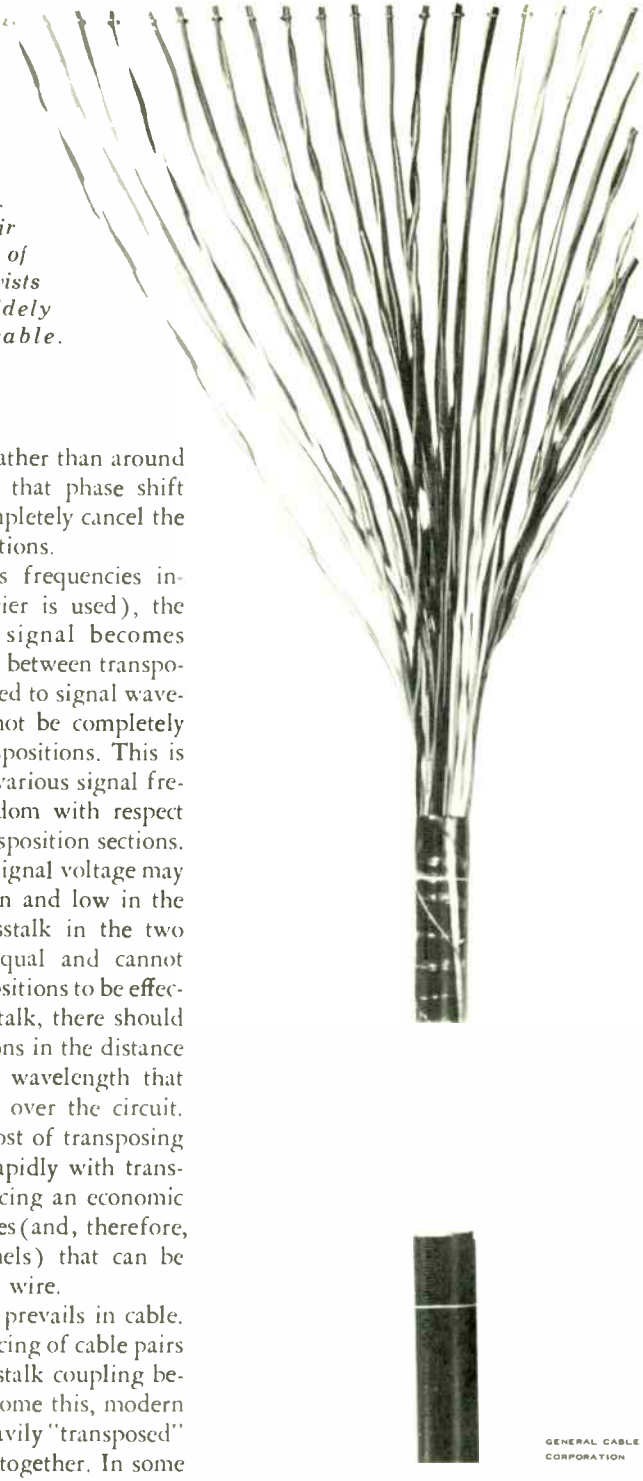


Figure 3. Typical modern plastic-insulated telephone cable. Note that each pair has a different rate of twist. Pairs having twists nearly alike are widely separated in the cable.

cel crosstalk locally, rather than around the whole circuit, so that phase shift won't partially or completely cancel the effect of the transpositions.

As communications frequencies increase (as when carrier is used), the wavelength of the signal becomes shorter. If the spacing between transpositions is long compared to signal wavelength, crosstalk cannot be completely cancelled by the transpositions. This is because the phase of various signal frequencies will be random with respect to the location of transposition sections. At any given instant, signal voltage may be high in one section and low in the next. Obviously, crosstalk in the two sections will be unequal and cannot cancel out. For transpositions to be effective in reducing crosstalk, there should be several transpositions in the distance equal to the shortest wavelength that might be transmitted over the circuit. For this reason, the cost of transposing pairs goes up quite rapidly with transmitted frequency, placing an economic limit on the frequencies (and, therefore, the number of channels) that can be transmitted over open wire.

A similar situation prevails in cable. The close physical spacing of cable pairs tends to increase crosstalk coupling between pairs. To overcome this, modern cable pairs are very heavily "transposed" by twisting each pair together. In some

GENERAL CABLE CORPORATION

World Radio History

cable, two pairs are twisted to form a "quad" and the quads in each layer are spiraled around the center in opposite directions. In all modern cables, the pitch or rate of twist of each pair will be different from other pairs in its group. This is necessary because where two adjacent pairs have the same twist rate, the wires in each pair maintain the same relationship over the entire length of cable. Any unbalance in either pair will permit crosstalk to build up. By varying the twist, the relationship continuously changes so that any coupling between the two pairs at one point will be reversed farther down the cable.

In conventional paper-insulated cables, crosstalk may be further reduced by "random splicing" so that pairs will not be adjacent to each other in successive splicing sections, thus reducing the coupling. Newer plastic-insulated cable, such as shown in Figure 3, employs increased numbers of pair twist lengths and careful location of the pairs and groups of pairs within the cable, so that there is minimum coupling between pairs of similar or near-similar twist lengths. As a result, there may be less crosstalk advantage in random-splicing these newer cables.

Near-end Crosstalk

Since crosstalk may result from both capacitive and inductive coupling, each provides an independent disturbing signal voltage. As shown in Figure 4, capacitively-coupled crosstalk may be represented by a signal source or generator connected *across* the disturbed pair. Inductive coupling, however, can be represented by a signal source connected *in series* with one of the conductors in the disturbed pair. The direction of current flow from the two sources is such that they add or reinforce each other at the "near" end of the disturbed circuit (the same end as that from which the disturbing signal starts), but oppose each other at the far end. The two types of coupling (inductive and capacitive) vary with frequency and spacing between pairs. The closer the spacing, the greater the capacitive coupling. In modern cable, capacitively-coupled crosstalk currents in *adjacent pairs* will be about ten times as great as the inductively-coupled currents at voice frequencies. At 10 kc, the two types of current will be equal, and at 1 mc, inductively-coupled currents will be twice as great as the capacitively-coupled crosstalk currents. If the pairs are not adjacent, or if the distance between pairs is increased, inductively-coupled currents are predominant at all frequencies above the voice band.

Frequency Staggering

Near-end crosstalk occurs primarily in voice-frequency circuits and between pairs transmitting carrier channels of the same frequency. Near-end crosstalk may be greatly reduced if different trans-

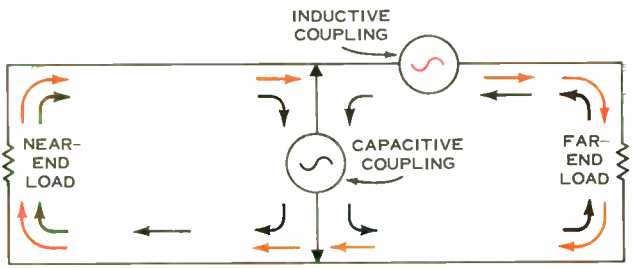


Figure 4. Where inductive and capacitive coupling are equal, they cancel at the far-end, add at the near-end. Inductive coupling predominates at high frequencies.

World Radio History

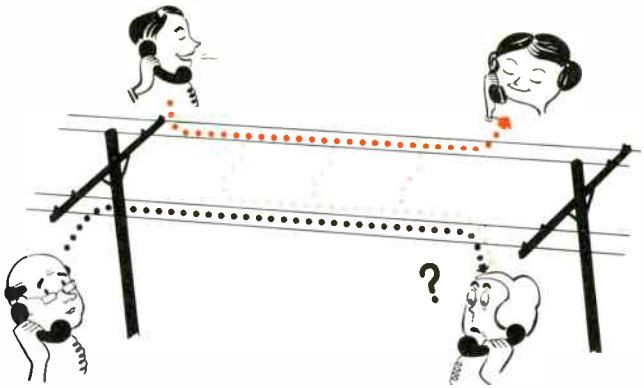


Figure 5. Transverse far-end crosstalk reaches far end without reversing direction in process, is due mostly to inductive coupling.

mission frequencies are used for each direction of transmission. Even though signal energy is coupled from one pair to another, the crosstalk is in the wrong portion of the frequency spectrum to pass the carrier channel filters. This "frequency staggering" not only reduces the effective crosstalk coupling between circuits, but it changes the nature of whatever crosstalk does get through, to a form that is much less annoying than the crosstalk between non-staggered channels.

Another way of coping with near-end crosstalk is to use separate cables for each direction of transmission. All the signals in the cable go in the same direction, so that high-level signals at the output of a west-east repeater are not physically adjacent to the low-level signals just entering the east-west repeater. Even if there is near-end crosstalk, it cannot be heard at the near-end because it terminates at the output of the transmitting amplifier, which is a one-way device. The use of separate cables for each direction, of course, is undesirable as a general practice because of the duplicate facilities required.

Far-end Crosstalk

As stated above, inductive and capacitive coupling tend to cancel each other at the far end of a circuit at low frequencies. In carrier systems, however, transmission frequencies are much higher than in voice-frequency circuits, and inductive coupling becomes much greater than the capacitive coupling. In addition, at these higher frequencies, overall coupling becomes greater, thus providing much higher chance for far-end crosstalk. Since near-end crosstalk is rather easily controlled by frequency staggering, far-end crosstalk is more of a problem in carrier communications. An exception to this may be found in high-speed pulse systems, such as in PCM (pulse code modulation) carrier systems. In such systems no carriers are used; the required bandwidth is a function of the pulse rate, which is determined by the number of channels, the sampling rate for each channel, and the number of code pulses or "digits" for each sample. A practical pulse code carrier system "uses up" the bandwidth provided by ordinary exchange cable and leaves no room for such techniques as frequency staggering. As a result, near-end crosstalk may be quite troublesome.

Several types of far-end crosstalk commonly occur. Where the disturbing signal is coupled inductively and appears at the far end without reversing

World Radio History

direction, it is known as *transverse far-end crosstalk*. One way of reducing this type of crosstalk is to use balancing coils between pairs. These balancing coils are actually dual transformers connected between the interfering pairs. The coils are connected so that they oppose each other in their action. They may be adjusted so that one dominates the other to any degree required. By adjusting them so that the coupling provided by the coils just equals and opposes the crosstalk coupling between the pairs, the crosstalk is cancelled out. In a similar fashion, the residual capacitive coupling between pairs may be cancelled out by using small variable capacitors. This balancing method was devised for the first carrier system for use over cables. With the advent of compandors, it has received little use.

Far-end crosstalk may be increased if transmission levels in adjacent circuits are not equal, or if repeater sections are too long. For instance, if one circuit is operated at a level 5 db below a paralleling circuit, crosstalk will be 5 db greater than if both circuits were operated at the same level. Crosstalk is coupled from the disturbing circuit to the disturbed circuit *in proportion to the level of the disturbing signal*. The lower the level of the disturbed signal, the less the difference between the signal and the crosstalk. When additional amplification brings the signal up to the level required at the terminal, the crosstalk is also amplified this additional amount.

If repeater sections are unusually long, more amplification will be required at each repeater. Crosstalk is usually increased where fewer repeaters but higher repeater gain is employed. This results from the greater difference in signal level at the input and output of each repeater. Since signal level is very low at the input of the repeater, the circuit is more vulnerable to crosstalk. The great relative difference between input levels and output levels supports the chance of crosstalk between the output of repeaters and the inputs of other repeaters. Where "frequency frogging" is used, i.e., high signal frequencies are translated to low frequencies, and low to high, at each repeater, this problem is avoided because the high-level and low-level signals are always in different frequency bands, thus providing a form of frequency staggering.

Still another type of far-end crosstalk is known as *reflected near-end crosstalk*.

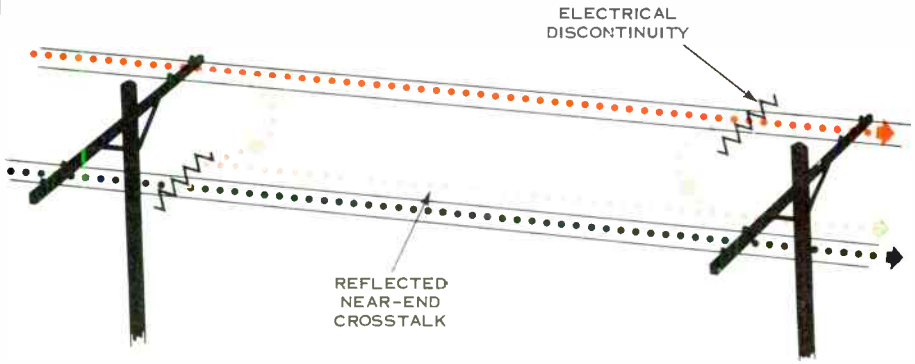


Figure 6 Two types of far-end crosstalk resulting from near-end coupling and reflection from an electrical discontinuity, such as an impedance mismatch.

World Radio History

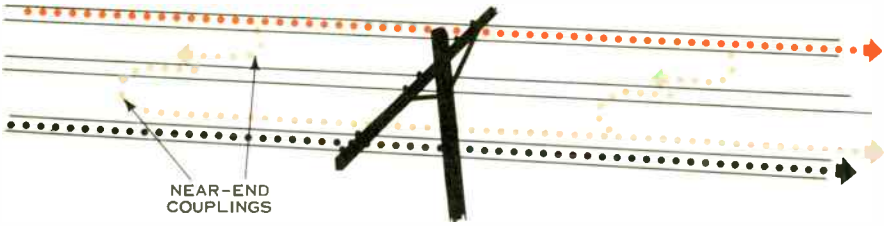


Figure 7. Simple interaction crosstalk, where no repeaters are involved. Far-end crosstalk results from near-end coupling to intermediate ("tertiary") circuit, then to disturbed circuit by similar near-end coupling. Two direction reversals are involved.

If the disturbed circuit has some sort of electrical discontinuity, such as an impedance mismatch, ordinary near-end crosstalk may be reflected toward the far end. This type of crosstalk may also occur if the far-end receiver does not match the impedance of the line. In this case, part of the received signal energy is reflected, then coupled back into the far end of the disturbed circuit by near-end coupling, as illustrated in Figure 6.

Interaction Crosstalk

Several important forms of far-end crosstalk are labeled *interaction crosstalk* because more than one coupling is involved. As shown in Figure 7, the disturbing signal appears in a third or "tertiary" circuit by *near-end* coupling, and is then transferred to the disturbed circuit by another near-end coupling. The crosstalk signal appears at the far end of the disturbed signal, but has reversed direction twice in so doing.

Runaround Crosstalk is a special case of interaction crosstalk, and refers to the signal from the output of a repeater "running around" to the input circuits of the same or other repeaters. As in the case above, the high-level output from the repeater appears in a tertiary circuit by near-end coupling. The tertiary circuit, which may not have a repeater, then couples the signal to the low-level *input* side of circuits with repeaters. This type of crosstalk usually requires more than one type of circuit (such as voice circuits and carrier circuits) in the same cable or open wire path. If all circuits have repeaters at the same location, there is no tertiary path by which the signals can "run around" from the output to the input. Where the required dissimi-

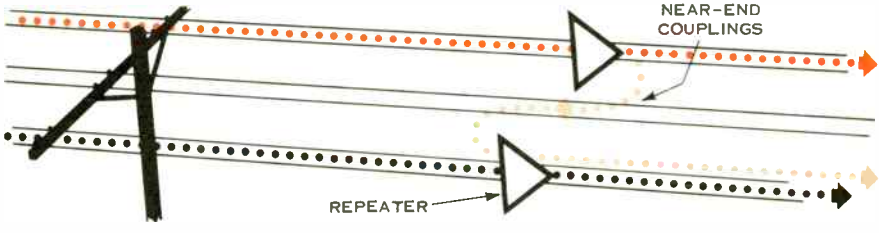


Figure 8. "Runaround" crosstalk is similar to interaction, but may be more troublesome because of couplings from high-level repeaters output to low-level input of same or other repeaters. Less gain per repeater, more repeaters reduce this type of crosstalk.

World Radio History

ar circuit exists, runaround crosstalk may be avoided by inserting a suitable "crosstalk suppression" filter so that the crosstalk frequency is blocked but the desired signal is passed with little loss.

Measuring Crosstalk

In designing a communications system, it is important that all known disturbing factors be taken into account so that they can be corrected or avoided. Crosstalk is one such factor, one that requires considerable effort to control. Like noise, special units of measurement are required for specifying crosstalk effects. Crosstalk, however, is more complicated than noise, since various types of crosstalk are more disturbing than others, and may result from more diverse causes. As a result, several units of measurement have been used to express crosstalk, its net effect, or the electrical coupling from which it results. All the units are related, since they refer to some aspect of how much the signal in one circuit will interfere with that in an adjacent circuit.

The *Crosstalk Unit* is the oldest unit used for expressing crosstalk coupling, and is abbreviated *cu*. It is one million times the ratio of the induced crosstalk voltage or current to the disturbing crosstalk voltage or current, where the impedances of the two circuits are equal. Where the impedance of the disturbing circuit differs from the impedance of the circuit in which the crosstalk appears, cu is one million times the square root of the ratio of crosstalk power to disturbing signal power, or

$$cu = 10^6 \cdot \sqrt{\frac{disturbing\ signal\ power}{crosstalk\ signal\ power}}.$$

Crosstalk units provide a direct measure of the coupling between two circuits. Larger *cu* values mean more crosstalk.

Another way of expressing coupling is as a "loss" between the disturbing and disturbed circuits. The term *coupling loss* refers to the fact that there is a *fixed attenuation* between any two circuits. Where the coupling (and crosstalk) is great, crosstalk coupling loss is low. Thus, if the coupling loss between two pairs is 50 db, a signal having a level of −5 dbm will appear in the other pair at a level of −55 dbm. If the disturbing signal is raised to +3 dbm, the crosstalk level will then be −47 dbm.

Db above Reference Coupling, or *dbx*, takes into account the different interfering effects of different frequencies. This term was invented to permit crosstalk measurements using a standard noise measuring set (such as described in the DEMODULATOR, *April*, 1960). The "reference coupling" is taken as a coupling loss of 90 db between the disturbing and disturbed circuits. Thus, if a 90 dba test-tone were inserted on the disturbing circuit, and the same noise weighting network were used in measuring the level in both circuits, the reference coupling would give an indication of 0 dba. Crosstalk coupling in dbx is equal to 90 minus the coupling loss in db. Figure 9 shows the relationship between coupling loss in db, coupling in dbx, and crosstalk units (cu).

Crosstalk Index

An attempt has been made to evaluate the crosstalk performance of a given transmission facility in terms of the actual disturbing effect of crosstalk. Since the disturbing effect of crosstalk is a function of intelligibility or syllabic pattern, many factors can reduce the actual disturbance that a listener experiences.

Ordinary noise in the system reduces the annoying effect of crosstalk. Even background noise at the listener's location can mask out crosstalk. Several car-

World Radio History

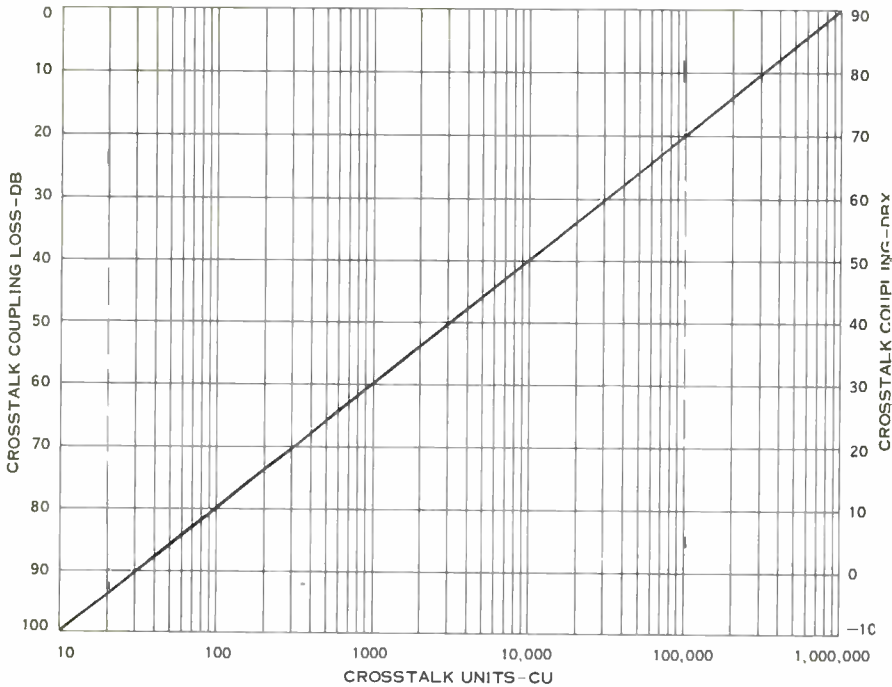


Figure 9. Relationship between crosstalk coupling loss in db, crosstalk units, and crosstalk coupling in dbx. Crosstalk units are little-used now, having given way to coupling loss and dbx coupling for measurements of circuit characteristics, and to crosstalk index for subjective performance rating.

rier systems have been designed which take advantage of this masking effect by including noise generators for use if crosstalk is otherwise excessive. The larger the number of disturbing circuits which contribute to the crosstalk in a circuit, the less annoying the crosstalk tends to be. In this case, the disturbing signals become more and more random as the number of disturbers increases, so that the crosstalk becomes more and more like noise. Offsetting this, however, is the fact that the total power of the crosstalk increases with the number of disturbers. A final comparison of actual annoying effect would depend on the nature of the crosstalk, staggering advantage, and background noise. Because so many variables are involved, the principal value of a crosstalk index is to provide a single numerical value which indicates the overall crosstalk performance. One such index that is widely quoted uses the following scale of merit:

Index	*Quality of Performance*
.01	Excellent
.1	Very Good
1.0	Good
5.0	Fair
10.0	Poor
20.0	Very Poor

World Radio History

These values result from experimental studies which take into account the amount of time that crosstalk exceeds a reference level, and the opinions of observers as to how much annoyance the crosstalk provides under various traffic conditions.

Compandors

One of the most effective ways of coping with crosstalk is the *compandor*, a device which doesn't actually reduce crosstalk, but does reduce its apparent effect. The compandor takes advantage of the fact that crosstalk is not very noticeable during speech, but becomes objectionable during pauses or other silent periods.

The compandor consists of a speech compressor at the transmitting end of the circuit, and an expandor at the receiving end. The amplitude range of the transmitted signal is "compressed" so that soft speech sounds are amplified greatly, while louder sounds are amplified less. Very loud sounds may actually be reduced in level. By reducing the amplitude range of the signal, even the softest sounds are substantially stronger than the crosstalk and noise acquired during transmission, yet the louder sounds are restricted from overloading amplifiers, modulators, and repeaters.

At the receiving end, the amplitude-compressed signal is "expanded" to its original amplitude range. The softer sounds are reduced in level, and the louder sounds may be amplified. Noise and crosstalk are by far the weakest sounds present and receive the greatest attenuation. Compandors usually provide a 20-28 db advantage over noise and crosstalk.

Other devices may be used in a similar fashion to reduce receiver gain during periods when there is no speech. A level-sensitive "gate" raises receiver gain to the level necessary for clear reception whenever speech is present. Between speech sounds, gain is reduced enough that noise and crosstalk are less noticeable. Response time and detector characteristics have an important bearing on the effect produced.

