THE BELL SYSTEM
TECHNICAL JOURNAL

VOLUME XXXVIII NOVEMBER 1959 NUMBER 6

An Experimental Transistorized Artificial Larynx

By H. L. BARNEY, F. E. HAWORTH and H. K. DUNN

(Manuscript received July 23, 1959)

A new experimental artificial larynx, which makes use of transistors and miniaturized components to provide a voice for those who have lost the use of their vocal cords by surgical removal or paralysis, is described. The larynx operates by introducing a substitute for the sound of the vocal cords into the pharyngeal cavity by means of a vibrating driver held against the throat. The acoustic principles of normal and artificial speech production that were followed in arriving at the new design are presented, along with descriptions of the transistor circuit and its operating characteristics.

I. INTRODUCTION

It is sometimes necessary, for the health of an individual, to remove his entire larynx by surgery. His trachea is then terminated at an opening (stoma) in the throat, and no connection between the lungs and the vocal tract remains. Since the normal source of energy for the speech process is provided by the lungs, such an individual loses his natural means of speaking.

These persons are usually advised by their surgeons and speech therapists to learn esophageal speech, and classes for this purpose are set up in various centers. In producing esophageal speech, the upper end of the esophagus serves as the substitute larynx and provides the necessary complex tone at an appropriate point in the vocal tract — the bottom of the pharynx. The esophageal speaker must learn to swallow air, or force air into the esophagus and then control its escape, in such a manner as to cause sustained vibrations of tissues at the upper end of the esophagus. Not all patients can do this successfully. In fact, surveys have shown that about a third of all larnygectomized patients are unable to master esophageal speech for one reason or another.[1] In addition, the quality of speech produced by this method is generally rather unpleassant — to such a degree that, in a comprehensive comparison test, listeners were unanimous in their preference for speech produced by a reed-type artificial larynx rather than esophageal speech.[2]

The use of an artificial larynx is therefore frequently desirable, and is often a necessity if the laryngectomized patient is to communicate by speaking. At the present time, there are several different artificial larynges available, including the Western Electric reed-type which has been distributed by the Bell System since 1930. However, both doctors and users are generally agreed that there are various deficiencies in the performances of all the available models, and that none are really efficient in their function. In the past few years, suggestions for the improvement of the Western Electric reed-type larynx have been received with increasing frequency, along with suggestions that a totally different design making use of transistors could provide improved performance. Accordingly, it was decided to investigate the problem further to see how modern .components and techniques might be used to make an improved artificial larynx.

The experimental artificial larynx to be described here is a result of these studies. Its characteristics are such that it provides an efficient means of communication for laryngectomized patients, while being more convenient and less conspicuous in use than the Western Electric Model 2 or other available larynges. It includes, in one small hand-held unit, a modified telephone receiver used as a vibrating driver that is held against the throat, a transistorized pulse-generating circuit and a battery power supply. When the pulse generator is switched on, vibrations are transmitted through the throat wall into the pharynx cavity and transformed into speech by the normal use of the articulatory mechanisms of the vocal tract. The loudness of the speech obtained with this unit is comparable with that of a normal person speaking conversationally. The artificial speech so produced sounds somewhat mechanical, but it is quite intelligible. By the use of an easily operated inflection control, a degree of naturalness heretofore unobtainable in artificial larynges may be achieved.

II. HISTORY OF BELL SYSTEM ARTIFICIAL LARYNX WORK

There is substantial evidence that artificial larynges were used as early as 1874, but it was not until 1925 that the Bell System became concerned with this area of communications. F. B. Jewett, who was president of Bell Telephone Laboratories at that time, suggested the development of an artificial larynx. His suggestion was prompted by discussions with a friend who had been laryngectomized and had impressed him with the need for a device that was more satisfactory than any then obtainable.

The Laboratories' initial efforts resulted in an instrument that employed rubber bands stretched in a manner to simulate the vocal cords, and was designated type 1A. These rubber bands deteriorated rapidly and were a source of considerable dissatisfaction. Consequently, during 1929 a new larynx, designated type 2A, was developed that incorporated several refinements,[3] including the substitution of a vibrating metallic reed for the elastic bands. It is this model, with a few minor changes, that is currently being manufactured by the Western Electric Company and distributed by the Bell System operating companies. The method of operation of the 2A artificial larynx is illustrated by the sagittal section view of the head in Fig. 1. The metallic reed is connected by tubing to the stoma in the throat, so that the user's breath can actuate the reed. The sound of the vibrating reed is conducted through another tube into the mouth, and this sound is used in the production of artificial speech sounds with normal tongue, lip and jaw movements.

In all, about 200 of the Model 1A larynges were made between 1926 and 1930, and about 5500 of the Model 2A larynges have been made to date. Since about 1950, the demand has remained constant at approximately 300 per year, although the number of laryngectomies performed annually has increased steadily. This leveling-off has occurred partly because there has been a marked increase during the last ten years in the use of esophageal speech, with the establishment of many speech clinics for the purpose of training laryngectomized patients in this method of speaking.

However, as noted previously, about a third of the total number of laryngectomized patients are unable to use esophageal speech, and consequently the need for an improved artificial larynx has become more urgent. In response to this need an advisory committee on artificial larynges was set up in 1956 by the National Hospital for Speech Disorders in New York, and its recommendations have provided helpful stimulation and guidance in the development of the new experimental model.

III. DESIGN OBJECTIVES FOR AN IMPROVED ARTIFICIAL LARYNX

In determining objectives toward which artificial larynx experimentation should be directed, preliminary discussions were held with the committee mentioned above, whose members include several surgeons, speech therapists and postlaryngectomized patients. To supplement the information obtained from them, all of the artificial larynges that were commercially available were studied and analyzed to ascertain their individual advantages and deficiencies.

The primary requirements, of course, were that the artificial speech be loud enough and natural enough so that the speaker could be easily understood. For the speech to sound natural, it should have pitch inflection, and, like the natural voice, should have a suitable fundamental pitch accompanied by harmonics that can be used to produce the various vowel sounds. These objectives were discussed in some detail in a recent paper.[4]

Secondary to the above, but still of great importance to the user, were the objectives that the device be inconspicuous and hygienic. It is in

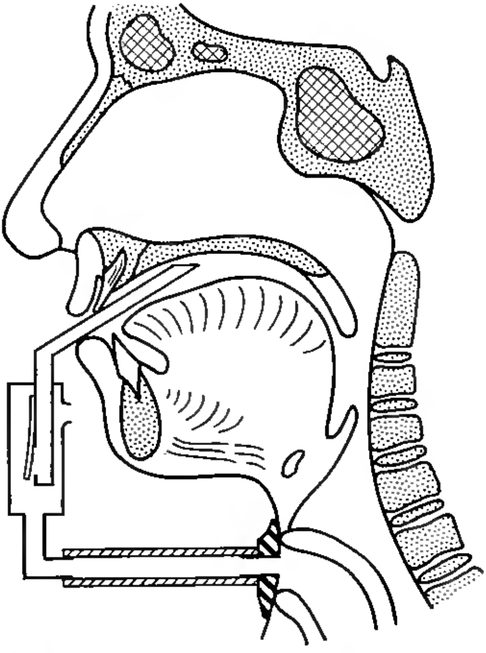


Fig. 1 — Sagittal section showing method of operation of reed-type larynx.

these respects that most of the presently available devices are deficient. If the user has to insert a tube into his mouth, it not only calls attention to his disability, but is also hygienically undesirable. Any connection at the opening in the throat, such as that required for the Western Electric Model 2, leaves much to be desired from the hygienic standpoint. In electrical devices, dangling wires leading to battery cases are also embarrassing and are liable to become entangled with other objects. The importance of making prosthetic devices inconspicuous may be inferred from the great efforts that hearing aid manufacturers have expended to make their product less noticeable in use.

Other desirable characteristics were simplicity of operation, reliability and low cost. Simplicity of operation is very desirable so that the patient will require only a minimum of training and, as soon as possible, gain the psychological benefits of vocal communication with his family and friends. Reliability and low cost can probably be attained most easily by the use of components that are already available commercially.

The design objectives, therefore, can be listed as follows:

(1) having output speech volume equal to that of a normal speaker,

(2) having output speech quality and pitch inflection like that of normal speech,

(3) inconspicuous,

(4) hygienically acceptable to the user,

(5) simple to operate,

(6) reliable,

(7) inexpensive.

IV. ACOUSTIC FACTORS IN PRODUCTION OF ARTIFICIAL SPEECH

4.1 *Types of Sound Source Needed*

In the production of normal speech, two types of sound energy are involved. One is a periodic tone produced by the vocal cords. It is variable in frequency and rich in harmonics, and is introduced into the pharyngeal cavity of the vocal tract. Except in whispered speech, this tone is always used in vowels and semivowels, including the nasal consonants, and is present in some other "voiced" consonants.

This normal vocal cord tone is completely lost when the larynx is removed. The esophageal speaker has learned to substitute the vibration of membranes at the mouth of the esophagus. But, if this sound cannot be produced and controlled adequately, another tone source must be supplied, and this is the chief function of the artificial larynx. For intelligibility of the speech produced, it is essential that the tone contain a wide range of harmonics, and, for naturalness, the harmonic amplitudes should fall off toward higher frequencies at the same rate that the real vocal cord tone does. Also, the tone should match that of the normal larynx in fundamental frequency and in frequency variability.

The second type of sound energy in speech is random noise, which is produced when the breath stream passes through a constriction formed by tongue or lips. It is present in stop and sibilant consonants, sometimes alone and sometimes in combination with the vocal-cord tone. These sounds are vital for the intelligibility of speech.

The normal means of generating random noise by the breath stream is also lost in the usual laryngectomy. However, it is not necessary to supply a substitute in an artificial larynx. Air trapped in throat and mouth can be forced out in such a way as to take the place of the normal breath stream in forming most of these sounds. Some deficiencies occur, such as a shortening of continuants like "s" and "sh", due to an insufficient volume of the trapped air; and the sound "h" is usually completely lost. The Western Electric reed-type artificial larnyx improves the ability to make some of these sounds, by allowing some breath stream to pass through the reed chamber into the mouth.

4.2 *Point of Application of Substitute Tone*

To match as nearly as possible the natural speech process, the artificial tone should be applied in the pharyngeal cavity. This requirement is not met in the Western Electric reed-type artificial larynx, yet understandable speech is produced. It is of interest to see just what changes in the quality of speech sounds result from a change of source application from pharynx to mouth.

It may be shown theoretically that a change from throat to mouth application, keeping the vocal tract configuration constant for a given vowel, does not change the resonant frequencies characteristic of that vowel. It does, however, change the relative amplitudes of the different resonances. The extent of the change depends upon the degree of constriction imposed by the tongue, which is different for different vowels. Another manifestation of the change is the appearance of antiresonances, which are not present when the source is in the throat.

To confirm these conclusions, two experiments were performed. In the first, an artificial tone was introduced into the pharynx of a human subject. A tube was attached to a transducer that produced the tone, and passed through the nose of the subject and into his pharynx until the opening of the tube was not far from his vocal cords.

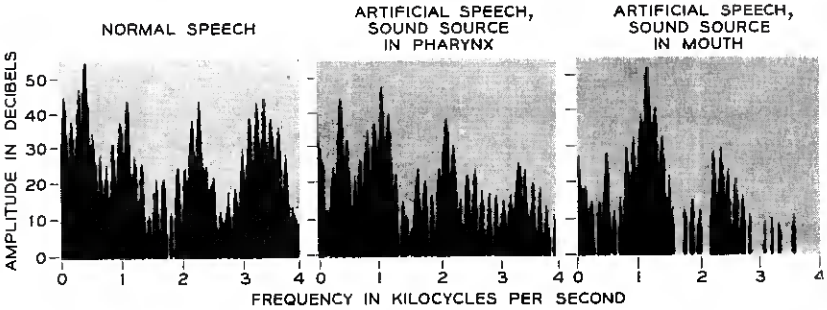


Fig. 2 — Sound spectra of vowel sound "oo" (as in "book") spoken normally and as produced artificially with sound source in pharynx, and in mouth.

Fig. 2 shows spectra taken as the subject made the vowel sound of "book". The first was made with his own vocal cords. The same vowel made with the artificial source yields the second spectrum. Although some pains were taken to make the spectrum at the output of the tube approach that of the real cord tone, some differences can be seen. The third spectrum was produced with the artificial source withdrawn from the pharynx and placed in the mouth of the subject. Particularly to be noticed are the change in relative amplitudes of the first two resonances and the "holes" in the mouth spectrum due to the antiresonances.

The second experiment made use of the Electrical Vocal Tract,[5,6] an analog of the vocal tract in which cavities are represented by lengths of transmission line and constrictions by inductances placed in series with the line (the tongue), or at its termination (the lips). An electrical complex tone can be applied easily to either throat or mouth cavity. Settings of such a device can be held constant more easily than a human subject can maintain a particular vocal tract configuration. On listening to the output of the artificial tract, it was found that vowel sounds changed considerably in character when the source was moved to the mouth. However, some but not all of their original naturalness could be restored by manipulation of the settings. It seems likely that the reed-type larynx user makes these readjustments naturally under the guidance of his own hearing, and that this accounts for the fact that his speech is still very intelligible.

Fig. 3 shows transmission measurements made with a sine-wave input on the Electrical Vocal Tract, in the three settings determined by previous listening tests: (1) the vowel "oo" (as in "fool") with a source in throat, (2) the same settings with source in mouth and (3) with controls readjusted to restore "oo" as nearly as possible.

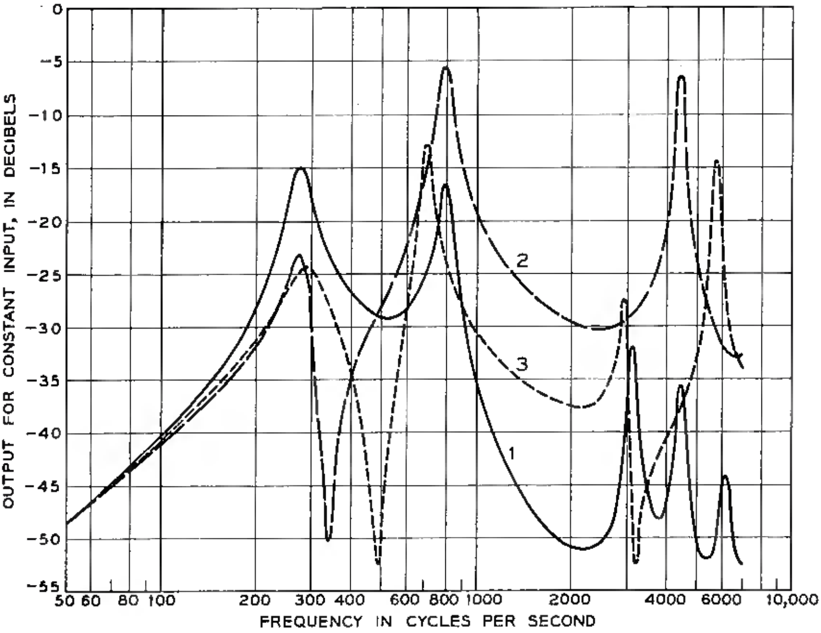


Fig. 3 — Transmission vs. frequency characteristic of electrical vocal tract: (1) adjusted for vowel "oo" (as in "fool") with source in throat position; (2) same settings, but with source in mouth position; (3) with source in mouth, but with controls readjusted to restore the "oo" sound as nearly as possible.

Although it would seem that excitation in the mouth is not as disadvantageous as it at first appears, pharynx excitation is still preferable. It is, of course, not practicable to introduce the sound through the nasal cavities as was done for the subject in the experiment described. In fact, insertion of any outside bodies into throat and mouth tends to be unhygienic. However, sound can be introduced into the throat from outside by transmission through the throat wall. This principle was used in an artificial larynx designed by Wright.[7] In the present development, it has been found possible to produce an adequate spectrum in the pharynx by this method, while at the same time limiting to a reasonable level the sound radiated directly from the device.

4.3 *The Use of a Throat Vibrator to Provide Substitute Vocal-Cord Tone*

The experiments just described indicate that the preferred position for the sound source is in the pharynx. Some thought was given to the use of a transducer surgically embedded in the throat. However, this would require a second operation for those who already had been laryngectomized, and the opinions of doctors consulted on the subject were divided as to its advisability. Accordingly, it was considered to be outside the scope of the present artificial larynx project. The problem then became one of transmitting through the flesh and cartilage around the pharynx a complex signal with a broad frequency spectrum. In order to obtain natural-sounding speech, the source spectrum must have strong low-frequency components. The total frequency range required extends from about 100 to several thousand cycles per second.

Experiments were conducted using a variety of vibrating devices held against the outside of the throat. Some of these were constructed especially for these tests and the rest were devices obtainable commercially. Of all these, the HA-1 telephone receiver used in the type 300 telephone sets proved the most promising.[8] However, when pressed against the throat, the loading on the diaphragm was far different from what it is when working into air, since the characteristic mechanical impedance of flesh is some 4000 times that of air. This heavy loading made desirable a number of modifications in the receiver to enable it to give a greater amplitude of vibration into the throat. These modifications are described in Section 5.4.

V. CIRCUIT AND MECHANICAL CONSTRUCTION

The circuit of the new experimental artificial larynx uses a highly efficient arrangement of transistors powered by mercury batteries to provide a compact, self-contained unit. In its design, an objective was to use commonly available, inexpensive components wherever possible. Fig. 4 illustrates the cylindrical configuration of the unit, with the combined on-off switch and pitch-inflection control knob arranged for operation by thumb or forefinger.

5.1 *Transistor Circuit*

A schematic diagram of the circuit is shown in Fig. 5. It is essentially a two-stage relaxation oscillator followed by a power stage that works into a transducer. The relaxation oscillator uses a p-n-p transistor, Q_1, and an n-p-n transistor, Q_2, coupled together with regenerative feedback. The frequency of oscillation is determined by the pitch-control resistance, R_1, in combination with capacitance C_1. The output of the relaxation oscillator appears across resistance R_5 as a series of short periodic pulses. The width of these pulses is determined by resistance R_2 and capacitance C_1. The values shown in Fig. 5 give a pulse width of 0.0005 second.

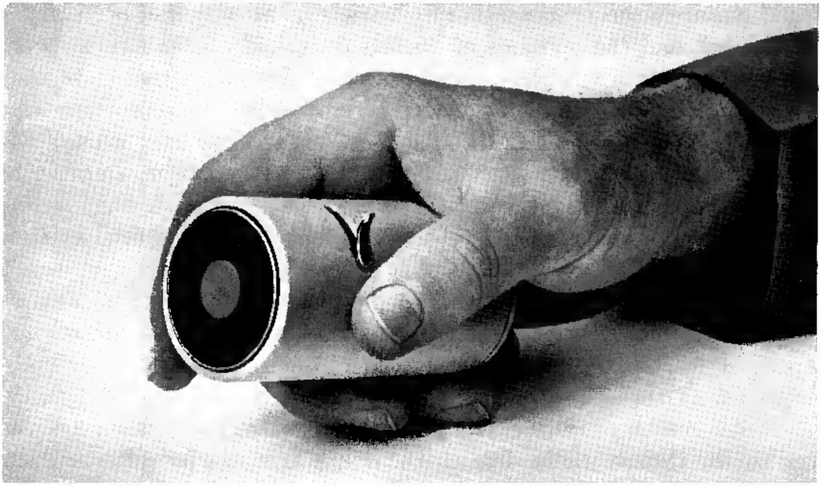


Fig. 4 — Picture of artificial larynx showing thumb-operated on-off switch and inflection control.

The periodic pulses generated in the relaxation oscillator are transmitted through the semiconductor diode, CR_1, to the base of power transistor Q_3. The HA-1 receiver is connected in the collector circuit of Q_3, and receives short periodic current pulses of about 0.45 ampere peak value at the oscillation frequency.

The range of oscillating frequency may be adjusted by changing the range of resistance R_1 available in the pitch-control rheostat, to simulate either a man's or a woman's pitch range. For men, the range is from 100 to 200 cycles and, for women, it is from 200 to 400 cycles. This is an octave range in either case, and is sufficient to duplicate the pitch inflection used in normal speech. The on-off switch and pitch-control rheostat are arranged so that the switch is closed at the lowest oscillating frequency, and further movement of the control causes the frequency to increase. The control knob is spring-loaded so as to return it to the off position when released.

Two 5.2-volt mercury batteries in series provide the necessary power to operate the circuit. Although the peak is 0.45 ampere, the pulse duty factor is so small that the average current drain from the batteries is only about 22 milliamperes. The rating of the batteries is 250 milliampere hours.

As an alternative to the self-contained mercury batteries, a small rectifier operated from 115-volt, 60-cycle line voltage may be substitued. This arrangement may be useful at an office desk or other fixed location.

When the rectifier power supply is plugged into the auxiliary power jack shown in Fig. 5, the batteries are disconnected from the circuit.

5.2 *Selection of Pulse Duty Cycle*

The average current drain on the batteries, the spectrum of the acoustic output of the artificial larynx and the loudness of the output, are all functions of the pulse width, assuming a fixed supply voltage. For widths of a few tenths of a millisecond, the average current drain would be low, and the spectrum would have a wide frequency band with strong harmonics running up to several thousand cycles per second, but the acoustic output would be weak. Fig. 6 shows the relation between acoustic output and pulse width, and also the relation between current drain and pulse width. The acoustic outputs displayed were obtained by measuring the output from a single subject saying "ah" at a distance of 3 feet from the sound level meter. Pulse widths of 0.5 to 0.6 millisecond gave near-maximum output. For wider pulses, the acoustic output decreased, and the speech became somewhat muffled and nasal in quality. A pulse width of 0.5 millisecond was adopted. Correspondingly, the average current drain was 22 milliamperes at a frequency of 100 pulses per second. Sound spectrograms of speech using the 0.5-millisecond pulse width indicated a satisfactory spectrum.

5.3 *Mechanical Construction*

For simplicity of construction, a cylindrical container was chosen to house the artificial larynx. The dimensions of the experimental model are

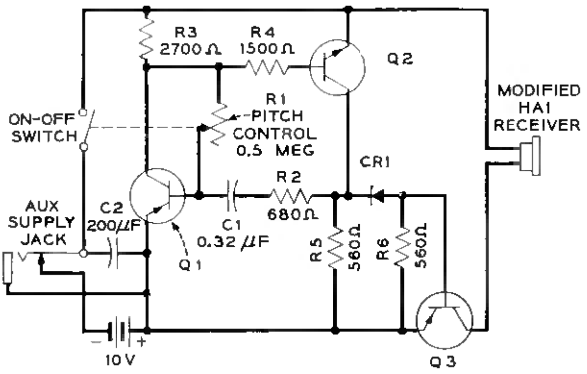


Fig. 5 — Schematic of artificial larynx circuit.

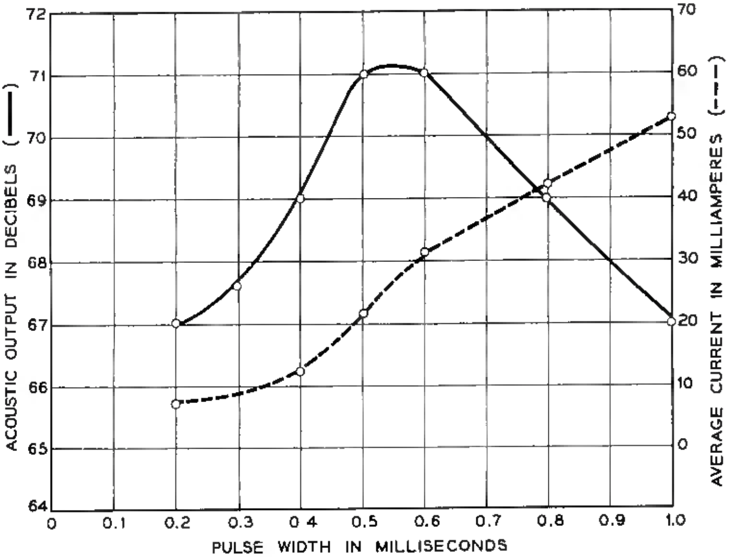


Fig. 6 — Characteristics of acoustic output vs. pulse width, and average battery supply current vs. pulse width.

$1\frac{3}{4}$ inches in diameter and $3\frac{1}{4}$ inches long. The weight, including batteries, is 8 ounces. To package all the components in this volume, a modular type of construction was used, as shown in the exploded view in Fig. 7.

The HA-1 receiver is at the front end of the unit, with the diaphragm flush with the end of the cylinder. The back of the receiver is wrapped with sponge rubber, and two discs of sponge rubber and one of thin brass sheet are placed between it and the adjacent components to attenuate the backward radiation of sound. Were it not suppressed, this direct radiation back through the shell and into the surrounding air would tend to mask the speech sounds and contribute a buzzy, mechanical quality to the over-all effect.