Conclusion

Crosstalk has become more of a problem as the various transmission media have become more congested. The battle against crosstalk is becoming more difficult as more and more new communications services are required by an increasingly complex society. This problem is bound to become even greater in the future because of growing populations and the need for even more elaborate communications services. Improved designs, new techniques, and unrelenting research will be required to prevent crosstalk and other detriments from setting a limit on this growth. •

BIBLIOGRAPHY

1. G. S. Eager, Jr., L. Jachimowicz, I. Kolodny, and D. E. Robinson, "Transmission Properties of Polyethylene Insulated Telephone Cables at Voice and Carrier Frequencies," *Communication and Electronics*, p. 618; Nov., 1959.
2. A. J. Aikens and C. S. Thaeler, "The Control of Noise and Crosstalk on N1 Carrier Systems," *Communication and Electronics*, p. 605; Nov., 1953.
3. A. G Chapman, "Open-Wire Crosstalk," *Bell System Technical Journal*, p. 19; Jan.-April, 1934.
4. C. M. Hebbert, "The Transmission Characteristics of Toll Telephone Cables at Carrier Frequencies," *Bell System Technical Journal*, p. 293; July, 1941.

World Radio History

World Radio History

the Lenkurt Demodulator

VOL. 11, NO. 10 OCTOBER, 1962

SHIELDING and GROUNDING

A trend in modern communications toward greater circuit density and higher operating frequencies has increased the possibility of interference or crosstalk between adjacent wires or cables — even when they are shielded. Many long-established shielding and grounding practices may not be fully adequate with sensitive, modern equipment. Strong sources of interference such as nearby radio transmitters, can severely test shielding effectiveness. This article reviews basic shielding and grounding methods for reducing interference.

In electrical communication, virtually any type of disturbance which impairs communication by obscuring the signal has come to be termed "noise," a very broad term which now includes hum, crosstalk, spurious signals, and of course, thermal noise. Any of these tend to limit the quality and permissible length of a communications circuit

One of the important ways of reducing some of this noise is to intercept the interference energy with shielding, and carry it away by grounding the shield. When properly done, this technique is very effective. However, many standard practices established long ago to cope with the interference problems of that time have not changed significantly with time. As a result, some shielding may be inadequate, others may be greatly "over-engineered" — using extremely costly materials, when a simpler approach might do better.

In a conventional telephone system, speech or carrier signals are carried on a "balanced" transmission line consisting of two conductors. Signals are transmitted as a current which travels down one conductor, returning on the other, thus forming a *transverse* or "metallic" circuit. Both conductors are at the same electrical potential above ground (hence the term "balanced"). A second type, the "unbalanced" transmission line, normally uses a single conductor to carry the signal, with ground providing the return path. "Ground" may take the form of another conductor which is grounded, or common to several circuits. All early telegraph and telephone lines were of the unbalanced type, since only half as much wire was

World Radio History

required. Because the unbalanced line is very vulnerable to interference, however, its use has been generally limited to coaxial circuits.

A balanced line is basically free of external interference. Signal currents travel in opposite directions on the two conductors as they complete the loop. Ideally, interference acts equally on the two conductors, inducing equal voltages which travel in the *same* direction on the two wires and thus oppose and cancel each other.

Although externally induced voltages tend to cancel their flow around the transverse circuit, they do seek a return path through ground, thus forming a so-called *longitudinal* circuit, as shown in Figure 1. If the interfering voltages induced in the two parallel conductors are unequal, perhaps because of circuit imbalance, or because the source of the disturbance is closer to one conductor than the other, a net transverse current may result, thus introducing interference into the circuit. Most interference enters a circuit in this way.

Sources of Interference

When a current flows through a conductor, it sets up two distinct fields around the conductor, the electrostatic or "electric" field, and the magnetic field. Both are capable of inducing longitudinal voltages in adjacent conductors, and both increase in proportion to the power and frequency of the current from which they result. They differ greatly, however, in how they affect nearby circuits.

The voltage resulting from magnetic induction varies inversely with the impedance of the line. That is, the higher the line impedance, the less voltage that can be induced by a magnetic field. Line impedance is usually determined by the reactance of the line itself and the nature of the terminating equipment. The voltage induced by the electric field, however, increases in direct proportion to line impedance. Thus, the higher the impedance, the greater the electric or capacitive pickup, and the lower the magnetic or inductive coupling. Figure 2 shows equivalent circuits for both types of coupling. Note that magnetic coupling is equivalent to being in series with the line, thus requiring a low impedance to be effective. Electric coupling is capacitive and requires a very high impedance in order to develop maximum potential.

Another source of interference is radio transmission which occurs in the same frequency ranges used for carrier transmission. Certain naval transmitters operate at very low frequencies which happen to coincide with some open-wire carrier channels. LORAN navigational transmissions may cause similar problems. This type of interference is

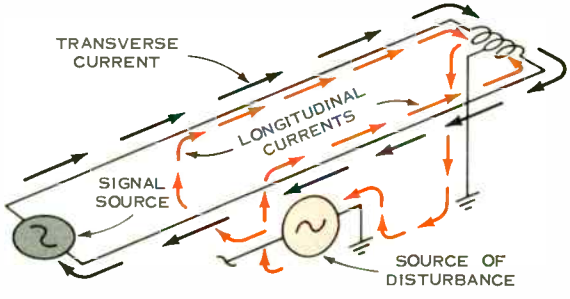


Figure 1. In a perfectly balanced line, longitudinal currents (red) can cel out in transverse circuit, but seek a ground return. Any imbalance converts some longitudinal current to transverse, producing interference.

World Radio History

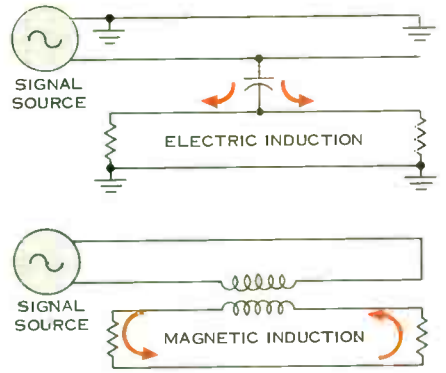


Figure 2. Comparison of electromagnetic ("magnetic") and electrostatic ("electric") coupling. Magnetic coupling is effectively "in series" with line, thus is suppressed by high impedance. Electric coupling is essentially capacitive, is reduced by low impedance line.

difficult to cope with, since measures designed to block reception of the radio transmission may interfere with the carrier signal. If this kind of interference is encountered in the central office, it may be possible to reduce it by shielding. If the interference is picked up on open wire lines, however, it may only be possible to reduce it, then only by careful balancing or special shielding techniques.

Shielding Techniques

If it were possible to keep all telephone circuits perfectly balanced with respect to ground, most interference would be eliminated. Unfortunately, a perfectly balanced circuit is all but impossible to obtain, except under laboratory conditions, and impracticable to maintain.

The most effective method of eliminating unwanted coupling is to isolate the disturbed circuit from the source of interference by some form of shielding. In principle, either the disturbing circuit, the disturbed circuit, or both, are surrounded by a metallic covering which intercepts the interfering fields and provides an alternate path for the longitudinal currents which are induced.

The nature of the shielding varies greatly, depending on the nature of the interference, its strength, and frequency. A good shield against electric fields may be ineffective against a magnetic field. Essentially, some sort of magnetic material such as iron or steel is required to block interference due to magnetic fields. Electric fields are best shielded by excellent conductors such as aluminum or copper.

At the lower (audio) frequencies, magnetic coupling can be minimized by enclosing the conductors in a braid or tape covering of steel, which tends to absorb magnetic fields. Electric coupling is usually negligible at these frequencies, so that magnetic shielding is usually all that is needed.

At higher (carrier) frequencies, braid, tape, or solid sheathing of copper, aluminum, or lead is somewhat effective against magnetic fields, due to induced eddy currents within the shield which oppose and partially neutralize the fields that produce them. The effectiveness of these materials against electric fields increases with frequency and with their conductivity and thickness. As frequency rises, the electric coupling remains much less than the magnetic coupling across the whole frequency range.

Another class of shielding uses a covering of steel wool, or a paper or fabric impregnated with carbon or some powdered metallic substance, which converts radio or other electromagnetic radiation into heat, and effectively dissipates these interfering fields. This form of shielding is particularly effective

World Radio History

against high-powered, high-frequency radio interference, which may be capable of filtering through the gaps on braided or overlapped shielding materials. In this respect, a laminated shield of two different materials is a far more effective shield than a solid shield of the same thickness made of only one material, due to reflective losses introduced at each interface. At 1 mc, a double braid of copper provides about 25 db better shielding than single braid, and a triple braid is 30 db better than the double. Coaxial cable with conventional copper braiding encased in a steel braid, and in turn encased in an outer copper braid provides a very effective shield for coaxial circuits from 10 kc up to several megacycles.

Although these generalities concerning shielding materials can be helpful, it is necessary to identify the exact nature of the interference in order to achieve maximum shielding effectiveness. Since magnetic and electric coupling cause different effects, shielding will have to be tailored to the type of interference encountered in the specific application.

By means of a simple test, it is possible to identify the nature of the coupling between two lines, as shown in Figure 3. A variable-frequency oscillator is connected to one circuit, and a sensitive voltmeter is connected to the other to measure the induced voltage. The oscillator is adjusted to a normal operating frequency and the far ends of both lines are short-circuited. Since voltages induced by magnetic coupling are *inversely* proportional to the line impedance, the low line impedance caused by the short circuit will result in a much reduced voltage if the principal coupling is electric. Conversely, if the principal fields are electromagnetic, the induced voltage will increase. The test should be repeated with the

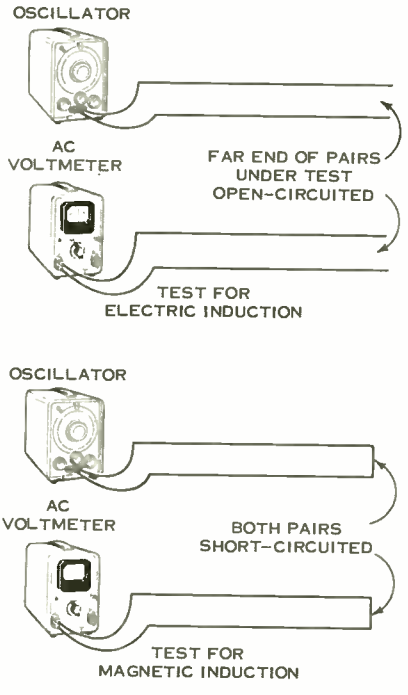


Figure 3. Type of coupling can be identified by simple tests. By opening or short-circuiting pairs at far end, relative degree of magnetic and electric coupling can be determined.

far ends of both lines open. The induced voltage will be the result of electric (or capacitive) coupling. Once the type of field has been identified, the most appropriate method shielding can be specified.

Very often, a change in line impedance can reduce interference (since coupling is frequency sensitive). Thus, if the impedance of a line is doubled, any voltage resulting from magnetic induction will be halved (at the same time, the voltage due to electrostatic coupling is doubled). If both fields are of equal strength, a change of imped-

World Radio History

ance will have no effect, since the voltage increase from one source is cancelled by the decrease from the second source.

Unbalanced Circuits

Unbalanced circuits present special shielding problems, since the shielding serves as the return path for the signal, and thus helps carry the signal. Even though part of the signal return may travel through other ground paths, a portion of it will travel through the shield, where it may be subject to interference caused by induced longitudinal voltages.

When coaxial cables are grounded at both ends, a longitudinal path or "ground loop" may be established which makes the cable vulnerable to interference from external magnetic fields. Even though the inner conductor may be shielded from the external field, longitudinal currents induced into the shield are, in effect, "in series" with the signal current flow through the shield, thus directly affecting the signal. Note that the low impedance of the ground loop restricts its effect to magnetic pickup. Interference primarily consists of lower frequencies, such as a-c hum from power mains, transformers, and other low frequency inductive components.

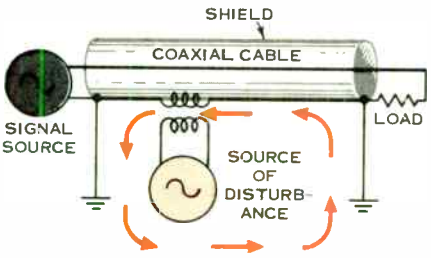


Figure 4. When coaxial cable used in unbalanced circuit is grounded at both ends, ground loop may introduce interference despite quality of shielding. Longitudinal currents induced magnetically share outer conductor with signal return.

In a run consisting of two or more coaxial cables, the possibilities of intercable coupling are great, simply because all shields are normally grounded at each end of the cable, and the shield of one cable could quite easily become the return path for the inner conductor of its neighbor, at least for some part of its full length. This can be reduced by enclosing the coaxial cable in heavy copper tubing or braiding, which serves as a ground return path for neighboring coaxial circuits.

Grounding

The objective of any grounding technique is to provide a path to earth of as low a resistance or impedance as possible. This is because the flow of current causes a voltage drop that is directly proportional to the resistance of the path to ground. Any resistance results in an unwanted difference of potential, the source of coupling to other circuits.

In a telecommunications system, it is necessary to provide a good ground for a wide variety of currents that may range from dc to very high radio frequencies. This is not easy to accomplish, since a low-resistance path for direct current is not necessarily a low-resistance path at radio frequencies. For example, the d-c resistance of any conductor is inversely proportional to its cross-sectional area, and directly proportional to its length. However, because of the self-inductance of the conductor, alternating currents tend to concentrate near the surface of the wire, rather than flowing uniformly through the whole conductor as in dc. This effect, diagrammed in Figure 5, which is called the "skin effect," increases as frequency and wire size become greater.

World Radio History

As a result, a typical solid copper conductor of 24 gauge shows a resistance of approximately 26 ohms per 1000 feet to direct current, and 37 ohms per 1000 feet at 500 kc. At 10 mc, the same wire shows a resistance of 170 ohms per 1000 feet.

Since any resistance to a flow of current creates a difference of potential, and a standing difference of potential contributes noise to the communications circuit, it is advantageous to provide the greatest possible conducting surface for any ground path. Thus, for direct current, a solid copper ground conductor is chosen which has a total current-carrying capacity substantially higher than the sum of all internal currents that share this ground. For alternating current, stranded conductors are preferred, since the perimeter of each strand provides a separate path for high-frequency currents, resulting in a much greater surface, or "skin" area for the whole conductor, and hence less resistance.

Since we must accept the fact that every ground path has some resistance or impedance, no matter how small, it is apparent that as more paths are added, the total resistance to ground can be reduced by adding parallel paths. Two basic grounding techniques may be used in achieving a physical ground. One technique, known as the single-point or common ground, uses a single driven ground rod, or some similar buried electrode, to which each ground wire is connected. The second approach, known as a distributed ground, employs a very heavy grounded buss, to which individual ground circuits are connected.

The same techniques also apply to grounding within electronic equipment. Several conductors requiring a ground connection can be joined at a single grounded "tie point," or can be connected at intervals along the chassis or a ground buss. In any multiple grounding procedure, however, it is extremely important to maintain all ground paths as near to one resistance as possible in order to maintain a fairly constant difference of potential. Should one path have a substantially larger difference of potential, there is additional danger of creating a ground loop.

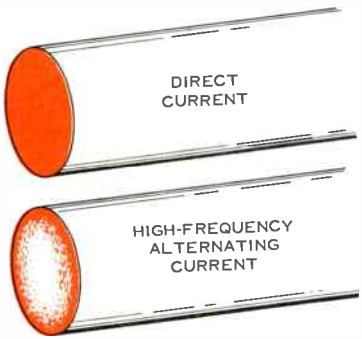


Figure 5. Increasing inductance of conductor at high frequencies forces electrons to flow near surface where flux density is less. This "skin effect" reduces available conduction path, increases a-c resistance.

Good Practice

The first requirement for a new installation (or for "quieting" an existing one) is to provide a well-bonded, low-impedance ground plane, which can normally be accomplished by electrically bonding all metal structures such as equipment racks, overhead rack supports, cable troughs, chassis, and equipment cabinets. The overall ground plane can usually be improved by spot-welding metallic structures at all joining surfaces. If at all possible, the ambient RF noise level within the building should be evaluated, and an attempt made to reduce it. In this respect, many items of test equipment often found in communications facilities are particu-

World Radio History

Figure 6. Typical HF or carrier frequency patch panel in telephone toll office. Shielded twisted pair is used to minimize coupling between adjacent circuits.

larly guilty of generating radio interference — among these are counters, oscillators, digital recorders, and similar instruments.

Finally, the appropriate coaxial cable or shielded wire should be selected for the prevailing signal frequencies and powers. The wide variety of shielded conductors available today makes it possible to achieve adequate isolation on either balanced or unbalanced transmission lines under most circumstances. However, at typical "HF" or carrier baseband frequencies, *shielded twisted pair* will usually prove superior to coaxial cable, particularly when the load must be isolated from ground. The improved balance obtained by the transposition or twisting of the conductors, plus the final protection of the outside shield, provides superior isolation. Thus, ground loops and RF pickup in areas of high signal level, can be controlled. In addition, there are certain applications where the frequency response limitations of transformers deny the use of a balanced line, and it is necessary to resort to unbalanced circuits. However, in many cases, the final selection may depend on the input or output configuration of the terminating equipment, and require some experimentation with the nature and location of the grounding. •

BIBLIOGRAPHY

1. F. H. Gooding & H. B. Slade, "Shielding of Communication Cables," *Communications and Electronics*: July, 1955.
2. O. K. Coleman, "Why Ground?," *Electrical Engineering*: May, 1956.
3. I. H. Pollack, "Skin Effect Resistance Data," *Electrical Design News*: December, 1956.
4. I. M. Newman and A. L. Albin, "An Integrated Approach to Bonding, Grounding, and Cable Selection," *Seventh Conference on Radio Interference Reduction and Compatibility*, IRE; November, 1961.

World Radio History

World Radio History

SECTION V
SEMICONDUCTOR DEVICES

World Radio History

World Radio History

the Lenkurt® Demodulator

VOL. 9 NO. 5 MAY, 1960

THE TUNNEL DIODE

In October, 1958, Dr. Leo Esaki, a Japanese physicist, published an account of a new phenomenon discovered in junction diodes which had been prepared in a special way. These diodes, now called "tunnel" diodes or Esaki diodes, exhibit remarkable properties which have unusual significance in the field of communications. Because the tunnel diode operates differently than transistors, and eliminates many of the transistor's inherent limitations, they are expected to have as great an effect on communications and electronics as did the discovery and development of the transistor itself. This article discusses some of the general characteristics of tunnel diodes and what may be expected from them.

It has been only about a year since the tunnel diode was announced in this country, and since then the tunnel diode has proved to be one of the most exciting advances in electronics since the announcement of the transistor. The reason is that the tunnel diode operates on new and different principles than the transistor, and is not subject to some of the limitations which restrict transistor performance at extremely high frequencies or in difficult environments.

Tunnel diodes, unlike transistors and conventional diodes, have no inherent frequency limitations. These diodes operate at the speed of light limited only by the inductance, capacitance, and resistance provided by their physical structure. For this reason, switching time of a diode used for computers or pulse communications, for instance, could readily be made as short as 10^{-12} second — that is, one-millionth of one-millionth of a second!

Transistors do not operate well while exposed to nuclear radiation — the radiation interferes with the lifetime of current carriers within the transistor. The tunnel diode uses a different type of current carrier and is not nearly as handicapped.

Temperature extremes do not interfere with the operation of the tunnel diode until the material becomes so hot

World Radio History

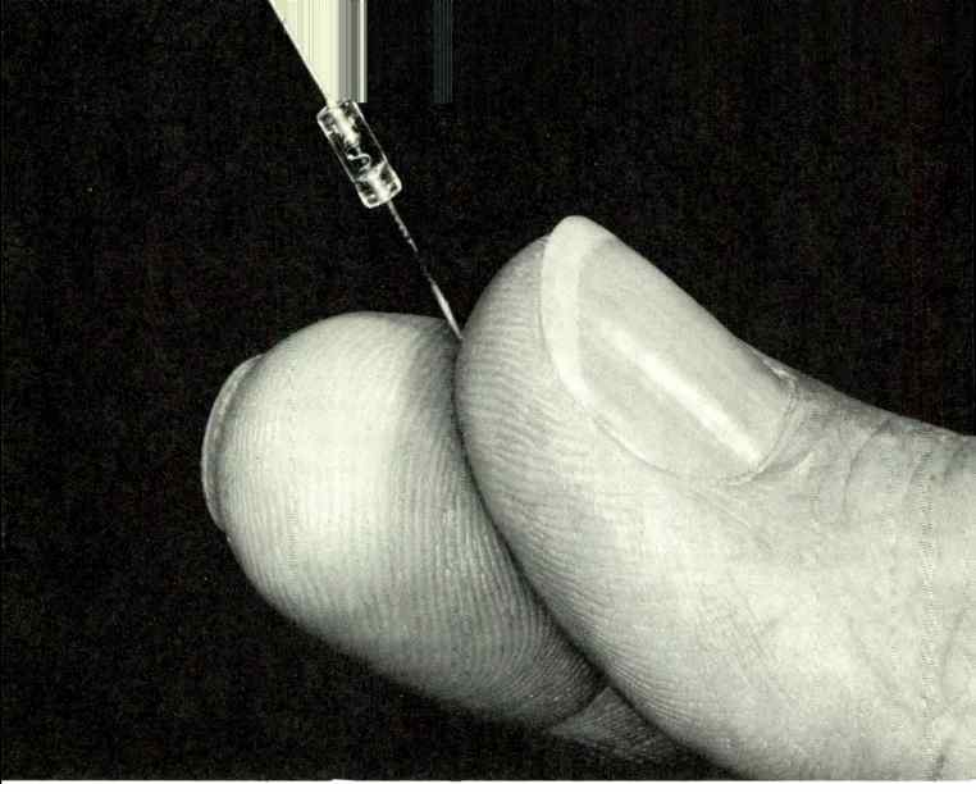


Figure 1. Tiny tunnel diode, shown here enlarged 2.5 times, actually takes up only about 1/100 the total volume of its glass "package." Physical dimensions are mostly determined by available mechanical techniques for sealing and making suitable connections to the device. Actual diode is thin wafer at upper end of S-shaped spring within glass. Since tunnel diodes require less than 1/100 the power of a typical transistor, and operate well in environments where transistors fail, unusually high numbers of such components may be crowded into tiny space for equipment of the future. An entire electronic telephone exchange might well fit into a coat closet or small vehicle.

as to lose its character. Tunnel diodes have operated quite successfully while immersed in liquid helium (-452°F.) and at temperatures above 750°F. By contrast, conventional transistors operate best at temperatures between the boiling point and freezing point of water.

Negative Resistance

How can any diode—which is a two-terminal device—amplify? The answer is provided by an examination of the E-I or voltage-current characteristics of the tunnel diode. Figure 2 compares the characteristics of a resistor, a conventional diode and a tunnel diode. Note that current through the resistor increases in direct proportion to the voltage applied across the resistor—just as defined in Ohm's law. The only time that this curve is not linear is when so much current is passed through a resistor as to change its resistance by heating.

The conventional junction diode will have an E-I curve similar to that shown in Figure 2-b. Note that current in-

World Radio History

creases almost linearly with applied voltage in the forward direction. In the reverse direction, little or no current flows until the reverse potential is so great as to cause breakdown of the junction.

The tunnel diode exhibits entirely different characteristics. It is highly conductive for reverse voltage. Current increases almost linearly with the application of forward voltage until the curve suddenly reverses. At this point, increased forward voltage causes a *reduction* in current, while decreased voltage *increases* current. This negative resistance effect occurs over a relatively narrow voltage range in presently available experimental diodes. Decreasing temperatures extends the voltage range over which the negative resistance effect occurs. If applied voltage is increased further, current again increases, as in other devices.

It is only the negative resistance portion of the curve that permits the diode to amplify. Negative resistance devices are not new, but have never been so simple, cheap, or convenient as the tunnel diode. For instance, negative resistance repeaters have been used for many years to amplify messages traveling in both directions on two-wire telephone lines. Figure 3 shows a simplified schematic diagram of such a repeater. The device requires two amplifiers (either tubes or transistors) plus many other components. A signal traveling in either direction is amplified and reapplied to the line in the proper phase required to reinforce the signal. A single, properly biased diode shunted across the line would have the same effect, as diagrammed in Figure 4. Practical tunnel diode amplifiers operating on this principle have been made. Figure 5 shows a suggested radio fre-

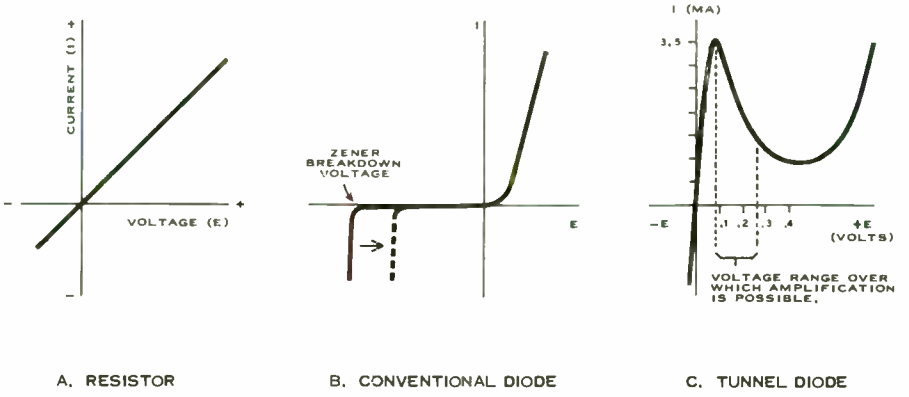


Figure 2. Comparison of the voltage-current relationships of an ordinary resistor, conventional junction diode, and a tunnel diode. As voltage across the resistor (A) is increased, current goes up proportionately. "Forward" voltage across junction diode (B) causes near-linear increase in current, high resistance to reverse voltage until breakdown voltage is reached. Increasing "doping" of junction results in lowered back resistance as indicated by dotted curve. Tunnel diode (C) is highly conducting for both reverse and forward voltages, but shows unique negative resistance over a narrow range of forward voltage.

World Radio History

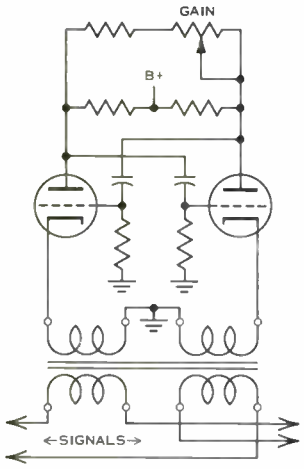


Figure 3. Typical negative impedance repeater such as long used for two-way amplification of telephone messages. Signal voltage in either direction is amplified and reinserted on line in correct phase to strengthen signal.

quency amplifier using a tunnel diode. Note the similarity to the telephone-type negative resistance repeater.

Negative resistance devices have been studied for at least thirty years, and many fascinating applications for the effect were proposed. The use of negative resistance to neutralize positive resistance in networks was described and sought, although researchers had difficulty in accomplishing it, due to instability of the resulting circuits.

As early as 1934, the use of negative resistance in voltage-controlled tuning was demonstrated, and circuits were devised to reduce or eliminate grid-plate capacitance in electron tubes, using negative resistance. Even the cumbersome negative resistance circuits then available were shown to provide significant advantages as heterodyne frequency meters and signal generators, and could readily be used as modulators, detectors, and oscillators.

How It Works

In order to explain the operation of tunnel diodes, let us first review briefly the operating principles of conventional junction diodes.

Tunnel diodes, transistors, and conventional junction diodes are composed of *semiconductor* material, that is, material having electrical properties midway between those of insulators and conductors. Whether a material is a conductor, semiconductor, or insulator is determined by its atomic structure.

All materials are composed of atoms consisting of a central nucleus having a positive charge, surrounded by negatively-charged electrons. In a complete atom, the number of electrons always matches the positive charge of the nucleus. Various materials will have different numbers of electrons. The electrons orbit around the nucleus in one or more "rings" or shells which repre-

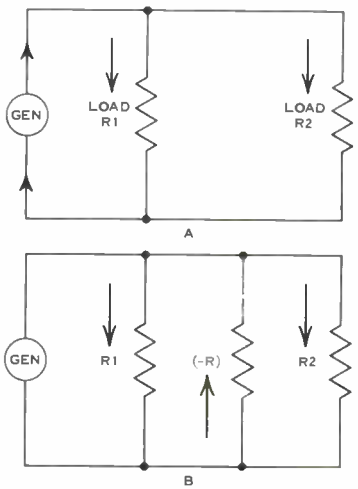


Figure 4. Generator in (A) causes flow of current through shunt loads as indicated by arrows. If negative resistance is inserted in circuit as shown in (B) negative conduction actually increases flow of current through R1 and R2.

World Radio History

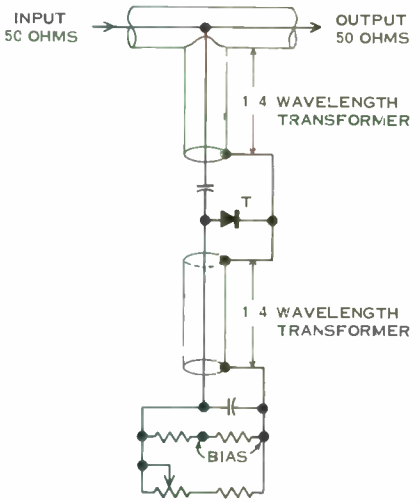


Figure 5. Tunnel diode equivalent of circuit shown in Figure 3. This amplifier is designed for UHF service, requires transmission-line transformers for impedance match between 50-ohm line and much lower impedance tunnel diode.

sent different energy levels of the electrons. Each ring has a definite quota of electrons that it can hold. For instance, there is room for only two electrons in the first ring, eight in the next, eight in the third ring, 18 in the fourth, and so forth.

Atoms always strive to complete their outer ring. Once the ring is completed, much greater amounts of energy are required to dislodge an electron than if the ring is incomplete. For this reason, atoms which have just enough electrons to completely fill one or more rings (such as helium, with two electrons, and neon with 10) are completely inert and will not react chemically with other elements. On the other hand, if the ring is incomplete, relatively little energy is required to attach or remove an electron from its incomplete ring. The actual energy required to free electrons depends on such factors as the number of electrons in the outer ring and how nearly complete it is.

Electrical current always requires the flow of electrons. These, and their positive counterparts ("holes") are often referred to as "current carriers" or just "carriers" in the language of solid-state electronics. Conductors have electrons so loosely bound to the outer ring that even a slight electrical potential will dislodge them, freeing them to conduct current. In the case of insulators, electrons are much more tightly bound, perhaps because the ring is very nearly complete. Much higher electrical potentials are required to loosen them. Higher temperatures may supply some of the energy necessary to dislodge electrons from the atoms of insulators. For instance, glass, which is an excellent insulator at normal temperatures, becomes a very good conductor at high temperatures.

Electrons must achieve a specific energy level before they can be freed

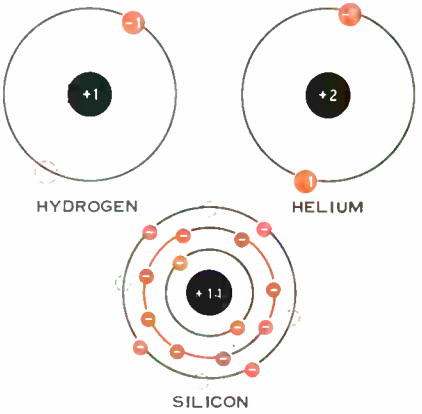


Figure 6. Hydrogen atom with one electron has incomplete outer ring. Helium atom has filled outer ring and is completely inert. Silicon atom has four of the eight possible electrons required to complete its outer ring. Relatively little energy is required to free electrons.

World Radio History

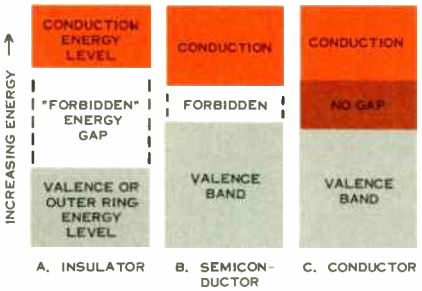


Figure 7. Electrons cannot occupy certain energy levels ("forbidden gap"), but must absorb enough energy to go "all the way" from valence band (energy level possessed by outer ring electrons) to conduction level, the level achieved by free electrons. Levels and forbidden gap vary from material to material, and according to applied voltage.

from their position in the outer shell of an atom. Sometimes there is a large gap between the normal energy of the outer-ring electrons and the energy required for freedom (conduction). In conductors, the gap is non-existent; that is, no extra energy is required for conduction. In semiconductors, the gap is very small; in insulators, it is quite large. Thus, only conductors have an abundance of electrons which are free to carry current.

In semiconductors, the application of external energy, such as heat, light, or electricity will free electrons for duty as current carriers. The number of electrons increases with the amount of energy applied. The number is never very large, however, compared to a conductor. Conventional semiconductor diodes make use of this characteristic to control the flow of current.

First, the pure semiconductor is modified to increase the number of easily-freed electrons. This is done by adding small amounts of selected impurities to the semiconductor; and the semiconductor material is crystallized. The impurities are selected for their ability to substitute for one of the semiconductor atoms in the crystal. In addition, after "locking arms" with the semiconductor atoms, the impurity atoms must have one electron left over, or be deficient by one electron. The presence of surplus electrons separate from the crystalline structure increases the number of carriers available. It also lowers the resistance of the material,

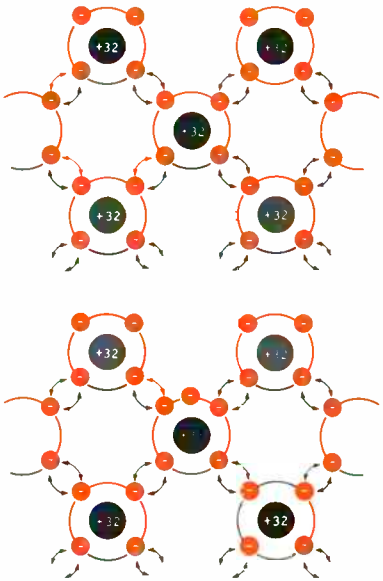


Figure 8. Crystalline structure of a semiconductor, showing how electrons are shared by atoms to complete their outer rings. If material is "doped" with impurity having the same valence, one outer electron is unused by crystal structure, thus increasing number of free electrons available for conduction. This type of doping produces n-type semiconductor. P-type semiconductor is produced by adding material which is one short of the number of electrons required by strong crystalline structure. Electron deficiency produces net positive charge ("hole") which may be transferred from atom to atom by shift of free electrons.

World Radio History

and makes it more sensitive to applied electrical potential. Just enough impurity is added to improve the sensitivity of the material, but not enough to create so many free electrons that they are impossible to control by applied voltage.

If the impurity provides a *surplus* electron, the semiconductor is called *n* type material because the current carriers are negative (electrons). If the impurity provides a *deficiency* of electrons, the material is called *p* type semiconductor because it behaves as though there were positive carriers for the current. Although no positive carriers physically exist, the *hole* caused by the absence of one electron exactly simulates a positively-charged particle.

PN Junctions

If samples of *p* and *n* semiconductor materials are intimately joined, a very efficient rectifier is produced. A situation similar to that diagrammed in Figure 10 exists. The *n* material consists of *donor* atoms (atoms which furnish an electron) and loosely-bound or free electrons. Donors have a positive charge because of the loss of the free electrons. The *p* material consists of acceptor atoms having a negative charge, and holes, which simulate mobile positive charges.

In the *n* material, the positively-charged donor atoms (ions) are locked in place by the crystalline structure. The free electrons, however, are able to migrate under the influence of electrical fields. A similar, but reverse, condition exists in the *p* material. At the junction, the positively-charged donor atoms create a field which extends across the junction and repulses the positively-charged holes. Similarly, the negative charges of the acceptor atoms in the *p* type semiconductor repulse the free electrons in the *n* material. This mutual repulsion creates a sort of "no man's land", called a *depletion zone* because it is depleted of current carriers.

The repulsive force is equivalent to a small voltage potential which must be overcome by the current-carrying holes and electrons before they can approach and cross the junction. Only carriers which have acquired enough energy to climb this electrical potential hill can cross the junction. Thus, the potential barrier may be symbolized as a small battery connected across the junction, or as a small voltage hill that must be "climbed" by carriers crossing the junction.

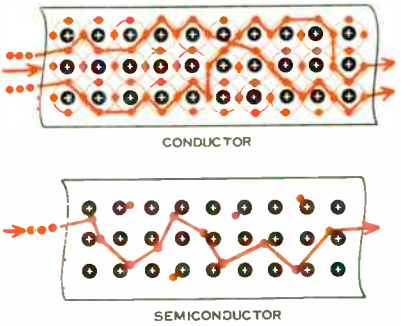


Figure 9. Flow of electricity through conductor results from entering electrons crowding free electrons out other end. In semiconductor, current carriers must hop from atom to atom in crystal lattice, requiring much more time than in conductor.

If an external battery is connected to the *pn* junction, as shown in Figure 11, the voltage or potential hill is reinforced, because the external voltage pulls the carriers even farther away from the depletion zone. This increases the electrical resistance of the junction by reducing the number of available carriers. If the polarity of the external battery is reversed, the applied voltage forces the carriers toward the junction,

World Radio History

lowering the potential hill. This lowers the resistance of the junction to a low value.

If the impurity concentration of an ordinary *pn* junction is increased, the breakdown voltage is decreased, as shown in Figure 2-b. If the diode is heavily doped with impurities (roughly 10^{19} atoms per cc.), it becomes highly conductive when reverse-biased, and

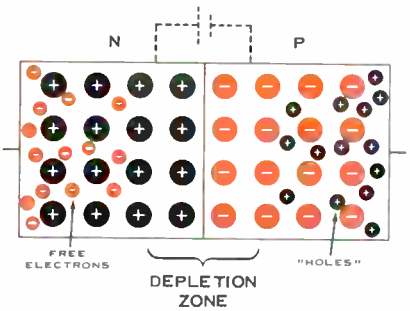


Figure 10. Field from donor atoms repulses similarly-charged carriers across junction. Simulated battery indicates potential that must be overcome to cause electron to diffuse from n *to* p *material.*

the negative conductance appears at certain values of forward bias. Apparently this occurs because the energy levels of the *p* and the *n* type materials shift with respect to each other so that the outer-ring electrons in the *p* material have the same energy level required for conduction in the *n* material. When this occurs, carriers "tunnel" through the depletion region at the speed of light. If the forward bias is further increased, the energy levels of the two materials are shifted so that the energy of outer-ring electrons of the *p* material no longer matches the conduction level of the *n* material. Therefore, tunneling is no longer possible and carriers can only cross the junction barrier by absorbing additional energy from the applied potential, as in ordinary diodes.

In conventional diodes and transistors, carriers must take a roundabout path, leaping from atom-to-atom in the crystal structure of the semiconductor. Because of this, carriers travel quite slowly through semiconductors, holes traveling only about 16 meters per second for each volt-per-centimeter of potential gradient. Electrons travel about 36 meters per second under the same conditions. It is this slow propagation rate that limits the frequency performance of transistors and diodes. The tunneling phenomenon, however, occurs at the speed of light, and frees the tunnel diode from such frequency limitations.

Unlike ordinary diodes and transistors, the tunnel diode is little affected by extremes of temperature. Since the shifting of the *relative* energy levels of

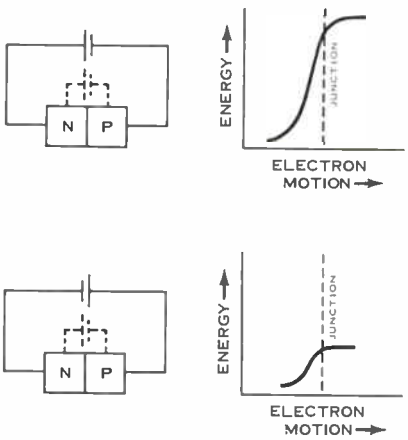


Figure 11. Potential "hill" to be overcome for conduction is increased by external battery connected as shown in (A), but is lowered when battery is connected as in (B). This accounts for high reverse-current resistance, low forward-current resistance of diodes.

World Radio History

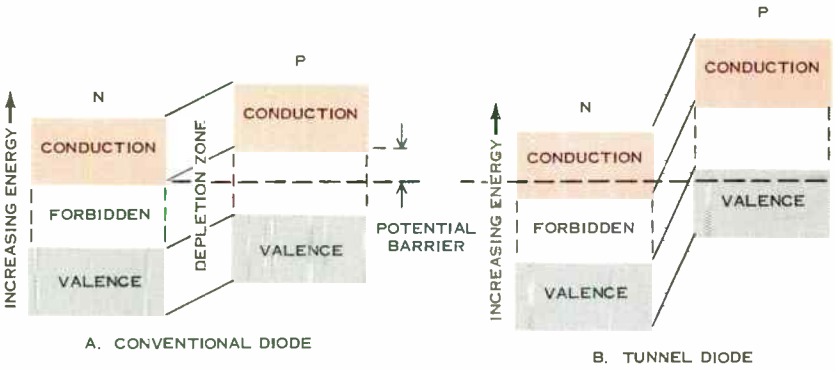


Figure 12. Very heavy doping of n *and* p *semiconductors shift their respective energy levels. Applied forward bias tends to reduce difference between two. When valence band of* p *material is directly opposite conduction level of* n *material, current carriers may "tunnel" through potential barrier at speed of light. Increased forward bias shifts* n *valence band opposite* p *forbidden zone, restoring normal diode behavior.*

the two types of material is primarily based on bias across the diode rather than temperature, tunneling is generally independent of temperature. In transistors and conventional diodes, temperature directly affects the number of carriers that are produced in both *p* and *n* materials, and this determines the current density and the performance of the device. Although various transistor and diode materials will have different sensitivities to temperature, they cannot avoid being influenced by temperature, so long as they continue to employ present operating principles.

Conclusions

The tunnel diode, for all its promise, presents many problems. It is a two-terminal device and will amplify equally well in either direction. This makes circuit design quite difficult. For instance, isolation between stages is difficult to obtain. Because of the negative resistance characteristics of the device, it tends to oscillate with almost any stray capacitance or inductance present—the value of the capacitance and inductance determining the frequency of oscillation. By using tunnel diodes with other components such as transistors and ordinary diodes, the disadvantages of both types may be overcome and the advantages of each reinforced. Just as the characteristics of transistors required new circuit designs, the tunnel diode presents similar design problems that will most certainly be overcome. •

BIBLIOGRAPHY

1. L. Esaki, "New Phenomenon in Narrow Germanium p-n Junctions," *Physical Review,* Vol. 109, p. 603; 1958.

2. E. W. Herold, "Negative Resistance and Devices for Obtaining It," *Proceedings of The IRE,* Vol. 23, No. 10, p. 1201; October, 1935.

3. H. S. Sommers Jr., "Tunnel Diodes as High Frequency Devices," *Proceedings of The IRE,* Vol. 47, No. 7, p. 1201; July, 1959.

4. I. A. Lesk, N. Holonyak Jr., U. S. Davidsohn, M. W. Aarons, "Germanium and Silicon Tunnel Diodes—Design, Operation, and Application," *IRE Wescon Convention Record,* p. 9; August, 1959.

5. Coblenz and Owens, *Transistors—Theory and Applications,* McGraw-Hill, New York; 1955.

World Radio History

World Radio History

the Lenkurt® Demodulator

VOL. 12, NO. 4 APRIL, 1963

The Varactor Diode

The total elimination of electron tubes from equipment is almost an obsession in some parts of the communication industry. More and more equipment is being "transistorized" or converted to "solid state" in the hope of reaching new levels of reliability or of reducing the space and power required. A new step in this direction has been made possible by the varactor diode, a device with some highly unusual characteristics. This article reviews some of the general properties and behavior of varactors and how they can be used in communications equipment.

The whole field of microwave communication was made possible by the invention of the klystron just before World War II. Despite great progress since then in almost every branch of electronics and physics, microwave communication still depends on the klystron. The extreme simplicity of the reflex klystron and its steadily growing life expectancy have made it hard to surpass as a source of easily modulated microwave energy.

The varactor diode, a relative newcomer in electronics, may eventually end the klystron's monopoly on the economical generation of microwaves suitable for communications. The varactor can be an extremely efficient — even irrepressible — generator of harmonics of the signals or waveforms applied to it. This particular characteristic lends itself particularly well to use in frequency multipliers.

The varactor is a simple *p-n* junction diode which is used as a voltage-controlled capacitor. The capacitance of the device can be made to vary not only with the applied bias voltage, but also with the instantaneous values of signal voltage. The result is a non-linear device which produces a wealth of harmonics of the applied signals, as well as a profusion of frequencies representing the sum and difference of the original signals and many of their harmonics. Although most diodes possess the re-

World Radio History

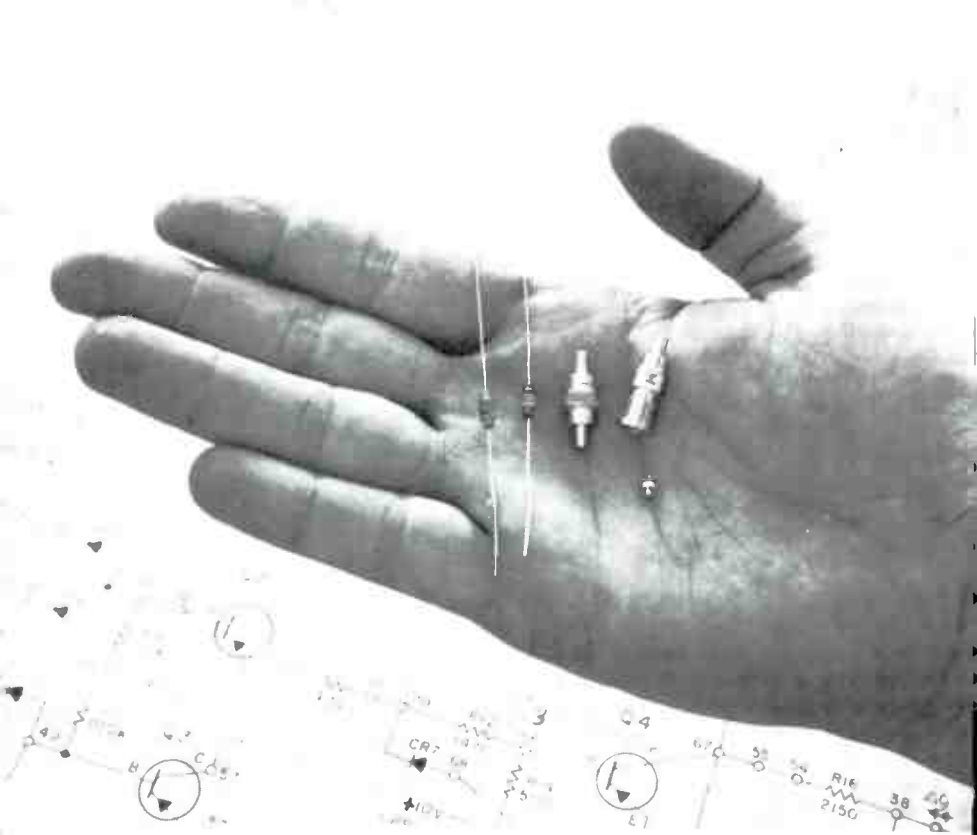


Figure 1. Varactor diodes take many forms, of which only a few are shown here. Diodes with leads are for low frequency circuits using "lumped" circuit elements. The tiny "pill" diode is used with stripline or waveguide circuits at very high frequencies. The cartridge types may be used for all frequencies, in both lumped-element or waveguide multipliers.

quired properties to some extent, diodes which have been designed to enhance certain characteristics provide the best performance.