The next two modules back of the receiver contain the pitch-control rheostat, the transistors and associated circuit elements. The last module contains the two mercury batteries in a plastic shell and the jack for the external power supply. The back plate may be removed by unscrewing a single machine screw which has a slot large enough so that a thin coin may be used in place of a screwdriver. This permits convenient access to the mercury batteries for changing them without disturbing the rest of the circuit.

While the experimental model is a compact unit, some further miniaturization could be achieved by the use of printed circuit techniques and closer component spacing.

5.4 *HA-1 Receiver Modifications*

The HA-1 receiver as normally used in a telephone set has a protective metal grid and cloth cover over the diaphragm. For use in the artificial larynx these are removed. And, in order to achieve greater efficiency in terms of output volume of artificial speech for a given battery supply power, several additional modifications were made.

The permanent magnets were magnetized to full strength, instead of being only partially magnetized. The diaphragm was correspondingly shimmed out from the pole pieces, so that it would not be pulled into contact with them. The spacing between diaphragm and pole pieces in this condition measures between 0.002 and 0.003 inch, and a slight push on the diaphragm is sufficient to make it adhere to the pole pieces. The electrical pulses from the transistor circuit are so poled as to oppose the permanent magnet field and release the diaphragm to spring outward. This driving polarity gives higher speech volumes than the opposite one.

In order to obtain sufficient current from the 10.4-volt supply to counteract the permanent magnetization, it was necessary to decrease the impedance of the receiver winding by connecting its two coils in parallel instead of in the usual series arrangement. In order to improve the match of mechanical impedances between the receiver and the throat, a diaphragm of 0.0083-inch permendur was used in place of the standard 0.011-inch thickness provided in the HA-1. A series of tests was made with a range of thicknesses from 0.0065 to 0.011 inch, and it was found that the highest speech volumes were obtained with a thickness in the order of 0.0083 inch.

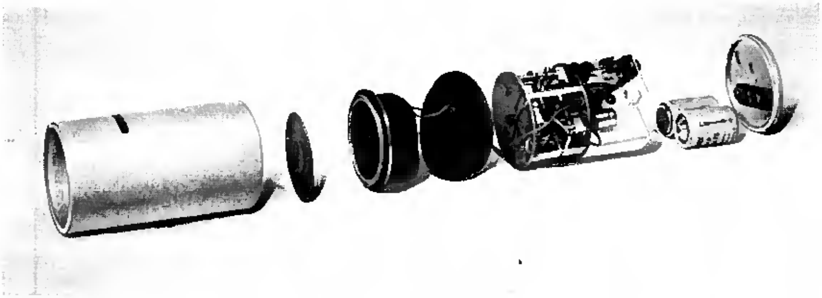


Fig. 7 — Exploded view of artificial larynx, showing modular construction.

In order to reduce the magnetic saturation of that area of the diaphragm between the pole pieces, a small center patch of permendur, 0.54 inch in diameter and 0.011 inch thick, was spot-welded to the diaphragm before heat treatment. This addition improves the magnetic circuit and does not materially affect the stiffness of the diaphragm.

The HA-1 receiver normally has a small resonance damper provided by a cloth-covered hole in the plastic back under the diaphragm. This damps the natural diaphragm resonance, which in air falls at about 3000 cycles. The cloth covering this hole is removed in the artificial larynx, and diaphragm damping is obtained by contact with the flesh of the throat. Removal of the cloth damping patch slightly increases the output of high-frequency harmonics in the artificial speech.

5.5 *On-Off Switch and Inflection Control*

The arrangement of the on-off switch and inflection control was designed for ease of manipulation. With practice on the present arrangement, either rising or falling inflection can be achieved at the beginning or ending of voicing.

Several other methods of control were tried. One made use of a rack and pinion gear arrangement, in which a button was pushed straight into the shell of the unit. Precise control of frequency was not easily obtained with that method. It was found more satisfactory to push the control sideways over a distance of a half-inch or more. Another early version depended for control on application of pressure along the longitudinal axis of the artificial larynx. This seemed satisfactory from the functional standpoint, but was more difficult to implement mechanically than the arrangement finally adopted.

VI. ACOUSTIC PERFORMANCE

Tests of the acoustic performance of the new artificial larynx have been made to find how nearly it meets the original design objectives with respect to output volume and speech quality.

6.1 *Loudness*

A little practice is required to find the proper pressure and placement on the throat that yield the best results. Output volume measurements on subjects who have acquired a moderate amount of proficiency show sound pressure levels on the vowel peaks of 70–75 db above 0.0002 microbars at a distance of three feet from the speaker's mouth. This is approximately a normal conversational level. However, in an environment so noisy as to require a speaker to raise his voice appreciably above the normal level, this volume would limit the separation between talker and listener to shorter distances than those possible for a normal speaker.

6.2 *Frequency Spectra*

Speech quality has been checked by comparisons of frequency spectra, and by measurement of the ratio of speech signal to directly radiated buzz. Spectrograms[9,10] of the words "artificial larynx" and amplitude sections of ten vowel sounds were made from the speech of one subject, using both the new artificial larynx and his natural voice. These are reproduced in Figs. 8 and 9 respectively. In Fig. 8, it may be seen that the "f" and "sh" sounds in the word "artificial" are shorter in duration for the artificial larynx speech than for the normal speech. With the artificial larynx, the speaker must make such sounds by means of the air trapped in his mouth and pharynx since his normal air supply is cut off.

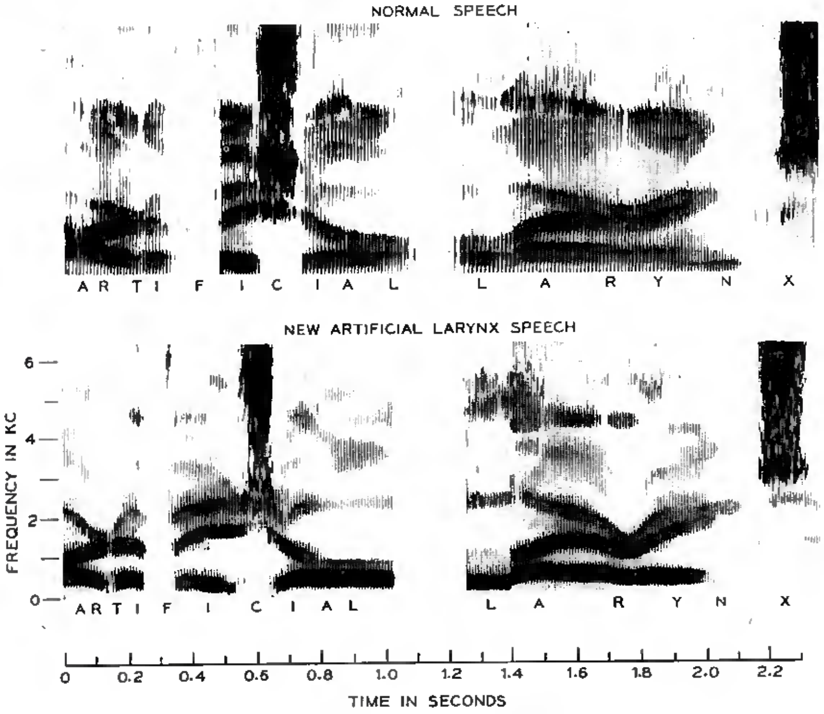


Fig. 8 — Sound spectrograms of the words "artificial larynx" as spoken normally and with the new experimental artificial larynx.

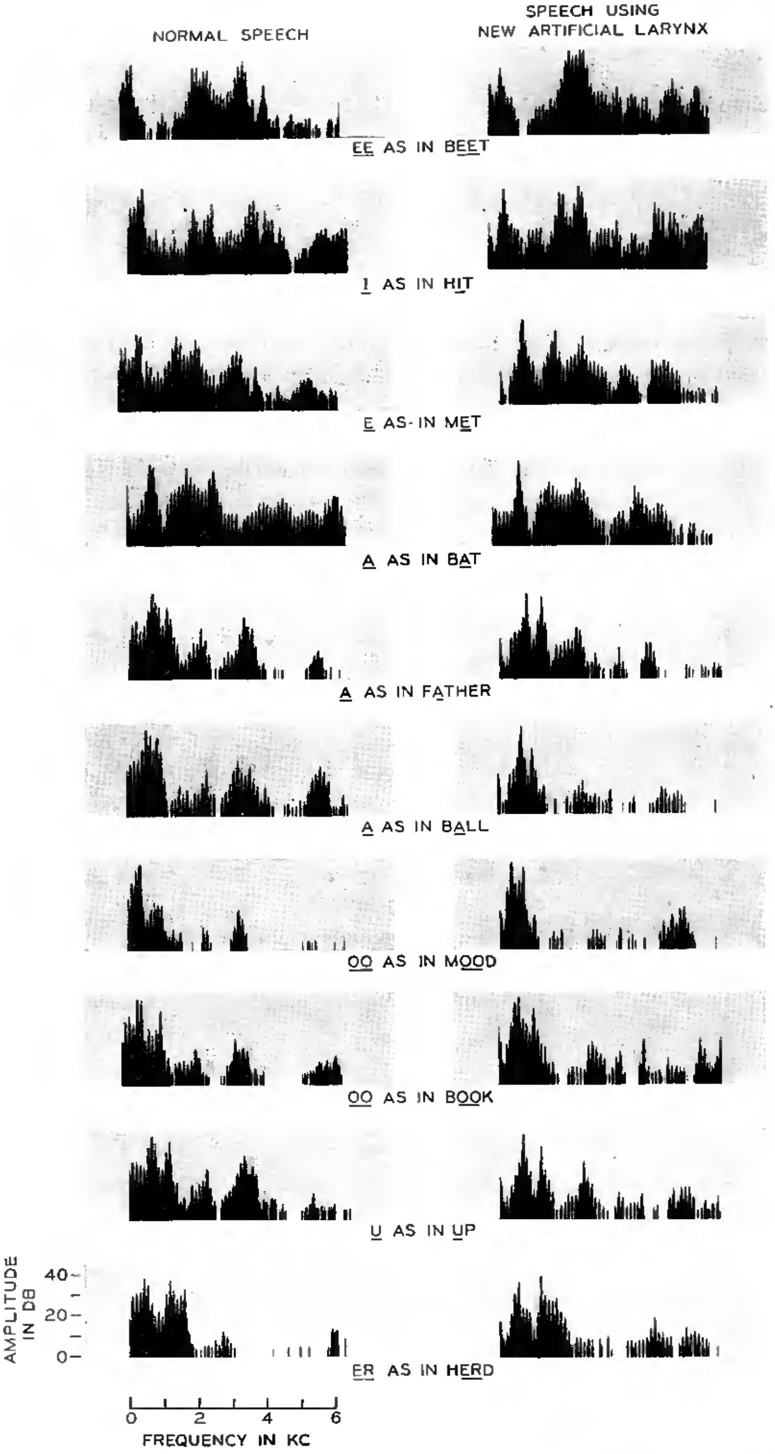


Fig. 9 — Sound spectra of ten sustained vowel sounds as spoken normally and with the new experimental artificial larynx.

This small air supply thus tends to shorten fricatives and sibilants, but the spectrogram indicates that they can be made satisfactorily. Some practice was required to make the "sh" sound in "artificial" as long as that which is shown.

In using the artificial larynx it is more convenient to leave it turned on through several syllables or words than to turn it on and off as one does in natural voicing. That this does not make the speech as unnatural as one might expect is indicated also in Fig. 8. The instrument was turned off between the two words, but it can be seen that, for the "t", "f", "sh" and "x" sounds, very little of the voicing comes through, although the device was operating while those sounds were being produced. For the unvoiced fricatives and stop consonants, the sound transmission path from the pharynx is evidently nearly closed off.

In the comparison of the vowel spectra shown in Fig. 9, it is apparent that the new artificial larynx is able to transmit sufficient power into the pharynx throughout the spectrum to permit satisfactory development of the high-amplitude regions (formants) of the vowel sounds. It has been indicated[4] that the harmonics in the source spectrum of the natural voice are strongest at the low frequencies, dropping in amplitude toward the high frequencies at about the inverse 1.5 power of the harmonic number. A cross comparison in Fig. 9 shows that, for some vowels, the difference in the high- and low-frequency amplitudes is greater for the natural source, and, for others, it is greater for the artificial source. This observation leads to the conclusion that, on the average, the artificial source has approximately the right spectrum.

6.3 *Externally Radiated Interference*

Some of the sound produced by the vibrating diaphragm does not pass through the speaker's throat but is radiated directly by the instrument itself or from areas of the throat around the place where the unit is pressed. This external radiation, of course, would interfere with the intelligibility of the speech if it were not well suppressed. Measurements taken in an anechoic chamber with the unit pressed against the throat but with the mouth closed showed an intensity level for this interference 20–25 db below the level when the vowel "ah" was being voiced. When the unit is operated with the vibrating end working into a sound-absorbing cavity, the level is about 6 db lower still, indicating that most of the interfering sound is from the throat areas immediately adjacent to the artificial larynx rather than from the instrument itself. If it should be desirable to reduce this noise still further, the end held against the throat might be specially shaped to reduce the external vibration of the throat tissues.

6.4 *Reactions of Laryngectomized Users*

In collaboration with the Advisory Committee on Artificial Larynges tions. Four units of the new design were used in the field test; two were assigned to two laryngectomized patients for the entire period of four weeks, and the other two units were used for shorter periods by several other patients. In all cases, favorable comments were made on the speech quality and the lack of externally radiated buzz. Comments of friends and relatives of the patients using the new model generally indicated that they liked the intelligibility and speech quality of the artificial speech produced by it.

One comment was made to the effect that, for optimum comfort in use, the diameter of the unit should be somewhat less. Adoption of this suggestion would preclude the use of the HA-1 receiver. The Advisory Committee on Artificial Larynges of the National Hospital for Speech Disorders felt that, for nearly all patients, the present diameter of $1\frac{3}{4}$ inches would be satisfactory, and did not recommend such a change.

Battery life was indicated to be satisfactory in these tests. The new units were used alternately with other models by the two patients who had them for the entire test period, and it is not known just what their cumulated operating times were. One of the two patients estimated that he had used the new unit for about half of his talking. None of the four units in the limited field test required a change of batteries during the four-week period.

VII. ARTICULATION TESTS

Articulation tests using speech produced by practiced talkers with previously available artificial larynges have been carried out. These tests were intended as a guide in the development of the new instrument. A second set of tests was made after the new experimental model was available, comparing it with previous types.

7.1 *Tests with Previous Types*

For the first test, it was possible to obtain two experienced users of esophageal speech, of the reed-type artificial larynx, and of an

Table I — Percentages of PB Words Heard Correctly, from Natural and Substitute-Larynx Speech

Natural voices	97.3 96.6
Esophageal speech	79.0 64.1
Reed-type artificial larynx	63.4 40.3
Throat-type artificial larynx	58.1 40.3

available type using throat application. These individuals were asked to read five of the Harvard PB (phonetically balanced) lists of 50 monosyllabic words.[11] Their utterances were recorded on tape and presented later, in a suitably mixed order, to a crew of seven listeners who recorded their responses. Two speakers with normal voices were included for comparison. The percentages of words heard correctly are given in Table I.

To understand the significance of these scores, it has been found that a 60 per cent articulation from such isolated words corresponds to a sentence intelligibility of more than 95 per cent, and that even 40 per cent in the word score means that more than 90 per cent of sentences would be understood.

In a test supplementary to the above, it was found that the articulation score with the throat type tested could be improved to about 70 per cent if the directly radiated sound were reduced about 20 db.

The number of individuals in the tests was too small to indicate an over-all ranking for the different types. It can be concluded, however, that either the reed type (with mouth application) or the external throat type could be sufficiently intelligible to give good conversational ability. The choice between these types could therefore be made by other criteria.

7.2 *Comparison of New with Best of Older Types*

The second set of tests was abbreviated, and was intended to provide a comparison between the new larynx and the previous types. Thus, only the higher-scoring individuals using the reed and throat types in the previous tests, with two PB lists (100 words) each, were incorporated. These utterances were compared with 100 words from the new experimental model. Because of changed conditions (principally the use of a crew of listeners who were less familiar with laryngectomized speech) the results shown in Table II are not directly comparable with the previous tests. They are comparable with each other, however.

With regard to population averages, these figures cannot be considered indicative. The differences, however, are favorable for the new model.

VIII. CONCLUSIONS

An artificial larynx has been developed that is hygienic, convenient and inconspicuous. It has a fundamental tone that is similar in pitch

TABLE II — ARTICULATION SCORES FROM NEW EXPERIMENTAL ARTIFICIAL LARYNX AND FROM THE MORE SUCCESSFUL USERS OF OLDER TYPES

Older throat type	43 per cent
Reed type	52 per cent
New experimental model	59 per cent

range and variability to the real voice, and near enough in spectrum to produce natural-sounding speech. The loudness of the speech produced with it is comparable to that used in normal conversation, and the speech is generally free of masking effects of directly radiated noise. The essential characteristics and performance of this experimental model will be incorporated into a commercial design to be manufactured by the Western Electric Company. Distribution of the new model, beginning some months hence, will be through the Bell System operating companies, following procedures similar to those used with the Model 2 reed-type larynx for the past 30 years.

IX. ACKNOWLEDGMENTS

We wish to thank Dr. L. H. Whitney, medical director of the American Telephone & Telegraph Company, and L. C. Withers of the Western Electric Company for their continued interest and encouragement, and Dr. Hayes Martin of New York and Dr. Lynwood Heaver, medical director of the National Hospital for Speech Disorders, for assisting us in finding expert laryngectomized talkers. Some ten of the latter were very cooperative and generous of their time in recording our test sentences and lists of words with esophageal speech and various models of artificial larynges. Our thanks are also due to D. J. MacLean for his valuable aid in many technical phases of the work, to Miss E. A. Klarmann for assistance in preparation and scoring of the articulation test and to J. H. Kronmeyer for aid in mechanical design and construction of the new model.

REFERENCES

1. Heaver, L., White, W. and Goldstein, N., Clinical Experience in Restoring Oral Communication to 274 Laryngectomized Patients by Esophageal Voice, J. Am. Geriatrics Soc., **3,** September 1955, p. 678.
2. Hyman, M., An Experimental Study of Artificial-Larynx and Esophageal Speech, J. Speech and Hearing Disorders, **20,** September 1955, p. 291.
3. Riesz, R. R., Description and Demonstration of an Artificial Larynx, J. Acoust. Soc. Amer., **1,** January 1930, p. 273.
4. Barney, H. L., A Discussion of Some Technical Aspects of Speech Aids for Postlaryngectomized Patients, Ann. Otology, Rhinology and Laryngology, **67,** June 1958, p. 558.
5. Dunn, H. K., The Calculation of Vowel Resonances and an Electrical Vocal Tract, J. Acoust. Soc. Amer., **22,** November 1950, p. 740.
6. Schott, L. O., An Electrical Vocal System, Bell Lab. Record, **28,** December 1950, p. 549.
7. Wright, G. M., U. S. Patent No. 2,273,077, February 17, 1942.
8. Jones, W. C., Instruments for the New Telephone Sets, B.S.T.J., **17,** July 1938, p. 338.
9. Koenig, W., Dunn, H. K. and Lacy, L. Y., The Sound Spectrograph, J. Acoust. Soc. Amer., **18,** July 1946, p. 19.
10. Kersta, L. G., Amplitude Cross-Section Representation with the Sound Spectrograph, J. Acoust. Soc. Amer., **20,** November 1948, p. 796.
11. Egan, J. P., Articulation Testing Methods, Laryngoscope, **58,** September 1948, p. 955.

Ideal Binary Pulse Transmission by AM and FM

By E. D. SUNDE

(Manuscript received May 18, 1959)

In binary pulse transmission by carrier amplitude or frequency modulation it is ordinarily desirable, both for efficient bandwidth utilization and for improved performance under adverse noise conditions, to use bandpass channels of the minimum practicable bandwidth, as determined by considerations of intersymbol interference and filter design. It is shown that intersymbol interference can be avoided in binary pulse transmission by FM without the need for a wider channel band than in double-sideband AM, for equal pulse transmission rates. Explicit general expressions are derived for the appropriate shaping of the bandpass channel and for the shapes of received pulses, for cases in which rectangular binary pulses are transmitted by FM, without premodulation or postdetection pulse shaping by low-pass filters. Illustrative comparisons are made of binary pulse transmission by AM and FM for two special cases of general interest in communication theory and pulse-system design. The more general case of partial pulse shaping by premodulation and postdetection low-pass filters is also considered.

The performance of FM and AM systems in the presence of noise depends on the division of channel shaping between transmitting and receiving filters. The optimum division with FM and AM is determined for random noise, and comparisons are made of signal-to-noise ratios for optimized FM and AM systems. It is shown that there is a single universal relation between error probability and signal-to-noise ratio, applying to an infinite universe of optimized baseband systems and optimized AM systems with ideal synchronous detection, and that this relation is the same as for baseband transmission over an idealized flat channel of minimum bandwidth. The analysis indicates that, with binary FM and appropriate postdetection low-pass filters, it is possible in principle to realize an improvement in signal-to-noise ratio over bipolar double-sideband AM with synchronous detection (phase reversal), for equal channel bandwidths, average signal power and pulse transmission rates, although this may not be feasible with practicable filters.

TABLE OF CONTENTS

I. INTRODUCTION

Transmission of digital or analog information by binary rather than by multilevel pulses offers significant advantages in systems design. For one thing, it simplifies the implementation of regenerative repeaters and various kinds of terminal equipment, such as carrier modulators and demodulators, and devices for timing-wave provision, coding and storing of messages and automatic error-checking or correction. For another thing, binary pulse transmission imposes less severe requirements on the transmission medium with respect to signal-to-noise ratio, amplitude and phase deviations over the channel band, and tolerable transmission-level variations. Because of these advantages, binary rather than multilevel pulse transmission is ordinarily the more practical and economical method, even in existing channel facilities designed primarily for voice or other analog transmission, where consideration of the rather high signal-to-noise ratio alone would permit a much greater number of pulse amplitudes and, thus, a substantially greater channel capacity than could be economically realized.

The three principal methods of binary pulse transmission by carrier modulation now in use are double- and vestigial-sideband AM, in the form of "on-off" keying with envelope detection, and FM in the form of "frequency-shift" keying. With synchronous or homodyne detection in AM, other methods are feasible that afford a bandwidth saving or improved signal-to-noise ratio or, like FM, have the advantage over "on-off" AM that they facilitate rapid automatic compensation of transmission-level variations. Among these binary methods are bipolar double-sideband AM, also referred to as phase reversal or two-phase modulation, and bipolar vestigial-sideband AM. Another method is bipolar double-sideband AM on each of two carriers at quadrature, also referred to as quadrature double-sideband AM or four-phase modulation.

For optimum performance in binary pulse transmission by AM or FM it is essential that the transmission-frequency characteristics of the channels be appropriately shaped with respect to amplitude and phase, so that intersymbol interference is avoided or at least reduced to a practicable minimum. A second requirement for optimum performance in the presence of noise is an appropriate division of channel shaping between transmitting and receiving filters. In addition, there are various other requirements not pertaining to the channel, such as exact timing in the transmission and reception of pulses and ideal AM and FM modulators and demodulators.

The purpose of this presentation is a determination of these optimum channel characteristics and the optimum signal-to-noise ratios for various error probabilities in binary pulse transmission by AM and FM, with particular emphasis on FM.

The analysis of both analog and digital pulse transmission for FM is more complex than it is for AM, since FM is a nonlinear modulation method. For this reason, the sideband spectrum of a given signal is wider than it is in double-sideband AM, and a wider bandpass channel is required for analog signal transmission without distortion.

In analog transmission it is possible to realize improved performance in the presence of noise in exchange for the increased bandwidth. In binary pulse transmission, however, it is ordinarily desirable, both for optimum performance under adverse noise conditions and for efficient bandwidth utilization, to use the minimum channel bandwidth practicable from the standpoint of intersymbol interference and filter design.

While a wider bandpass channel than in AM is required for distortionless analog transmission, this does not preclude the possibility that, under appropriate conditions, pulses can be transmitted by FM with no intersymbol interference, without the need for a greater channel band than is required in double-sideband AM. This depends on the possibility of controlling pulse distortion resulting from bandwidth limitation so that zero points in the received pulses occur at uniform intervals from the peak pulse amplitude. When pulses are transmitted by AM, this can be accomplished by appropriate shaping of the transmission-frequency characteristic of the channel, as shown elsewhere.[1,2] The analysis is extended herein to ideal binary pulse transmission by FM.