How it Works

A junction diode consists of two layers of semiconductor material (such as silicon) which has been "doped" with impurities so that one layer has a deficiency of electrons, and the other a surplus. These layers are usually designated as the *p* and *n* layers, respectively. (For a fuller description of *p-n* junctions, see DEMODULATOR, May, 1960). In the *n* layer, the surplus electrons are free to migrate under the influence of a field or applied voltage, thus leaving a net positive charge. Similarly, in the *p* layer, the deficiency of electrons leaves "holes" which behave as though they were mobile positive charges. Holes also are free to leave the area under the influence of external fields, thus leaving a negative charge.

Where the two materials meet, their respective charge fields extend across the junction and influence the current

World Radio History

carriers (mobile electrons or holes) in the opposite material. The positive charge from ions in the *n* crystal tends to repulse the holes in the *p* material, and the negative charge of the *p* material pushes back the free or loosely-bound electrons in the *n* material. By thus removing the current carriers from the junction area, a *"depletion zone"* is created through which current cannot flow until the mutual repulsion of carriers is overcome by the application of a suitable voltage of the correct polarity. As shown in Figure 3, the depletion zone behaves as though it were a battery of a certain voltage. In order for current to flow through the diode, it is necessary to overcome this voltage or "contact potential" by a greater voltage of the opposite polarity.

If the external voltage is connected so as to *reinforce* the contact potential rather than oppose it, the depletion zone is made larger as the current carriers in both layers are drawn even farther from the junction. Such a reverse-biased diode behaves just like a capacitor. The depletion zone becomes the dielectric, and the boundaries of the current carriers simulate the two plates or electrodes of the capacitor. The capacitance value will vary with the area of the junction, the nature of the semiconductor material and — most important of all — the applied voltage. As the reverse bias increases, the two capacitor "plates" are pushed farther apart, thus reducing the effective capacitance across the junction. As bias is lowered, the two plates draw closer together, and capacitance increases. The result is a new class of "capacitor" in which capacitance can be varied by changing the applied voltage. Figure 5 shows a typical voltage-versus-capacitance curve for a varactor diode.

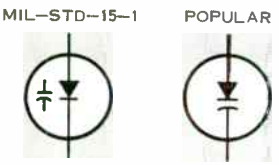


Figure 2. Two widely used symbols for the varactor diode. Varactors are semiconductor junction diodes which are used as variable capacitors.

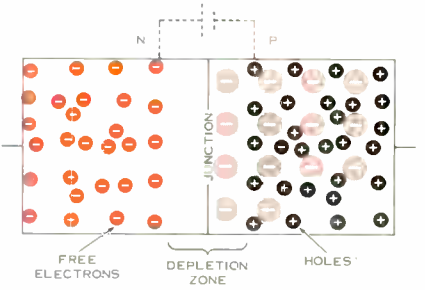


Figure 3. P-n junction consists of semiconductor material which has been suitably "doped" with donor atoms. Field from charged donors extends across junction to repulse similarly charged current carriers (electrons and "holes"). Simulated battery represents the "contact potential" which must be overcome before current can flow. Current carriers act as capacitor plates, depletion zone serves as dielectric.

This unique electrical control of capacitance opens the door to a great number of applications. Perhaps the simplest and most obvious is the electrical control of tuned circuits for automatic frequency control. Another is a "parametric" amplification. Still another use is harmonic generation. Unlike a simple capacitor, the varactor capacitance varies continuously as the signal wave itself changes value. The effect on the signal is very complex. Not only does the changing signal value cause the circuit

World Radio History

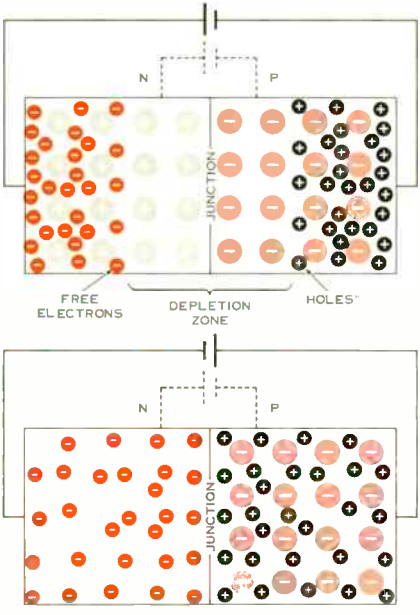


Figure 4. If contact potential is reinforced by external voltage, carriers are forced back from junction area. This enlarges depletion zone, lowers capacitance. If external voltage opposes contact potential, current carriers are forced closer to junction, changing its capacitance.

reactance to vary continuously, but the amount of signal energy absorbed and then returned to the circuit by the capacitance also changes. The result is a highly distorted output wave that is extremely rich in harmonics.

The extreme non-linearity of the varactor diode lends itself admirably to use in frequency multiplying circuits similar to those long used in FM transmitters for mobile communications and broadcasting. An important difference between conventional amplifier-type multipliers and varactor multipliers, however, is that the latter are entirely "passive." That is, they require no power other than the input signal to be multiplied. Not only that, they are often remarkably efficient, converting as much as 90% of the input signal into the desired higher harmonic. By contrast, conventional multipliers rarely exceed 30-40% efficiency.

Figure 6 illustrates a typical varactor frequency doubler. It consists essentially of two resonant circuits coupled together through a common impedance, the varactor itself. The input circuit is series resonant to the fundamental frequency, F, which is to be multiplied, thus assuring maximum energy transfer to the diode. Similarly, the output circuit is tuned to the $2F$ harmonic, thus assuring maximum output. Series resonant frequency "traps" block the flow of F currents in the output circuit or $2F$ currents in the input circuit. It is also possible to tune the circuits to obtain the third, fourth, or higher harmonic to achieve higher-order multiplication. However, this tends to complicate the circuit and lower efficiency. Figure 7 shows a typical tripler circuit, one that multiplies the input frequency by three. Note that it is necessary to provide an

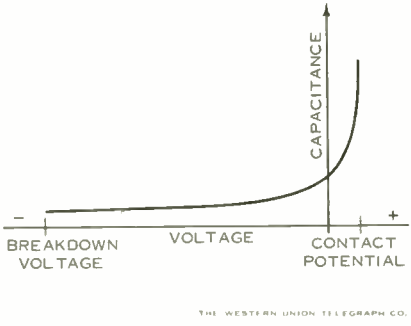


Figure 5. A typical varactor diode voltage-capacitance relationship. Note that capacitance increases very rapidly as applied voltage approaches the contact potential. Diode conducts above contact potential or below breakdown voltage.

World Radio History

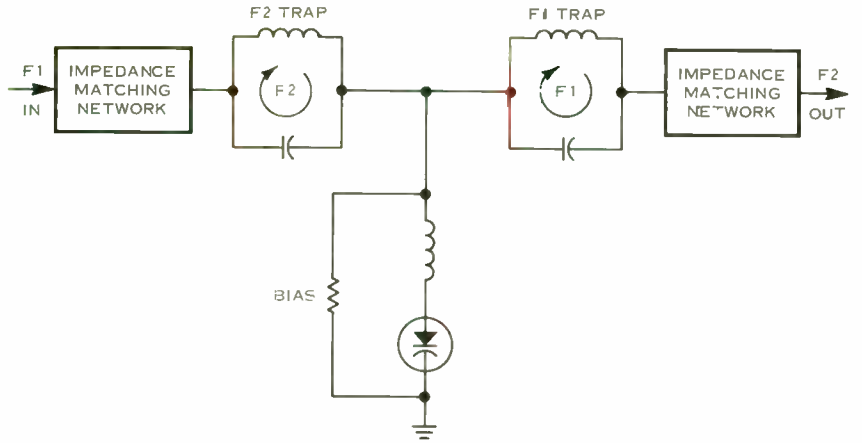


Figure 6. Typical varactor frequency doubler circuit. Frequency F is coupled to diode by series resonance of diode, inductor, and 2F frequency trap elements. The 2F harmonic is coupled efficiently to output, but blocked from input by trap.

additional shunt-resonant "idler" circuit which is resonant at $2F$, followed by the circuit resonant at the desired $3F$ harmonic.

Efficiency

The very high efficiency that may be attained in a varactor multiplier is achieved because the device is essentially reactive rather than resistive. Theoretically, a purely reactive (in this case, capacitive) frequency multiplier will be 100% efficient. In reality, however, small losses inevitably occur in the conductors and circuit components and in the series resistance of the diode itself.

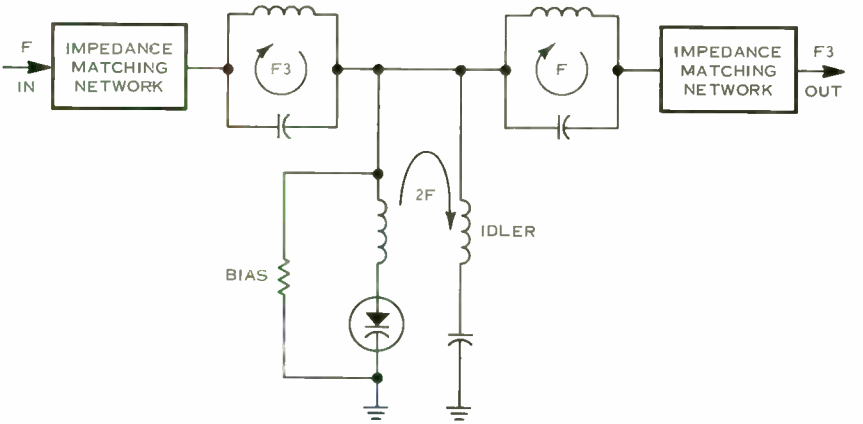


Figure 7. Frequency tripler is essentially the same as doubler, but requires a so-called "idler" circuit resonant at 2F harmonic.

World Radio History

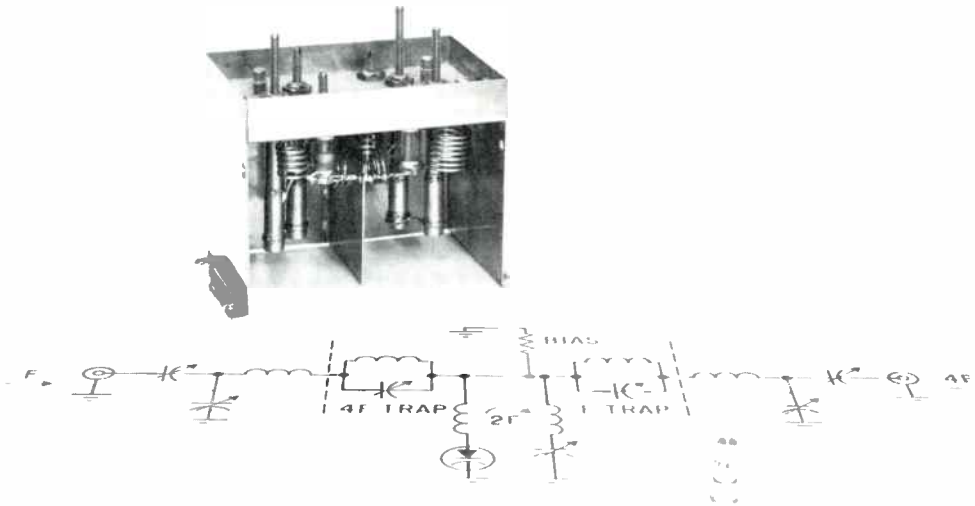


Figure 8. Actual frequency quadrupler and its schematic. This prototype device is of the same configuration as shown in Figure 7, also requires idler resonant at 2F.

Diode resistance is perhaps the most significant source of loss since it largely determines the amount of power that the diode will be able to handle.

Most of the diode series resistance occurs in the semiconductor material itself. For high frequency operation, it is necessary that the diode capacitance, and therefore the junction area, be quite small. For instance, for operation at 6000 mc, junction capacitance should be about 0.2 picofarad (micromicrofarad), thus dictating an almost microscopic junction area. The volume of semiconductor material involved is extremely small, yet it must dissipate whatever loss results. The power loss in the tiny semiconductor wafer will equal I^2R, where I is current and R is resistance. Even though series resistance may be low, typically only a few ohms, it provides most of the loss in the diode. Even where loss is small, it is still concentrated in a very small volume of matter. This results in a rather severe local temperature rise which will quickly ruin the device unless the heat is carried away very rapidly or the applied power is carefully limited. Because of this temperature problem, manufacturers almost exclusively use silicon as the semiconductor material because of its excellent thermal conductivity and low temperature sensitivity, compared to other semiconductors.

It follows that greater power can be accommodated if the loss is reduced. The power handling capability of the varactor is proportional to its breakdown voltage value, since the applied voltage must be limited to the range between the diode contact potential and the reverse voltage at which the diode breaks down.

World Radio History

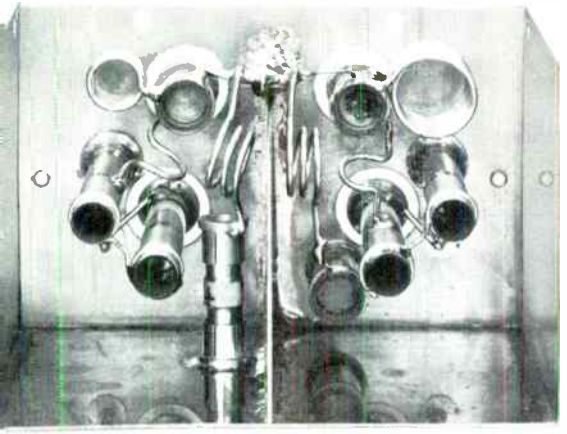


Figure 9. Detailed view of the quadrupler. Varactor is the cylinder attached to the wall in the left chamber. Sole source of power for such circuits is the input signal. Many such stages, plus special tuners may be required to reach microwave frequencies.

Practical Devices

Although these restrictions limit the amount of power that may be handled at high frequencies, it has been possible to generate as much as a watt at 6000 mc with laboratory units, and about half a watt with commercially available varactors. Power at these frequencies is achieved by cascading several multipliers in tandem. Although the efficiency of each stage may be high, depending on the degree of multiplication achieved, the total efficiency of the chain is much lower. For instance, if a chain of four frequency triplers were employed to convert a 95-mc signal to one of 6080 mc, and each tripler were 50% efficient, only 6% of the 95-mc input power would be converted to the 6080 microwave signal. If it were necessary to achieve a full watt of microwave output for transmission, at least 16 watts of input power would be required. Actually, even more input power is normally required, for at higher frequencies varactor efficiency becomes less.

Figure 10 shows a simplified block diagram of an actual multipler used to provide local oscillator energy in a microwave receiver. In this application, required power was 10 milliwatts or less. To achieve this, it was necessary to amplify the output of a crystal oscillator to provide 1½ watts of driving power for the varactor multiplier chain. A tripler and two quadruplers yielded the desired 6400-mc signal at a power of 15 milliwatts — an efficiency of 1%.

This is directly comparable to the performance of reflex klystrons used as local oscillators, except that the multiplier chain is more complicated and far more difficult to adjust. Offsetting this difficulty is the fact that the power supply requirements for a multiplier chain are usually quite modest. Dc power is applied only to the crystal oscillator and the transistor power amplifiers which drive the multiplier chain. Unlike most solid-state local oscillators which have appeared in commercial microwave equipment, this one introduced slightly

World Radio History

less noise than a typical local oscillator klystron.

The performance quality that can be obtained from multiplier chains is probably a function of the engineering refinement invested in the device. For all their efficiency in converting a frequency to its multiple, varactors have certain other characteristics which may cause only trouble.

For one, varactors are excellent mixers as well as efficient frequency multipliers. In fact, they appear to be slightly better mixers than multipliers. In addition to yielding a multiple of the input frequency, they also produce frequencies which are the sum and difference of the desired multiple and the frequencies which have appeared in preceding circuits.

For instance, if a varactor chain is driven by a 6-mc signal as the first stage in reaching a higher frequency, the 6-mc component will be sharply attenuated in the first multiplier by the presence of frequency "traps" and other filtering. However, with each successive multiplication, 6-mc sidebands will be present above and below the desired multiple. As multiplication increases, these sidebands gain strength relative to the desired frequency. In the worst case, this enhancement of the spurious frequencies is equal to 20 log R, where R is the degree of multiplication. Assume, for instance, that the 6-mc signal is doubled, and that the filtering in the doubler reduces the 6-mc signal 50 db in the doubler output. If the 12-mc output is multiplied an additional 64 times to yield 768 mc, the 6-mc sidebands will appear on either side of the carrier at each stage. Assuming that the filters in these multipliers are too broad to attenuate these sidebands further, they will grow in strength relative to the desired frequency. After a multiplication of 64 times, the sidebands will be only 13 db lower in amplitude than the desired 768-mc signal. Further multiplication would enhance these spurious sidebands even more. In addition to the 6-mc sidebands, other frequencies which occur in the chain would also be pres-

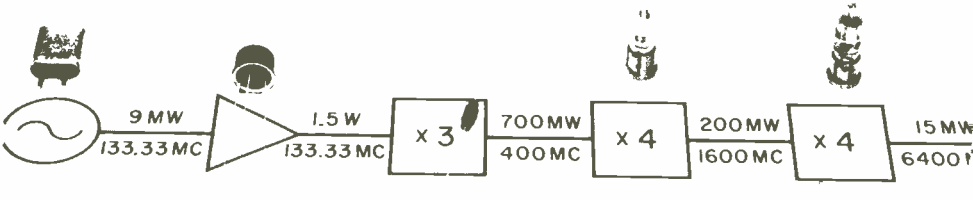


Figure 10. Simplified block diagram of an experimental microwave local oscillator. Actual circuit is much more complex than indicated here. At microwave frequencies, coaxial or stripline circuits are required. Greater multiplying factors may be achieved in each stage, but these tend to be less efficient than low-order multipliers. Principal advantage is to reduce the number of expensive diodes and other components required.

World Radio History

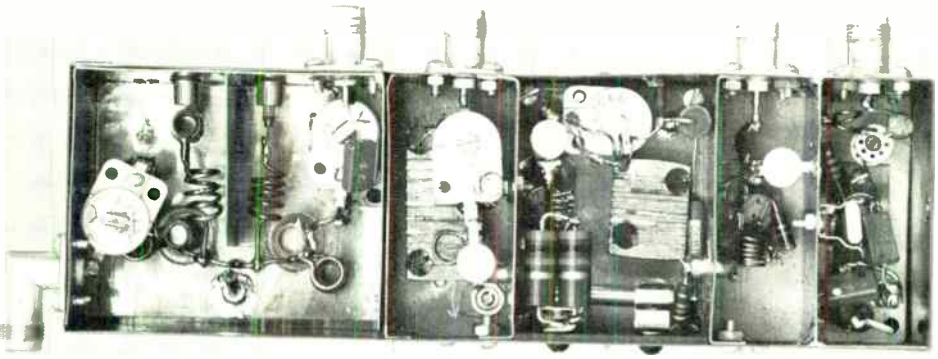


Figure 11. Typical solid-state multiplier circuit. This "breadboard" device employs crystal oscillator to drive varactor multipliers. Output is 450-mc signal which is used to drive more multipliers.

ent. However, since these have undergone less multiplication and lie farther from the desired frequency they will probably be attenuated much more. In order to obtain reasonably pure frequencies, therefore, it is vital that filtering be extremely effective, particularly in the earliest multiplier stages. This, however, tends to increase the cost and complexity of the equipment and reduces bandwidth. This is one of the reasons why varactor multipliers are generally used for generating single frequencies rather than for broadband modulated signals.

Other Difficulties

Varactors show a variety of effects which complicate the design of frequency multipliers. These include hysteresis, tuning difficulties, parametric amplification and oscillation, and others. Hysteresis may be present when the input signal to a multiplier is smoothly increased and the output signal is observed to make a large, sudden jump. As input power is reduced gradually, the output signal will drop sharply, but not at the same point at which it previously jumped. In some cases, a circuit may have two possible operating conditions at certain signal levels. When the circuit is first energized it will operate in one or the other of the two modes, but may switch spontaneously to the other under the influence of strong driving signals. When this occurs, it is usually the result of extreme detuning

World Radio History

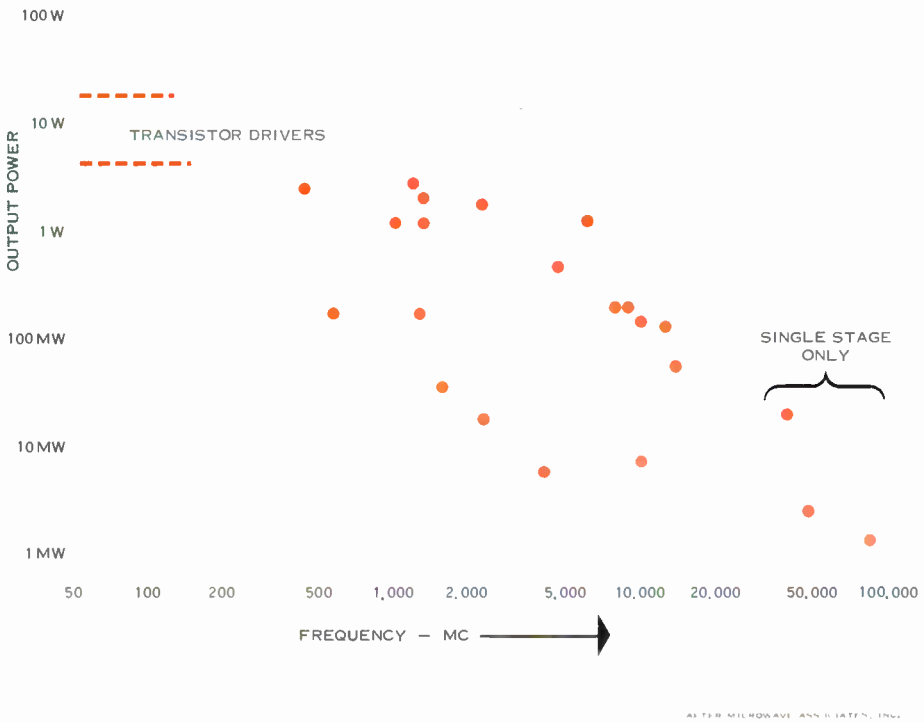


Figure 12. Typical performance capabilities of present-day production varactors. In general, the greater the power-handling ability, the lower the maximum operating frequency. These limits are imposed by materials and techniques used in manufacturing the devices.

of the circuit or of changes in impedance caused by the variable capacitance of the varactor. With suitable care in circuit design and adjustment these effects may be avoided.

Parametric amplification within the varactor may be another source of trouble. Parametric amplifiers are those which use a varying "parameter" such as reactance to take power from one source of energy and use it to build up or amplify another. Parametric amplifiers are sometimes used as the first stage of microwave receivers where receiver thermal noise must be low, as in satellite communications. In general, parametric amplifiers require a "pump" frequency, usually at twice the frequency of the signal to be amplified, which provides the energy necessary to amplify the input signal.

Varactors often show a tendency to behave as parametric amplifiers. Energy from the input signal may be used to

World Radio History

nourish parametric amplification of noise, subharmonics, low-level fundamental frequencies, or other signals which may be present. Spurious, chance resonances in the circuitry may increase these effects, leading to unstable operation.

A characteristic difficulty in varactor multiplier chains is the tuning and adjustment of the circuit. Varactor characteristics change with signal amplitude, and this, in turn, varies with circuit tuning. Because of the mutual interaction between the signal, the varactor, and the rest of the circuit, most multipliers must be tuned in "one direction" only. If some element is changed too far, tuning cannot be corrected by merely returning the adjustment to a previous setting. The varactor reactance will have altered in response to the improper setting and will not be the same as it was at the previous setting. When many tuning elements are involved, tuning may become a most delicate and precarious undertaking.

Conclusions

Although varactors introduce a number of design problems, these can be overcome with suitable insight and care. The varactor diode permits many applications which have not heretofore been practicable. Better and simpler ways of using varactors will be developed. Progress in semiconductor techniques will undoubtedly yield varactors with lower series resistance and higher breakdown voltage, thus extending the cut-off frequency and power handling capability. This can be expected to permit completely solid-state microwave systems at frequencies where they are not now practicable.

One aspect of the current interest in varactors as a source of microwave energy is the hope that they will have greater reliability. Although semi-conductor devices are, in general, forging well ahead of electron tubes in reliability, this varies from device to device, and is also a function of how they are used. After the superior reliability of varactors has been thoroughly established, they will still have to be used with other varactors in complex circuits. Reliability, like efficiency, is the product of all the individual "reliabilities" of the components in a given piece of equipment. If a multiplier uses six varactors which are each 90% reliable, the entire multiplier will have a reliability of only 59%, assuming that all the other components in the circuit—power transistors, crystals, resistors, and the like — have perfect reliability. Thus, for a time, at least, the increased complexity of varactor devices may tend to offset their still-unproven reliability advantage. Better circuits and better components are bound to come with time and experience, however, and microwave users can look forward to substantial benefits. •

BIBLIOGRAPHY

1. P. Penfield and R. P. Rafuse, *Varactor Applications*, The M.I.T. Press, Cambridge, Massachusetts; 1962.
2. F. P. Storke, "High Order Harmonic Generation with Varactor Diodes," *1962 International Solid-State Circuits Conference*, Philadelphia, Pennsylvania; February, 1962.
3. R. L. Ernst and J. K. Fitzpatrick, "Varactor Diode" Parts I and II, *Western Union Technical Review*; January and April, 1963.
4. D. O. Fairley, "Practical Design Techniques for Solid-State Microwave Generators," presented at the *National Symposium, Professional Technical Group on Microwave Theory and Techniques*, Santa Monica, California; May 21, 1963.

World Radio History

World Radio History

the Lenkurt Demodulator

VOL. 10 NO. 12 DECEMBER, 1961

New knowledge about
TRANSISTOR RELIABILITY

The transistor, born of telephone research, has been long-delayed in appearing in telephone and other multi-channel communications equipment. Despite the difficulties of supplying power across an ocean, designers of the long underseas cables turned to electron tubes in order to assure themselves that the built-in repeater amplifiers would continue to operate over a twenty-year period. Why has the transistor, once hailed as the final answer to tube failure problems, been so conspicuously absent, until now, from most communications equipment? This article discusses some of the causes of transistor failure and the techniques that are restoring hope that the transistor may soon be the most reliable of all electronic components.

An urgent United States missile development program has focused new attention on the problem of component reliability. The "Minuteman" missile, dedicated to providing incentive for continued world peace, is designed to rest unattended in buried launching tubes in widely-dispersed locations. Missiles are so extremely complicated and have so many interdependent systems that satisfactory operation in the absence of constant care and maintenance requires a level of reliability that is incredible by conventional standards.

The significance of this missile program to the communications industry lies in the radical new approach to designing a reliable system in a field which is notorious for its vulnerability to the failure of individual components. Actually, there is a remarkable similarity between the reliability requirements of

World Radio History

PHILCO LANSDALE DIV.

Figure 1. Typical junction transistor (from which upper case has been removed), shown here greatly enlarged. The square plate is only about 1/50 inch (½ mm.) on each side; the junctions which comprise the central spot and the ring are extremely small.

missiles such as "Minuteman" and large communications systems. Both are extremely complex, thus requiring close control of the performance of individual parts; both are important, since large-scale failure of either would have dire effects on general well-being. In both, it is vital to have a realistic appreciation of system dependability so that enough standby equipment can be provided to guarantee the required service. Obviously, the more reliable the system, the better the performance that can be provided at a reasonable cost.

In order to assure the astonishingly high degree of reliability required for the success of the "Minuteman" project, intensive reliability programs have been established in all the industries which supply parts for the missile. One of the most important of these programs is centered in the electronics industry, since the success of a solid-fueled missile is more dependent on electronics than almost any other single factor. As a result of this and other urgent research programs, all users of transistors benefit from the improved devices which result.

Status of Transistors

Compared to electron tubes, transistors are still a very new type of component. In the past few years, diligent research has rapidly broadened the range of applications for which transistors are suited. Transistor frequency response and power handling capability have been increased greatly, while noise generated within the device has been steadily reduced. Inherent difficulties in using transistors for certain applications are rapidly being overcome by the development of new circuit techniques (see DEMODULATOR, *September* and *October*, 1961). Of all the benefits expected of transistors, only reliability has lagged behind its potential.

This estimate of transistor reliability is relative, of course. Despite the immaturity of transistor technology compared to electron tube techniques, many types of transistors are about ten times as reliable as most tubes, and are on a par with wire-wound resistors and foil capacitors (assuming, of course, that the transistors are properly used).

Progress is quite rapid in improving transistor reliability. Only about two years ago, several studies showed that electron tubes and transistors showed about the same rate of failure in military equipment. Although progress in increasing transistor reliability is excellent, there is plenty of opportunity for improvement, since the basic transistor

World Radio History

action does not wear out, and is basically impervious to shock.

When quality control engineers speak of reliability, they refer to the *probability* that a device will continue to function within certain design limits for a given period of time. It is usually expressed as % failures per 1000 hours of service or as the Mean Time Between Failures (MTBF). In some applications such as pocket radios, performance limits may be very broad, thus allowing lower grade units to be used. In military and communications equipment, however, requirements are usually much more stringent in order to maintain suitable performance reserves. Transistors are considered to have failed once their performance characteristics drift beyond a certain specified limit. Some typical estimates of electronic component reliability are given in Table 1.

Table 1.
Typical Failure Rates

Component	Present %/1000 hr	Minuteman Objective %/1000 hr
Transistors		
Computer	0.07	0.0007
General	0.1	0.001
Power	0.4	0.004
Diodes		
Computer	0.02	0.0002
Other	0.1	0.001
Capacitors		
Solid	0.5	0.001
Foil	0.1	0.001
Glass	0.05	0.0006
Paper	0.001	0.0006
Resistors		
Carbon Composition	0.001	0.0001
Metal film & grid	0.04	0.0004
Wire Wound	0.1	0.001

Why Transistors Fail

The successful manufacture of transistors requires rather incredible precision and control at every step in the process. The basic quality of transistor action depends on the precision with which specific impurities or "dopants" are added to the silicon or germanium in making the *p* and *n* layers which form the transistor junctions. The basic raw material must be so pure that impurities are measured in parts-per-billion. If

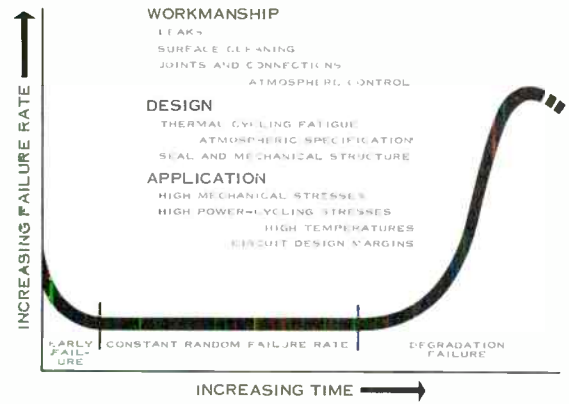


Figure 2. Generalized transistor failure rate pattern, with major influencing factors. Most sources of transistor failure are now being controlled by using automatic machinery almost exclusively for fabricating the devices.

World Radio History

these requirements are not adhered to scrupulously, performance may be badly degraded or even unattainable.

Transistors are remarkably small devices, sometimes consisting of a tiny slab of base material less than 0.02 inch on a side, and onto which are bonded or deposited tiny spots of material which become the collector and emitter. These tiny junctions, which are sometimes smaller in diameter than a human hair, are the gates through which the signal must pass and which must be capable of handling the full range of signal power. Figure 1 shows a typical transistor with the case removed.

Transistor failures are usually broken down into two basic categories: *catastrophic* failures and *degradation* failures. Catastrophic failures are those in which the device completely stops functioning, often the first time it is used or tested. Most failures of this type are caused by mechanical defects such as open circuits, short circuits, improper bonding of leads, and the like. Usually this type of defect is relatively easy to discover by inspection during assembly. A smaller number of catastrophic failures occur a little later in the life of the transistor from such causes as thermal cycling which may release contaminating material from the case, loosen improperly soldered joints, or volatilize a contaminant so that it is deposited on the junction. Many manufacturers now subject all new transistors to short, severe trial runs which are designed to eliminate the "weaklings" at the very outset.

Much more insidious are the degradation failures, since they may not appear for many thousands of hours and may even then only consist of degraded performance, rather than absolute failure. Surprisingly, the performance characteristics of almost all transistors tend to degrade with time, whether in service or not. In fact, because of the thermal gradient at the junction, some transistors appear to maintain more stable characteristics in service than during storage. Figure 2 summarizes the causes for most transistor failures.

Two basic performance characteristics are usually chosen as indicators of tran-

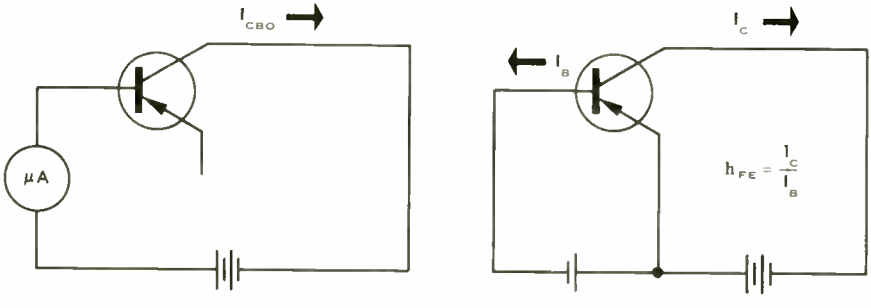
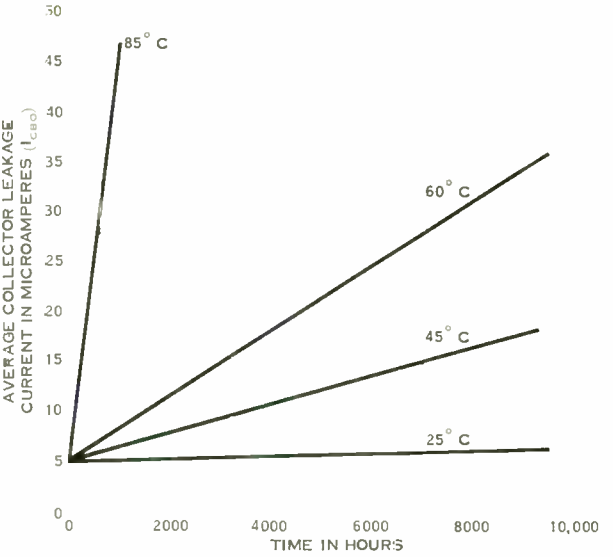


Figure 3. Typical method of measuring transistor parameters. Collector leakage current (I_{CBO}) is measured with reverse-bias, as shown in sketch at left. Direct current gain (h_{FE}) is ratio of output current to input current in common-emitter circuit.

World Radio History



BELL LABORATORIES RECORD (REF. 2)

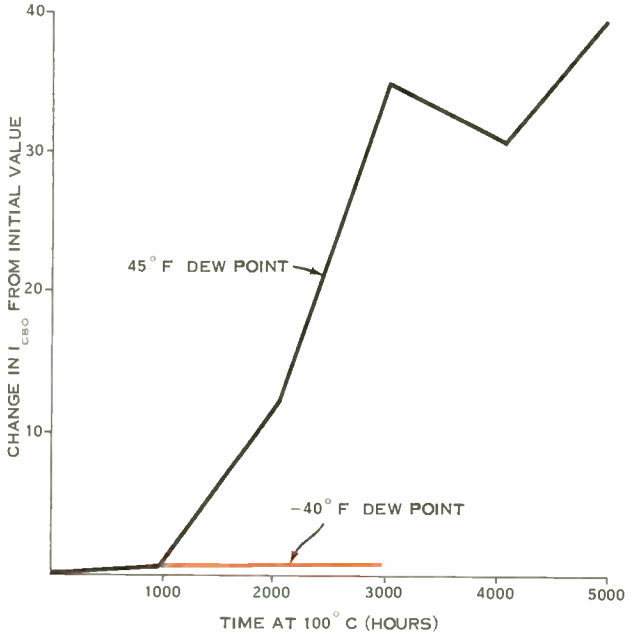
Figure 4. Effect of storage temperature on performance degradation of an experimental germanium transistor. Even at 25° storage temperature, some drift in performance is experienced.

sistor performance stability because they are the parameters most likely to change with time: one is the leakage current from collector to base when the collector is reverse-biased and the emitter is left unconnected; this leakage current is usually designated I_{CBO}. The other is the d-c current gain in the common emitter arrangement, with the emitter grounded. This is usually designated h_{FE}, and is the ratio of collector (output) current to base (input) current. These basic relationships are diagrammed in Figure 3. A low value of I_{CBO} leakage current is desirable, but in most transistors tends to increase with time, particularly when the device is stored or operated at elevated temperatures, as indicated in Figure 4. Increased leakage current is often accompanied by a corresponding reduction in h_{FE} and generally poorer performance. When these characteristics finally exceed the specification limits on which circuit designs are based, the transistor is considered to have failed, even though it may still function to some degree.

Contamination

Transistor degradation appears to result from three basic factors: the presence of fantastically tiny amounts of foreign substances or contaminants within the transistor case, moisture, and elevated temperatures. Apparently, all three work together, altering the nature of the transistor material and creating leakage paths across its surface. Since the action is progressive and becomes greater with higher temperatures and longer periods of time, it seems likely that the moisture helps ionize the contaminating substance so that it can enter into a chemical reaction with the transistor material, thus changing its nature. Some transistor degradation by moisture is reversible, however, as evidenced by the fact that intermittent operation of the transistor, or continued operation of the transistor at low power may result in

World Radio History



SYLVANIA ELECTRIC PRODUCTS

Figure 5. The effect of moisture in transistor case on leakage current. Transistors manufactured in dry-box (dewpoint −40°F) showed I_{CBO} drift indicated by red line. Black line indicates degradation experienced by identical transistors assembled and sealed in moist air (dewpoint 45°F). Both groups included moisture-absorbing dessicants.

less degradation than would occur on the shelf. Evidently, the heat generated at the transistor junction tends to drive the moisture away from the junction. However, if the transistor becomes sufficiently hot, the increased operating temperature accelerates any degrading chemical action.

Contamination of the transistor material is a very serious problem, and manufacturers go to remarkable lengths to eliminate it. Purity of materials is maintained at an extremely high level. Water used for cleaning during the manufacture of the transistor is, for instance, normally deionized so completely that it has a resistivity of 16 to 18 megohms per centimeter. Almost all manufacturing operations are conducted in so-called "white rooms" in which dust, air purity, and humidity are very closely controlled.

Moisture

Probably the most important single factor in the reliability of transistors (other than how it is used) is the amount of moisture remaining in the case after it is sealed. One way of restricting the amount of moisture and

World Radio History

contaminating substances that reach the transistor junction is to include encapsulating materials in the transistor case. These may be desiccants for taking up moisture, "getters", or molecular sieves which trap contaminating molecules of many types. Both liquid and solid types are used, frequently in combination with each other. Even these devices vary considerably in their effectiveness, according to how carefully they are prepared. It should be noted, however, that oil alone is a "barrier" material and may only delay the onset of degradation. This may actually impair reliability by giving false indication of actual quality as indicated by measurement of I_{CBO}.

Figure 5 gives the results of a test designed to show the effect of tiny amounts of moisture left within the case. The transistors represented by the black curve were encapsulated and sealed in a room in which the dewpoint was 45° F. A similar group of transistors (red curve) was encapsulated and sealed in an extremely dry atmosphere in which the dewpoint was −40° F. Identical desiccants were used in both. Since handling and processing were identical in every respect for both groups, the rather dramatic increase in I_{CBO} for the group processed in the moist room can only be accounted for by the effect of the moisture that the desiccant was unable to take up.

Although the moisture often has a direct effect on transistor degradation, it more likely serves as a vehicle for other contaminating substances, creating ions which enter into chemical reaction with the junction material. The cleaner the transistor material, the case, and the leads, the less likelihood there is that the finished transistor will exhibit severe degradation.

The remarkable sensitivity of transistors to efficient cleaning was demonstrated in another test in which one group of transistors was washed in a single tank of turbulent, ultra-pure water for one hour, and another group was tank-washed for twenty minutes and then spray-washed for five minutes. After three thousand hours storage at 100° C, the spray-washed transistors showed a reduction in I_{CBO} leakage current, while the I_{CBO} of the tank-washed devices increased after storage at 100° C, as shown in Figure 6.

Another experimental technique which was effective in delaying the onset of degradation was to add special substances to an etching bath used in the transistor manufacture, so that the contaminating substances were "snapped up" to form complex ions which had less contaminating effect. Eventually, transistor degradation reached approxi-

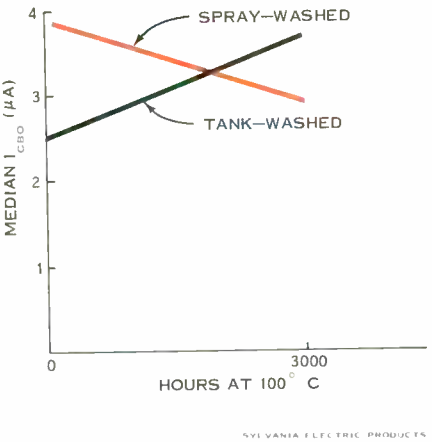


Figure 6. Even using purest water, minor differences in washing technique have important effect on ultimate performance of transistors. Spray-washed transistors were adjudged to have better stability.

World Radio History

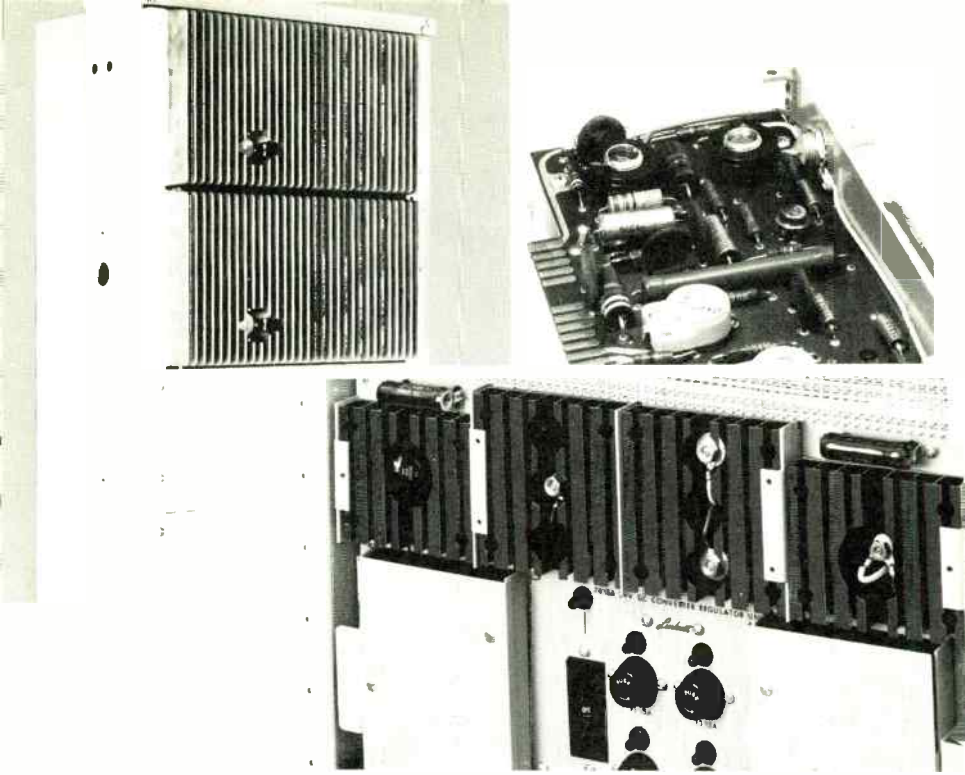


Figure 7. Upper left illustration shows massive finned heat sink used to dissipate heat from transmit klystron in transistorized 76A microwave transmitters. Radiating fins are attached to power amplifier transistors in 76A modulating amplifier (upper right). Note low-level transistor without special cooling. Special cooling fins are used for power transistors used in transistorized power regulator, shown in lower illustration.

mately the same value as in transistors made without the complexing substance, although the onset of degradation was delayed several thousand hours. This suggests that the complex ions thus formed were either unstable, or that they were slower in entering the junction area.

By using these and related techniques, transistor manufacturers have been able to produce devices which may be stored 30,000 to 100,000 hours at 100° C before the onset of performance degradation.

Reliable Equipment

Once the transistor has been sealed and shipped, the life expectancy and performance of the device is in the hands of the equipment manufacturer. Although extreme measures are taken to make each transistor as nearly perfect as possible, there is a great likelihood that this objective will not quite be reached. Accordingly, transistor performance will tend to degrade with time.

Transistor degradation will be accelerated or delayed by the manner in which the devices are used, since degra-

World Radio History

dation becomes a function of the transistor junction temperature. In general, transistor life is reduced by about half for every ten-degree increase in junction temperature. The more power handled by the transistor, the greater the temperature, and the shorter the potential life of the device. Although the transistor may not generate many calories, the junction area is usually so minute that even a small amount of heat will result in a rather large temperature rise at the junction.

Equipment designers have a number of techniques at their disposal for dissipating the heat generated by transistors. First, of course, is to limit the power handled by each transistor by designing circuits conservatively. Another important technique is to design the equipment to provide natural routes for the heat to escape from the transistor. Such means include radiating fins on the transistor, mounting the transistor on some sort of heat sink that conducts the heat away, or which increases the surface area from which the heat may radiate. In addition, transistor equipment should be designed to isolate sources of heat from the transistors themselves, in order to permit the most rapid transfer of heat away from the junction as possible.

Figure 7 illustrates several ways of dissipating heat so that transistors will operate at a lower temperature. By using fins with large areas, heat is rapidly transferred to the surrounding air. The temperature of the surrounding air thus controls the junction temperature, since the transfer of heat is directly proportional to the difference between the temperatures of the transistor junction and the air. The larger the radiating surface, the more nearly the junction can approach the temperature of the outside air.

By employing such techniques judiciously, engineers are now producing equipment which takes full advantage of the tremendous longevity available from modern transistors. Future equipment will carry this trend further, perhaps even using such techniques as thermo-electric cooling devices within the transistors themselves! •

BIBLIOGRAPHY

1. Richard G. Stranix, "Minuteman Reliability . . . Guide for Future Component Manufacturing," *Electronic Industries*: December, 1960.
2. M. C. Waltz, "Semiconductor Reliability Studies," *Bell Laboratories Record*: March, 1960.
3. Harold J. Sullivan and Lowell L. Scheiner, "Observations on Semiconductor Device Reliability," *Semiconductor Products*; April, 1961.
4. Arnold J. Borofsky, "Relationship of Germanium Alloy Transistor Reliability to Etching, Washing and Encapsulation Processing," *Paper Presented At The Conference on Reliability of Semiconductor Devices*, New York; January, 1961.
5. Elvin D. Peterson, "Calculating 'Worst-Case' Transistor Leakage Current," *Electronics*: October 6, 1961.