A basic criterion for performance in digital transmission through noise is the error probability as related to the signal-to-noise ratio, which has been dealt with elsewhere for baseband transmission,[3] "on-off" double-sideband AM with envelope detection,[4,5] bipolar AM or carrier phase modulation with synchronous detection[5,6] and frequency modulation.[5,7] In the above analyses random noise is assumed and the signal-to-noise ratio is stated in terms of signal power during a steady "mark" or "space," which, in bandlimited channels, is usually not equal to average signal power, even in binary phase or frequency modulation systems. Moreover, ideal flat baseband or bandpass channels of the minimum bandwidth required to avoid intersymbol interference are assumed or implied in AM, although they are not practicable in actual systems. In the case of FM, no consideration is given to bandwidth requirements and channel shaping for optimum performance.

As an aid in systems design, specific consideration is given in Sections II through XI of this presentation to appropriate bandwidths and channel shaping for AM and FM systems. The remainder of the analysis is concerned with signal-to-noise ratios as related to channel shaping and to appropriate filter shaping for optimum performance in the presence of random noise.

II. FREQUENCY AND AMPLITUDE MODULATION BY BINARY PULSES

The original signal ordinarily would consist of rectangular baseband pulses of duration T, with a negative polarity to indicate "space" and a positive polarity to indicate "mark," or conversely. These rectangular pulses may be applied directly to the frequency modulator, or they may be applied to a premodulation low-pass filter for preshaping.

The modulator output is applied to a bandpass channel with a certain transmission-frequency characteristic, which, in the ideal case, would be symmetrical about the midband frequency and have a linear phase characteristic. The envelope of the received pulses at the channel output and the frequency modulation of the carrier within the envelope depend on the shape of the modulating pulses and on the transmission-frequency characteristic of the bandpass channel. With an ideal detector at the receiving end, the demodulated signal is proportional to the time derivative of the phase of the carrier within the envelope, i.e., to the "instantaneous frequency deviation."

Detection of the phase derivative is facilitated by conventional frequency discriminators or zero-crossing detectors when the channel bandwidth is narrow in relation to the carrier frequency, but this is not a basic theoretical requirement if appropriate detectors are postulated. Nor, with appropriate ideal balanced FM detectors as assumed herein, is a limiter necessary for elimination of amplitude modulation effects, although it is highly desirable with unbalanced detectors and for prevention of undesirable effects of sudden level changes.

At the detector output, a postdetection low-pass filter may be used for final pulse shaping, but it is not essential for this purpose. Such a filter may be required to eliminate unwanted demodulation products (carrier ripple), but the bandwidth required for this can be much greater than that of the bandpass channel, particularly when the channel bandwidth is small in relation to the midband frequency. This condition can always be realized by frequency translation before demodulation, which may also be required for optimum performance with conventional frequency discriminators or zero-crossing detectors.

A more important function of the postdetection filter in conventional analog signal transmission is elimination of higher frequency noise components, in order to realize the inherent FM noise advantage. In binary pulse transmission with a bandpass channel of no greater bandwidth than is required to avoid intersymbol interference, as considered here, the noise advantage that can be derived from the use of a low-pass filter may be rather limited. To realize a significant noise advantage the filter must, in this case, have the appropriate shape, depending on its bandwidth, to avoid excessive intersymbol interference, as will be shown later.

In frequency modulation, the signal applied to the input of a bandpass communication channel is of the general form

$$E_i(t) = \sin [\omega_0 t + \varphi + \psi_i(t)], \tag{1}$$

where ω_0 is the radian frequency of the unmodulated carrier, φ is the carrier phase and $\psi_i(t)$ is related to the modulating voltage $V_i(t)$ by

$$\psi_i(t) = \bar{\omega}_1 \int_0^t V_i(t)\, dt, \tag{2}$$

with $\bar{\omega}_1$ the frequency deviation in radians per second per volt.

In the case of bipolar binary pulse transmission, the original signal, $V_0(t)$, is ordinarily in the form of rectangular pulses of amplitude V_0 and duration T, of either positive or negative polarity. In this case, the original signal is constant during a signal interval of duration T and is given by

$$V_0(t) = \pm V_0 \,. \tag{3}$$

In general, with a premodulation low-pass filter, the carrier modulating voltage, $V_i(t)$, differs from $V_0(t)$. Without such a filter, as assumed here, $V_i(t) = V_0(t) = \pm V_0$ and

$$\begin{aligned} \psi_i(t) &= \pm\bar{\omega}_1 \int_0^t V_0 \, dt \\ &= \pm\bar{\omega}t, \end{aligned} \tag{4}$$

where the frequency deviation $\bar{\omega}$ is

$$\bar{\omega} = \bar{\omega}_1 V_0 . \tag{5}$$

The voltage applied to the input of the bandpass channel in this case is, in accordance with (1),

$$E_i(t) = \sin [(\omega_0 \pm \bar{\omega})t + \varphi]. \tag{6}$$

Equivalent performance could accordingly be obtained if the outputs of two oscillators of frequency $\omega_0 + \bar{\omega}$ and $\omega_0 - \bar{\omega}$ were gated by the voltage $V_0(t)$, so that carrier step pulses of duration T and one or the other of the above two frequencies would be applied directly to the bandpass channel. If the latter method is actually used, the two oscillators must be interlocked to avoid excessive phase discontinuities and resultant transmission impairments that would otherwise be likely to occur in switching from one oscillator to the other.

With phase modulation, rather than frequency modulation as considered above, $\psi_i(t)$ in (1) is related to the modulating voltage by

$$\psi_i(t) = \psi_1 V_i(t), \tag{7}$$

where ψ_1 is the phase modulation in radians per volt.

In the case of bipolar binary pulse transmission, (1) becomes

$$\begin{aligned} E_i(t) &= \sin [\omega_0 t + \varphi \pm \psi_1 V_i(t)] \\ &= \sin (\omega_0 t + \varphi) \cos [\psi_1 V_i(t)] \pm \cos (\omega_0 t + \varphi) \sin [\psi_1 V_i(t)], \end{aligned} \tag{8}$$

where the negative sign is used for a space and the positive sign for a mark, or conversely.

The first component in (8) is independent of the pulse polarity. In an optimized system this component must be minimized and the second component, which depends on the pulse polarity, be maximized. The optimum condition is obtained when ψ_1 is so chosen that $\psi_1 \hat{V}_i = \pi/2$, where $\hat{V}_i$ is the peak amplitude of $V_i(t)$. In the particular case of rectangular modulating pulses, $V_i(t) = \hat{V}_i = V_0$ and (8) becomes

$$E_i(t) = \pm\cos (\omega_0 t + \varphi), \tag{9}$$

which represents a sudden phase reversal from space to mark.

In amplitude modulation, the signal applied to the bandpass channel is

$$E_i(t) = [a_0 + a_1 V_i(t)/\hat{V}_i] \cos(\omega_0 t + \varphi), \tag{10}$$

where a_0 and a_1 are constants that determine the degree of modulation, as discussed below for two special cases.

In unipolar or "on-off" binary pulse transmission, $a_1 = a_0$ and, in the particular case of rectangular modulating pulses,

$$\begin{aligned} E_i(t) &= a_0(1 \pm 1) \cos(\omega_0 t + \varphi) \\ &= 0 \qquad\qquad\qquad\quad \text{for space} \\ &= 2a_0(\cos \omega_0 t + \varphi) \quad \text{for mark.} \end{aligned} \tag{11}$$

In bipolar AM, $a_0 = 0$ and, for rectangular modulating pulses, (10) becomes

$$E_i(t) = \pm a_1 \cos(\omega_0 t + \varphi), \tag{12}$$

which is identical to (9) with $a_1 = 1$.

With phase reversal or bipolar AM, the signal can be recovered with the aid of a product demodulator, i.e., by homodyne or synchronous detection. To this end, a synchronous demodulating carrier, $\cos(\omega_0 t + \varphi)$, must be derived from or controlled by the signal, which may entail more complicated instrumentation at the receiving end than is required with frequency modulation. Unipolar AM permits the use of simple envelope detection in exchange for a sacrifice in signal-to-noise ratio compared to the other methods. A further disadvantage of unipolar AM is that it is more susceptible to errors during sudden level changes than is bipolar AM or FM.

With any of the above modulation methods the shape of the received pulses depends on that of the modulating pulses and on the transmission-frequency characteristic or "shaping" of the bandpass channel. The appropriate shaping for avoiding intersymbol interference is well known[1,2] for baseband transmission and amplitude and phase modulation systems, and is determined in the following sections for binary FM.

The particular case of rectangular modulating pulses will be considered in detail, and explicit expressions will be derived for appropriate channel shaping to avoid intersymbol interference. The more complicated cases of premodulation and postdetection pulse shaping will be discussed later.

III. TRANSMITTED FREQUENCY-SHIFT WAVE

Let a continuing "space" be represented by a steady-state transmitted wave

$$E_s^{\circ}(t) = \sin[(\omega_0 - \bar{\omega})t + \varphi] \tag{13}$$

and a continuing "mark" by

$$E_m{}^o(t) = \sin [(\omega_0 + \omega)t + \varphi], \tag{14}$$

where t is the time from the beginning of a signal element of duration T and is related to the time with respect to the midpoint of a signal element by

$$t = t_0 + T/2. \tag{15}$$

With (15) in (13) and (14):

$$\begin{aligned} E_s{}^o(t_0) &= \sin [(\omega_0 - \bar{\omega})(t_0 + T/2) + \varphi] \\ &= \sin [(\omega_0 - \bar{\omega})t_0 + \varphi_0 - \bar{\omega}T/2], \end{aligned} \tag{16}$$

$$E_m{}^o(t_0) = \sin [(\omega_0 + \bar{\omega})t_0 + \varphi_0 + \bar{\omega}T/2], \tag{17}$$

where

$$\varphi_0 = \varphi + \omega_0 T/2. \tag{18}$$

It will be assumed that

$$\bar{\omega}T = \pi, \tag{19}$$

in which case the frequency difference between mark and space in cycles per second is $2\bar{\omega}/2\pi = 1/T$, or equal to the bit-rate. This assumption need not be made at this point, but it turns out later to be a condition for avoiding intersymbol interference and simplifies the analysis. With the above assumption,

$$E_s{}^o(t_0) = -\cos [(\omega_0 - \bar{\omega})t_0 + \varphi_0], \tag{20}$$

$$E_m{}^o(t_0) = +\cos [(\omega_0 + \bar{\omega})t_0 + \varphi_0]. \tag{21}$$

Assume that a single mark of duration T is preceded and followed by a continuing space. The resultant transmitted wave can be regarded as made up of two components. One is a steady-state component given by the following expression applying for $-\infty < t_0 < \infty$:

$$E_s{}^o(t_0) = -\cos [(\omega_0 - \bar{\omega})t_0 + \varphi_0]. \tag{22}$$

The other is a transient $E_{sm}{}^o = -E_s{}^o + E_m{}^o$ given by the following expression applying for $-T/2 < t_0 < T/2$:

$$\begin{aligned} E_{sm}{}^o(t_0) &= \cos [(\omega_0 - \bar{\omega})t_0 + \varphi_0] + \cos [(\omega_0 + \bar{\omega})t_0 + \varphi_0] \\ &= \cos (\omega_0 t_0 + \varphi_0) 2 \cos \bar{\omega}t_0 . \end{aligned} \tag{23}$$

The spectrum of $E_{sm}{}^{o}$ is given by

$$\begin{aligned} S_{sm}{}^{o} &= 2\int_{-T/2}^{T/2} \cos\,(\omega_0 t_0 + \varphi_0)\cos\bar{\omega}t_0\, e^{-i\omega t_0}\, dt_0 \\ &= e^{i\varphi_0}\int_{-T/2}^{T/2} \cos ut_0 \cos\bar{\omega}t_0\, dt_0 \\ &\qquad + e^{-i\varphi_0}\int_{-T/2}^{T/2} \cos\,(2\omega_0 + u)t_0 \cos\bar{\omega}t_0\, dt_0\,, \end{aligned} \tag{24}$$

where $u = \omega - \omega_0$ is the frequency from midband.

When the bandwidth of the channel is small in relation to the midband frequency, the spectrum need only be considered for $u \ll \omega_0$. The second integral in (24) can then be disregarded in comparison with the first, and the amplitude of the spectrum becomes independent of φ_0, i.e., independent of the phase of the carrier with respect to the modulating pulse. On this assumption, the amplitude of the spectrum of the carrier envelope at the frequency u from the midband frequency ω_0 is given by the first integral in (24), and becomes

$$S_{sm}{}^{o}(u) = S^{o}(u) = \frac{T}{2}\left[\frac{\sin\,(\bar{\omega} - u)T/2}{(\bar{\omega} + u)T/2} + \frac{\sin\,(\bar{\omega} + u)T/2}{(\bar{\omega} + u)T/2}\right]. \tag{25}$$

With $\bar{\omega}T = \pi$ in accordance with (19), (25) becomes

$$S^{o}(u) = S^{o}(-u) = \frac{T}{2}\,\frac{4}{\pi}\,\frac{\cos\,(\pi u/2\bar{\omega})}{1 - (u/\bar{\omega})^2}\,. \tag{26}$$

For $u = \pm\bar{\omega}$, the value of (26) is*

$$S^{o}(\pm\bar{\omega}) = \frac{T}{2}\,. \tag{27}$$

IV. FREQUENCY-SHIFT PULSE TRANSMISSION CHARACTERISTIC

Let the phase characteristic of the channel be assumed to be linear, and let $A(u)$ be the amplitude characteristic of the channel as a function of the frequency $u = \omega - \omega_0$. At the channel output, the spectrum of the received wave resulting from $E_{sm}{}^{o}(t_0)$ given by (23) is then

$$S(u) = A(u)S^{o}(u). \tag{28}$$

If the amplitude characteristic is symmetrical, i.e., if $A(-u) = A(u)$, the spectrum is also symmetrical; i.e.,

$$S(-u) = S(u). \tag{29}$$

* This result is obtained by determining the value of the limit 0/0 as $u \to \bar{\omega}$.

When the above conditions are satisfied, the shape of a received pulse in response to $E_{sm}^{o}(t_0)$ is given by[2]

$$E_{sm}(t_0) = \cos [\omega_0 t_0 + \varphi_0]\bar{E}_{sm}(t_0), \tag{30}$$

where $\bar{E}_{sm}$ is the envelope of the received pulse. The shape of the envelope is the same as that of a demodulated pulse in AM, when the pulse spectrum is $S(u)$, and is given by[2]

$$p(t_0) = \bar{E}_{sm}(t_0) = \frac{1}{\pi}\int_{-\omega_0}^{\infty} S(u) \cos ut_0 \, du, \tag{31}$$

where the lower limit can, for practical purposes, be replaced by $-\infty$, since $S(u) \approx 0$ for $u = -\omega_0$ when $\bar{\omega} \ll \omega_0$. The received wave in response to the steady-state component $E_s^{o}(t_0)$ given by (22) is

$$\begin{aligned} E_s(t_0) &= -A(-\bar{\omega}) \cos [(\omega_0 - \bar{\omega})t_0 + \varphi_0] \\ &= -A(-\bar{\omega})[\cos (\omega_0 t_0 + \varphi_0) \cos \bar{\omega}t_0 + \sin (\omega_0 t_0 + \varphi_0) \sin \bar{\omega}t_0], \end{aligned} \tag{32}$$

where $A(-\bar{\omega}) = A(\bar{\omega})$ is the amplitude of the transmission-frequency characteristic $A(u)$ at $u = \mp\bar{\omega}$, i.e., at the frequencies $\omega_0 \mp \bar{\omega}$.

The resultant received wave when a mark of duration T is preceded and followed by a continuing space is

$$E(t_0) = E_s(t_0) + E_{sm}(t_0) \tag{33}$$

$$\begin{aligned} &= -\cos (\omega_0 t_0 + \varphi_0)[A(-\bar{\omega}) \cos \bar{\omega}t_0 - p(t_0)] \\ &\qquad -A(-\bar{\omega}) \sin (\omega_0 t_0 + \varphi_0) \sin \bar{\omega}t_0 . \end{aligned} \tag{34}$$

This can be written in the form

$$E(t_0) = \bar{E}(t_0) \cos (\omega_0 t_0 + \varphi_0 + \psi_0), \tag{35}$$

where the envelope is given by

$$\begin{aligned} \bar{E}(t_0) &= A(-\bar{\omega})\{[\cos \bar{\omega}t_0 - \mu p]^2 + \sin^2 \bar{\omega}t_0\}^{1/2} \\ &= A(-\bar{\omega})\{1 + \mu^2 p^2 - 2\mu p \cos \bar{\omega}t_0\}^{1/2} \end{aligned} \tag{36}$$

and the phase modulation ψ_0 is given by

$$\tan \psi_0(t_0) = -\frac{\sin \bar{\omega}t_0}{\cos \bar{\omega}t_0 - \mu p}, \tag{37}$$

where $p = p(t_0)$ and

$$\mu = 1/A(-\bar{\omega}) = 1/A(\bar{\omega}). \tag{38}$$

With an ideal detector, the received signal is proportional to $\psi_0' = d\psi_0/dt$, which becomes

$$\psi_0' = -\bar{\omega}\,\frac{1 - \mu p \cos \bar{\omega} t_0 + (\mu p'/\bar{\omega}) \sin \bar{\omega} t_0}{\sin^2 \bar{\omega} t_0 + [\cos \bar{\omega} t_0 - \mu p]^2}, \tag{39}$$

where $p' = dp(t_0)/dt_0$.

Relation (39) gives the frequency deviation from the midband frequency ω_0, in which case a continuing space is represented by a frequency deviation $-\bar{\omega}$ and a continuing mark by $\bar{\omega}$. If the frequency $\omega_0 - \bar{\omega}$ is instead used as reference, the frequency deviation at the receiving end is

$$\psi'(t_0) = \bar{\omega} + \psi_0'(t_0). \tag{40}$$

The ratio $\psi'(t_0)/2\bar{\omega}$ represents the pulse transmission characteristic of the channel in response to a sudden frequency shift $2\bar{\omega}$ of duration T, from $\omega_0 - \bar{\omega}$ to $\omega_0 + \bar{\omega}$, and is given by

$$P(t_0) = \frac{\mu}{2}\,\frac{\mu p^2 - \cos \bar{\omega} t_0 - (p'/\bar{\omega}) \sin \bar{\omega} t_0}{\sin^2 \bar{\omega} t_0 + (\cos \bar{\omega} t_0 - \mu p)^2}. \tag{41}$$

From (39) or (41) the conditions for binary pulse transmission without intersymbol interference can be established, as is discussed in the next section.

V. IDEAL FREQUENCY-SHIFT TRANSMISSION CHARACTERISTICS

In order to transmit binary pulse trains without intersymbol interference, it is necessary that the transmission characteristic $P(t_0)$ for a single mark or pulse as considered in the preceding section be zero at sampling instants $t_0 = \pm mT$, $m = 1, 2, 3$, etc., and that, at $t_0 = 0$, $P = 1$.

In view of (19), $\cos(\pm m\bar{\omega}T) = (-1)^m$ and $\sin(\pm m\bar{\omega}T) = 0$. Hence at sampling points (41) becomes, for $m \neq 0$,

$$P(mT) = \frac{1}{2}\,\frac{\mu p(mT)}{\mu p(mT) + (-1)^m} \tag{42}$$

and

$$P(0) = \frac{1}{2}\,\frac{\mu p(0)}{\mu p(0) - 1}. \tag{43}$$

Thus $P(mT) = 0$ provided

$$p(mT) = 0, \tag{44}$$

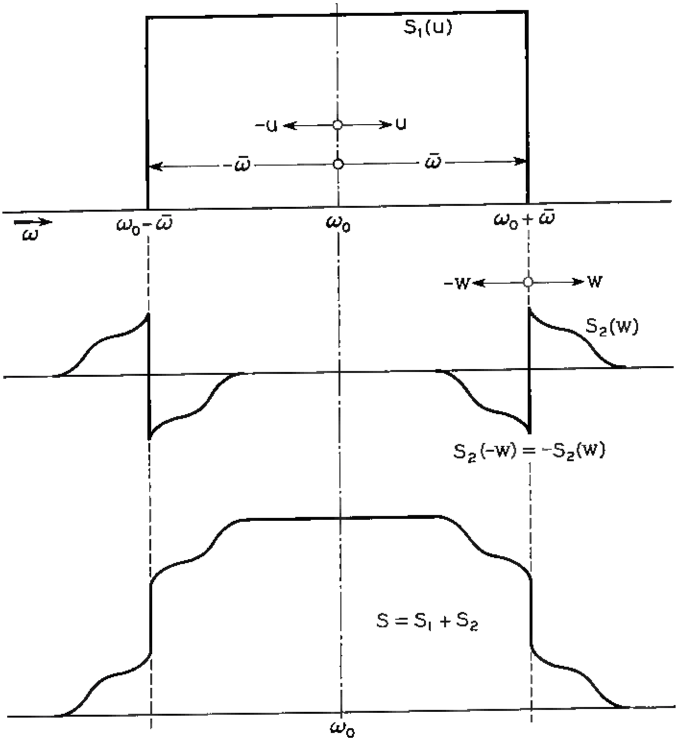


Fig. 1 — Properties of spectra at detector input such that, in double-sideband amplitude modulation, $p(0) = 1$, $p(mT) = 0$, $T = \pi/\bar{\omega}$, $m = 1, 2, 3, \cdots$.

and $P(0) = 1$ provided

$$\mu p(0) = 2. \tag{45}$$

The above conditions (44) and (45) can be satisfied provided that $S(u)$, in addition to being symmetrical as required by (31), has the further property illustrated in Fig. 1. This property is the same as that required for double-sideband transmission of pulses at intervals T without intersymbol interference,[1,2] and can be satisfied by an infinite variety of spectra. Among these it is convenient from the standpoint of theoretical evaluation of $p(t_0)$ to assume spectra of the form shown in Fig. 2 given by the following expressions.

In the range $0 < u < \bar{\omega} - \omega_x$:

$$S(u) = S(-u) = \frac{T}{2}. \tag{46}$$

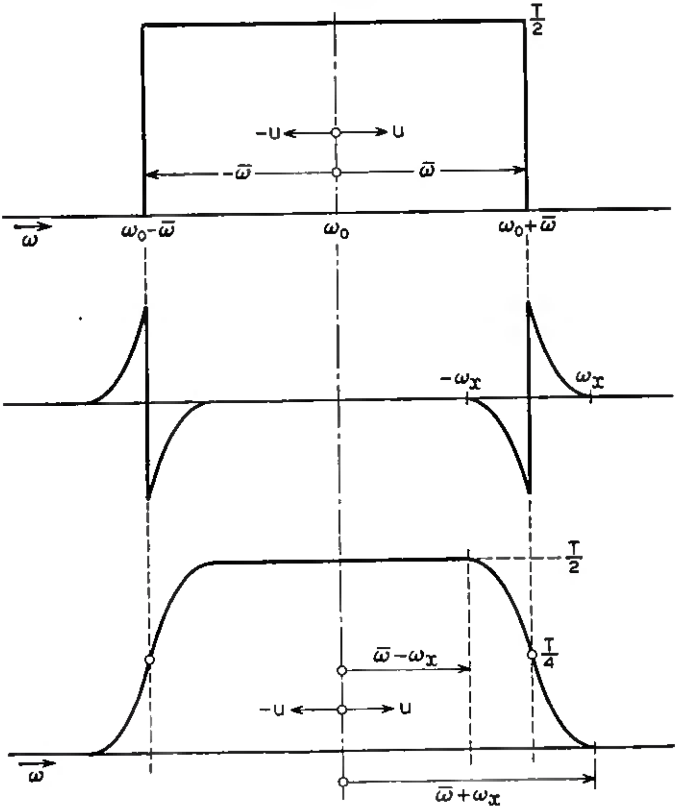


Fig. 2 — Special case of spectra with properties indicated in Fig. 1.

In the range $\bar{\omega} + \omega_x > u > \bar{\omega} - \omega_x$:

$$S(u) = S(-u) = \frac{T}{2} \cos^2 \left(\frac{\pi}{4} \frac{u - \bar{\omega} + \omega_x}{\omega_x} \right). \tag{47}$$

At $u = \bar{\omega}$, (47) becomes

$$S(\bar{\omega}) = S(-\bar{\omega}) = \frac{T}{4} . \tag{48}$$

The required amplitude characteristic of the channel obtained from (28) is

$$A(u) = \frac{S(u)}{S^o(u)} . \tag{49}$$

In view of (27) and (48), $A(-\bar{\omega}) = \frac{1}{2}$, so that (38) gives

$$\mu = 2. \tag{50}$$

When the spectrum is given by (46) and (47), evaluation of (31) gives[2]

$$p(t_0) = \bar{E}_{sm}(t_0) = \frac{T\bar{\omega}}{\pi} \frac{\sin \bar{\omega} t_0}{\bar{\omega} t_0} \frac{\cos \omega_x t_0}{1 - (2\omega_x t_0/\pi)^2}, \tag{51}$$

where $T\bar{\omega}/\pi = 1$, in view of (19), so that

$$\begin{aligned} p(t_0) &= \frac{\sin \bar{\omega} t_0}{\bar{\omega} t_0} \frac{\cos \omega_x t_0}{1 - (2\omega_x t_0/\pi)^2} \\ &= 1 \qquad \text{for} \qquad t_0 = 0 \\ &= 0 \qquad \text{for} \qquad t_0 = mT, \qquad m = 1, 2, 3, \cdots . \end{aligned} \tag{52}$$

Hence, $\mu p(0) = 2$ and $p(mT) = 0$ so that conditions (44) and (45) are satisfied, and single binary pulses can be transmitted without intersymbol interference. This is also the case for pulse trains, as is shown in Section XI.