World Radio History

World Radio History

the Lenkurt Demodulator

VOL. II, NO. 3 MARCH, 1962

Surge Protection of Transistorized Circuits

The increased use of transistorized equipment in telecommunications has imposed new problems in protecting the equipment from voltage and current surges caused by lightning and other sources. The low power drain and small size of semiconductor circuits encourages their use in remote, exposed locations; they are also becoming much more widely used in switching centers. Unfortunately, neither type of environment is "healthy" for semiconductors, due to their sensitivity to voltage surges. This article reviews traditional methods of protecting communications equipment and discusses more recent methods of protecting transistorized equipment.

Open-wire and cable transmission equipment is subject to dangerous "overvoltages" from several sources. Perhaps the most common is induction, which can occur when a telephone or telegraph line passes through the electromagnetic field of a nearby power line. In most cases, induced currents are held within safe limits by maintaining good line balance so that induced currents in each wire of a pair cancel each other. Similarly, balance of the power transmission line helps reduce induction. However, serious induced surges can result when one wire or phase of a power line is accidentally short-circuited or grounded, thereby destroying the balance of the transmission lines. The amplitude of the voltage induced into nearby communications circuits will depend on the current of the power source and the degree of coupling.

Another possible source of overvoltage results from accidental direct contact of communications lines and power wires. Such contact usually occurs, de-

World Radio History

spite all precautions, as a result of sagging conductors or structural failures during storms.

It is lightning, however, with its terrible destructive force, that causes greater damage to transmission facilities than any other source. In areas where electrical storms are common, lightning is also the most frequent cause of overvoltage difficulties.

Foreign voltages caused by direct contact with a power transmission line or by induction, are normally rather constant in value. By contrast, a surge caused by lightning is *impulsive* in nature; both types of overvoltage may range from a few volts to many thousands of volts. The current in a lightning stroke is primarily dependent on meteorological conditions and is not greatly affected by the earth's impedance at the point of entry. Most strokes average around 12,000 amperes, although crest values over 200,000 amperes have been recorded.

A lightning stroke has a wide range of effects on a transmission facility, depending upon whether it is a direct hit or a near miss. A direct hit on an open-wire line, of course, places an unusually high voltage directly on the conductors. Conventional lead-sheathed aerial cables are quite vulnerable to a stroke, since they attract lightning from a distance that is approximately equal to their height from ground. Aerial cable can usually withstand a minor stroke, since the current flows along the cable to a ground point where it is readily dispersed. However, even a small stroke can induce potentially dangerous transients into the inner conductors of a cable.

Buried or underground cable is not completely immune to the effects of a lightning stroke, although it is much less vulnerable than aerial cable. Once the lightning discharge strikes the ground, the current fans out in all directions, and any buried metallic object,

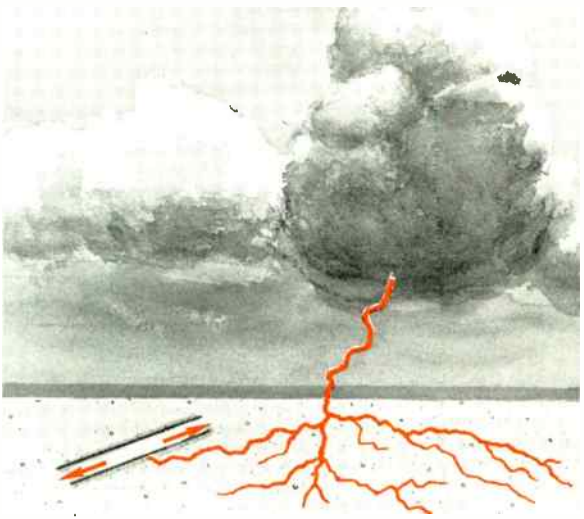


Figure 1. Tremendous voltage and current in lightning strokes can cause dangerous surges in wire and cable, even at great distances from the point of strike. Even if cable sheath acts as a shield, substantial currents may be induced to flow in the pairs within.

World Radio History

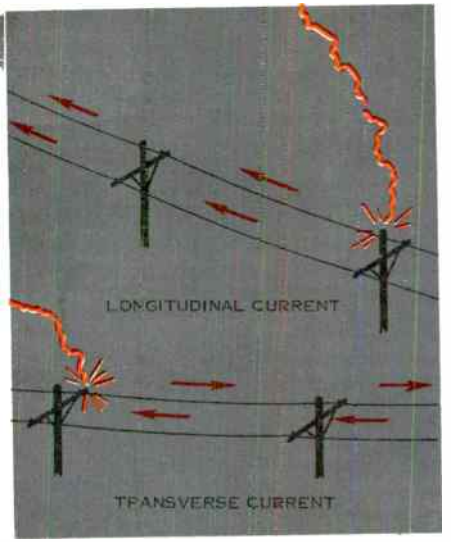


Figure 2. "Near miss" by lightning strokes may induce large equal voltages in both wires of a pair to produce "longitudinal" currents. When surge is not equal in both wires, a "transverse" current flows around the loop. Many other causes for both types also prevail.

such as a telephone cable, presents a low-impedance path to the current flow. If the lightning strikes close to the cable, the voltage drop developed between that point and the cable may break down the insulation of the intervening soil. For a 20,000-ampere stroke, the range of attraction in a clay soil is about five feet. Once soil resistance has broken down, the stroke travels in both directions along the cable sheath, attenuated more and more as the distance increases. If the earth resistance is high, the current will travel farther along the cable, since loss due to leakage to ground is relatively low.

Surges or overvoltages appear on a transmission line as so-called *longitudinal* currents or as *transverse* or "metallic" currents. Both types can be caused by direct contact, induction, or lightning.

Longitudinal currents result from equal voltages appearing on both wires of a balanced pair. Since the circuit must be completed by some path other than the conductors themselves, they usually appear as a voltage between the line and ground.

Transverse (metallic) currents result when the currents in the pair flow in opposite directions. The currents complete their path by flowing down one wire and returning by the other. Even if currents in the separate wires flow in the same direction, a net transverse current can result if one current is larger than the other. Normal voice or carrier currents are transverse.

Surge damage is not confined to cable and open-wire communications lines. Many remote microwave radio and telephone carrier installations receive their a-c operating power over aerial power distribution lines, a notorious collector of lightning strokes. The power supplies which convert the power to a useful form, are often vulnerable to voltage surges in the power line.

Whatever the nature and origin of voltage surges, communications facilities must be protected from them. It follows that any protective device or circuit must be able to cope with a very wide range of impulsive or standing voltages.

Protection Methods

Conventional practices for reducing surge damage include the use of lightning arresters at intervals on an open-wire transmission line, or a grounded parasitic line placed parallel to and above an open-wire line to divert lightning. Cable is protected by careful bonding and grounding of the sheath, and by using special types of cable

World Radio History

armor to suit varying degrees of lightning exposure. To minimize power line induction, both the communication and power lines are transposed at predetermined intervals, thus improving balance and reducing the inductive effect.

Various devices are also available to protect equipment from surge damage. Most are a variation of the air-gap or carbon-block protector, which has been in use for many years. It usually consists of two carbon electrodes separated from ground by an air gap as shown in Figure 3. Each electrode is

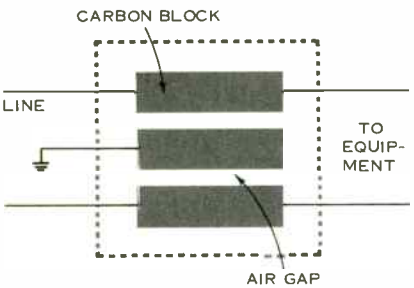


Figure 3. Simplified diagram of carbon block or air gap protector. Gap between blocks is set at time of manufacture. Slight differences between gaps may cause one to arc over before the other when equal voltages are present, thus causing a transverse surge.

connected to one conductor of the wire pair. A voltage that exceeds the protector rating will ionize the air in the gap and arc across to ground. This provides a low-impedance path away from the equipment to be protected, but which disappears once the excess voltage has been dissipated.

The arc-over ratings of carbon protectors may range from 450 volts to 3,000 volts. In the telephone industry, an air gap of about 0.003 inch has been established as a standard for central office protection. This spacing results in a peak breakdown voltage of about 500 volts. Once the arc has been established, it will continue to conduct with a much lower voltage differential. Protectors designed for less than 400 volts are not considered to be practical, since more closely spaced electrodes tend to short-circuit permanently, thus requiring excessive maintenance.

Air gap protectors can also be used to reduce surges that enter an installation on the a-c service mains. In such an application, a calibrated gap is used in series with a non-linear resistor, across the a-c line. A surge exceeding the protector rating will arc across the gap, thus connecting the non-linear resistor directly across the line or equipment to be protected. The non-linear resistor offers a low resistance to this high voltage, and the surge is equalized and absorbed on the transmission line. Although system current also flows through the conductive gap, this flow ceases once the a-c voltage waveform passes through zero, due to the high resistance of the resistor at low voltages.

Another protective device, now usually restricted to local telephone trunks is the *heat coil*. Heat coils are similar to the "slow-blow" fuses in some electronic equipment, and are used in series with each wire of a transmission pair. They protect against fairly heavy standing currents such as could be caused by induction or accidental contact. The device consists of a small coil of wire, wound around a tube. A metal pin within the tube is spring-loaded and held in place with solder of low melting point. If enough current flows through the coil to melt the solder, the pin is released and either opens the line or grounds it. Heat coils are usually combined with carbon air gap protectors

World Radio History

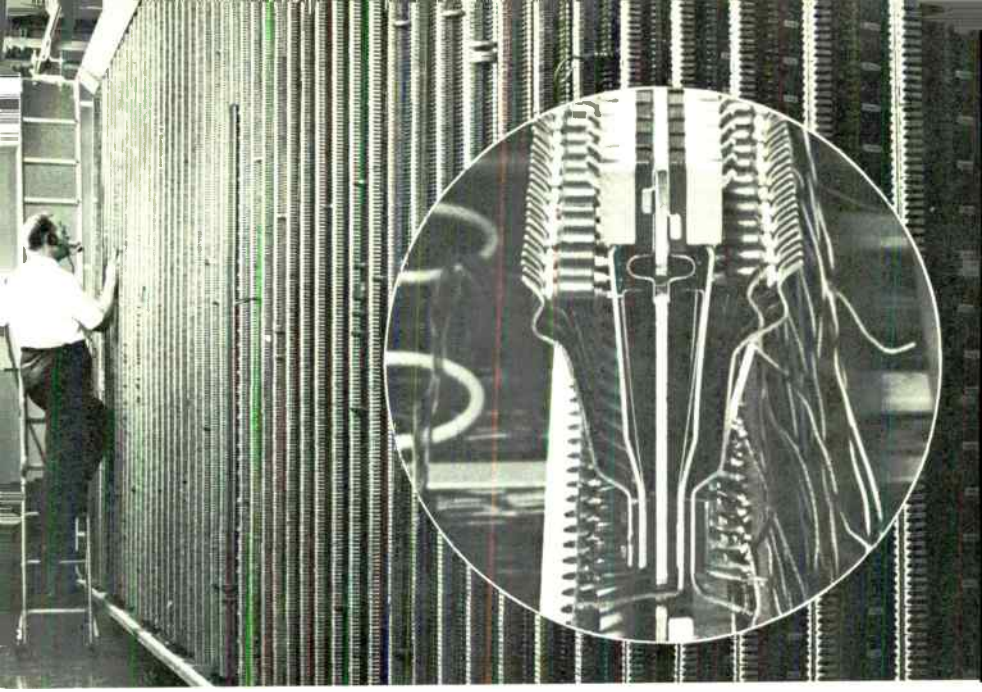


Figure 4. Great numbers of protectors are required in telephone office. Heat coils are mounted adjacent to carbon blocks. Only grounded blocks are exposed: side blocks and air gaps are concealed by porcelain insulators.

to form an integrated device such as shown in Figure 4. The heat coil is not generally used on carrier circuits because of its inductance. In addition, it cannot be conveniently used in circuits which employ repeaters powered over the line itself.

Semiconductor Vulnerability

In most present-day circuits using electron tubes, carbon block protectors are useful in protecting against excessive longitudinal voltages, but are ineffectual against voltages too low to arc across the carbon protectors. Fortunately, electron tubes and their associated circuitry have sufficient current-carrying capability and dielectric strength to be insensitive to these smaller voltages. By contrast, a relatively minor surge may destroy a transistor. One manufacturer's specification for a typical high frequency transistor quotes a breakdown rating of ½ volt across the base and emitter, and 20 volts across the collector and emitter. These values, while quite conservative, provide some indication of the limited surge tolerance of some transistors. For this reason, older protection methods are often inadequate for transistorized equipment, and an additional stage of protection is needed to reduce surges to a level that the transistors can withstand.

One simple way of protecting transistorized equipment from longitudinal surges on open wire and cable is to use a balanced transformer between the

World Radio History

equipment and the line, as shown in Figure 5. Since longitudinal voltage appears on both wires of the pair, these tend to cancel each other and are grounded through the center-tap on the transformer winding.

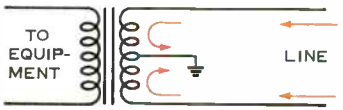


Figure 5. Line isolation transformer or "longitudinal coil" dissipates longitudinal voltages to ground through center tap of winding, but couples desired signal through to equipment. This method provides little protection against transverse surges.

In systems employing repeaters powered over the line itself, the return path for the power may be provided by the grounded center-tap of the line-isolation transformer. Under these circumstances, the power sent through the line can vary considerably or include voltage transients caused by lightning or earth currents. This can be overcome, as in Lenkurt's Type 81A exchange cable carrier system, by returning the repeater power through one of the other pairs in the cable, thus allowing the system to "float" above ground potential and avoid surges that might harm the equipment.

Transverse voltages are more difficult to overcome since they may pass through the line-isolation transformer as though they were part of the signal. Some degree of protection can be built into the circuit itself by taking advantage of the properties of the components used. For instance, line isolation transformers may be designed to have poor response to frequencies below the signal frequencies, thus sharply attenuating 60-cycle power voltages acquired by induction or contact.

Ironically, carbon block protectors themselves may be an important source of harmful transverse voltages. When a longitudinal surge builds up the potential between the two side carbons and the center ground, one side usually arcs before the other, thus unbalancing the pair and causing a sharp transverse voltage. For this reason, in certain types of installations, it may be undesirable to use carbon protectors with transistor equipment. It may be possible to omit carbon block protectors, provided some sort of alternate protection is available. In the case of multi-pair cable in which only a few pairs are used for carrier, and the rest used for voice-frequency circuits, carbon protectors on the voice-frequency circuits may be able to dissipate the voltage surge adequately for *all*

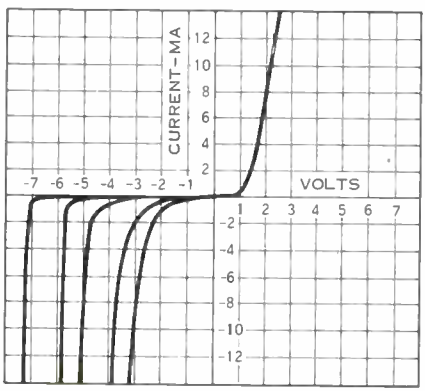
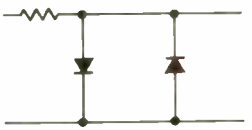
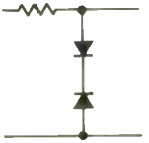


Figure 6. Typical forward and reverse conductivity characteristics of several Zener diodes. Note that breakdown is much sharper in diodes designed for greater negative voltages. Forward characteristic can also be used for low-voltage limiting.

World Radio History



FORWARD CONDUCTION LIMITING



ZENER BREAKDOWN LIMITING

Figure 7. Typical circuits which use forward or reverse conduction characteristics of Zener diodes for limiting voltages in pair of wires.

pairs, thus permitting protectors to be omitted from those pairs used for carrier transmission.

Diode Protectors

Semiconductors themselves afford one of the most effective means of protecting other semiconductor circuits. A special type of semiconductor diode called a *Zener* or *avalanche breakdown* diode is used to shunt voltages exceeding a certain critical value. The Zener diode is designed to have a very sharp reverse breakdown characteristic, as shown in Figure 6. Note that the Zener diode has extremely low conductivity in the reverse direction over a wide range of voltages. However, when the reverse voltage reaches the critical "breakdown" value, the diode's ability to hold back the current virtually disappears and the diode can pass large values of reverse current without damage (unless the diode junction temperature becomes excessive). When the voltage drops below the breakdown voltage, the diode again becomes nonconductive. Zener diodes are available with breakdown ratings ranging from about −3 to −150 volts.

It is also possible to use the forward-conduction characteristics of a semiconductor diode to protect circuits where the normal circuit voltage levels are quite low. At approximately +0.5 volt, a silicon diode undergoes a significant change in forward junction resistance. Although the transition from high to low resistance is less sharp than in the reverse direction, it is still enough to provide effective voltage limiting at very low voltages. Figure 7 shows typical voltage-limiting circuits using forward- and reverse-conduction characteristics of semiconductor diodes.

In protecting lower frequency transistor circuits, diodes are much faster than conventional mechanical or electrical protection methods. Diodes, however, have a significant limitation in high-frequency applications. A high-frequency transistor has a much smaller junction area (and consequent faster switching time) than many Zener diodes. As a result, the high speed transistor may react to a voltage transient and be damaged before the protection diode is able to "turn on" and protect it.

Typical Arrangement

A typical plan for protecting transmission equipment on cable or open-wire lines is shown in Figure 8. In this arrangement, air-gap protectors rated at 2000 to 3000 volts are used at intervals from the midsection of the cable or open-wire link, toward the end terminal or repeater location. At the pole or access point adjacent to the office, a 700-volt protector is used. The terminal equipment itself is protected by conventional 500-volt air-gap protectors. Current-limiting resistors are often placed

World Radio History

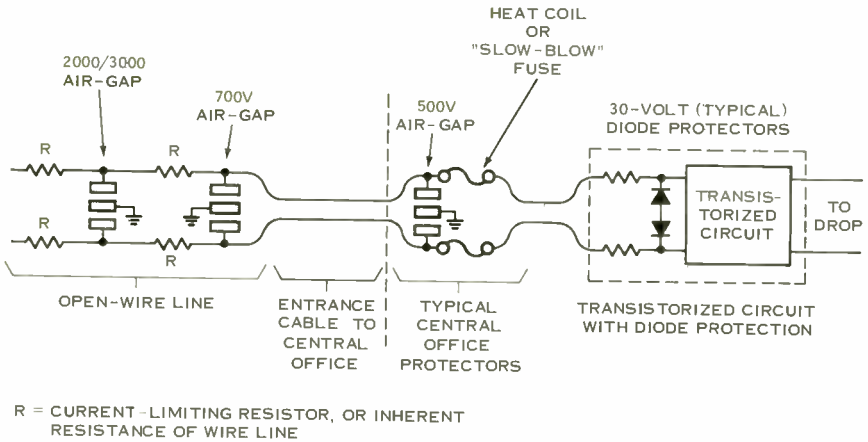


Figure 8. Typical arrangement for reducing surges in wire or cable to safe value in gradual steps. Fuses or heat coils are not normally used if repeaters are powered through cable or if high frequencies are used. In some applications, air gap protectors are dispensed with on selected pairs in order to avoid transverse surges.

between the source of interference and the individual protectors to limit surge currents. The inherent resistance of the wire or cable pair often provides sufficient limiting. In this way, most heavy surges are sharply restricted before reaching the equipment, since lightning strokes or other sources of foreign voltage will usually occur at some distance from a protector. However, nearby or direct hits will usually damage protective devices, and an occasional destructive hit from time to time must be expected.

Another semiconductor protection circuit employs a silicon controlled rectifier ("SCR") to shunt excessive voltages. The SCR has the advantage that it can be made to conduct at any predetermined voltage, and is therefore the equivalent of an adjustable air gap. It has, however, the disadvantage of requiring a relatively complicated control circuit.

In normal operation, the SCR will not conduct in the forward direction unless triggered by a small positive pulse on its control lead. Once triggered, it will continue to conduct until the voltage across its cathode and anode has disappeared, or until a reverse voltage is applied briefly. When conducting, it has a very high current capacity and very low internal resistance, thus making it particularly useful as a protector. (For a more detailed account of the SCR, see DEMODULATOR, *January*, 1961).

A typical circuit designed for low voltage d-c lines is shown in Figure 9. In this circuit, capacitor C1 permits impulsive line surges to pass, but isolates the protector from any standing voltages on the line. The SCR is connected across the line to be protected, but passes no current until triggered. The voltage at which the SCR will be triggered is determined by the

World Radio History

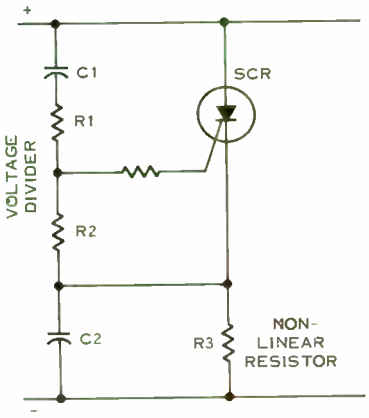


Figure 9. Surge protector using a silicon controlled rectifier protects d-c line against voltage transients. Values of R1 and R2 determine at what voltage SCR "turns on." Capacitor C2 and non-linear resistor R3 provide means for turning SCR off after surge has ended.

values assigned to the voltage divider, R1 and R2. Capacitor C2 and non-linear resistor R3 provide the mechanism for turning the SCR off after the voltage surge has passed.

Once the rectifier is in full conduction, shunting the excess current to ground, C2 charges to a potential higher than the voltage normally standing on the line. As the overvoltage dissipates, non-linear resistor R3 (which has a low resistance to high voltages) increases its resistance. The charge on C2, now higher than the voltage at the cathode of the SCR, appears across R3, and effectively places a reverse voltage across the SCR, thus turning it "off."

Conclusions

Logical extensions of the protection methods described should prove to be adequate for most communications applications for some time to come. The trend in new transistor design is toward higher operating voltages, therefore, transistors used in the normal frequency and power environment of low-frequency communications may prove more rugged and surge-resistant than ever before. However, new, ultra-high-frequency transistors present another problem. Transistors capable of operating at frequencies of 70 mc and above, are still so sensitive that even minute surges can cause damage, due to the small junction area. The future, therefore, demands continuing research toward improving protection techniques, and in increasing the surge tolerance of high-frequency transistors. •

BIBLIOGRAPHY

1. M. O. Williams, "Lightning Protection," *Automatic Telephone and Electrical Journal*, January, 1956.
2. J. W. Phelps, "Electrical Protection for Transistorized Equipment", *Bell Laboratories Record;* July, 1958.
3. D. W. Bodle and J. B. Hays, Jr., "Electrical Protection for Transistorized Equipment", *Electrical Engineering;* August, 1959.
4. R. W. Blackmore and B. A. Pickering, "Lightning Protection for Transistor Repeaters", *Automatic Telephone and Electrical Journal;* October, 1960.
5. P. Chowdhuri and L. J. Goldberg, "Surge Protection of Low-voltage DC Circuits", *Electrical Engineering;* May, 1961.

World Radio History

World Radio History

the Lenkurt® Demodulator

VOL. 13 NO. 4 APRIL, 1964

MICROELECTRONICS

The physical size of electronics equipment was reduced drastically and dramatically by the invention of the transistor a number of years ago. Now another size-reducing revolution — microcircuitry — is producing a similar impact on the industry. Like the transistor, microelectronics promises more than mere size reduction. It also offers the potential for increased reliability, lower manufacturing costs, and lower power consumption. But microcircuitry is still in its infancy, and major problems remain to be solved. This article presents a survey of microelectronics and considers its impact on the future of the telecommunications industry.

Powerful factors are accelerating the improvement of electronic packaging. One of the strongest influences, of course, is the aerospace industry. This gigantic industry uses some of the most complex electronic devices ever developed. Furthermore, reliability, size, and weight are of the utmost importance. Every extra pound of load placed in a rocket adds many extra pounds to the gross weight, in such things as additional structural material and more fuel required for lift off — and the failure of a single component worth only a few cents may mean that millions of dollars and months of effort have been wasted. The field of aircraft electronics is only slightly less demanding than the space industry in its size and weight requirements.

Some of the same factors that are so vitally important to the aerospace industry are also the ones that have traditionally concerned telephone companies and other operators of communications equipment: reliability, cost, maintainability. Size also is important to the earth-bound communications industry, although often it is overlooked or outweighed by other factors. The actual cost of a square foot of floor space in a telephone office building, for example, would astonish many people. In addition to the original cost of the building, several other things such as light, heat, insurance, and maintenance for a period of perhaps thirty or more years must be considered.

Thus, it is evident that the whole electronics industry benefits by better and smaller circuit packages. Initial efforts in re-packaging were directed primarily toward improving vacuum tubes. Then along came the transistor, which virtu-

World Radio History

ally revolutionized the field. And now, barely fifteen years after the transistor, microelectronics may make "conventional" concepts of electronics obsolete.

Before discussing the advantages and the disadvantages of the various approaches to microelectronics, it is necessary to clarify and define some terminology to provide a common meeting ground. Over the past year some terminology has come to be quite generally accepted. Agreement is by no means unanimous, but the terms used here are those enjoying fairly general usage.

The field of microcircuitry is in its infancy, so naturally the terminology changes almost from day to day. Even the term "microelectronics" does not have a precise meaning. It is a loosely defined term which covers several methods of achieving smaller circuit packages than have heretofore been possible. These approaches to microelectronics range from the use of conventional components in a very high-density packaging arrangement to actual integrated circuitry where discrete components do not exist, and the circuit can be identified only by the function it performs.

High-Density Packaging

As might be expected, the earlier approaches to microelectronics have been extensions of conventional packaging techniques. One method is to take a conventional circuit package and simply shrink its size by reducing both the size of the discrete components and the space between them. This approach is simple in concept — the problems are primarily mechanical.

Component manufacturers are constantly striving to reduce the size of their discrete components; but even so, the space required for mounting and for the interconnecting wiring eventually approaches a certain irreducible minimum. Thus, there are limits to how far this approach can go toward miniaturization. Efforts so far have indicated that perhaps a two-to-one or three-to-one reduction in size is possible.

Furthermore, this approach contributes nothing to reliability. There are just as many components as there ever were, and should one fail the problems of replacing it are complicated by the lack of working space. Nor does the high-density method offer much reduction in

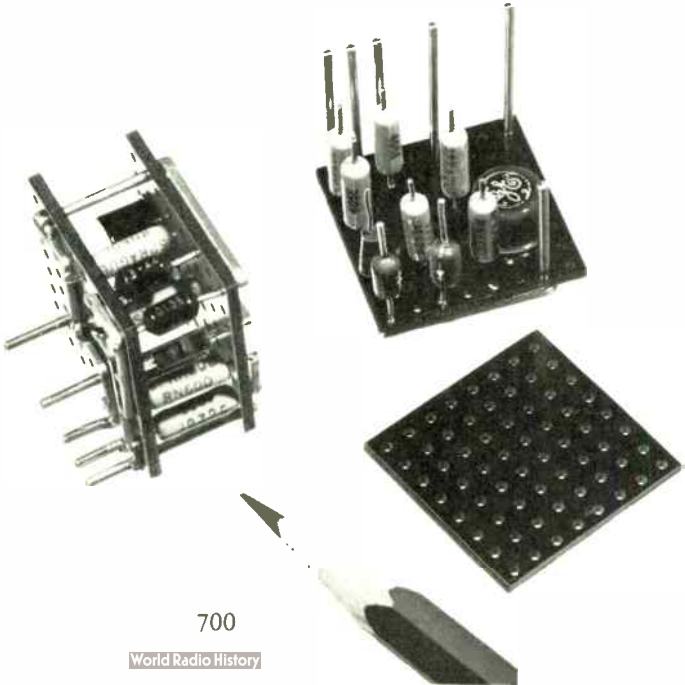


Figure 1. "Cordwood" construction is one method of high-density packaging using discrete components. It reduces size but does not improve reliability or maintainability.

World Radio History

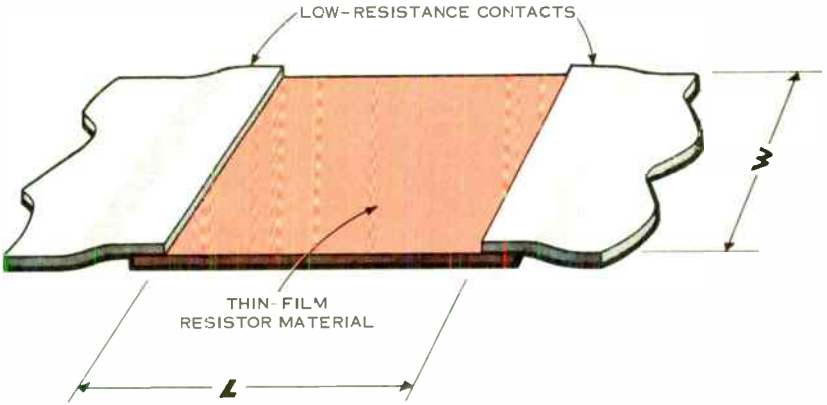


Figure 2. Since the resistance of a thin-film resistor depends only on shape, not on size, such resistors can be considered in terms of sheet resistance, measured in ohms per square.

power consumption. This in itself is a problem because essentially the same amount of power is dissipated in perhaps one-half or one-third of the space of conventional packaging. Heat dissipation may therefore become the limiting factor on this type of packaging.

The next logical step in miniaturization is to form a complete circuit, consisting of several elements, on or in a single block of material to provide an *integrated circuit*.

Thin-film Techniques

One method which shows considerable promise in integrated-circuit design is that of thin-film deposition. In this technique circuits are made by depositing passive elements such as resistors and capacitors and their associated wiring on inert or passive "substrates." These substrates, usually made of glass or ceramic, simply provide a foundation for the circuitry. The thin-film method can be viewed as merely another way of making and interconnecting components which are electrically similar to conventional components. One of the chief advantages of this technique is the close control of tolerances. Since the thickness of the film can be controlled quite precisely, a good portion of the manufacturing problem is two dimensional. Five-percent resistor tolerances are achieved routinely, and one percent or better can be achieved. Capacitor tolerances are somewhat more difficult, with 15 to 20 percent being common values.

As an example of the thin-film technique, Figure 2 indicates the method of forming a resistor. When two opposite sides of a square section of the thin-film material are connected to low-resistance contacts, the value of the resistor depends only on its shape, not on its size. A long narrow film provides more resistance than a short wide one in the same way that a long thin wire has a higher resistance. Thus, a resistor consisting of 10 squares in a row has 10 times the resistance of a single square — if the contacts are at the ends. If, however, the contacts are along the sides, the resistance is only 1/10 that of a single square.

While it is not particularly difficult to deposit resistors, capacitors, and wiring by thin-film techniques, inductors and transformers present an entirely different problem. The thin-film method simply does not lend itself well to dupli-

World Radio History

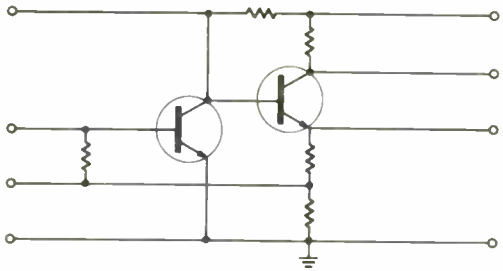


Figure 3. One solid-state circuit may perform the functions of many discrete components. Small disk at left provides performance equal to that of the two-transistor amplifier shown in the inset.

cating the effect of the traditional inductor. However, considerable research effort is being expended in attempts to find a satisfactory method of depositing inductors as spirals of thin-film material. These efforts have met with some success, but much more work remains to be done in this field. For the communications industry this is perhaps one of the most serious drawbacks to the use of thin-film techniques. Inductors are a vital part of the filters used in modern telecommunications systems. Furthermore, in terms of physical size, inductors often form the largest part of the electronic equipment. Hence, this is the area of the largest potential savings in size and weight.

Another major drawback to the use of thin-film techniques is the difficulty encountered in attempting to deposit active devices such as transistors and diodes. The usual method, at present, is to add conventional active devices to a completed thin-film circuit. However, progress is being made in this area. A field-effect device called a thin-film transistor has been developed. In this device, current flows through a channel in a semiconductor between two electrodes called the "source" and the "drain." The voltage applied to an insulated gate controls the current flow through this channel. This is one of several approaches which show satisfactory performance in the laboratory. Quantity production, however, presents a different situation because of the mechanical fabrication problems involved. In spite of these problems, however, the thin-film technique does show considerable promise.

Solid-State Circuits

Another form of integrated circuit is the "solid-state circuit," which refers to a fabrication method in which a single semi-conductor block contains all of the circuit elements: transistors, diodes, resistors, capacitors, etc. In this technique, different portions of the semiconductor block represent various circuit elements. In other words, all the elements are constructed from the same block and are physically inseparable.

For example, each of the single semiconductor chips shown in Figure 3 performs the same functions as a two-stage transistor amplifier. Input and output parameters can be measured, gain can be calculated, and even a feedback loop can be connected, but the transistors or other components cannot be identified individually. The circuit must be considered as a whole, identifiable only by its function.

World Radio History

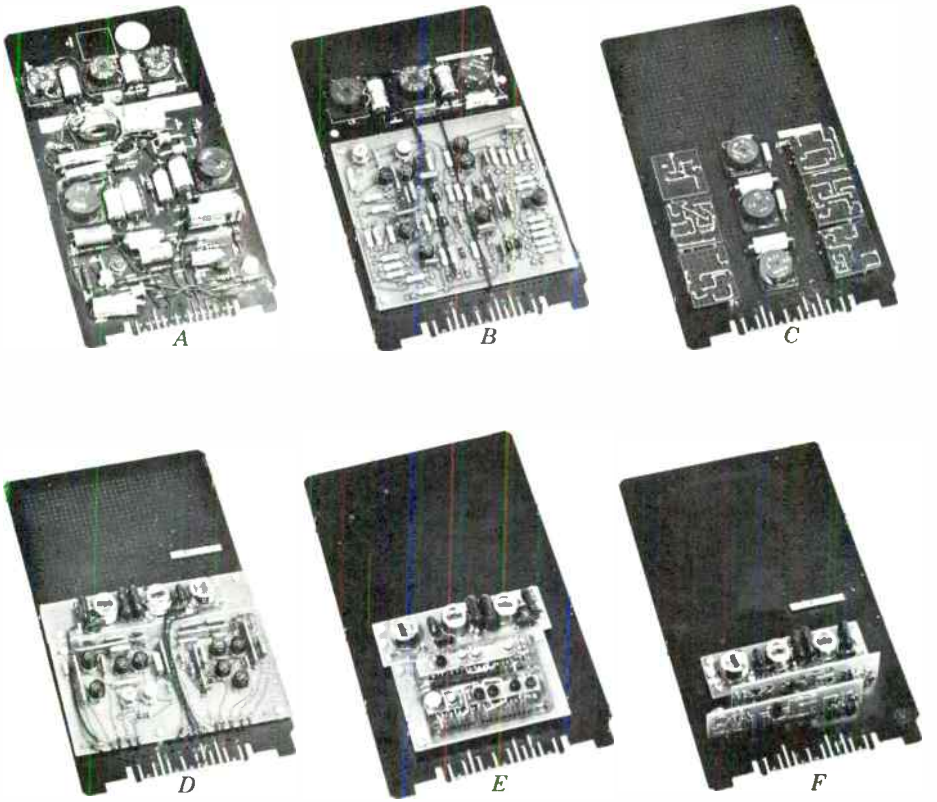


Figure 4. Several approaches to miniaturization are illustrated by the same circuit constructed by six different methods: (A) conventional components and wiring, (B) miniature components, (C) cordwood, (D) thin film-discrete component hybrid, (E) pellet pack, (F) Swiss cheese. Production of such circuits is primarily an economic problem, since (F), for example, costs several times as much as (A).

While solid-state circuits provide perhaps the most dramatic example of space-saving in the microelectronic field, they do have their disadvantages. One of the chief disadvantages is the lack of close control over component tolerances. While tolerances of 5% and even 1% are quite readily attainable in thin films, solid-state circuits typically have tolerances of 20%.

Another disadvantage of solid-state circuits is their high susceptibility to parasitic capacitance between the components. This parasitic capacitance may amount to several picofarads, whereas a comparable thin-film circuit would have virtually none. Because of this capacitance, considerable care is necessary in designing solid-state circuitry.

Hybrids

Often two or more approaches to microelectronics can be combined in a so-called "hybrid" circuit. As an example, such a hybrid circuit might be constructed by using individual active components in combination with thin-film passive components and wiring. A good example of this approach occurs in Lenkurt's recently introduced four-

World Radio History

wire terminating set. This unit includes strappable loss pads which require a total of 40 one-percent resistors. By using thin-film resistors deposited on a substrate and encapsulated in ceramic material, reliability was improved, size was reduced, and component tolerances were easily met. The rest of the four-wire terminating unit consists of conventional circuitry attached to the thin-film pads by means of several leads projecting through the ceramic material.

Another hybrid design is to combine the thin-film and solid-state circuit approaches. Thus, the components which are better formed by semiconductor technologies can be "built in" to the substrate. The substrate is then either insulated or rendered passive, and other components are deposited on its surface by thin-film techniques. Thus, the active components are in the block, the passive components are on the surface, and the whole circuit is fabricated by the most effective techniques.

The so-called "pellet" method offers a technique which is easily adaptable to automation. Each circuit element is manufactured in a standard component size. These uniform-size pellets are then placed in rows built around a structure of interconnecting wires. The wires are cut anywhere an interconnection is not desired. Figure 4(E) shows an experimental circuit which uses conventional transistors connected to pellet-type passive components to form a discrete-component circuit. However, active components and integrated circuits can also be formed as pellets, permitting a hybrid arrangement.

Another manufacturing technique using pellet components is the "Swiss cheese" approach. In this method, the pellets are set in holes drilled in a circuit board as shown in Figure 5. Interconnections are then made in much the same manner as in conventional circuitry. The chief disadvantage to either pellet approach is that the cost is several times that of conventional circuitry.

Pros and Cons

The goal of all these approaches to microelectronics can be stated quite simply. It is to improve reliability and increase maintainability while reducing cost, size and weight. (Of course there may also be secondary benefits, such as reduced power consumption.) Therefore, any particular miniaturization technique must be judged by how well it meets these established goals. For example, high-density packaging of con-

Figure 5. "Swiss cheese" approach uses discrete components in high-density packaging. Unlike some other techniques, such as "cordwood," it improves reliability as well as reducing size.

World Radio History

ventional components does reduce size somewhat and it may reduce weight. But it does not increase reliability, because there are as many components and as many connections as there were before. Furthermore, maintenance and heat dissipation are both complicated by the reduced space. Therefore, this "interim" approach to miniaturization could have applications where there is an immediate need for small size — if the requirement is urgent enough to warrant sacrificing the other qualities.

When a way is found to eliminate inductors from circuits, the thin-film approach will show even greater promise. One possibility is the use of active filters composed of capacitors, resistors, and transistors. This would eliminate or drastically reduce the need for the space-consuming inductors required by conventional passive filter networks.

Many people feel that integrated circuits will provide the "ultimate" in microelectronics. Certainly they provide one of the most dramatic size reductions. Furthermore, because there are no interconnections and no separable circuit elements, the potential for reliability approaches that of a single component. But integrated circuits do have disadvantages. The difficulty in achieving close tolerances in solid-state circuits is one of these. Also, the limited power-handling capacity of integrated circuits is much more restrictive than that of other types of microcircuits.

The Future of Microelectronics

The microelectronics industry has grown tremendously since its start only about three years ago — and its rate of growth is increasing. Conservative estimates indicate annual sales of perhaps 500 million dollars by 1970.

But such figures can be deceptive without interpretation. For example, by far the largest segment of the microelectronics market is in digital circuitry. Computers, of course, are in the forefront because they probably have the most to gain from size reduction. Microelectronic logic circuits are much easier to build than are linear circuits, and often these logic circuits are used in greater quantity, which helps to offset the high tooling cost for each circuit. Furthermore, the reductions in power consumption are much greater for logic circuitry than for linear circuits.

Thus, analog applications are lagging behind digital applications both in technology and in cost. These two factors, of course, are inextricably intertwined. As the solutions to technical problems are found, the cost will come down, allowing microelectronics to compete with conventional components in more and more applications.

In addition to the economic and technical problems which must be overcome, new equipment must be compatible with existing conventional equipment. Microcircuitry generally requires low voltages and low power. This may seriously reduce the advantages of microelectronics by requiring conventional components for interfaces with existing equipment.

Thus, although the trend toward microcircuitry is clear, it is equally clear that the telecommunications industry will not be "revolutionized." The change-over will take years, gradually coming about as a dynamic and growing industry finds ways to adapt the latest technological advances to its needs •

BIBLIOGRAPHY

1. Michael Wolff, "Equipment Makers Push Thin-Film Microcircuits," *Electronics;* February 7, 1964.
2. C. D. Simmons, "The Design of Thin-Film Circuits," *Semiconductor Products and Solid-State Technology;* March, 1964.
3. D. E. McElroy, "New Developments in Microcircuit Packaging," *Electronic Industries;* April, 1964.
4. P. J. Klass, "Panel Appraises Outlook for Microcircuits," *Aviation Week and Space Technology;* April 6, 1964.

World Radio History

World Radio History

the Lenkurt Demodulator

VOL. 10 NO. 1 JANUARY, 1961

Power Conversion

Transistors and related semi-conductor components have brought revolutionary changes to the design of power supplies, just as they have to more "glamorous" types of equipment. Semiconductor or "solid-state" components are noted for having far greater reliability than electron tubes, which they are gradually displacing. Nowhere is this additional reliability more vital than in power supplies. Some of the recent techniques used to enhance the performance and reliability of modern power supplies are described in this article.

In the sometimes exotic field of communications, power supplies rarely receive credit as being among the most important elements in a system. So humdrum and routine the subject of power supplies may seem, that new contributions toward reliability made by power supply engineering may go unnoticed.

No matter how cleverly engineered the transmission equipment, or how ingenious the modulation plan, the system goes "off the air" if the power supply fails for any reason. The source of power and the means of converting it to a useful form comprise the vital heart of any communications system.

Obviously, the prime requisite must be *reliability*: the power supply must be engineered so that there is hardly *any* chance whatsoever that it can fail. To achieve this end, components and circuit techniques are employed which, of themselves, have little or no chance of failure. For instance, communications power supplies should avoid the use of electron tubes. Instead, voltage regu-

World Radio History

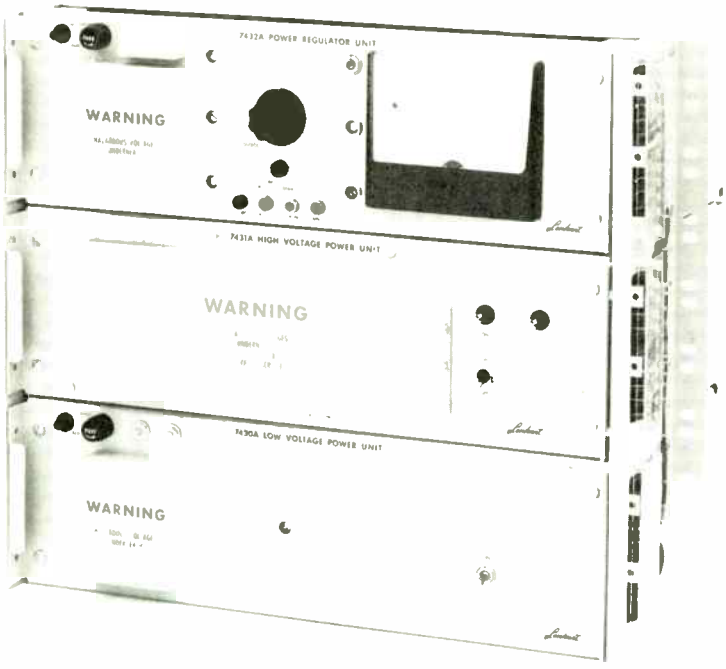


Figure 1. Reliability is the most important consideration in designing communications power supplies, followed by stability, regulation, and freedom from hum and ripple. Microwave power supply shown employs semiconductor rectifiers, magnetic amplifiers for voltage regulation. Plug-in construction permits each unit to be replaced instantly, provides superior access to internal components.

lation may be achieved with magnetic amplifiers, rugged, passive devices which fail only by burning out a winding.

A second important quality of power supplies is that they must be efficient. As always in communications, economics must enter the picture. A high degree of reliability is not difficult to obtain if bulkiness, weight, and cost are of no concern. Under these conditions, savings gained in reliability might be more than offset by high operating costs or excessive plant facilities. Since power supplies may handle and modify *all* the power used in the equipment, relatively small differences in power supply efficiency may make a large difference in the total power consumed by the system. The wasted energy appears as heat which may be costly to control or dissipate in large installations.

Last, but hardly least, the power supply must be well-disciplined, delivering the exact voltages and currents demanded by the equipment—not approximations or variations of these needs. Transmission equipment may be "finicky" or relatively sensitive to such external variables as temperature and

World Radio History

operating voltages. Some equipment might exhibit frequency drift or vary its signal power as a result of such factors. Accordingly, the power supply must not magnify this difficulty by adding to it with its own sensitivity to ambient conditions.

Transmission equipment has enough difficulty combatting unavoidable noise and interference in the transmission path. If the power supply introduces ripple, hum, or noise, it is like sabotage. Fortunately, most hum and ripple are easily controlled by various conventional filtering techniques, such as described in a previous issue (DEMODULATOR, *April*, 1958).

Converters and Inverters

The most important recent changes in power supply design have occurred in converters and inverters which have a d-c input. Direct-current conversion has always presented an engineering problem, the solution of which has usually been far short of ideal. This problem has been greatly increased with the rapid growth of electronics. Most electronic circuits must operate from d-c power. In many applications, the most logical source of power provides only dc, but at the wrong voltage for use by the equipment. To cope with this problem, many types of converters and inverters have been developed.

The terms converter and inverter are often used in different and inconsistent ways. In general, a *converter* is a device or circuit for changing d-c power to a different voltage, or a-c power to a different frequency—the latter device being mostly used in the power transmission industry. A converter does not change ac to dc or dc to ac; thus, a dc-dc converter might convert 24-volt d-c power to 500 volts dc.

An *inverter* does change one type of power (ac or dc) to the other. Thus, an inverter is a device for changing direct current to alternating current, regardless of the voltage involved. By convention, the term "inverter" is not usually applied to devices which change alternating current to direct current. Such units are usually called "power supplies" or "rectifiers" in the United States. The term "inverter" normally is applied to any device used to produce a-c power from a direct-current source.