To find the shape of the received pulses, it is necessary to employ (41) for other values of t_0 than $t_0 = 0$ and mT, as is illustrated in the next sections for two limiting cases of general interest.

VI. SPECIAL CASE OF FLAT SPECTRUM AT DETECTOR INPUT

The amplitude characteristic $A(u)$ of the channel and the frequency-shift pulse transmission characteristic $P(t_0)$ will be determined here for a channel of minimum bandwidth, in which case the spectrum $S(u)$ will be flat for $0 < u < \bar{\omega}$, and will be zero for $u > \bar{\omega}$.

With sharp cutoffs at $u = \pm\bar{\omega}$, ω_x will be zero in (46) and (47), so that

$$S(u) = S(-u) = \frac{T}{2} \qquad 0 \leqq u \leqq \bar{\omega} \tag{53}$$

$$= 0 \qquad\qquad u > \bar{\omega}. \tag{54}$$

The required amplitude characteristic of the channel, as obtained from (49), is

$$\begin{aligned} A(u) &= \frac{\pi}{4} \frac{1 - (u/\bar{\omega})^2}{\cos (\pi u/2\bar{\omega})} \\ &= \frac{\pi}{4} \qquad \text{for} \qquad u = 0 \\ &= 1 \qquad \text{for} \qquad u = \pm\bar{\omega} \\ &= 0 \qquad \text{for} \qquad |u| > \bar{\omega}_0. \end{aligned} \tag{55}$$

In the case of double-sideband AM, the spectrum of a pulse of duration T with respect to the midband frequency is

$$s^{o}(u) = \frac{T}{2}\frac{\sin (uT/2)}{uT/2}, \tag{56}$$

where, as before, it has been assumed that the channel bandwidth is small in relation to the midband frequency.

To obtain a flat spectrum of amplitude $T/2$ at the detector input, the required amplitude characteristic of the channel is

$$\begin{aligned} a(u) &= \frac{uT/2}{\sin (uT/2)} = \frac{\pi u/2\bar{\omega}}{\sin (\pi u/2\bar{\omega})} \\ &= 1 \qquad \text{for} \quad u = 0 \qquad\qquad (57) \\ &= \frac{\pi}{2} \qquad \text{for} \quad u = \bar{\omega}. \end{aligned}$$

The amplitude characteristics $A(u)$ and $a(u)$ are shown in Fig. 3.

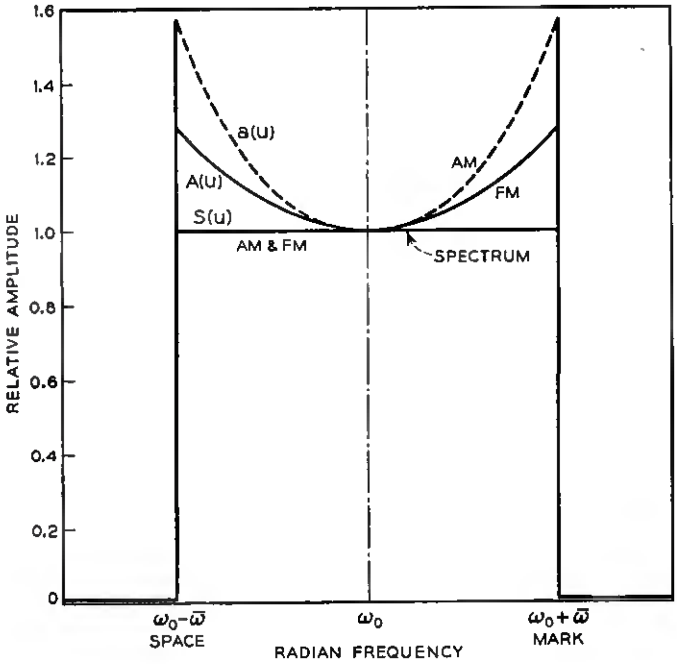


Fig. 3 — Ideal transmission-frequency characteristics of bandpass channels in AM and FM for rectangular modulating pulses and flat spectrum of minimum bandwidth at detector input.

For purposes of direct comparison, the value of $A(u)$ as obtained from (55) has been normalized to $A(u) = 1$ for $u = 0$, by multiplication by $4/\pi$.

With a channel characteristic as given by (55), $p(t_0)$, as obtained from (52) with $\omega_x = 0$, is

$$p(t_0) = \frac{\sin \bar{\omega} t_0}{\bar{\omega} t_0}, \tag{58}$$

$$p'/\bar{\omega} = \frac{1}{\bar{\omega} t_0} \left(\cos \bar{\omega} t_0 - \frac{\sin \bar{\omega} t_0}{\bar{\omega} t_0} \right). \tag{59}$$

With (58) and (59) in (41), the frequency-shift pulse transmission characteristic of the channel becomes

$$P(t_0) = \frac{\sin \bar{\omega} t_0}{\bar{\omega} t_0} \; \frac{3 \dfrac{\sin \bar{\omega} t_0}{\bar{\omega} t_0} - 2 \cos \bar{\omega} t_0}{1 + 4 \dfrac{\sin \bar{\omega} t_0}{\bar{\omega} t_0} \left(\dfrac{\sin \bar{\omega} t_0}{\bar{\omega} t_0} - \cos \bar{\omega} t_0 \right)}. \tag{60}$$

The function $P(t_0)$ is given in Table I and shown in Fig. 4.

The term $\sin \bar{\omega} t_0 / \bar{\omega} t_0$ in (60) represents the pulse-transmission characteristic for double-sideband transmission over a channel with an amplitude characteristic $a(u)$. In Fig. 4 the double-sideband AM and the frequency-shift transmission characteristics are compared. It will be noted that they differ appreciably in shape, but have the common properties of unit amplitude at $t_0 = 0$ and zero amplitude at intervals such that $\bar{\omega} t_0 = m\pi$, $m = 1, 2, 3, \cdots$.

VII. SPECIAL CASE OF RAISED COSINE SPECTRUM AT DETECTOR INPUT

In actual communication systems, channels with sharp cutoffs as assumed in the previous example are impracticable for various reasons, such as excessive phase distortion near the band edges and relatively

TABLE I — FREQUENCY-SHIFT TRANSMISSION CHARACTERISTIC $P(t_0)$ FOR FLAT SPECTRUM OF MINIMUM BANDWIDTH

$\bar{\omega} t_0/\pi$	0	0.20	0.40	0.60	0.80	1.0
0	1.0	0.7860	0.5468	0.4182	0.2746	0
1	0	−0.3026	0.0072	0.1629	0.1454	0
2	0	−0.1540	−0.0330	0.0940	0.0985	0
3	0	−0.1023	−0.0339	0.0646	0.0743	0
4	0	−0.0766	−0.0303	0.0488	0.0597	0
5	0	−0.0611	−0.0267	0.0391	0.0499	0
10	0	−0.0304	−0.0160	0.0194	0.0274	0
19	0	−0.0160	−0.0091	0.0101	0.0151	0

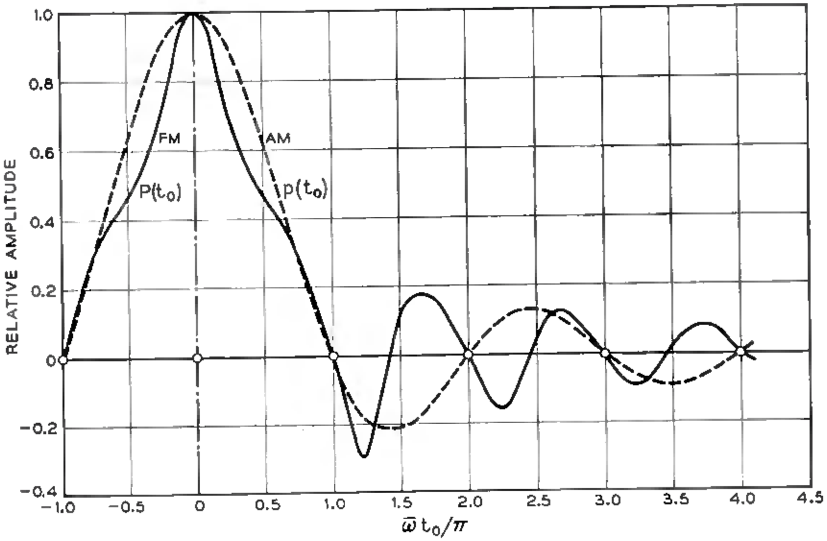


Fig. 4 — Shapes of demodulated pulses in FM and AM for rectangular modulating pulses and for flat spectrum of minimum bandwidth at detector input, as in Fig. 3.

large oscillations in the received pulses that entail precise synchronized sampling at fixed intervals to avoid intersymbol interference. A preferable type of channel characteristic is one that results in a "raised cosine" spectrum of the received wave in response to the transmission of a single rectangular pulse of duration T, as considered below. With this type of channel characteristic, the bandwidth is twice the minimum possible considered in the previous limiting case.

With $\omega_x = \bar{\omega}$ in (47),

$$\begin{aligned} S(u) = S(-u) &= \frac{T}{2}\cos^2(\pi u/4\bar{\omega}) \qquad u \leqq 2\bar{\omega} \\ &= 0 \qquad\qquad\qquad\qquad\quad u > 2\bar{\omega}. \end{aligned} \tag{61}$$

The required amplitude characteristic of the channel as obtained from (49) is

$$\begin{aligned} A(u) &= \frac{\pi}{4}\frac{1-\cos(\pi u/2\bar{\omega})}{2\cos(\pi u/2\bar{\omega})}[1-(u/\bar{\omega})^2] \\ &= \frac{\pi}{4} \qquad \text{for} \quad u = 0 \\ &= \tfrac{1}{2} \qquad \text{for} \quad u = \bar{\omega} \qquad (\mu = 2) \\ &= 0 \qquad \text{for} \quad u \geqq 2\bar{\omega}. \end{aligned} \tag{62}$$

The amplitude characteristic required for double-sideband transmission without intersymbol interference is, in this case, given by

$$
\begin{aligned}
a(u) &= \frac{\cos^2 (\pi u/4\bar{\omega})\pi u/2\bar{\omega}}{\sin (\pi u/2\bar{\omega})} \\
&= \frac{\pi u/4\bar{\omega}}{\tan (\pi u/4\bar{\omega})} \\
&= 1 \qquad \text{for} \quad u = 0 \\
&= \frac{\pi}{4} \qquad \text{for} \quad u = \bar{\omega} \\
&= 0 \qquad \text{for} \quad u \geqq 2\bar{\omega}.
\end{aligned}
\tag{63}
$$

The amplitude characteristics $A(u)$ and $a(u)$ are shown in Fig. 5. For convenient direct comparison, the value of $A(u)$ obtained from (62) has been normalized to $A(u) = 1$ at $u = 0$.

With $\omega_x = \bar{\omega}$ in (52),

$$
p(t_0) = \frac{\sin 2\bar{\omega}t_0}{2\bar{\omega}t_0[1 - (2\bar{\omega}t_0/\pi)^2]}, \tag{64}
$$

where the relation $\cos \bar{\omega}t_0 \sin \bar{\omega}t_0 = \frac{1}{2} \sin 2\bar{\omega}t_0$ has been used.

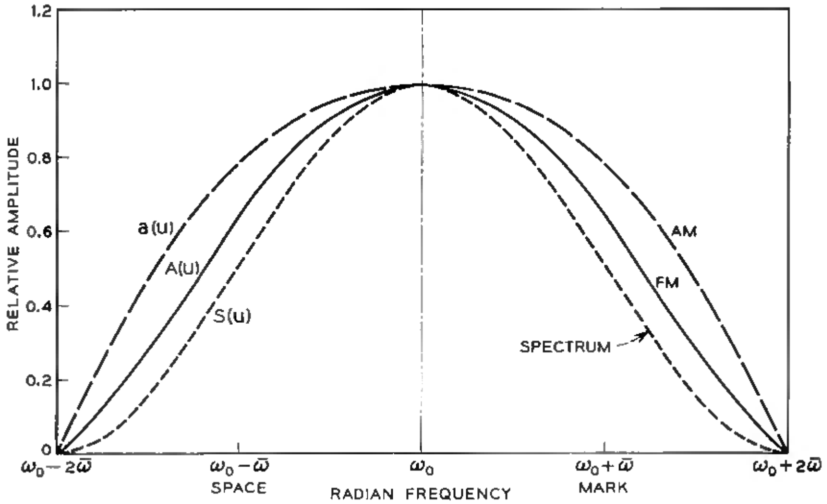


Fig. 5 — Ideal transmission-frequency characteristics of bandpass channels in FM and AM for rectangular modulating pulses and raised cosine spectrum at detector input.

TABLE II — FREQUENCY-SHIFT TRANSMISSION CHARACTERISTIC $P(t_0)$ FOR RAISED COSINE SPECTRUM AT DETECTOR INPUT

$\bar{\omega}t_0/\pi$		0.2	0.4	0.5	0.6	0.8
0	1.0	0.8053	0.5769	0.4887	0.4030	0.2035
1	0	−0.0175	0.0301	0.0265	0.0109	−0.0108
2	0	0.0035	−0.0048	−0.0053	−0.0027	0.0025
3	0	−0.0012	0.0016	0.0019	0.0011	−0.0009
4	0	0.0005	−0.0007	−0.0009	−0.0005	0.0004

Differentiation of (64) yields

$$\frac{p'}{\bar{\omega}} = 2\,\frac{\cos 2\bar{\omega}t_0}{2\bar{\omega}t_0[1 - (2\bar{\omega}t_0/\pi)^2]} - 2\,\frac{\sin 2\bar{\omega}t_0[1 - 3(2\bar{\omega}t_0/\pi)^2]}{(2\bar{\omega}t_0)^2[1 - (2\bar{\omega}t_0/\pi)^2]^2}. \qquad (65)$$

For $\bar{\omega}t_0 = \pi/2$, $p'/\bar{\omega} = -3/(2\pi)$.

With (64) and (65) in (41) the frequency-shift transmission characteristic $P(t_0)$ given in Table II and shown in Fig. 6 is obtained, for a channel with an amplitude characteristic $A(u)$ as shown in Fig. 5.

For comparison with $P(t_0)$, Fig. 6 also shows the pulse-transmission characteristic $p(t_0)$ obtained from (64) for double-sideband transmission over a channel with the amplitude characteristic $a(u)$ shown in Fig. 5. In both cases, the oscillations in the received pulses are quite small, and for this reason the transmission-frequency characteristics shown in Fig. 5 are preferable to those in Fig. 3 in practicable AM and FM systems.

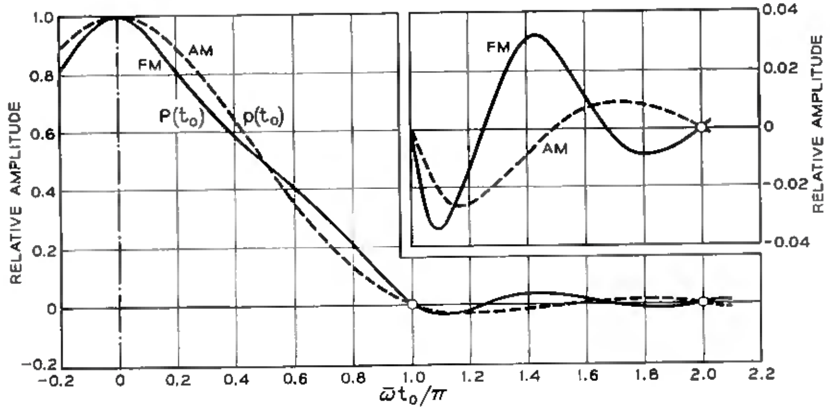


Fig. 6 — Shapes of demodulated pulses in FM and AM for rectangular modulating pulses and raised cosine spectrum at detector input, as in Fig. 5.

VIII. PREMODULATION PULSE SHAPING

In the preceding analysis the modulating pulses were assumed to be rectangular in shape and of duration T, in which case the modulator input in response to a change from space to mark was given by (23) and the corresponding spectrum by (24). In the more general case of premodulation pulse shaping, these equations are replaced by

$$E_{sm}^{o}(t_0) = 2\cos(\omega_0 t_0 + \varphi_0)\cos\psi_i(t_0), \tag{66}$$

$$S^{o}(u) = 2\int_0^{\infty}\cos\psi_i(t_0)\cos ut_0\,dt_0, \tag{67}$$

where, as before, the second component in (24) has been neglected, and the phase modulation ψ_i is related to the modulating voltage $V_i(t_0)$ by (2), or

$$\psi_i(t_0) = \bar{\omega}_1\int_0^{t_0} V_i(t_0)\,dt_0. \tag{68}$$

The above relations apply provided a continuing space is represented by a constant frequency deviation $-\bar{\omega}$ and a continuing mark by $\bar{\omega}$. To this end, it is necessary that the individual modulating pulses $V_i(t_0)$ overlap and be of such form that

$$\sum_{n=-\infty}^{\infty} V_i(t_0 + nT) = V_i(0). \tag{69}$$

For example, the latter relation is satisfied when impulses are applied to a flat low-pass filter of bandwidth $\bar{\omega}$ and linear phase, resulting in a modulating voltage $V_i(t) = V_i(0)(\sin\bar{\omega}t_0)/\bar{\omega}t_0$. The simpler case of overlaps between adjacent pulse intervals only is illustrated in Fig. 7.

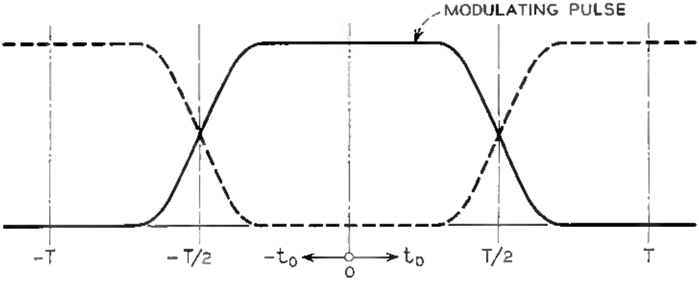


Fig. 7 — Modulating pulses with overlap between adjacent pulse intervals.

The amplitude characteristic of the channel required for a spectrum $S(u)$ at the detector input is, as before, given by

$$A(u) = \frac{S(u)}{S^o(u)}, \tag{70}$$

To avoid intersymbol interference, it is necessary to satisfy (45); i.e., $\mu p(0) = 2$. To this end, it is necessary, in accordance with the discussion in Section V, that $S(\pm\bar{\omega}) = T/4$ and that $A(\bar{\omega}) = \frac{1}{2}$. Hence, a requirement imposed on $S^o(u)$ as given by (67) is that

$$S^o(\bar{\omega}) = \frac{T}{2}. \tag{71}$$

There is an infinity of pulse shapes, $V_i(t_0)$, that satisfy (69), and the corresponding $\psi_i(t_0)$ can be determined formally from (68) and, in turn, $S^o(u)$ can be determined from (67). The principal problem is to determine pulse shapes, other than the rectangular shape considered previously, which also have a spectrum that satisfies (71). The solution of this problem will not be attempted here, but two pulse shapes of general interest in pulse transmission theory will be considered.

A familiar example of modulating pulses that overlap into an infinite number of pulse intervals is represented by the idealized pulse shape obtained by applying impulses to an ideal flat low-pass filter with linear phase characteristic. It will be assumed that the filter bandwidth is $\bar{\omega}$, in which case the modulating voltage is

$$V_i(t_0) = V_i(0)\,\frac{\sin \bar{\omega}t_0}{\bar{\omega}t_0}. \tag{72}$$

With (72) in (68), and with $\bar{\omega}_1 V_i(0) = \bar{\omega}$,

$$\psi_i(t_0) = \mathrm{Si}\,(\bar{\omega}t_0) \tag{73}$$

where Si is the sine integral function.

In Fig. 8 the phase modulation, ψ_i, obtained from (73) is compared with that for rectangular modulating pulses, for which $\psi_i = \bar{\omega}t_0$ for $\bar{\omega}t_0 \leqq \pi/2$ and $\psi_i = \pi/2$ for $\bar{\omega}t_0 > \pi/2$.

With (73) in (67) and $\bar{\omega}t_0 = \tau$, $\bar{\omega}T = \pi$, $a = u/\bar{\omega}$:

$$S^o(u) = \frac{T}{2}\,\frac{4}{\pi}\int_0^\infty \cos \mathrm{Si}\,(\tau) \cos a\tau \, d\tau. \tag{74}$$

For $\tau > 17$ the following approximation applies:

$$\mathrm{Si}\,(\tau) \cong \frac{\pi}{2} - \frac{\cos \tau}{\tau}. \tag{75}$$

The peak amplitudes, $\hat{p}(0)$, required to produce an error when an impulse occurs midways between sampling points is greater than when they occur at a sampling point by a factor of $2^{1/2}$ in FM and a factor of 2 in AM. With a Gaussian amplitude distribution of the pulses, the probability of an error from a pulse midway between two sampling points is in the order of 1 per cent of the probability of an error from a pulse at a sampling point in the case of FM, and is substantially smaller for AM. Hence, virtually all the errors will be caused by pulses that occur near sampling points. The AM advantage over FM for equal error probability is 3 db for impulses that occur at sampling points, and would be expected to be only slightly greater, about 4 db when impulses occurring at all instances with respect to a sampling point are considered.

The above comparisons apply without a postdetection low-pass filter in FM. With an optimum bandpass receiving filter characteristic in FM, the reduction in peak impulse noise afforded by low-pass filter would be expected to be about the same as the reduction in average random noise.

APPENDIX C

Optimum Receiving Filter Characteristic

The optimum receiving filter characteristic in AM and in FM without a postdetection low-pass filter can be determined from the solution of the more general case considered here, of FM with a postdetection filter.

In the latter case, the optimum $R(u)$ is obtained when the product of the two integrals in (125) is a minimum, or for the minimum value of the product:

$$J = J_1 J_2 , \tag{256}$$

where J_1 and J_2 are functions of $R(u)$ given by

$$J_1 = \int_{-\infty}^{\infty} L^2(u) R^2(u) (1 + u/\bar{\omega})^2 \, du , \tag{257}$$

$$J_2 = \int_{-\infty}^{\infty} \frac{S^2(u)}{R^2(u)} \, du = 2 \int_0^{\infty} \frac{S^2(u)}{R^2(u)} \, du . \tag{258}$$

In (257), $L(-u) \neq L(u)$, so that it is convenient to resolve the integrand into one component with even symmetry with respect to u and one with odd symmetry. The integral of the latter component vanishes and that of the component with even symmetry becomes

$$J_1 = \int_0^{\infty} H^2(u) R^2(u) \, du , \tag{259}$$

raised cosine pulse, which overlaps into adjacent pulse intervals, and is considered in the next section. With the latter type of pulse it turns out that $S^{o}(\bar{\omega}) = 0.994\ T/2$. It can thus be conjectured that there is some intermediate shape of overlapping modulating pulses for which (71) is satisfied. For practical purposes, this is the case with modulating pulses given by (64) or for the raised cosine pulses considered in the next section.

IX. RAISED COSINE MODULATING PULSES

Raised cosine modulating pulses have the shape indicated in Fig. 9, and can be derived conveniently by appropriate gating of a biased steady-state sine wave, with the total pulse duration, $2T$, equal to one cycle of the sine wave, which may have advantages from the standpoint of instrumentation.

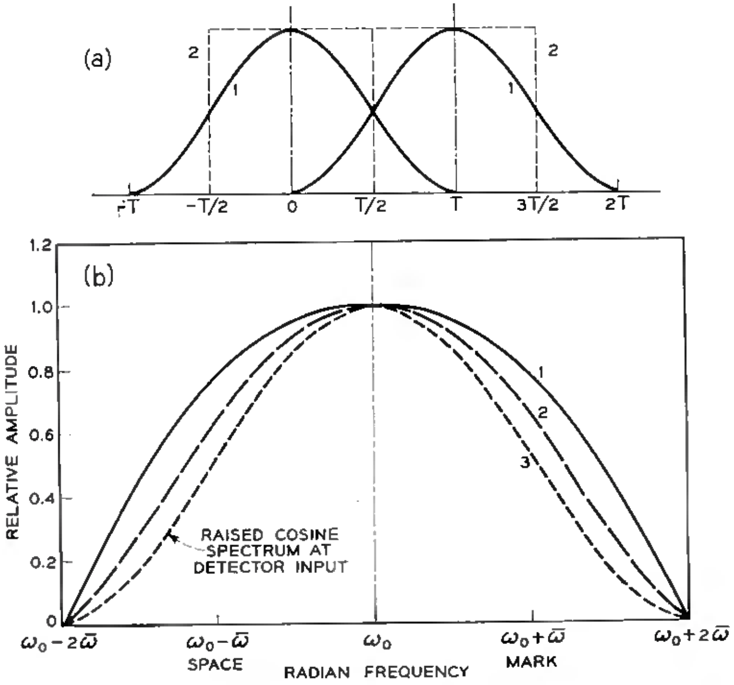


Fig. 9 — Comparison of transmission-frequency characteristics of bandpass channel in FM for raised cosine modulating pulses (solid curve) and rectangular modulating pulses (dashed curve), with raised cosine spectrum at detector input (dotted curve) in both cases. Pulse shapes are same as those for FM in Fig. 6.

With a raised cosine modulating voltage,

$$\begin{aligned} V_i(t_0) &= \frac{V_0}{2}[1 + \cos(\pi t_0/T)] \qquad -1 < t_0/T < 1 \\ &= 0 \qquad\qquad\qquad\qquad\qquad -1 > t_0/T > 1. \end{aligned} \tag{78}$$

In accordance with (2), the resultant phase modulation is, for $\bar{\omega} t_0 \leqq \pi$,

$$\begin{aligned} \psi_i(t_0) &= \frac{\bar{\omega}_1 V_0}{2}\int_0^{t_0}[1 + \cos(\pi t_0/T)]\,dt_0 \\ &= \tfrac{1}{2}(\bar{\omega} t_0 + \sin \bar{\omega} t_0). \end{aligned} \tag{79}$$

For $\bar{\omega} t_0 \geqq \pi$, $\psi_i(t_0) = \psi_i(T) = \pi/2$.

The spectrum of $\cos \psi_i(t_0)$ is obtained from (67) with the upper limit equal to T, since $\cos \psi_i(t_0) = 0$ for $t_0 > T$. The following relation is thus obtained:

$$S^o(u) = 2\int_0^T \cos \tfrac{1}{2}(\bar{\omega} t_0 + \sin \bar{\omega} t_0) \cos u t_0 \, dt_0 \,. \tag{80}$$

With $\bar{\omega} t_0 = x$, $\bar{\omega} T = \pi$ and $u/\bar{\omega} = a$, (80) becomes

$$\begin{aligned} S^o(u) &= \frac{2}{\bar{\omega}}\int_0^\pi \cos(x/2 + \tfrac{1}{2}\sin x)\cos ax\,dx \\ &= \frac{T}{\pi}\int_0^\pi \cos(\nu_1 x + \tfrac{1}{2}\sin x)\,dx \\ &\quad + \frac{T}{\pi}\int_0^\pi \cos(\nu_2 x + \tfrac{1}{2}\sin x)\,dx\,, \end{aligned} \tag{81}$$

where

$$\nu_1 = \tfrac{1}{2} - a, \qquad \nu_2 = \tfrac{1}{2} + a.$$

The above relation can be written

$$S^o(u) = T\,[J_{\nu_1}(-\tfrac{1}{2}) + J_{\nu_2}(-\tfrac{1}{2})], \tag{82}$$

where $J_\nu(z)$ is a so-called Anger function, which is associated with Bessel functions and is defined by[8]

$$J_\nu(z) = \frac{1}{\pi}\int_0^\pi \cos(\nu x - z \sin x)\,dx, \tag{83}$$

$$
\begin{aligned}
J_\nu(z) = \frac{\sin \nu\pi}{\nu\pi} \Bigg[1 &- \frac{z^2}{2^2 - \nu^2} + \frac{z^4}{(2^2 - \nu^2)(4^2 - \nu^2)} \\
&- \frac{z^6}{(2^2 - \nu^2)(4^2 - \nu^2)(6^2 - \nu^2)} + \cdots \Bigg] \\
+ \frac{\sin \nu\pi}{\pi} \Bigg[& \frac{z}{1^2 - \nu^2} - \frac{z^3}{(1^2 - \nu^2)(3^2 - \nu^2)} \\
&+ \frac{z^5}{(1^2 - \nu^2)(3^2 - \nu^2)(5^2 - \nu^2)} - \cdots \Bigg].
\end{aligned}
\tag{84}
$$

The spectrum as a function of a obtained from (82) and (84) is given in Table III, together with that for a rectangular modulating wave, as obtained from (26).

TABLE III — VALUES OF $(2/T)S^o(u)$ FOR RAISED COSINE AND RECTANGULAR MODULATING PULSES

$a = u/\bar{\omega}$	0	0.5	1	1.5	2
Raised cosine.	1.58	1.39	0.994	0.544	0.282
Rectangular.	1.28	1.20	1.000	0.720	0.425

To obtain a raised cosine spectrum at the detector input, as given by (61), the required amplitude characteristic of the bandpass channel obtained from (70) is

$$
A(-u) = A(u) = \frac{\cos^2 (\pi u/4\bar{\omega})}{2[J_{\nu_1}(-\frac{1}{2}) + J_{\nu_2}(-\frac{1}{2})]}, \tag{85}
$$

which gives for various values of $a = u/\bar{\omega}$

$a = u/\bar{\omega}$:	0	0.5	1	1.5	2
$A(u)$:	0.63	0.61	0.503	0.275	0.

Since $A(\bar{\omega}) = 0.503$, rather than 0.50, the factor $\mu = 2$ in (43) is replaced by $\mu = 2.012$. The peak pulse amplitude at $t_0 = 0$ as obtained from (43) is, in this case,

$$
P(0) = \frac{1}{2} \frac{2.012}{2.012 - 1} = 0.994.
$$

Thus, intersymbol interference in this case results in a slight reduction in the peak amplitude of a pulse.

In Fig. 9 the above channel characteristic is compared with that for rectangular modulating pulses. In both cases the shape of the demodulated pulses would be as shown in Fig. 6 for FM, aside from the slightly smaller peak amplitude, 0.994 rather than 1.00. It can be shown that, with raised cosine modulating pulses, the error involved in neglecting the second integral in (24) is not appreciable, even when the total channel band width is equal to the midband frequency. Hence, the channel shaping shown in Fig. 9 for raised cosine modulating pulses applies without restriction on channel bandwidth relative to midband frequency.

X. POSTDETECTION PULSE SPECTRA AND FILTERING

At the detector output a low-pass filter may be desirable for final pulse shaping, elimination of unwanted demodulation products or higher-frequency noise components, as noted in Section II. The appropriate transmission-frequency characteristic of such filters depends on the spectra of the demodulated pulses, as discussed here.

Let $P(t_0)$ be the shape of the pulses at the detector output in FM as given for rectangular modulating pulses by (41) and illustrated in Figs. 4 and 6 for two special cases. These pulses have a baseband spectrum given by

$$S_0(\omega) = \int_{-\infty}^{\infty} P(t_0)e^{-i\omega t_0}\, dt_0 \tag{86}$$

or, since $P(-t_0) = P(t_0)$, by

$$S_0(\omega) = 2\int_0^{\infty} P(t_0) \cos \omega t_0 \, dt_0 . \tag{87}$$

In accordance with the definition in Section IV, $P(t_0) = \psi'(t_0)/2\bar{\omega}$. Hence, for $\omega = 0$, (87) becomes

$$\begin{aligned} S_0(0) &= \frac{1}{2\bar{\omega}} \int_{-\infty}^{\infty} \psi'(t_0)\, dt_0 \\ &= \frac{1}{2\bar{\omega}} [\psi(\infty) - \psi(-\infty)]. \end{aligned} \tag{88}$$

In view of (19), the phase change caused by transmission of a mark preceded and followed by a continuing space, is $\psi(\infty) - \psi(-\infty) = 2\bar{\omega}T = 2\pi$. Hence (88) becomes

$$S_0(0) = T. \tag{89}$$

In the case of AM, the spectrum of the baseband pulses is equal to the spectrum at the detector input above the midband frequency multiplied by a factor of 2, because of the direct addition of the two sideband spectra. From Fig. 2 it follows that, in this case, $S_0(\mathbf{0}) = T$. Relation (89) thus shows that the dc component of the demodulated pulses is the same in FM as in AM. That is, the areas under the FM and AM pulses shown in Figs. 4 and 5 are equal.

For equal spectra at the detector input in AM and FM, the pulse shape $P(t_0)$ has a nonlinear relation (41) to the pulse shape $p(t_0)$ in AM. For this reason, the bandwidth of the demodulated pulses will be greater in FM than in AM, and, in view of the appearance of $p(t_0)$ in the denominator of (41), the bandwidth is theoretically infinite. For this reason, part of the spectrum will be eliminated by any postdetection low-pass filter, and intersymbol interference is thereby introduced unless the filter has an appropriate amplitude characteristic.

With the aid of a postdetection low-pass filter having an amplitude characteristic $A_0(\omega)$ and a linear phase characteristic, it is possible to modify the spectrum S_0 into a desired spectrum S_m with such properties that intersymbol interference is absent. To this end, the amplitude characteristic would be so chosen that

$$A_0(\omega)S_0(\omega) = S_m(\omega). \tag{90}$$

For example, by appropriate choice of $A_0(\omega)$ the pulse shape shown in Fig. 4 for FM could be modified into that shown for AM, or into other shapes. The principal difficulty resides in the determination of the spectrum $S_0(\omega)$ from (87), which entails numerical integration, in view of the fairly complicated expressions for $P(t_0)$.

The spectrum obtained by numerical integration of (87) is given in Table IV for the special case of a flat spectrum of minimum bandwidth at the detector input, as considered in Section VI. In the numerical integration, contributions to the integral were neglected for $\bar{\omega}t_0 = \tau > 20$.

The above spectrum is shown in Fig. 10, together with the baseband

TABLE IV — $S_0(\omega)/T$ FOR FLAT SPECTRUM AT DECTECTOR INPUT

$u/\bar{\omega}$	0	0.25	0.5	0.75	0.9
0	1*	0.8779	0.7561	0.6254	0.5385
1	0.4755	0.2932	0.0626	−0.2370	−0.4609
2	−0.1321	0.2871	0.2031	0.4186	−0.0696
3	0.0396	−0.0233	−0.0529	−0.0220	0.0418
4	0.1105	0.0650	0.0270	0.0003	—

* Based on (89); compares with computed value of 0.99999654.

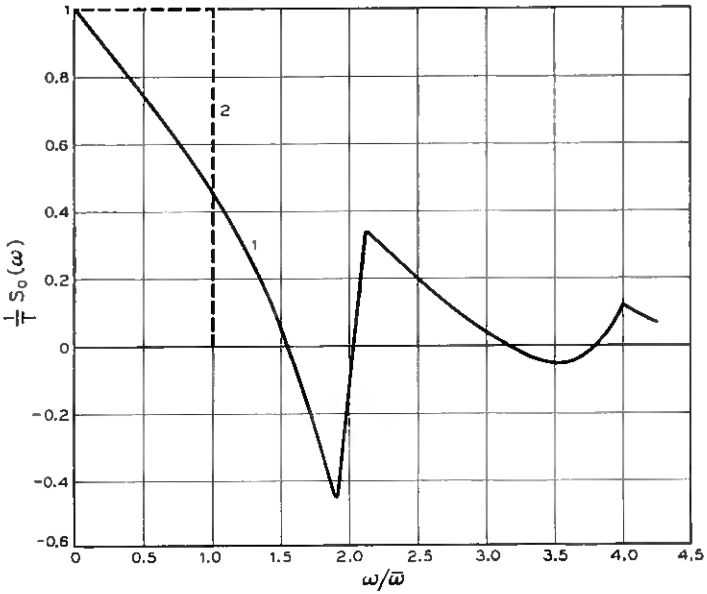


Fig. 10 — Spectra, $S_0(\omega)$, of demodulated FM (solid curve) and AM (dashed curve) pulses shown in Fig. 4 for flat spectrum at detector input.

spectrum of the AM pulses. It will be noted that the spectrum is negative in certain ranges for $\omega > 1.6\bar{\omega}$. This places a restriction on the choice of the modified spectrum S_m and on the filter bandwidth, if $A_0(\omega)$ is to be positive and finite for all values of ω in the filter band. To this end, it is necessary that the filter bandwidth be less than $1.6\bar{\omega}$.

In Fig. 11 is shown the amplitude characteristic $A_0(\omega)$ of the post-detection low-pass filter obtained from (90) when $S_m(\omega)$ is assumed to be equal to the AM baseband spectrum. With the amplitude characteristic $A_0(\omega)$ shown in Fig. 11, the FM pulses shown in Fig. 4 would be converted into pulses of the same shape as shown for AM.

With a raised cosine spectrum at the detector input, as considered in Section VII, the spectrum of the demodulated pulses obtained by numerical integration of (87) is given in Table V. In the numerical integration, contributions to the integral for $\bar{\omega}t_0 = \tau > 5$ were disregarded.

The above spectrum is shown in Fig. 12, together with that of the AM pulses.

The circumstance that $S_0(\omega)$ is negative in the approximate range $1.9\bar{\omega} < \omega < 2.1\bar{\omega}$ in this case limits the filter bandwidth to less than $1.9\bar{\omega}$, if $A_0(\omega)$ is to be positive and finite for all values of ω in the filter band.

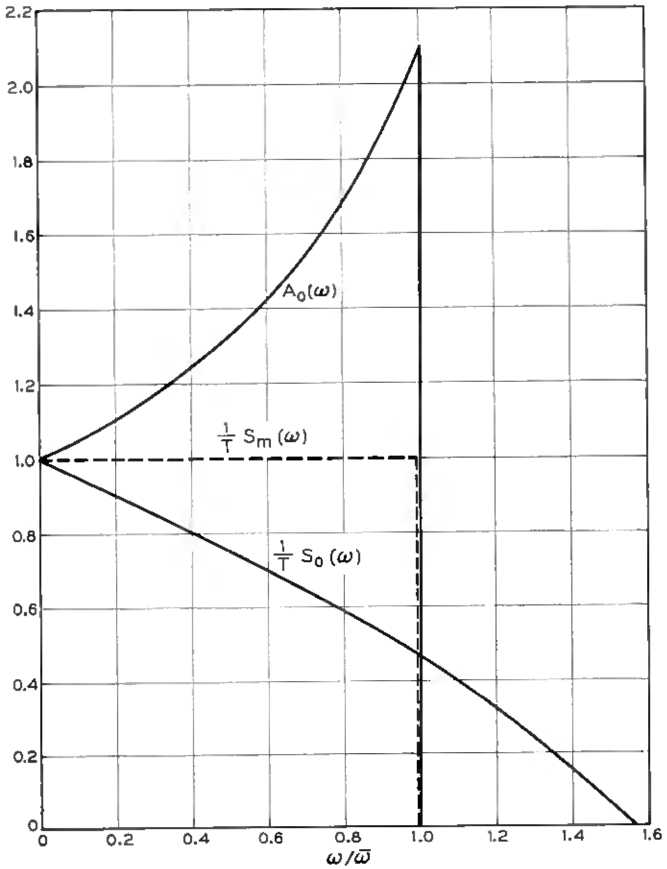


Fig. 11 — Amplitude characteristic, $A_0(\omega)$, of postdetection low-pass filter required to convert spectrum $S_0(\omega)$ into modified spectrum $S_m(\omega)$.

TABLE V — $S_0(\omega)/T$ FOR RAISED COSINE SPECTRUM AT DETECTOR INPUT

$u/\bar{\omega}$	0	0.25	0.5	0.75	1.0
0	1*	0.9492	0.8092	0.6147	0.4159
1	0.4159	0.2438	0.1063	0.0208	−0.0069
2	−0.0069	0.0131	0.0582	0.0926	0.0742
3	0.0742	0.0302	0.0093	0.0032	0.0052
4	0.0052	0.0096	0.0121	0.0114	0.0084

* From (89); actual computed value = 0.99999951.

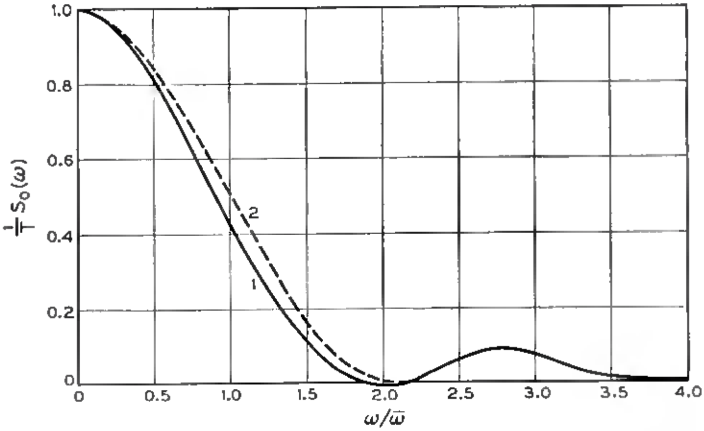


Fig. 12 — Spectra $S_0(\omega)$ of demodulated FM (solid curve) and AM (dashed curve) pulses shown in Fig. 6 for raised cosine spectrum at detector input.

By way of illustration, an appropriate type of modified spectrum of bandwidth $1.75\bar{\omega}$ is given by the following expressions.

For $0 < \omega/\bar{\omega} < \frac{1}{4}$:

$$S_m(\omega) = T. \tag{91}$$

For $\frac{1}{4} < \omega/\bar{\omega} < \frac{7}{4}$:

$$S_m(\omega) = T\cos^2\left(\frac{\pi}{4}\,\frac{4\omega/\bar{\omega} - 1}{3}\right). \tag{92}$$

The above spectrum S_m is shown in Fig. 13, together with the spectrum $S_0(\omega)$ and the amplitude characteristic $A_0(\omega)$ of the low-pass filter obtained from (90). The shape of the pulses at the output of the filter for the above spectrum $S_m(\omega)$ can be obtained from (52) with $\omega_x = \frac{3}{4}\bar{\omega}$, but it does not differ significantly from that shown in Fig. 6 for AM.

With a low-pass filter having a linear phase and an amplitude characteristic $A_0(\omega)$ as shown in Fig. 13, intersymbol interference is avoided, and some improvement in signal-to-noise ratio is realized by elimination of higher-frequency noise components in the detector output, as will be shown later.

XI. PULSE TRAINS

In the preceding analysis, transmission of a single mark of duration T was assumed. When a pulse train consisting of a sequence of marks and spaces is transmitted, (33) is modified into the following expression

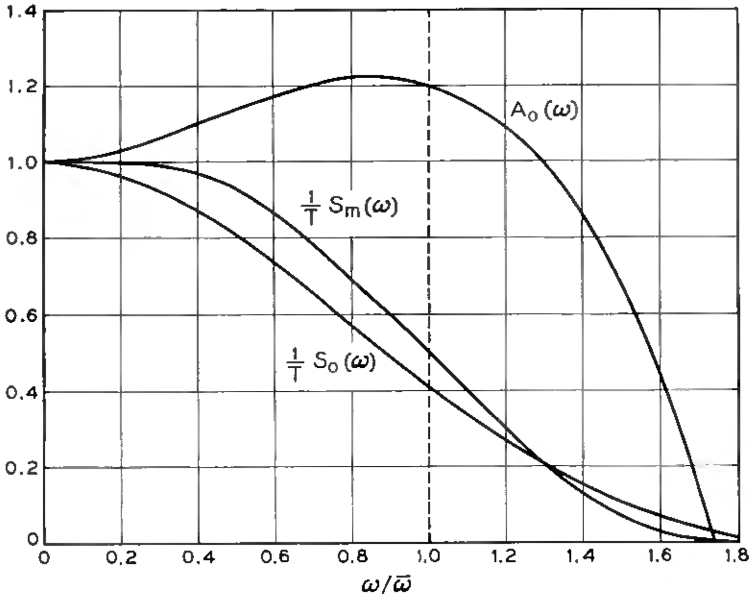


Fig. 13 — Amplitude characteristic, $A_0(\omega)$, of postdetection low-pass filter required to convert spectrum $S_0(\omega)$ of Fig. 12 into a modified spectrum $S_m(\omega)$.

for the received wave at the detector input:

$$\begin{aligned} W_i(t_0) = &-\cos(\omega_0 t_0 + \varphi_0)\,[A(-\bar{\omega})\cos\bar{\omega}(t_0 - mT) \\ &- \sum_{m=-\infty}^{\infty} a_m p(t_0 - mT)] - A(-\bar{\omega})\sin(\omega_0 t_0 + \varphi_0)\sin\bar{\omega}(t_0 - mT), \end{aligned} \tag{93}$$

where $a_m = 0$ for a space and 1 for a mark, and m is an integer. The above expression can be written in the form

$$W_i(t_0) = \bar{W}_i(t_0)(\cos\omega_0 t_0 + \varphi_0 + \Psi_0), \tag{94}$$

where the pulse train envelope is given by the following expression in place of (36):

$$\begin{aligned} \bar{W}_i(t_0) = A(-\bar{\omega})[1 &+ \mu^2 \sum a_m^2 p^2(t_0 - mT) \\ &- 2\mu \sum a_m p(t_0 - mT)\cos\bar{\omega}(t_0 - mT)]^{1/2}, \end{aligned} \tag{95}$$

where Σ indicates summation between the limits $m = -\infty$ and $m = \infty$, as in (93). Expression (37) for the phase Ψ_0 is replaced by

$$\tan\Psi_0 = \frac{\sin\bar{\omega}(t_0 - mT)}{\cos\bar{\omega}(t_0 - mT) - \mu\sum a_m p(t_0 - mT)}. \tag{96}$$

Expression (39) for the time derivative of Ψ_0 is replaced by

$$\begin{aligned}\Psi_0' = -\frac{\bar{\omega}}{D}\Big[1 - \mu \sum a_m p(t_0 - mT) \cos \bar{\omega}(t_0 - mT) \\ + \frac{\mu}{\bar{\omega}} \sum a_m p'(t_0 - mT) \sin \bar{\omega}(t_0 - mT)\Big],\end{aligned} \tag{97}$$

where

$$D = \sin^2 \bar{\omega}(t_0 - mT) + [\cos \bar{\omega}(t_0 - mT) - \mu \sum a_m p(t_0 - mT)]^2. \tag{98}$$

Expression (41) is replaced by the following expression for the demodulated pulse train:

$$\begin{aligned}W(t_0) = \frac{\mu}{2D}\Big\{\mu\Big[\sum a_m p(t_0 - mT)\Big]^2 - \sum a_m p(t_0 - mT) \\ \cdot \cos \bar{\omega}(t_0 - mT) - \frac{1}{\bar{\omega}} \sum a_m p'(t_0 - mT) \sin \bar{\omega}(t_0 - mT)\Big\}.\end{aligned} \tag{99}$$

At the sampling points $t_0 = 0$, $\sin \bar{\omega} mT = 0$ and $\cos \bar{\omega} mT = (-1)^m$, so that (99) becomes

$$W(0) = \frac{\mu}{2} \frac{\sum a_m p(mT)}{\mu \sum a_m p(mT) - (-1)^m}. \tag{100}$$

Since $p(mT) = 0$ except for $m = 0$, (100) becomes

$$W(0) = \frac{\mu}{2} \frac{a_0 p(0)}{\mu a_0 p(0) - 1}, \tag{101}$$

where $\mu = 2$ and $p(0) = 1$, so that

$$\begin{aligned} W(0) &= \frac{a_0}{2a_0 - 1} \\ &= 1 \qquad \text{for} \qquad a_0 = 1 \text{ (mark)} \\ &= 0 \qquad \text{for} \qquad a_0 = 0 \text{ (space)}. \end{aligned} \tag{102}$$

There is thus no intersymbol interference when a pulse train is transmitted.

In (99) the denominator D depends on the composition of the pulse train. For this reason the shape of the demodulated pulse train between sampling points cannot be obtained by direct superposition of individual demodulated pulses, such as those shown in Fig. 4 and Fig. 6.

XII. AVERAGE SIGNAL POWER

The preceding analysis was concerned with the ideal shaping of the over-all transmission-frequency characteristic of the bandpass channel required to avoid intersymbol interference in FM and AM systems. This over-all shaping will ordinarily be divided between bandpass filters at the transmitting and of the receiving ends. While the division between the two ends is immaterial from the standpoint of intersymbol interference, it does affect signal power at the transmitting end and interference of various kinds at the receiving end. In order to determine the optimum proportioning and shaping between transmitting and receiving ends, it is thus necessary to consider both signal power and interference.