The simplest and most reliable device for converting a low voltage to a higher voltage is the transformer. Unfortunately, however, a transformer will not function unless the applied voltage is constantly or periodically changing. This requirement is satisfied by alternating current, of course, but not by direct current of a constant voltage.

In order to use a transformer with direct current, some means must be found for varying it periodically. One way of accomplishing this is to use some sort of interrupter or "chopper" to break the smooth direct current into a series of intermittent pulses. When one of these pulses is applied to a winding of a transformer, the flow of current causes a magnetic field to build up in the transformer core. When the pulse ends, this field collapses. The changing magnetic flux causes a current to be induced in the various windings of the transformer—flowing in one direction during the flux build-up, then reversing direction when the field collapses. The secondary windings in the transformer thus yield an alternating current, the voltage of which is determined by the turns ratio of the primary and secondary windings.

One way of interrupting the direct current is to use an electro-mechanical vibrator. The contacts are so arranged that the energized magnet opens the contacts and de-energizes itself. The contacts again close, energizing the mag-

World Radio History

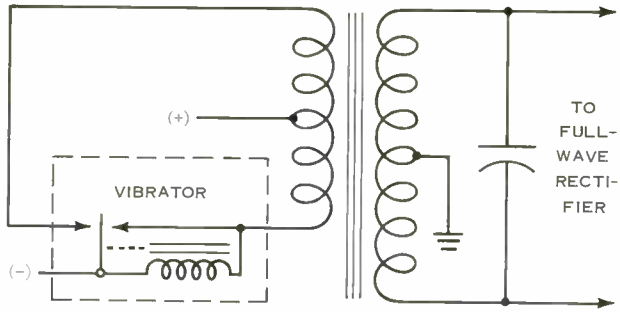


Figure 2. Typical vibrator inverter circuit, simplified. Vibrator circuits exhibit poor regulation, low reliability.

net, which again opens the circuit. Figure 2 shows a diagram of such an arrangement. Until quite recently, this method of power inversion was very widely used in commercial automobile radios because of its low cost.

An important objection to vibrator inverters is that they are relatively inefficient and highly unreliable. In addition, they exhibit poor voltage regulation. Since the interrupted dc is applied directly to the primary winding of the power transformer—a highly inductive load—considerable arcing occurs at the vibrator contacts. The arcing, in turn, generates a particularly troublesome type of electrical radio noise which is difficult to filter from the output power. In addition, arcing rapidly erodes the vibrator contacts, and builds up oxides and other impurities on the contact surface, increasing its resistance. As the surface of the contact becomes rougher, its effective area may be reduced (because electrical contact may occur only at the high points of the surface). Eventually, contact surface may be so reduced that the contacts weld together due to the high current density through the limited area of conduction. If this occurs, the contacts stick and may cause the electromagnetic coil to burn out or cause other damage in the power supply.

Another method of d-c conversion that has been widely used is the *dynamotor*. In principle, a dynamotor consist of a d-c motor driving a d-c generator to produce the desired voltage. In practice, both motor and generator share a common field winding, powered from the input power, and a common armature. Input and output circuits have separate armature windings and commutators, however. In fact, each different output voltage obtained from the dynamotor requires its own armature winding and commutator.

Although the dynamotor is simple, it is heavy and requires considerable maintenance. Bearings, commutators, and brushes suffer continuous wear and require inevitable repair. Like the vibrator, commutator arcing introduces noise which may be difficult to eliminate.

Still another way of achieving dc-dc conversion is the use of an RF oscillator. In a resonant circuit, such as the tank circuit of an RF oscillator, voltage and current are out of phase; where current is low, voltage is high. If there is little loss in the resonant circuit, very high a-c voltages may be obtained by tapping the resonant circuit at a high-voltage point. After rectifying and filtering the high voltage, it may be used for powering portable radiation detectors, providing accelerator voltage for television pic-

World Radio History

ture tubes, or in other applications where high voltage and little current are required.

Solid-State Inverters

With the invention of transistors, a new and better type of power inverter was made possible. The transistor proved to be an excellent switch, had no problems of mechanical wear, operated very nicely on low voltages (unlike the electron tube), and was quite efficient. Accordingly, many types of transistorized power inverters have been developed. Two basic types predominate: the *self-oscillating* push-pull inverter, and the *driven* push-pull inverter. Both are usually symmetrical. Although many single-transistor and other unbalanced circuits have been developed, these are not generally as efficient as the symmetrical arrangements.

In the driven inverter, two transistors arranged in a push-pull circuit are alternately turned on and off by an external control circuit. The control circuit may itself be controlled externally, or it may be independent. If tuning fork or crystal oscillators are used to provide the control, a very high degree of frequency stability may be obtained regardless of load variations. For this reason, driven inverters are widely used in missile and aircraft applications where variations in the a-c power frequency may affect the accuracy of gyroscopes powered by the alternating current.

Self-oscillating push-pull inverters are more commonly used in communications equipment than the driven inverters because they are simpler, and therefore more reliable. The typical inverter consists of two or more power transistors (or controlled rectifiers) connected in a symmetrical arrangement such as shown in Figure 4. As in the case of a vibrator inverter, the transistors act only as switches to turn the d-c power on and off. Unlike the vibrator, transistors have no mechanical wear, and

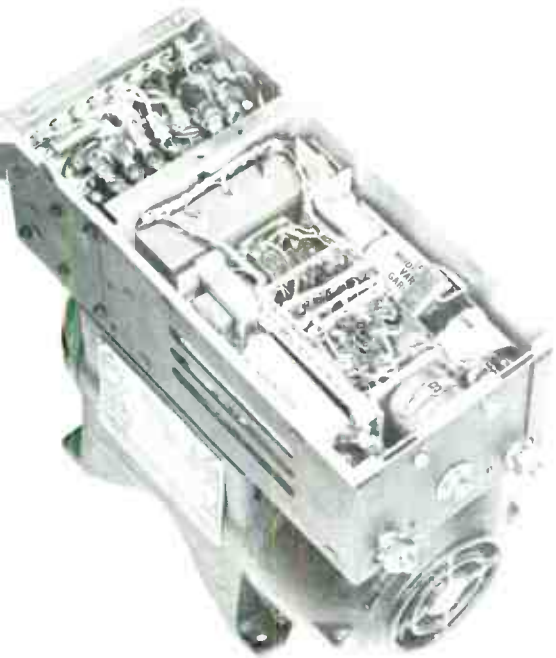


Figure 3. Typical rotary inverter such as used in some aircraft and missile applications. Cover has been removed to show circuitry required to stabilize output voltage and frequency. Newer solid state inverters eliminate mechanical problems inherent in rotating machinery, permit quicker response to regulating signals.

World Radio History

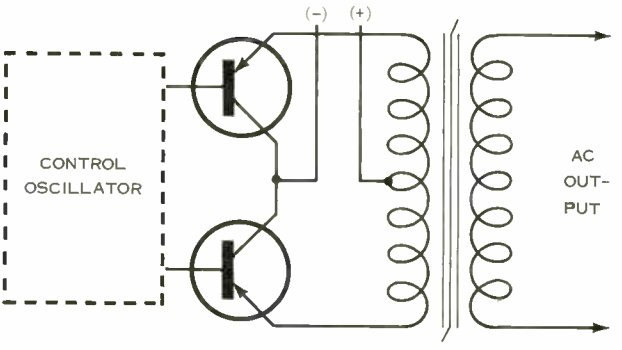


Figure 4. Simplified diagram of transistor inverter using external driving or control circuit.

are obviously free from such mechanical problems as contact arcing. They are fast, efficient, and have a life expectancy virtually unlimited, unless abused.

In the circuit shown in Figure 6, power is applied to the center tap of the transformer primary winding. From there it flows through the two transistors, Q1 and Q2, back to the power source. Because the two halves of the circuit cannot be *exactly* alike, one half of the primary winding will carry somewhat more current than the other. This causes a small voltage to be induced across the feedback winding. The polarity of this voltage is such as to bias the less conductive transistor Q1, for instance to conduct even less, and Q2 to conduct more. This biasing effect is self-amplifying, with the result that Q1 is rapidly cut off and Q2 conducts the maximum current permitted by the circuit impedance.

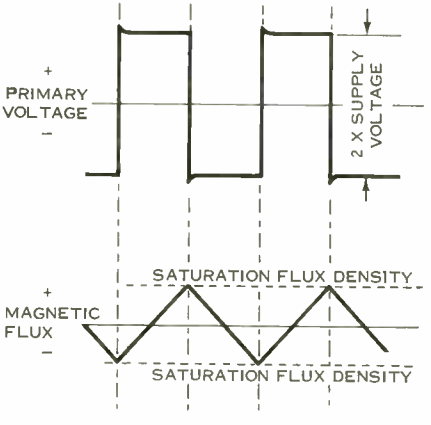


Figure 5. Comparison of voltage across primary winding of inverter transformer, and the change in flux through the saturable core.

Since the transistor bias is obtained by inductive coupling in the transformer, it can be maintained only while the magnetic flux is building up in the transformer core. When the core reaches magnetic saturation, the bias voltage on the two transistors disappears. Transistor Q1 begins to conduct again moderately, while Q2 has its conductivity reduced. This change causes the established flux to diminish. The changing flux again induces a voltage across the feed-back winding, but of the opposite polarity to that before. Now Q2 is rapidly cut off and the Q1 conducts heavily, thus building up the magnetic field in the opposite direction, and this alternating action continues automatically. As a result of this regular "flip-flop" action, a square-wave alternating current is induced in the secondary winding.

Note that the frequency of oscillation is determined by how long it takes to

World Radio History

saturate the transformer core. This is largely a function of the applied voltage, transformer characteristics, and the load across the secondary. The lower the input voltage or the larger the load on the inverter, the lower the frequency of oscillation.

Since the inverter output is usually rectified and filtered in communications applications, output frequency is not very critical.

The Controlled Rectifier

Within the past two years, a new class of semiconductor component has appeared which is quite superior to transistors for power switching and inversion. The new device, called a *controlled rectifier* or *silicon controlled rectifier* (after the principal substance from which it is made), is capable of handling tremendous currents for its size. Like a thyratron, the silicon controlled rectifier, or SCR, is a rectifier which, when "turned on", conducts freely in *one* direction, but will not conduct in *either* direction unless triggered by a control "gate." (SCR's will conduct without a gate signal if their forward "breakover" voltage is exceeded.) Once the SCR is conducting, the gate has no further effect, and cannot be used to turn the device off. To return the SCR to the non-conducting state, the applied voltage must be interrupted or reversed. To date, commercially available SCR's small enough to fit in the palm of a hand can switch currents up to 100 amperes, at peak inverse voltages of 400 volts.

Although SCR's are like thyratrons in that they can be turned on but not off by a control signal, they are far more efficient than thyratrons. For example, whereas a typical thyratron will exhibit a voltage drop between cathode and anode of about 10 to 15 volts when conducting, the SCR exhibits a voltage drop of one volt or less. This is very important when controlling power at lower voltages. At 24 volts, the thyratron would be only about 50% efficient, whereas the SCR is 95% efficient. Another point of superiority is the speed with which the SCR operates. The typical thyratron has an ionization and de-

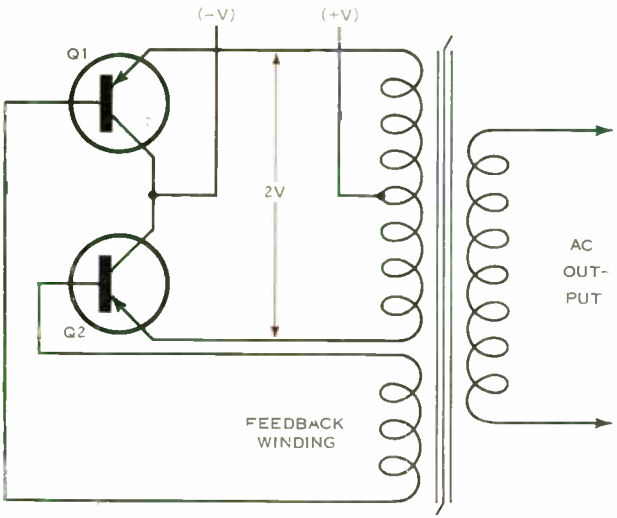


Figure 6. Simplified self-oscillating inverter circuit. Device functions as a "flip-flop" square-wave oscillator. Oscillation frequency is quite sensitive to load and input voltage.

World Radio History

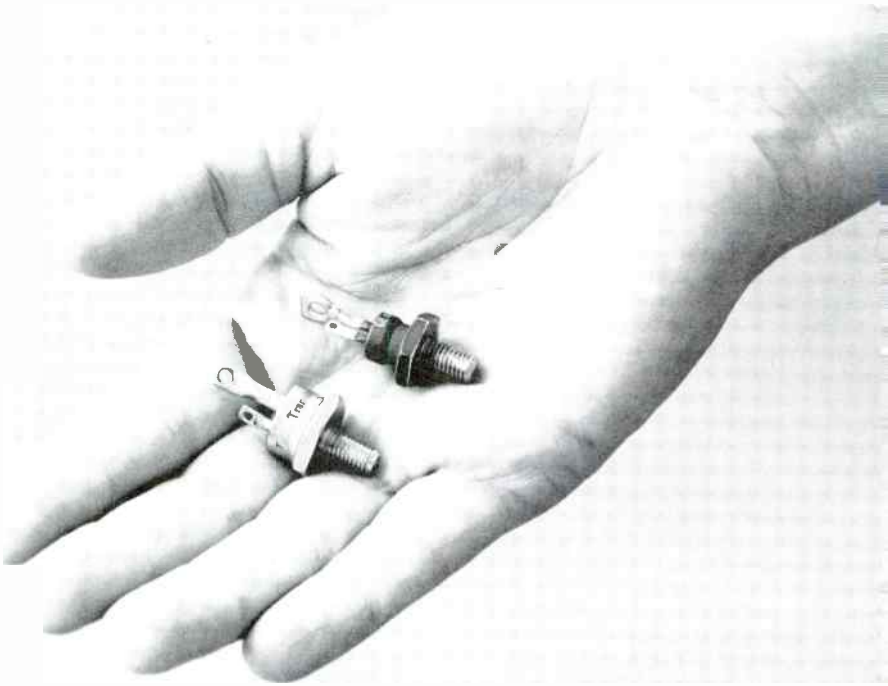


Figure 7. Tiny silicon controlled rectifiers can control surprising amounts of power. The smaller unit is able to switch 16 amperes at 400 peak inverse volts; larger unit is rated at 20 amperes. Parallel operation permits control of much larger currents.

ionization time (time required to establish or destroy conductivity) in the order of milliseconds, but the typical SCR turns on within a microsecond and turns off within one to twenty microseconds, depending on the size of the unit, and other conditions, such as the temperature of the unit.

How It Works

The controlled rectifier is a new version of a relatively old semiconductor device—the four-layer transistor. As indicated in Figure 8, it is made up of alternate *p* and *n* layers.* The SCR may be regarded as two separate transistors —a *npn* and a *pnp* which, in effect, overlap so that the center *n* and *p* layers are common to both. The center *p* section may then be regarded as the base of the *npn* transistor, and the center *n* section as the base of the *pnp* transistor. When "forward" voltage is applied to the SCR, conduction is blocked by the center junction, which acts like a back-biased diode.

If current is injected at either of the center sections (positive at the *p* layer or negative at the *n* layer), the center junction becomes conductive by transistor action, permitting the flow of current carriers through the device. In addition, this conduction biases the other center layer so that it also provides

* See DEMODULATOR, *May*, 1960 for a general discussion of semiconductors.

World Radio History

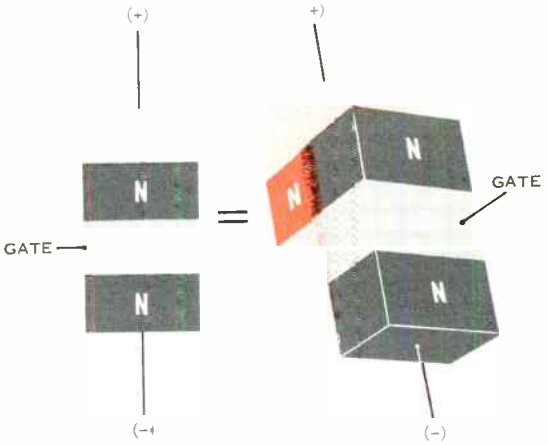


Figure 8. Four-layer controlled rectifier is analagous to two transistors sharing common layers. Injecting current at base of one "transistor" permits conductivity, and turns other transistor "on." Each biases the other to full conduction.

transistor-like current amplification. Because these two functions are self-supporting once started, the SCR reaches full conductivity very rapidly and the gate by which the process was started loses control.

One difficulty encountered with transistor inverters is that they have poor tolerance to voltage irregularities. Few transistors are available which can safely operate with more than 100-125 peak inverse volts across the transistor. However, in some communications installations where the local battery is nominally 48 volts, battery output may actually reach 60 volts. In the typical push-pull inverter where d-c power is applied to the center tap of a choke or transformer, the actual voltage applied to the transistors will be twice the battery voltage because the transformer winding acts as an auto-transformer, doubling the applied voltage. Thus, peak inverse voltages applied to the switching transistors

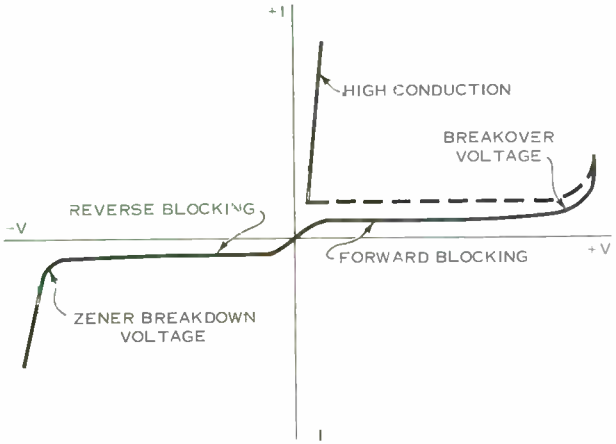


Figure 9. Typical controlled rectifier voltage-current characteristic. In absence of gate current, device will not conduct unless breakover voltage or zener breakdown voltages are exceeded. Gate current reduces breakover voltage, permitting forward conductivity.

World Radio History

will range from 96 to 120 volts—equal to or greater than their maximum ratings! Although some transistors are now becoming available which have higher voltage ratings, even these don't provide the performance reserve demanded by conservative engineering practice.

The silicon controlled rectifier provides an immediate solution to this problem because of its excellent voltage characteristics. An additional bonus provided by the SCR is that it requires much less driving power than a typical power transistor. Where a transistor might require half an ampere of driving current in order to conduct five amperes, a controlled rectifier requires a gate current of under 0.1 ampere to start conduction of fifteen amperes or more.

Practical SCR Circuit

The main difficulty in the use of SCR's is that they cannot be turned off

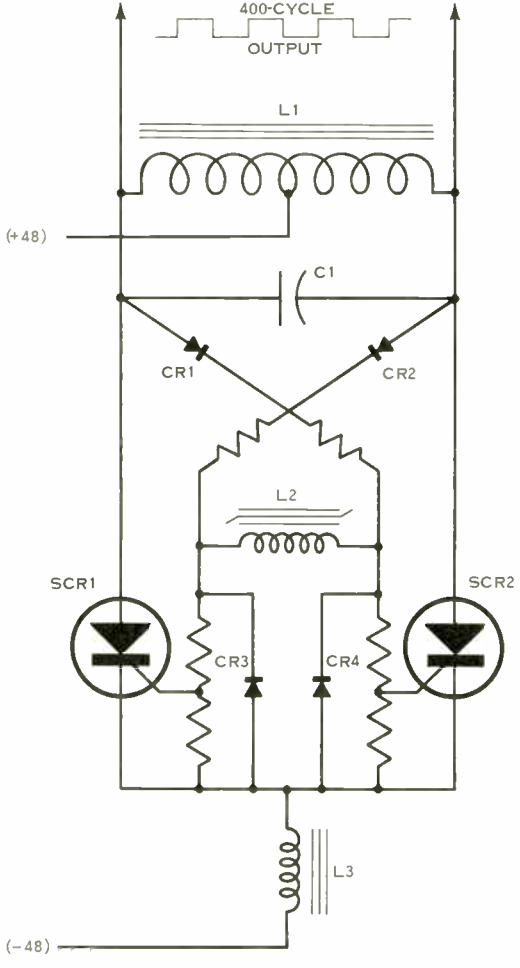


Figure 10. Simplified diagram of Lenkurt SCR inverter. Both SCR's are nonconductive until gate current from voltage divider causes SCR1 to conduct. When L2 saturates, gate current of SCR2 "fires" SCR2. Stored voltage in C1 surges through SCR2, is blocked by high impedance of L3, is applied to bottom of SCR1. Since C1 voltage is twice the forward voltage on SCR1, SCR1 is turned off. When L2 saturates in reverse direction, process reverses and continues automatically.

World Radio History

Table 1. Power supply weight vs. frequency (from Ref. 2)

	60 ~	400 ~	800 ~
TRANSFORMERS	61.0	19.0	8.6
FILTER CHOKE	20.2	1.8	0.8
CAPACITORS	20.0	4.7	2.2
HOUSING	8.0	5.0	4.0
TOTAL	109.2	30.5	15.6

by the control gate, as can the transistor. Normally, current flow through the SCR must be interrupted long enough for the non-conductive condition to be restored, or the flow of current must be reversed briefly. Figure 10 shows a simplified schematic diagram of a practical inverter now used in Lenkurt's dc-dc converter for powering microwave equipment from battery sources. This circuit is self-oscillating and does not vary its frequency appreciably with variations in load or input voltage. The frequency of the square-wave output is approximately 400 cycles per second, thus permitting the use of small but efficient transformers, magnetic amplifiers (for voltage regulation), and filter components.

It may be interesting to note how much the weight and size of equipment decreases as the frequency of the a-c power is increased. The comparison in Table 1 was compiled on the basis of airborne communications equipment, where weight savings are particularly important. As frequency increases, weight and size of the reactive components needed for control and filtering is reduced, but electrical losses begin to increase, so that eventually diminishing benefits result from higher frequencies unless special core materials are used for the cores of inductive components. Normally, this is only worthwhile in aircraft or missile applications.

Future Trends

We can expect power equipment to continue to become smaller, lighter, and more efficient as improved components and ways of using them are developed. The increasing use of transistors and other semiconductor components in all types of communication equipment is changing power requirements. Because of the steadily increasing reliability available from semiconductor components, power supplies may make more use of such techniques as active or electronic filtering instead of reactive filtering, thus reducing the size and weight of power supplies. •

Bibliography

1. D. A. Paynter, B. D. Bedford, J. D. Harnden, Jr., "Solid State Power Inversion Techniques," *Semiconductor Products;* March-April, 1960.
2. D. E. Fritz and C. K. Hooper, "Alternating Versus Direct Current for Aircraft-Radio Power Supply," *AIEE Transactions*, Vol. 63, p. 1227; 1944.
3. G. C. Uchrin and W. O. Taylor, "A New Self-Excited Square-Wave Transistor Power Oscillator," *Proceedings of the IRE;* January, 1955.
4. T. D. Towers, "Practical Design Problems in Transistor D.C./D.C. Convertors and D.C./A.C. Inverters," *The Institution of Electrical Engineers Paper No 2984E;* April, 1960.
5. *Silicon Controlled Rectifier Manual*, The General Electric Company, Auburn, New York; 1960.

World Radio History

World Radio History

World Radio History

World Radio History