In analog systems for voice transmission and other purposes, the peak signal power is ordinarily substantially greater than the average signal power, by 10 db or more, and is usually a limitation in systems design. In binary pulse systems with representative transmission-frequency characteristics,* however, peak signal power is not much greater than average signal power, and the latter is ordinarily a limitation, either from the standpoint of repeater design or interference with other systems. For this reason, average signal power will be considered here in comparing FM and AM binary pulse systems.

Let the amplitude characteristic of the transmitting filter be $T(u)$ and that of the receiving filter be $R(u)$, in which case $T(u)R(u) = A(u)$, where $A(u)$ is the over-all amplitude characteristic and u the frequency from midband.

Let the peak amplitude of the carrier for a continuous mark or space at the output of the transmitting filter be E, in which case the signal power for a continuous mark is

$$P_m = \frac{E^2}{2}. \tag{103}$$

In Appendix A it is shown that the average signal power for random pulse trains with FM and with bipolar AM are given by

$$P_{\text{FM}} = P_m \frac{4}{\pi T} R^2(\bar{\omega}) \int_{-\infty}^{\infty} \left[\frac{S(u)}{R(u)}\right]^2 du, \tag{104}$$

$$P_{\text{AM}} = P_m \frac{2}{\pi T} R^2(0) \int_{-\infty}^{\infty} \left[\frac{S(u)}{R(u)}\right]^2 du, \tag{105}$$

where T is the pulse interval or duration, and $S(u)$ is the spectrum at

* This excludes idealized flat channels of minimum bandwidth.

the channel output, or detector input, in response to the transmission of a single pulse, i.e., in response to the transmission of a mark preceded and followed by a continuing space.

By way of illustration, in the case of a raised cosine spectrum as given by (61)

$$P_{\text{FM}} = P_m \frac{T}{\pi} R^2(\bar{\omega}) \int_{-2\bar{\omega}}^{2\bar{\omega}} \frac{\cos^4 (\pi u/4\bar{\omega})}{R^2(u)}\, du, \tag{106}$$

$$P_{\text{AM}} = P_m \frac{T}{2\pi} R^2(0) \int_{-2\bar{\omega}}^{2\bar{\omega}} \frac{\cos^4 (\pi u/4\bar{\omega})}{R^2(u)}\, du. \tag{107}$$

If the receiving filter has a half-cosine shape as given by

$$\begin{aligned} R(u) &= \cos \pi u/4\bar{\omega} \qquad \text{for} \quad -2\bar{\omega} < u < 2\bar{\omega} \\ &= 0 \qquad\qquad\qquad \text{for} \qquad |u| > 2\bar{\omega} \end{aligned} \tag{108}$$

expressions (106) and (107) become, with $R(\bar{\omega}) = \frac{1}{2}^{1/2}$ and $R(0) = 1$,

$$\begin{aligned} P_{\text{FM}} = P_{\text{AM}} &= P_m \frac{T}{2\pi} \int_{-2\bar{\omega}}^{2\bar{\omega}} \cos^2 \pi u/4\bar{\omega})\, du \\ &= P_m. \end{aligned} \tag{109}$$

In this particular case the average signal power of a random pulse train in both FM and bipolar AM is equal to the signal power for a continuous mark or space.

Consider next a flat receiving filter, in which case

$$\begin{aligned} R(u) &= 1 \qquad \text{for} \quad -2\bar{\omega} < u < 2\bar{\omega} \\ &= 0 \qquad \text{for} \qquad |u| > 2\bar{\omega}. \end{aligned} \tag{110}$$

In this case (106) and (107) yield

$$\begin{aligned} P_{\text{FM}} &= P_m \frac{T}{\pi} \int_{-2\bar{\omega}}^{2\bar{\omega}} \cos^4 \pi u/4\bar{\omega}\, du \\ &= \tfrac{3}{2} P_m, \end{aligned} \tag{111}$$

$$P_{\text{AM}} = \tfrac{1}{2} P_{\text{FM}} = \tfrac{3}{4} P_m. \tag{112}$$

Finally, let the receiving filter have a raised cosine shape as given by

$$\begin{aligned} R(u) &= \cos^2(\pi u/4\bar{\omega}) \qquad \text{for} \quad -2\bar{\omega} < u < 2\bar{\omega} \\ &= 0 \qquad\qquad\qquad\quad \text{for} \qquad |u| > 2\bar{\omega}. \end{aligned} \tag{113}$$

In this case $R(\bar{\omega}) = \frac{1}{2}$ and

$$P_{\text{FM}} = P_m \frac{T}{4\pi} \int_{-2\bar{\omega}}^{2\bar{\omega}} du \tag{114}$$

$$= P_m ,$$

$$P_{\text{AM}} = 2P_{\text{FM}} = 2P_m . \tag{115}$$

The above relations show that, for equal signal power in FM and bipolar AM during transmission of a continuing mark (or space), there may be appreciable difference in average signal power for a random pulse train, depending on the shape of the transmitting filter. For this reason, signal power during a continuing mark, which is often used in specifying signal-to-noise ratio, may not be an appropriate reference signal power.

To determine the optimum division of channel shaping between transmitting and receiving ends, it is necessary to consider the effect of random noise or other interference, such as impulse noise. The effect of any particular type of interference depends on the shape of the receiving filter, as discussed in the following sections for random noise. Impulse noise is discussed briefly in Appendix B.

XIII. RANDOM NOISE IN FM AND AM SYSTEMS

Certain basic equations relating to noise and interference on FM and AM are given in Appendix B and applied to the particular case of single-frequency interference. In the case of a sinusoidal interfering voltage at a frequency u from midband and of amplitude $e(u)$ at the input of the receiving filter, rms interference in FM and bipolar AM taken in relation to the peak-to-peak signal amplitude at the detector output is given by

$$\bar{\eta}_{\text{FM}} \cong \frac{\bar{e}(u)R(u)}{2ER(\omega)} (1 + u/\bar{\omega}) , \tag{116}$$

$$\bar{\eta}_{\text{AM}} = \frac{\bar{e}(u)R(u)}{2ER(0)} \tag{117}$$

where $\bar{e}(u) = e(u)/2^{1/2}$, $R(u)$ is the amplitude characteristic of the receiving filter and E is the peak amplitude of the carrier for a continuing mark or space. The above relation for $\bar{\eta}_{\text{FM}}$ is a first-order approximation applying if $\bar{e}(u)$ is small in relation to E, as is required for transmission without excessive error rates.

The equations give the rms amplitude of the interfering voltage at the detector output, taken in relation to the peak-to-peak difference in pulse amplitudes for mark and space, considering all possible phases of the interfering voltage equally probable. The slicing or threshold level is ordinarily half this difference, and rms interference voltage taken in relation to the slicing level would be twice as great.

Random noise can be regarded as the sum of a very large number of sinusoidal waves of different frequencies, with both the amplitude and phase of each component wave varying with time. For a single sinusoidal wave at a frequency u from midband and peak amplitude e, the rms amplitude is $\bar{e} = e/2^{1/2}$. When the sinusoidal wave varies in amplitude with time there will be a certain rms amplitude $\mathbf{e}$ over a long interval, and a corresponding average noise power $\mathbf{e}^2$ for each sinusoidal component. The corresponding average noise power per unit of bandwidth at the receiving filter input will be designated $n(u)$. In the case of white thermal noise as assumed in the following, $n(u) = n$ is independent of u.

The ratio of noise power to signal power at the output of the detector can be obtained from (116) and (117) for single-frequency waves, by integration of the noise power density over the channel band. Thus, in the case of FM, the ratio of average output noise power N_0 to the output peak-to-peak signal power $\hat{S}_0$ between mark and space, as obtained by integration of (116), becomes

$$(N_0/\hat{S}_0)_{\text{FM}} = \frac{n}{8P_mR^2(-\bar{\omega})} \int_{-\infty}^{\infty} R^2(u)(1 + u/\bar{\omega})^2 \, du \qquad (118)$$

$$= \frac{n}{8P_mR^2(-\bar{\omega})} \int_{-\infty}^{\infty} R^2(u)(1 + u^2/\bar{\omega}^2) \, du, \qquad (119)$$

where the last expression follows since $R(-u) = R(u)$, so that the integral of $R^2 2u/\bar{\omega}$ vanishes.

In case of bipolar AM, the corresponding ratio obtained by integration of (117) is

$$(N_0/\hat{S}_0)_{\text{AM}} = \frac{n}{8P_mR^2(0)} \int_{-\infty}^{\infty} R^2(u) \, du. \qquad (120)$$

Relations (119) and (120) can be expressed in terms of average signal power in FM and AM with the aid of relations (104) and (105). The following expressions are thus obtained:

$$(N_0/\hat{S}_0)_{\text{FM}} = \frac{2n}{4\pi TP_{\text{FM}}} \left[\int_{-\infty}^{\infty} R^2(u)(1 + u^2/\bar{\omega}^2) \, du \right] \cdot \left[\int_{-\infty}^{\infty} \frac{S^2(u)}{R^2(u)} \, du \right], \qquad (121)$$

$$(N_0/\hat{S}_0)_{\text{AM}} = \frac{n}{4\pi T P_{\text{AM}}} \left[\int_{-\infty}^{\infty} R^2(u)\, du \right] \left[\int_{-\infty}^{\infty} \frac{S^2(u)}{R^2(u)}\, du \right]. \tag{122}$$

The above expressions for FM apply without a postdetection low-pass filter. As discussed in Section X, it is possible to modify the pulse shape in FM without causing intersymbol interference, with the aid of a low-pass filter of appropriate amplitude characteristic $A_0(\omega)$ that depends on the bandwidth. Such a filter reduces the noise power in a narrow band at ω by the factor $A_0^2(\omega)$. On a carrier basis, this is equivalent to multiplying the noise power in a narrow band at a frequency u from midband by $A_0^2(\bar{\omega} + u)$ during transmission of a space, as indicated in Fig. 14, or by $A_0^2(-\bar{\omega} + u)$ during transmission of a mark. This equivalent representation on a carrier basis is legitimate, provided the carrier power at the detector input is substantially greater than the noise power, in which case the ratio $N_0/\hat{S}_0$ obtained from (121) is much smaller than one.

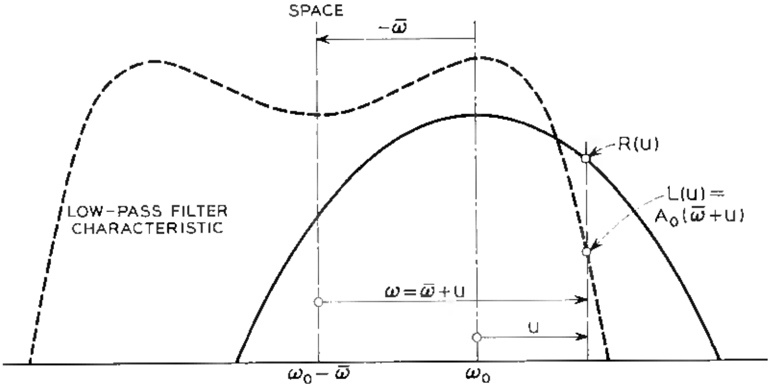

SPACE
$-\bar{\omega}$
R(u)
L(u) = $A_0(\bar{\omega}+u)$
LOW-PASS FILTER CHARACTERISTIC
$\omega = \bar{\omega} + u$
u
$\omega_0 - \bar{\omega}$
ω_0

Fig. 14 — Frequency modulation with postdetection low-pass filter having transmission characteristic $A_0(\omega)$. Noise power in the narrow band at frequency u from midband is reduced by the factor $L^2(u) = A_0^2(\bar{\omega} + u)$ during transmission of space with carrier at $\omega_0 - \bar{\omega}$, as assumed above.

The following notation indicated in Fig. 14 will be used:

$$L(u) = A_0(\bar{\omega} + u). \tag{123}$$

With such a filter, (118) is modified into

$$(N_0/\hat{S}_0)_{\text{FM}} = \frac{n}{8P_m R^2(-\bar{\omega})} \int_{-\infty}^{\infty} L^2(u) R^2(u)(1 + u/\bar{\omega})^2\, du, \tag{124}$$

and (121) is modified into

$$(N_0/\hat{S}_0)_{\text{FM}} = \frac{2n}{4\pi TP_{\text{FM}}}\left[\int_{-\infty}^{\infty} L^2(u)R^2(1 + u/\bar{\omega})^2\,du\right] \cdot\left[\int_{-\infty}^{\infty}\frac{S^2}{R^2}\,du\right], \tag{125}$$

where L, R and S are functions of u.

Comparison of (125) with (121) shows that, for a given $R(u)$, a post-detection low-pass filter reduces the ratio $(N_0/\hat{S}_0)_{\text{FM}}$ by the factor

$$\rho = \frac{\int_{-\infty}^{\infty} L^2R^2(1 + u/\bar{\omega})^2\,du}{\int_{-\infty}^{\infty} R^2(1 + u^2/\bar{\omega}^2)\,du}, \tag{126}$$

where L and R are functions of u.

XIV. OPTIMUM AM SYSTEMS

The minimum value of $(N_0/\hat{S}_0)$ for a given average power $P = P_{\text{AM}}$ is obtained when the product of the two integrals in (122) is a minimum. As shown in Appendix C, this is the case when $R(u)$ is such that the two integrals are equal, in which case the optimum $R(u)$ is given by

$$R^o(u) = cS^{1/2}(u), \tag{127}$$

where c is an arbitrary constant independent of u that can be chosen to give $R(u)$ the appropriate dimension.

With (127) in (122), the optimum ratio $N_0/\hat{S}_0$ becomes

$$(N_0/\hat{S}_0)^o = \frac{n}{4\pi TP}\left[\int_{-\infty}^{\infty} S(u)\,du\right]^2. \tag{128}$$

When $S(u)$ has the properties previously discussed and illustrated in Fig. 1 and Fig. 2, the integral in (128) is always equal to the area under the rectangle in the upper part of these figures, or $2\bar{\omega}T/2 = \bar{\omega}T = \pi$. Furthermore, with (127) in (105), it follows that $P_m = P_{\text{AM}} = P$. Hence, for all spectra $S(u)$ of the form previously assumed, (128) becomes

$$(N_0/\hat{S}_0)^o = \frac{N(\bar{\omega}T)^2}{4\pi TP} = \frac{N}{4P} = \frac{N}{4P_m}, \tag{129}$$

where

$N = n\bar{\omega} =$ average noise power in a flat band $\bar{\omega}$ at input of receiving filter,

$P =$ average signal power at receiving filter input,

$P_m =$ signal power for continuous mark (or space),

$N_0 =$ average noise power at detector output,

$\hat{S}_0 =$ peak-to-peak signal power at detector output.

In (129) N is the average noise power in a flat band equal to that of the minimum possible bandwidth $\bar{\omega}$ over which pulses can be transmitted at intervals $T = \pi/\bar{\omega}$ without intersymbol interference, i.e., the noise power in a band $1/2T$ cps. With the above definition of N and for the above pulse transmission rate, (129) applies for all spectra at the detector input with the properties illustrated in Figs. 1 and 2. There is thus no noise penalty, but only a bandwidth penalty, in modifying the spectrum that was indicated in Fig. 2 and discussed previously to obtain pulses whose shape is more appropriate shape than that of systems with the minimum possible bandwidth. Moreover, (129) also applies for optimum bipolar baseband systems with base band spectra equal to those shown on the right-hand side of ω_0 in Figs. 1 and 2. Thus there is no noise penalty in bipolar double sideband AM, but the bandwidth of the carrier channel is twice that of the baseband channel.

This two-fold increase in bandwidth for a given transmission rate can be overcome by providing two independent channels on two carriers at quadrature, a method sometimes referred to as four-phase transmission. With 3-db reduction in noise power, because of the two-fold reduction in bandwidth, and with 3 db less signal power per channel, so that the average signal power P is the same as for a single channel in bipolar AM, (129) applies with N defined as above An alternate means of avoiding the two-fold increase in bandwidth is to use bipolar vestigial sideband AM[1,2] with homodyne detection, in which case (129) also applies. The last two methods are thus equivalent to bipolar baseband transmission, both as regards bandwidth and signal-to-noise ratio.

At the detector output the pulses may be bipolar or may be biased into unipolar (on-off) pulses, and (129), in terms of the peak-to-peak signal power $\hat{S}_0$ at sampling instants, applies regardless of any bias.

In the particular case of a raised cosine spectrum as given by (71), the over-all amplitude characteristic of the channel is given by (63) or

$$a(u) = \frac{\pi u/4\bar{\omega}}{\tan(\pi u/4\bar{\omega})}. \tag{130}$$

The optimum receiving filter characteristic as obtained from (127) is

$$R^o(u) = \cos(\pi u/4\bar{\omega}). \tag{131}$$

The corresponding optimum transmitting filter characteristic is $T^o = a(u)/R^o$, or

$$\begin{aligned} T^o(u) &= \frac{\pi u/4\bar{\omega}}{\sin\,(\pi u/4\bar{\omega})} \\ &= 1 && \text{for} \quad u = 0 \\ &= \frac{\pi}{4}\,2^{1/2} \cong 1.12 && \text{for} \quad u = \bar{\omega} \\ &= \frac{\pi}{2} && \text{for} \quad u = 2\bar{\omega}. \end{aligned} \tag{132}$$

These characteristics are shown in Fig. 15.

In the case of vestigial-sideband AM, the carrier would be at $\omega_0 + \bar{\omega}$ or $\omega_0 - \bar{\omega}$ rather than at ω_0. Pulses can then be transmitted at twice the double-sideband rate, provided homodyne detection is used so that the effect of the quadrature component is eliminated.[2] The optimum shape of the receiving filter is again given by (131), but the shape of the associated optimum transmitting filter is modified, since the spectrum at the

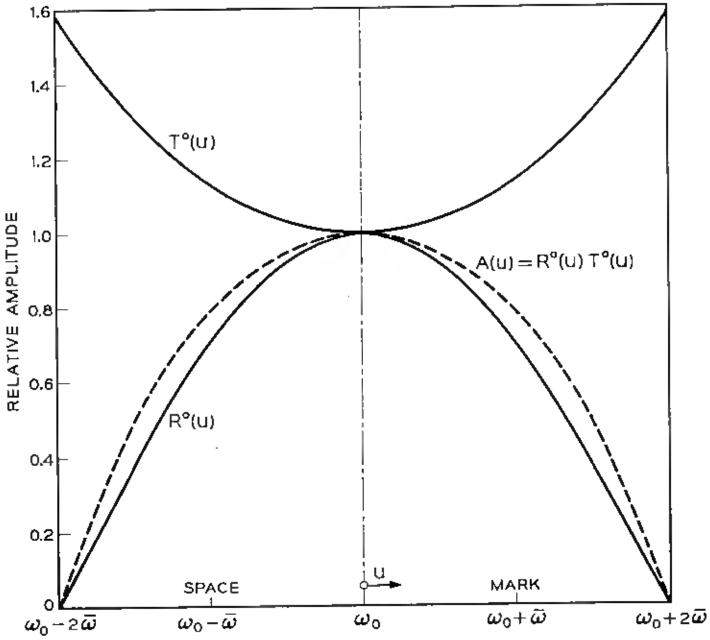


Fig. 15 — Double-sideband AM with raised cosine spectrum at detector input: R^o = optimum shape of receiving filter; T^o = optimum shape of transmitting filter; $A = R^oT^o$ = combined transmission characteristic.

channel input is now caused by a rectangular carrier pulse of frequency $\omega_0 \pm \bar{\omega}$ rather than ω_0. The optimum transmitting filter characteristic is given by

$$T^o(u) = \cos(\pi u/4\bar{\omega}) \frac{\pi(u \pm \bar{\omega})/4\bar{\omega}}{\sin[\pi(u \pm \bar{\omega})/4\bar{\omega}]}. \tag{133}$$

The positive signs apply when the carrier is at $\omega_0 - \bar{\omega}$ and the negative signs apply when it is at $\omega_0 + \bar{\omega}$; the expression applies for $-2\bar{\omega} \leqq u \leqq 2\bar{\omega}$.

If the carrier is assumed at $\omega_0 + \bar{\omega}$,

$$\begin{aligned} T^o(u) &= \cos(\pi u/4\bar{\omega}) \frac{\pi(u - \bar{\omega})/4\bar{\omega}}{\sin[\pi(u - \bar{\omega})/4\bar{\omega}]} \\ &= 0 \qquad \text{for} \quad u \leqq -2\bar{\omega} \\ &= \frac{\pi}{4} 2^{1/2} \qquad \text{for} \quad u = -\bar{\omega} \\ &= \frac{\pi}{4} 2^{1/2} \qquad \text{for} \quad u = 0 \\ &= \frac{1}{2} 2^{1/2} \qquad \text{for} \quad u = \bar{\omega} \\ &= 0 \qquad \text{for} \quad u \geqq 2\bar{\omega}. \end{aligned} \tag{134}$$

In actual systems, it may be expedient from the standpoint of design to employ transmitting and receiving filter characteristics that differ from the optimum characteristics shown in Fig. 15. This results in some penalty in signal-to-noise ratio, as shown below.

By way of illustration, it will be assumed that the transmitting filter has the shape shown in Fig. 5 for the channel transmission-frequency characteristic, in which case the receiving filter could be flat and given by

$$\begin{aligned} R(u) &= 1 \qquad \text{for} \quad -2\bar{\omega} \leqq u \leqq 2\bar{\omega} \\ &= 0 \qquad \text{for} \quad -2\bar{\omega} > u > 2\bar{\omega}. \end{aligned} \tag{135}$$

In this case, (122) for AM becomes

$$\begin{aligned} N_0/\hat{S}_0 &= \frac{n}{4\pi TP} \left[\int_{-2\bar{\omega}}^{2\bar{\omega}} du \right] \left[\int_{-2\bar{\omega}}^{2\bar{\omega}} \frac{T^2}{4} \cos^4(\pi u/4\bar{\omega})\, du \right] \\ &= \frac{3}{2} \frac{N}{4P} \qquad (136) \\ &= 2 \frac{N}{4P_m}, \qquad (137) \end{aligned}$$

where N, P and P_m are defined as in (129) and the last relation follows from (112) for the above case; i.e., $P = \frac{3}{4}P_m$.

Comparison of (136) with (129) shows that, for equal average signal power P, the ratio $N_0/\hat{S}_0$ is greater here than for optimum division of filter shaping by a factor of $\frac{3}{2}$ (or 1.8 db), while comparison of (137) with (129) shows that, for equal continuous mark power P_m, it is greater by a factor of 2 (or 3 db). This illustrates that the penalty incurred in departing from the optimum division of channel shaping between transmitting and receiving filters can be significant.

XV. OPTIMUM FM SYSTEMS

The minimum ratio $N_0/\hat{S}_0$ is obtained when the product of the two integrals in (121) is a minimum. As shown in Appendix C, this is the case when $R(u)$ is such that the two integrals are equal, resulting in the following expression for the optimum $R(u)$ without a postdetection low-pass filter:

$$R^o(u) = cS^{1/2}(u)(1 + u^2/\bar{\omega}^2)^{-1/4}, \tag{138}$$

where c is an arbitrary constant. The corresponding optimum $N_0/\hat{S}_0$ obtained with (138) in (125) is

$$(N_0/\hat{S}_0)^o = \frac{2n}{4\pi TP}\left[\int_{-\infty}^{\infty}(1 + u^2/\bar{\omega}^2)^{1/2}S(u)\,du\right]^2 \tag{139}$$

$$= \frac{N}{4P}\lambda^o, \tag{140}$$

where $P = P_{\mathrm{FM}}$, N is defined as in (129) and

$$\lambda^o = 2\left\{\frac{\left[\int_{-\infty}^{\infty}(1 + u^2/\bar{\omega}^2)^{1/2}S(u)\,du\right]^2}{\left[\int_{-\infty}^{\infty}S(u)\,du\right]^2}\right\}. \tag{141}$$

Comparison of (140) with (129) shows that, for equal average signal power, the optimum ratio $N_0/\hat{S}_0$ is greater in FM than in bipolar AM by the factor λ^o. Inspection of (141) shows that $\lambda^o > 2$, without a postdetection low-pass filter.

With (138) in (126), the factor ρ, by which the noise power is reduced by a postdetection low-pass filter, becomes

$$\rho = \frac{\int_{-\infty}^{\infty}L^2(u)S(u)\,\frac{(1 + u/\bar{\omega})^2}{(1 + u^2/\bar{\omega}^2)^{1/2}}\,du}{\int_{-\infty}^{\infty}(1 + u^2/\bar{\omega}^2)^{1/2}S(u)\,du}, \tag{142}$$

where $L(u)$ is defined by (123).

With a postdetection low-pass filter, (140) is replaced by

$$(N_0/\hat{S}_0)^o = \frac{N}{4P}\gamma, \tag{143}$$

where

$$\gamma = \lambda^o \rho \tag{144}$$

is the factor by which the optimum ratio $N_0/\hat{S}_0$ differs in FM from bipolar AM, for equal average signal power at the receiving filter input.

The above factor γ is not the minimum (optimum) factor with a postdetection low-pass filter, but rather the factor applying when a low-pass filter is applied to a system in which $R(u)$ is optimized without such a filter. If a postdetection low-pass filter of specified amplitude characteristic $A_0(\omega)$ is assumed, the optimum $R(u)$ is related to $A_0(\omega)$, as discussed below.

With a postdetection low-pass filter, the minimum ratio $N_0/\hat{S}_0$ in terms of average signal power is again obtained when $R(u)$ is such that the two integrals in (125) are equal. As shown in Appendix C, $R(u)$ is in this case given by

$$R^o(u) = c2^{1/4}S^{1/2}(u)[H(u)]^{-1/2}, \tag{145}$$

where

$$H(u) = [L^2(u)(1 + u/\bar{\omega})^2 + L^2(-u)(1 - u/\bar{\omega})^2]^{1/2} \tag{146}$$

$$= [A_0^2(\bar{\omega} + u)(1 + u/\bar{\omega})^2 + A_0^2(\bar{\omega} - u)(1 - u/\bar{\omega})^2]^{1/2}. \tag{147}$$

In the last relation, $A_0(\bar{\omega} \pm u)$ designates the amplitude characteristic of the low-pass filter at $\omega = \bar{\omega} \pm u$.

The optimum ratio obtained with (145) in (125) can be written

$$(N_0/S_0)^o = \frac{N}{4P}\gamma^o, \tag{148}$$

where

$$\gamma^o = \frac{\left[\int_{-\infty}^{\infty} H(u)S(u)\,du\right]^2}{\left[\int_{-\infty}^{\infty} S(u)\,du\right]^2} \tag{149}$$

$$= \frac{\left[\int_{0}^{\infty} H(u)S(u)\,du\right]^2}{\left[\int_{0}^{\infty} S(u)\,du\right]^2}, \tag{150}$$

where (150) follows from (149) since $H(-u) = H(u)$ and $S(-u) = S(u)$.

For the reasons discussed in Section X, the bandwidth of the low-pass filter must be less than $2\bar{\omega}$. Hence, for $u \approx \bar{\omega}$, $A_0(\bar{\omega} + u) = 0$, so that the first term in (147) vanishes and

$$\begin{aligned} H(u) &\cong A_0(0)(1 - u/\bar{\omega}) \qquad \text{for} \quad u/\bar{\omega} \cong 1 \\ &= 0 \qquad\qquad\qquad\qquad \text{for} \quad u = \bar{\omega}. \end{aligned} \tag{151}$$

Thus, for $u \cong \bar{\omega}$, (145) becomes

$$\begin{aligned} R^o(u) &\cong 2^{1/4}c\left[\frac{S(u)}{A_0(0)(1 - u/\bar{\omega})}\right]^{1/2} \qquad \text{for} \quad u/\bar{\omega} \cong 1 \\ &= \infty \qquad\qquad\qquad\qquad\qquad\qquad \text{for} \quad u = \bar{\omega}. \end{aligned} \tag{152}$$

The corresponding transmitting filter characteristic is $T^o(u) = A(u)/R^o(u) = 0$ for $u = \bar{\omega}$.

With $R^o(\pm\bar{\omega}) = \infty$, the noise power at the detector input would become infinite and the signal-to-noise ratio at the detector input would be zero. Hence, the basic premise of adequately high signal-to-noise ratios underlying the representation in Fig. 14 and expression (125) would be violated. In this case, $N_0/\hat{S}_0$, without a postdetection low-pass filter as given by (121), would become infinite. To limit the noise power at the detector input, so that (125) is a legitimate approximation, it is necessary to modify $R^o(u)$ near $u = \pm\bar{\omega}$ in such a way that $R^o(\pm\bar{\omega}) \neq \infty$ and $N_0/\hat{S}_0$, as obtained from (121) without a postdetection filter, becomes appropriately small. The value of $N_0/\hat{S}_0$ obtained from (125) after such modification of $R(u)$ will be greater than that obtained from (150), but may be smaller than that given by (144). The factor γ^o is thus to be regarded as a lower bound that cannot be fully realized but may be closely approached, at least for small ratios N/P in (148), by appropriate modification of $R^o(u)$ near $u = \pm\bar{\omega}$. (An example of such modification is indicated by the dotted curves in Fig. 17, to be discussed later.)

XVI. OPTIMUM FM SYSTEMS OF MINIMUM BANDWIDTH

In the limiting case of a channel of the minimum possible bandwidth, as considered in Section VI, the channel characteristic is given by (55). When normalized to unit amplitude for $u = 0$, this characteristic is

$$A(u) = \frac{1 - (u/\bar{\omega})^2}{\cos(\pi u/2\bar{\omega})}. \tag{153}$$

With a spectrum as given by (53) and (54), the optimum receiving filter characteristic obtained from (138) becomes, with c chosen as $(2/T)^{1/2}$,

$$R^o(u) = (1 + u^2/\bar{\omega})^{-1/4}. \tag{154}$$

The corresponding optimum receiving filter characteristic,

$$T^o(u) = A(u)/R^o(u),$$

is

$$T^o(u) = \frac{[1 - (u/\bar{\omega})^2](1 + u^2/\bar{\omega}^2)^{1/4}}{\cos(\pi u/2\bar{\omega})}. \tag{155}$$

The above expressions apply for $-1 < u/\bar{\omega} < 1$.

The factor λ^o defined by (141) becomes

$$\begin{aligned} \lambda^o &= 2\frac{\left[\int_{-\bar{\omega}}^{\bar{\omega}} (1 + u^2/\bar{\omega}^2)^{1/2}\,du\right]^2}{\left[\int_{-\bar{\omega}}^{\bar{\omega}} du\right]^2} \\ &= \tfrac{1}{2}[2^{1/2} + \log_e(1 + 2^{1/2})]^2 \\ &\cong 2.65. \end{aligned} \tag{156}$$

This corresponds to about 4.2 db disadvantage in signal-to-noise ratio for an optimum FM system without a postdetection low-pass filter, as compared to an optimum bipolar AM system, for equal average signal power.

In the above case the spectrum $S_0(\omega)$ of the demodulated pulses is as shown in Fig. 10. With a low-pass filter having the amplitude charactertic shown in Fig. 11, this spectrum is converted into a flat spectrum of the minimum permissible bandwidth. With the above type of filter, $L(u) = A_0(\bar{\omega} + u) = 0$ for $u > 0$. Since $S(u) = 0$ for $\bar{\omega} < u < -\bar{\omega}$, (142) becomes

$$\begin{aligned} \rho &= \frac{\int_{-\bar{\omega}}^{0} A_0^2(\bar{\omega} - u)\,\dfrac{(1 + u/\bar{\omega})^2}{(1 + u^2/\bar{\omega}^2)^{1/2}}\,du}{\int_{-\bar{\omega}}^{\bar{\omega}} (1 + u^2/\bar{\omega}^2)^{1/2}\,du} \\ &\cong 0.38 \text{ (by numerical integration).} \end{aligned} \tag{157}$$

With the above low-pass filter, (144) gives

$$\gamma = \lambda^o \rho \cong 1.0, \tag{158}$$

so that the signal-to-noise ratio is virtually the same as for an optimum bipolar binary AM system, for equal average signal power.

The factor γ^o given by (150) becomes

$$\gamma^o = \frac{\left[\int_0^{\bar{\omega}} A_0(\bar{\omega} - u)(1 - u/\bar{\omega})\, du\right]^2}{\left[\int_0^{\bar{\omega}} du\right]^2} \tag{159}$$

$$\cong 0.60 \text{ (by numerical integration).}$$

For reasons mentioned in Section XV, the minimum factor γ that can be realized will be less than that given by (158) but greater than that given by (159), which is to be regarded as a lower bound. Since this minimum factor γ is less than one, some advantage in signal-to-noise ratio can be realized with FM, as compared to an optimum bipolar baseband or AM system with synchronous detection, for equal average signal power at the input of the receiving filter. This advantage would be small, and is principally of theoretical interest as an indication that a noise advantage can be derived from the unavoidable two-fold increase in channel bandwidth with FM as compared to baseband transmission or equivalent AM methods.

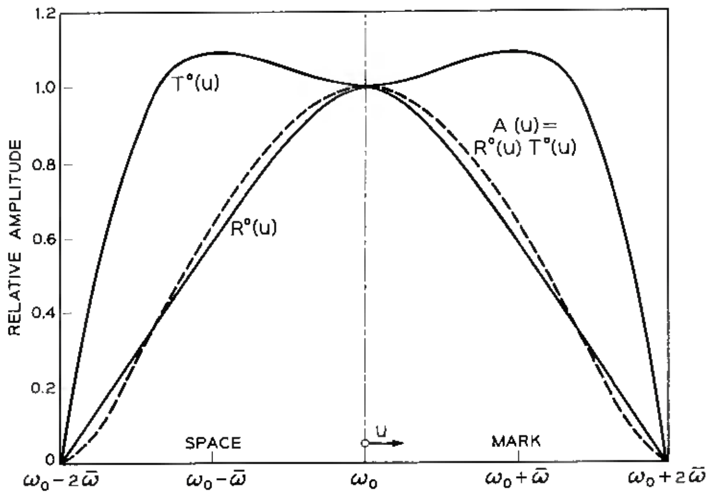


Fig. 16 — Frequency modulation with raised cosine spectrum at detector input and no postdetection low-pass filter: R^o = optimum shape of receiving filter; T^o = optimum shape of transmitting filter; $A = R^oT^o$ = combined transmission characteristic.

XVII. OPTIMUM FM SYSTEMS WITH RAISED COSINE SPECTRUM

With a raised cosine spectrum at the detector input, as considered in Section VII, the over-all amplitude characteristic of the bandpass channel is given by (62). When normalized to unit amplitude for $u = 0$, the characteristic is

$$A(u) = \frac{1 + \cos(\pi u/2\bar{\omega})}{2\cos(\pi u/2\bar{\omega})}(1 - u^2/\bar{\omega}^2). \tag{160}$$

The optimum characteristic of the receiving filter as obtained from (138) is, in this case,

$$R^o = \frac{\cos(\pi u/4\bar{\omega})}{(1 + u^2/\bar{\omega}^2)^{1/4}}. \tag{161}$$

The corresponding optimum characteristic of the transmitting filter is $T^o = A(u)/R^o$, or

$$\begin{aligned} T^o &= \frac{\cos(\pi u/4\bar{\omega})}{\cos(\pi u/2\bar{\omega})}(1 - u^2/\bar{\omega}^2)(1 + u^2/\bar{\omega}^2)^{1/4} \\ &= 1 \qquad\qquad\qquad \text{for} \quad u = 0 \\ &= \frac{4}{\pi}(\tfrac{1}{2})^{1/4} \cong 1.09 \qquad \text{for} \quad u = \bar{\omega} \\ &= 0 \qquad\qquad\qquad \text{for} \quad u = 2\bar{\omega}. \end{aligned} \tag{162}$$

The above filter characteristics are shown in Fig. 16.

With a raised cosine spectrum in (141),

$$\begin{aligned} \lambda^o &= 2\left\{\frac{\left[\int_{-2\bar{\omega}}^{2\bar{\omega}} (1 + u^2/\bar{\omega}^2)^{1/2}\cos^2(\pi u/4\bar{\omega})\,du\right]^2}{\left[\int_{-2\bar{\omega}}^{2\bar{\omega}} \cos^2(\pi u/4\bar{\omega})\,du\right]^2}\right\} \\ &= 2\left[\int_0^2 (1 + x^2)^{1/2}\cos^2(\pi x/4)\,dx\right]^2 \\ &\cong 2.8 \text{ (by numerical integration).} \end{aligned} \tag{163}$$

With the aid of a postdetection low-pass filter, the noise power in the output is reduced by the following factor ρ obtained from (142):

$$\rho = \frac{\int_{-2\bar{\omega}}^{2\bar{\omega}} G(u)\cos^2(\pi u/4\bar{\omega})\,du}{\int_{-2\bar{\omega}}^{2\bar{\omega}} (1 + u^2/\bar{\omega}^2)^{1/2}\cos^2(\pi u/4\bar{\omega})\,du}, \tag{164}$$

where

$$G(u) = L^2(u) \frac{(1 + u/\bar{\omega})^2}{(1 + u^2/\bar{\omega}^2)^{1/2}}. \tag{165}$$

The factor γ^o given by (150) in this case becomes

$$\gamma^o = \frac{\left[\int_0^{2\bar{\omega}} H(u) \cos^2 (\pi u/4\bar{\omega}) \, du \right]^2}{\left[\int_0^{2\bar{\omega}} \cos^2 (\pi u/4\bar{\omega}) \, du \right]^2}, \tag{166}$$

where, as in (147),

$$H(u) = [A_0^2(\bar{\omega} + u)(1 + u/\bar{\omega})^2 + A_0^2(\bar{\omega} - u)(1 - u/\bar{\omega})^2]^{1/2}.$$

The optimum receiving filter characteristic is, in accordance with (145),

$$R^o(u) = c2^{1/4} \cos (\pi u/4\bar{\omega})[H(u)]^{-1/2}. \tag{167}$$

With a raised cosine spectrum at the detector input, the spectrum of the demodulated pulses is as shown in Fig. 12. For a low-pass filter with a transmission-frequency characteristic as shown in Fig. 13, approximate values of $L(u) = A_0(\bar{\omega} + u)$, $G(u)$ and $H(u)$ are given in Table VI by way of illustration.

Table VI — Illustrative Values of L, G and H

	$-u/\bar{\omega}$							
	−0.75	−0.5	0	0.5	0.75	1.0	1.5	2.0
$L(u)$	0	0.65	1.2	1.15	1.05	1.00	1.15	1.2
$G(u)$	0	0.8	1.44	0.30	0.05	0	0.185	0.51
$H(u)$	0.27	1.1	1.7	1.1	0.27	0	0.58	1.2

For the particular low-pass filter above, the optimum receiving filter characteristic obtained from (167) and the corresponding transmitting filter characteristic $T^o(u) = A(u)/R^o(u)$ are as shown in Fig. 17, where they are normalized to unity for $u = 0$. It will be noted that, for $u = \pm\bar{\omega}$, $R^o = \infty$ and $T^o = 0$. Hence these optimum characteristics cannot be realized physically, though they can be approached, as indicated by the dotted lines near $u = \pm\bar{\omega}$.

For the above low-pass filter, numerical integration of (164) and (166) gives $\rho = 0.5$ and $\gamma^o = 0.84$.

The significance of the above various numerical results are, in summary, as follows.

The factor $\lambda^o = 2.8$ indicates about a 4.5-db disadvantage in signal-to-noise ratio, for an optimum FM system without a postdetection filter as compared to an optimum bipolar AM or baseband system, for equal average signal power at the input of the receiving filter. The factor $\rho = 0.5$ indicates a 3-db improvement in signal-to-noise ratio obtained with the aid of the particular postdetection low-pass filter assumed above, so that the above FM disadvantage is reduced to 1.5 db; i.e., $\gamma = \lambda^o \rho = 1.4$.

In accordance with the discussion in Section XV, a somewhat lower factor γ could be realized when the division of channel shaping between transmitting and receiving bandpass filters is optimized for this particular low-pass filter. The lower bound is represented by $\gamma^o = 0.84$, and the minimum value that could be realized with an appropriate modification in filter characteristics near $u = \bar{\omega}$ would be greater than 0.84 but less than $\gamma = 1.4$. With the modification indicated in Fig. 17, it turns out

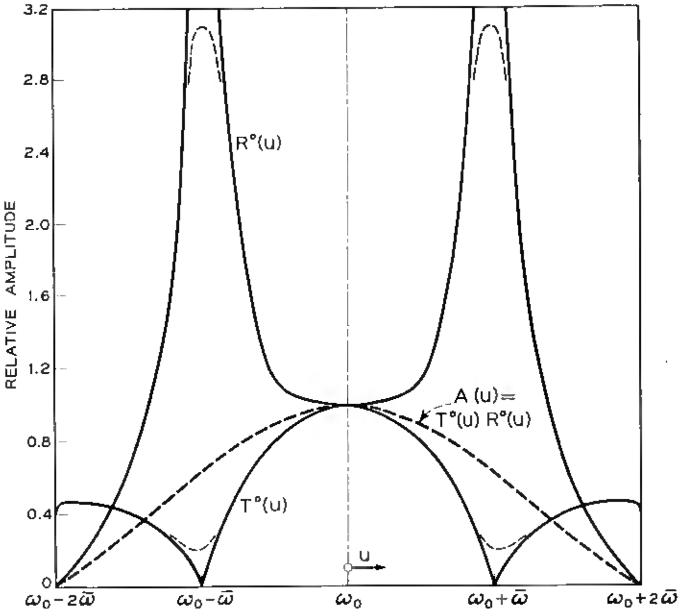


Fig. 17 — Frequency modulation with raised cosine spectrum and post-detection low-pass filter with transmission characteristic $A_0(\omega)$ shown in Fig. 13: R^o = optimum shape of receiving filter; T^o = optimum shape of transmitting filter; $A = R^oT^o$ = combined transmission characteristic.

that the ratio $N_0/\hat{S}_0$ obtained from (125) is smaller than that for bipolar AM by a factor $\gamma \cong 0.94$. Thus, with optimum design in FM, the signal-to-noise ratio would be very nearly the same as for an optimum bipolar AM or baseband system.

If the spectrum $S_0(\omega)$ shown in Fig. 12 is converted into a flat spectrum of bandwidth $2\bar{\omega}$ by an appropriate low-pass filter, it turns out that $\gamma^o \cong 0.64$, corresponding to about a 2-db advantage in signal-to-noise ratio over an optimum bipolar AM or baseband system.

As an illustration of the penalty incurred in departing from the optimum division of channel shaping between transmitting and receiving bandpass filters, it will be assumed that $R(u)$ is given by (135), as previously considered for AM. In this case, the receiving filter is flat between $-2\bar{\omega} < u < 2\bar{\omega}$, and (125) becomes

$$N_0/\hat{S}_0 = \frac{2n}{4\pi TP} \int_{-2\bar{\omega}}^{2\bar{\omega}} L^2(u)(1 + u/\bar{\omega})^2 \, du \cdot \int_{-2\bar{\omega}}^{2\bar{\omega}} \frac{T^2}{4} \cos^4 (\pi u/4\bar{\omega}) \, du, \tag{168}$$

which can be written as

$$N_0/\hat{S}_0 = \frac{N}{4P} \gamma, \tag{169}$$

with

$$\gamma = \left(\frac{1}{\bar{\omega}}\right)^2 \int_{-2\bar{\omega}}^{2\bar{\omega}} L^2(u)(1 + u/\bar{\omega})^2 \, du \cdot \int_0^{2\bar{\omega}} \cos^4 (\pi u/4\bar{\omega}) \, du \tag{170}$$

$$\cong 1.66 \text{ (by numerical integration)},$$

where the numerical result applies for a postdetection low-pass filter with the same amplitude characteristic as assumed previously.

The value $\gamma = 1.66$ corresponds to a 2.2-db disadvantage in signal-to-noise ratio as compared to an optimum bipolar AM system, and about a 0.7-db disadvantage compared to an optimum FM system ($\gamma = 1.4$). With the above type of flat receiving filter, about a 1.8-db penalty in signal-to-noise ratio was incurred in AM, as shown in Section XIV.

XVIII. SIGNAL-TO-NOISE RATIOS AND ERROR PROBABILITIES

In the case of baseband transmission or AM with homodyne detection, the probability of exceeding the rms noise amplitude by a specified factor follows the normal law.

If the polarity of the noise voltage is specified, (i.e., positive or negative) the probability of exceeding the rms noise amplitude by a factor k is

$$p = \tfrac{1}{2} \operatorname{erfc} (k/2^{1/2}), \tag{171}$$

where erfc is the error function complement.

If the noise amplitude at a sampling instant t is $[N_0(t)]^{1/2}$, an error will occur if the ratio $[N_0(t)/\hat{S}_0]^{1/2}$ exceeds $\frac{1}{2}$ in the presence of a space or negative pulse, or exceed $-\frac{1}{2}$ in the presence of a mark or positive pulse. The rms ratio, $(N_0/\hat{S}_0)^{1/2}$, must be held smaller by an appropriate factor $1/k$. The probability of an error in a digit is accordingly $p = p_e$, as given by (171), provided

$$(N_0/\hat{S}_0)^{1/2} = \frac{1}{2k}. \tag{172}$$

In accordance with (129),

$$(P/N)^o = \tfrac{1}{4}(\hat{S}_0/N_0). \tag{173}$$

From (172) and (173),

$$k = (P/N)^{1/2}. \tag{174}$$

Hence, for an optimum AM system as assumed above, the error probability in binary bipolar pulse transmission is, with (174) in (171),

$$p_e = \tfrac{1}{2} \operatorname{erfc} (P/2N)^{1/2}. \tag{175}$$

In Table VII the signal-to-noise ratios P/N obtained from (175) are shown for various probabilities of an error in the digit. In accordance with (129), under the optimum condition the average signal power P in bipolar AM is equal to the continuous mark power P_m.

Ideal synchronous detection, as assumed above, cannot be fully realized with symmetrical bipolar AM methods. Derivation of a demodulating carrier from the signal wave, or adequate phase control of a locally supplied carrier, entails some increase in either signal power, noise power or bandwidth, depending on the particular method used, and thus entails a somewhat greater ratio P/N than that given in Table VII.

In the case of unipolar or "on-off" baseband or double-sideband AM with homodyne detection, the maximum tolerable peak noise power is 6 db less than it is with bipolar transmission, but the average signal power is reduced by 3 db, so that the ratio P/N must be increased by 3 db, as indicated for unipolar AM in Table VII. The average signal power is, in this case, 3 db less than the signal power during a continuous mark.

Table VII — Optimum Signal-to-Noise Ratios P/N in Decibels for Binary AM and FM Systems[A]

Probability of an Error in a Digit	AM with Ideal Synchronous Detection[B]		FM with Ideal Frequency Discriminator Detection		
	Bipolar (Two-Phase)	Unipolar (On-Off)[C, D]	No Low-Pass Filter[E]	With Low-Pass Filter[E, F]	With Low-Pass Filter[F, G]
10^{-2}	7.3	10.3 (11.3)	11.8	8.8	6.5
10^{-4}	11.4	14.4	15.9	12.9	10.6
10^{-6}	13.6	16.6 (16.9)	18.1	15.1	12.8
10^{-8}	15.0	18.0	19.5	16.5	14.2
10^{-10}	16.0	19.0	20.5	17.5	15.2
10^{-12}	17.0	20.0	21.5	18.5	16.2

P = Average signal power at input of receiving filter
N = Average noise power in flat band $W = 1/2T$ cps at input of receiving filter
T = Interval between pulses, in seconds

Notes:

[A] Signal-to-noise ratios in terms of noise power $2N$ in double-sideband channel of bandwidth $2W$ are 3 db smaller than in table.

[B] Applies for baseband transmission, double-sideband, double-sideband on two carriers at quadrature and vestigial-sideband AM.

[C] Signal-to-noise ratios in terms of steady mark power are 3 db greater than in table.

[D] Values in brackets are for double-sideband with optimum envelope detection.

[E] Band-pass filter shaping as in Fig. 16.

[F] Low-pass filter shaping as in Fig. 13.

[G] Band-pass filter shaping as in Fig. 17.

For the reasons discussed in Section XIV, the optimum signal-to-noise ratios shown in Table VII for bipolar AM apply for bipolar baseband transmission, for bipolar double-sideband AM (phase reversal or two-phase transmission), for bipolar quadrature double-sideband AM (four-phase transmission) and for bipolar vestigial-sideband AM. Similarly, the ratios shown for unipolar AM with synchronous detection apply for unipolar baseband, double-sideband, quadrature double-sideband and vestigial-sideband transmission. Furthermore, with optimum division of channel shaping between transmitting and receiving filters, the above optimum signal-to-noise ratios apply for all spectra of the pulses at the detector input with the properties illustrated in Figs. 1 and 2. Moreover, the signal-to-noise ratios apply not only for pulses of duration T equal to the pulse interval, as assumed in the previous analysis, but also for pulses of shorter duration. When the duration of the pulses is less than T, the receiving filter characteristic remains unchanged, but the shape of the transmitting filter is modified because of the different spectrum of the modulating pulses. This also applies with other than rectangular shapes of the modulating pulses.

There are thus an infinite number of optimum AM systems with a performance from the standpoint of signal-to-noise for a given error probability equivalent to that of a baseband system of the minimum possible bandwidth, provided ideal homodyne (synchronous) detection is used, in which the pulse train is applied to a product demodulator together with a constant demodulating carrier of proper phase.

The ratios shown in Table VII for unipolar AM also apply in a first approximation to "on-off" double-sideband AM with envelope rather than homodyne detection. However, in the latter case both the average noise power and the probability distribution at the detector output differ between mark and space.[4] During a mark, the rms noise amplitude, and the probability that this noise amplitude is exceeded by a factor k, is virtually the same as it is with homodyne detection, for large signal-to-noise ratios. During a space, however, the rms noise amplitude is increased by a factor of $2^{1/2}$, and the probability that the rms amplitude is exceeded by a specified factor k follows the Rayleigh law

$$p = e^{-k^2} \tag{176}$$

rather than the Gaussian law (171).

For the above reason, the optimum slicing or threshold level is not one-half the peak pulse amplitude, but slightly greater, depending on signal-to-noise ratio and error probability. The optimum slicing levels with binary double-sideband AM and envelope detection and the corresponding optimum signal-to-noise ratios versus error probability have been determined elsewhere,[4] and are indicated in Table VII for two cases. For an error probability of 10^{-6} the optimum threshold level is about 52 per cent of the peak pulse amplitude and the signal-to-noise ratio is about 0.3 db greater than for unipolar AM with homodyne detection. Hence, for error probabilities in the range ordinarily considered acceptable, the difference in signal-to-noise ratio with envelope and homodyne detection is insignificant.

Comparison of binary FM and AM on the basis of signal-to-noise ratios is legitimate provided that, for a given ratio $N_0/\hat{S}_0$, the error probability is the same in FM and AM. For high signal-to-noise ratios, this is approximately the case, since the normal law (171) is then closely approximated in FM.[9] On this premise, comparison on the basis of signal-to-noise ratios is legitimate for small error probabilities.

In accordance with the discussion in Section XV, the optimum signal-to-noise ratio in binary FM without a postdetection filter is related to the optimum ratio in bipolar binary AM by

$$(P/N)^o_{\mathrm{FM}} = \lambda^o (P/N)^o_{\mathrm{AM}}, \tag{177}$$

where $\lambda^o \cong 2.65$ (4.2 db) for a flat spectrum as considered in Section XVI and $\lambda^o \cong 2.8$ (4.5 db) for a raised cosine spectrum at the detector input, as considered in Section XVII. In Table VII the latter case is assumed as being the more representative, and the signal-to-noise ratios are taken 4.5 db greater than they are in bipolar AM.

With the aid of an appropriate postdetection low-pass filter, the signal-to-noise ratio can be improved such that

$$(P/N)_{\text{FM}}^o = \rho\lambda^o(P/N)_{\text{AM}}^o. \tag{178}$$

With a postdetection filter having an amplitude characteristic as shown in Fig. 13, $\rho \cong 0.5$, corresponding to 3-db improvement in signal-to-noise ratio, and this case is assumed in Table VII.

As discussed in Sections XV, XVI and XVII, when a postdetection low-pass filter is used, it is possible with FM to realize some improvement in signal-to-noise ratio over bipolar AM. In this case, a lower bound is given by

$$(P/N)_{\text{FM}}^o = \gamma^o(P/N)_{\text{AM}}^o, \tag{179}$$

where $\gamma^o \cong 0.84$ with the type of filter shown in Fig. 13. This corresponds to about an 0.8-db improvement in signal-to-noise ratio over bipolar AM and entails transmitting and receiving bandpass filter characteristics as indicated in Fig. 17. This lower bound is given in the last column of Table VII, but it cannot be fully realized, for reasons discussed in Section XV.

As noted before, the optimum signal-to-noise ratios given in Table VII for synchronous AM are universal and apply to a variety of optimized systems, including the special case of ideal flat channels of minimum bandwidth assumed in other analyses of baseband,[3] synchronous AM or PM systems.[5,6] With appropriate allowance for different definitions of signal power (continuous mark versus average power) and of the bandwidth used in specifying noise power (flat single-sideband versus flat double-sideband), the results given in Table VII for synchronous AM conform with those in the above references.

In the case of FM, however, the optimum signal-to-noise ratio depends on several factors that need not be considered in AM, such as the shape of the spectrum at the detector input and the shape of the post-detection low-pass filter. There is thus no universal optimum signal-to-noise ratio for a given error probability in FM, and the ratios given in Table VII for FM apply for the particular conditions indicated. For this reason, significant comparisons cannot be made with signal-to-noise ratios for FM given elsewhere[5,7] that are based on simplified mathematical models that ignore the various factors above. One analysis[5] indicates about a 1-db advantage of bipolar AM (or phase reversal) over FM for an error probability of 10^{-4}.

XIX. SUMMARY

It has been shown that binary pulses can be transmitted without intersymbol interference by FM over a channel of the same bandwidth as is required for double-sideband AM. To this end, a first requirement is that the total frequency shift between space and mark be equal to the pulse transmission rate, for example 500 cps for 500 bits per seconds. A second basic requirement is that the pulses at the input of the frequency modulator have the appropriate shape, a condition that is met with rectangular modulating pulses of duration equal to the interval between pulses. A third requirement with rectangular modulating pulses is that the channel bandwidth be small in relation to the midband frequency. A fourth requirement is that the bandpass channel must have the appropriate amplitude-versus-frequency characteristic and a linear phase characteristic.

The appropriate amplitude characteristic of the bandpass channel is not the same as for AM, nor is the shape of the received pulses the same. By way of illustration, a comparison is made in Fig. 3 of the amplitude characteristic of the bandpass channels for FM and AM, for a channel of the minimum possible bandwidth, if intersymbol interference is to be avoided. With the channel characteristics shown in Fig. 3, the transmission of a single pulse, (i.e., transmission of a mark, preceded and followed by a continuing space) will give rise to a flat frequency spectrum at the detector input with both FM and AM, as indicated in Fig. 3. Such a flat spectrum at the detector input will give rise to a pulse at the detector output, but the pulse shape is not the same in FM and AM, as illustrated in Fig. 4. However, a common property of the pulse shapes shown in Fig. 4 is that they have zero points at intervals T equal to the duration of the rectangular modulating pulses. Thus, pulses can be transmitted at these intervals without intersymobl interference by FM or AM.

In actual pulse systems, channels of the minimum possible bandwidth are not practicable for various reasons. In Fig. 5 comparison is made of the appropriate amplitude characteristics with FM and AM for channels with twice the minimum bandwidth. In this case, transmission of a single pulse by FM or AM gives rise to a "raised cosine" spectrum at the detector input and to pulse shapes at the detector output, as shown in Fig. 6. Because of the small oscillations in the tails of the received pulses, the channel characteristics shown in Fig. 6 are desirable for FM and AM systems.

In the above illustrations, rectangular modulating pulses were assumed. With modulating pulses of other shapes that overlap between pulse intervals, certain restrictions are imposed on the shape of the pulses if intersymbol interference is to be avoided, which renders determination of the exact appropriate shapes difficult. For example, $(\sin x)/x$ modulating pulses, which are often considered in AM, are inappropriate in FM. However, with raised cosine modulating pulses, as shown in Fig. 9, it is possible by appropriate channel shaping to virtually avoid intersymbol interference. With such modulating pulses and a channel characteristic as shown in Fig. 9, the shape of the demodulated pulses will be virtually the same as those shown in Fig. 6 for FM. Although intersymbol interference cannot be avoided with raised cosine modulating pulses, it is small enough to be disregarded (less than 1 per cent).

The pulses at the output of the FM detector, such as those shown in Figs. 4 and 6, have baseband spectra of infinite bandwidth, as in Figs. 10 and 12. It is possible to modify the shape of the pulses in such a way that intersymbol interference is not introduced, with the aid of a postdetection low-pass filter having the appropriate amplitude characteristic. For example, the FM pulse of Fig. 4 has a spectrum as shown in Fig. 10. This spectrum can be converted into a flat spectrum of minimum bandwidth with the aid of a postdetection low-pass filter having the amplitude characteristic shown in Fig. 11 and a linear phase characteristic. The pulse shown in Fig. 4 for FM would thereby be converted into the same shape as shown for AM.

The ideal amplitude characteristics of the bandpass channels in FM, such as those exemplified in Figs. 3 and 5, are obtained with the aid of an appropriate combination of transmitting and receiving bandpass filters. From the standpoint of intersymbol interference, as considered above, the division of channel shaping between these filters is immaterial, but there is an optimum division from the standpoint of performance in the presence of noise.

By way of example, for the over-all amplitude characteristics of the bandpass channels shown in Fig. 5 for AM, the optimum division of channel shaping between transmitting and receiving filters is shown in Fig. 15 for the case of random noise. As discussed in Section XVIII, with optimum division of channel shaping and ideal synchronous (homodyne) detection, there is an infinity of optimum AM systems with performance equivalent to that of a baseband system of the minimum possible bandwidth, as regards signal-to-noise ratio for a given pulse transmission rate and error probability.

With an over-all channel characteristic as shown in Fig. 5 for FM, the optimum division of channel shaping is as shown in Fig. 16, assuming no postdetection low-pass filter. Such an optimum FM system has about a 4.5-db disadvantage in signal-to-noise ratio compared to an optimum bipolar AM or phase reversal system, for equal average signal power. By providing a postdetection low-pass filter with a transmission characteristic as shown in Fig. 13 and linear phase, the signal-to-noise ratio is improved about 3 db, and the FM disadvantage of 4.5 db is reduced to about 1.5 db. This assumes the same division of bandpass channel shaping as without a low-pass filter. However, the optimum division with the above low-pass filter is different, and is approximately as indicated in Fig. 17, except near $\omega_0 \pm \bar{\omega}$. With an optimum division, it appears possible, in principle, to realize an advantage in signal-to-noise ratio of at most 0.8 db over an optimum bipolar AM or baseband system, for equal average signal power.

With appropriate postdetection low-pass filters of the minimum permissible bandwidth, it is possible in principle to realize an advantage in signal-to-noise ratio of at most 2 db over an optimum bipolar AM or baseband system. However, the above FM advantages cannot be fully attained in practice. They are principally of theoretical interest in that they indicate that an advantage in signal-to-noise ratio can be derived from the unavoidable two-fold increase in channel bandwidth with FM as compared to baseband transmission, or equivalent AM methods.

XX. ACKNOWLEDGMENTS

The writer is indebted to H. O. Pollak for a critical review that resulted in a radical modification of Sections VIII and X, to Miss Marion C. Gray for supervision of computer programming carried out by Miss D. G. Becker to determine the pulse spectra given in Section X, to S. O. Rice for pointing out the Anger function relations used in Section IX and for helpful suggestions at other points in the analysis, and to G. Wascheck for many editorial comments.

APPENDIX A

Pulse Train Envelopes and Average Signal Power

In Section IV the transmission of a single mark or pulse was considered, and the resultant wave at the channel output (detector input) was, in this case, given by (35) and (36), or

$$E(t_0) = \cos(\omega_0 t_0 + \varphi_0 + \psi_0)\bar{E}(t_0), \tag{180}$$

where

$$\bar{E}(t_0) = A(-\bar{\omega})[1 + \mu^2 p^2(t_0) - 2\mu p(t_0) \cos \bar{\omega} t_0]^{1/2} \tag{181}$$

and

$$\mu = 1/A(-\bar{\omega}) = 1/A(\bar{\omega}). \tag{182}$$

When a train of pulses is transmitted, the resultant envelope is given by (95) or

$$\begin{aligned}\mathbf{W}(t_0) = A(-\bar{\omega}) \Bigg[1 + \mu^2 \sum_{m=-\infty}^{\infty} a_m^{\,2} p^2(t_0 - mT) \\ - 2\mu \sum_{m=-\infty}^{\infty} a_m p(t_0 - mT) \cos \bar{\omega}(t_0 - mT) \Bigg]^{1/2},\end{aligned} \tag{183}$$

where T is the interval between pulses and

$$\begin{aligned} a_m &= 0 \qquad \text{for space} \\ &= 1 \qquad \text{for mark.} \end{aligned} \tag{184}$$

In view of (184), the rms value of the envelope at a particular time, t_0, with respect to a sampling point is for equal probability of marks and spaces:

$$\begin{aligned}\bar{W}(t_0) = A(-\bar{\omega}) \Bigg[1 + \frac{\mu^2}{2} \sum p^2(t_0 - mT) \\ - \mu \sum p(t_0 - mT) \cos \bar{\omega}(t_0 - mT) \Bigg]^{1/2},\end{aligned} \tag{185}$$

where the limits of the summations are as in (183).

The mean squared value of the envelope taken over a pulse interval T is

$$\begin{aligned}\bar{W}^2 &= \frac{1}{T} \int_{-T/2}^{T/2} \bar{W}^2(t_0)\, dt_0 \\ &= A^2(-\bar{\omega}) \Bigg[1 + \frac{\mu^2}{2} \int_{-T/2}^{T/2} \sum p^2(t_0 - mT)\, dt_0 \\ &\qquad - \mu \int_{-T/2}^{T/2} \sum p(t_0 - mT) \cos \bar{\omega}(t_0 - mT)\, dt_0 \Bigg],\end{aligned} \tag{186}$$

which can be transformed into

$$\bar{W}^2 = A^2(-\bar{\omega}) \left[1 + \frac{\mu^2}{2T} \int_{-\infty}^{\infty} p^2(t_0)\, dt - \frac{\mu}{T} \int_{-\infty}^{\infty} p(t_0) \cos \bar{\omega} t_0\, dt_0 \right]. \tag{187}$$

In (32) for $p(t_0)$, the lower limit of integration, $-\omega_0$, can be replaced by $-\infty$, since $S(u) = 0$ for $u \leqq -\omega_0$, in which case (32) can be written

$$p(t_0) = \frac{1}{2\pi} \int_{-\infty}^{\infty} 2S(u) \cos ut_0 \, du \tag{188}$$

or, by inversion,

$$2S(u) = \int_{-\infty}^{\infty} p(t_0) \cos ut_0 \, dt_0 . \tag{189}$$

In view of (189), the last-bracket term in (187) with $u = \bar{\omega}$ becomes $-2\mu S(\bar{\omega})/T$. Since $S(\bar{\omega}) = T/4$, in accordance with (48), and $\mu = 2$, in accordance with (50), the last-bracket term becomes -1, and (187) simplifies to

$$\bar{W}^2 = A^2(-\bar{\omega}) \frac{\mu^2}{2T} \int_{-\infty}^{\infty} p^2(t_0) \, dt_0 . \tag{190}$$

In view of (189),

$$\int_{-\infty}^{\infty} p^2(t_0) \, dt_0 = \frac{1}{2\pi} \int_{-\infty}^{\infty} [2S(u)]^2 \, du. \tag{191}$$

Hence (190) can be written

$$\bar{W}^2 = A^2(-\bar{\omega}) \frac{\mu^2}{\pi T} \int_{-\infty}^{\infty} S^2(u) \, du. \tag{192}$$

This is the mean squared value of the envelope at the output of a channel with an over-all amplitude characteristic $A(u)$. If the transmitting filter is assumed to have an amplitude characteristic $T(u)$ and the receiving filter to have a characteristic $R(u)$, then

$$T(u)R(u) = A(u). \tag{193}$$

The spectrum at the transmitter output, or at the input of the receiving filter in a channel without transmission loss, is in this case $T(u)S(u) = S(u)/R(u)$. The mean squared value of the envelope at the transmitter output is obtained by replacing $A(\bar{\omega})$ by $T(\bar{\omega})$ and $S(u)$ by $S(u)/R(u)$ in (192), so that in this case

$$\bar{W}_0^2 = T^2(\bar{\omega}) \frac{1}{\pi T} \left[\frac{R(\bar{\omega})}{A(\bar{\omega})} \right]^2 \int_{-\infty}^{\infty} \left[\frac{S(u)}{R(u)} \right]^2 du, \tag{194}$$

where $A(\bar{\omega}) = \frac{1}{2}$.

When a continuous mark with unit amplitude of the carrier at the input of the transmitting filter is transmitted, the envelope of the carrier at the output is $T(\bar{\omega})$. It will be assumed that the carrier amplitude at the input is so chosen that the peak amplitude at the output is E for a continuous mark or space. With $T(\bar{\omega})$ replaced by E in (194), and with $A(\bar{\omega}) = \frac{1}{2}$,

$$\bar{W}^2 = E^2 \frac{4R^2(\bar{\omega})}{\pi T} \int_{-\infty}^{\infty} \frac{S^2(u)}{R^2(u)}\, du. \tag{195}$$

The average signal power within the envelope is, in view of the factor $\cos(\omega_0 t_0 + \varphi_0 + \psi_0)$ in (180), smaller by a factor of $\frac{1}{2}$, or

$$P_{\text{FM}} = \frac{E^2}{2} \frac{4R^2(\bar{\omega})}{\pi T} \int_{-\infty}^{\infty} \frac{S^2(u)}{R^2(u)}\, du. \tag{196}$$

In bipolar AM (183) is replaced by

$$\bar{W}(t_0) = A(0) \sum_{m=-\infty}^{\infty} a_m p(t_0 - mT), \tag{197}$$

where $a_m = -1$ or 1.

Equation (190) is replaced by

$$\begin{aligned} \bar{W}^2 &= A(0)^2 \frac{1}{T} \int_{-\infty}^{\infty} p^2(t_0)\, dt_0 \\ &= A^2(0) \frac{2}{\pi T} \int_{-\infty}^{\infty} S^2(u)\, du, \end{aligned} \tag{198}$$

and (194) is replaced by

$$\bar{W}_0^2 = T^2(0) \frac{2}{\pi T} \left[\frac{R(0)}{A(0)}\right]^2 \int_{-\infty}^{\infty} \frac{S^2(u)}{R^2(u)}\, du. \tag{199}$$

When the peak amplitude of the transmitted carrier for a continuous mark is E, the average signal power in this case is, with $A(0) = 1$,

$$P_{\text{AM}} = \frac{E^2}{2} \frac{2}{\pi T} R^2(0) \int_{-\infty}^{\infty} \frac{S^2(u)}{R^2(u)}\, du. \tag{200}$$

APPENDIX B

Interference in AM and FM Systems

B.1. *General*

B.1.1 *Frequency Modulation*

Let a carrier of frequency $\omega_0 - \bar{\omega}$ representing a space be transmitted, and let its peak amplitude at the input of the receiving filter be E. If the receiving filter has an amplitude characteristic $R(u)$ as a function of the frequency u from midband, the signal at the detector input is

$$e_s(t_0) = -ER(-\bar{\omega})[\cos(\omega_0 - \bar{\omega})t_0 + \varphi_0]. \tag{201}$$

An interfering voltage at the detector input can be written in the general form

$$e_i(t_0) = r_i(t_0)\cos(\omega_0 t_0 + \varphi_0) + q_i(t_0)\sin(\omega_0 t_0 + \varphi_0). \tag{202}$$

The signal at the detector input in the presence of interference is then

$$\begin{aligned} e_s + e_i = -\cos(\omega_0 t_0 + \varphi_0)[ER(-\bar{\omega})\cos\bar{\omega}t_0 - r_i(t_0)] \\ -\sin(\omega_0 t_0 + \varphi_0)[ER(-\bar{\omega})\sin\bar{\omega}t_0 - q_i(t_0)]. \end{aligned} \tag{203}$$

The phase at the detector input in the presence of interference is given by

$$\tan\psi_{0,i} = -\frac{\sin\bar{\omega}t_0 - \mu_i q_i}{\cos\bar{\omega}t_0 - \mu_i r_i}, \tag{204}$$

where

$$\mu_i = \frac{1}{ER(-\bar{\omega})}. \tag{205}$$

The demodulated signal in presence of interference is proportional to $d\psi_{0,i}/dt_0$, which becomes

$$\begin{aligned} \psi_{0,i}' = -\frac{\bar{\omega}}{D}\Big[1 - \mu_i(r_i\cos\bar{\omega}t_0 + q_i\sin\bar{\omega}t_0) \\ + \frac{\mu_i}{\bar{\omega}}(r_i'\sin\bar{\omega}t_0 - q_i'\cos\omega t_0) + \frac{\mu_i^2}{\bar{\omega}}(q_i'r_i - r_i'q_i)\Big], \end{aligned} \tag{206}$$

where $r_i' = dr_i/dt_0$, $q_i' = dq_i/dt_0$ and

$$\begin{aligned} D &= (\sin\bar{\omega}t_0 - \mu_i q_i)^2 + (\cos\bar{\omega}t_0 - \mu_i r_i)^2 \\ &= 1 + \mu_i^2(r_i^2 + q_i^2) - 2\mu_i(r_i\cos\bar{\omega}t_0 + q_i\sin\bar{\omega}t_0). \end{aligned} \tag{207}$$

The frequency deviation with respect to the frequency $\omega_0 - \bar{\omega}$ is $\psi_i' = \bar{\omega} + \psi_{0,i}'$, and the ratio $\eta_i = \psi_i'/2\bar{\omega}$ becomes

$$\begin{aligned} \eta_i(t_0) = \frac{\mu_i}{2D}\Big[\mu_i(r_i^2 + q_i^2) - r_i\cos\bar{\omega}t_0 - q_i\sin\bar{\omega}t_0 \\ - \frac{1}{\bar{\omega}}(r_i'\sin\bar{\omega}t_0 - q_i'\cos\bar{\omega}t_0) - \frac{\mu_i}{\bar{\omega}}(q_i'r_i - r_i'q_i)\Big]. \end{aligned} \tag{208}$$

This is the amplitude of the interfering voltage taken in relation to the peak amplitude of a demodulated pulse.

At sampling points $t_0 = mT$, $\sin \bar{\omega} t_0 = 0$ and $\cos \bar{\omega} t_0 = (-1)^m$. At these points

$$\eta_i(mT) = \frac{\mu_i}{2D}\left[\mu_i(r_i^2 + q_i^2) - (-1)^m\left(r_i - \frac{1}{\bar{\omega}} q_i'\right) - \frac{\mu_i}{\bar{\omega}}(q_i' r_i - r_i' q_i)\right], \tag{209}$$

$$D(mT) = -(-1)^m 2\mu_i r_i + \mu_i^2(r_i^2 + q_i^2). \tag{210}$$

If second-order terms are neglected, which is permissible for adequately low amplitudes of the interfering voltage, the above expression reduces to

$$\eta_i(mT) \cong -\frac{1}{2ER(-\bar{\omega})}\frac{r_i(mT) - q_i'(mT)/\bar{\omega}}{1 - (-1)^m r_i(mT)/ER(-\bar{\omega})} \tag{211}$$

$$\cong -\frac{1}{2ER(-\bar{\omega})}[r_i(mT) - q_i'(mT)/\bar{\omega}]. \tag{212}$$

B.1.2 *Bipolar AM or Two-Phase Modulation*

In bipolar AM with homodyne detection, let a carrier $-E \cos (\omega_0 t_0 + \varphi_0)$ represent a space and a carrier $E \cos (\omega_0 t_0 + \varphi_0)$ a mark. The signal plus interference at the detector input during a space in this case is

$$e_s + e_i = -\cos (\omega_0 t_0 + \varphi_0)[ER(0) - r_i] + \sin (\omega_0 t_0 + \varphi_0) q_i \tag{213}$$

The demodulated output after elimination of high-frequency demodulation products by low-pass filtering is

$$V_{s+i} = -\tfrac{1}{2}[ER(0) - r_i(t_0)]. \tag{214}$$

If a space is represented by 0 at the output, rather than by $-ER(0)/2$, the demodulated output is $V_{s+i} + ER(0)/2$, or $r_i/2$. The resultant interference taken in relation to the amplitude $ER(0)$ of a mark is

$$\eta_i = \frac{r_i(t_0)}{2ER(0)}. \tag{215}$$

B.1.3 *Unipolar AM with Envelope Detection*

In unipolar or "on-off" pulse transmission with envelope detection, the demodulated interference voltage in the presence of a space (zero carrier) is

$$V_i^{(s)} = (r_i^2 + q_i^2)^{1/2}. \tag{216}$$

The amplitude of the interfering voltage taken in relation to the amplitude of a demodulated mark is

$$\eta_i^{(s)} = \frac{(r_i^2 + q_i^2)^{1/2}}{ER(0)} . \tag{217}$$

When a mark is transmitted, the voltage at the detector input is

$$e_{s,i}^{(m)} = \cos(\omega_0 t_0 + \varphi_0)[ER(0) + r_i] + \sin(\omega_0 t_0 + \varphi_0) q_i \tag{218}$$

and at the detector output is

$$\begin{aligned} V_{s,i}^{(m)} &= \{[ER(0) + r_i]^2 + q_i^2\}^{1/2} \\ &\cong ER(0) + r_i . \end{aligned} \tag{219}$$

The amplitude of the interfering voltage taken in relation to the amplitude of a demodulated mark in this case is

$$\eta_i^{(m)} = \frac{r_i}{ER(0)} . \tag{220}$$

B.2 *Single Frequency Interference*

In the particular case of a sinusoidal interfering voltage of frequency $\omega_0 + u$ and amplitude $e(u)$ at the input of the receiving filter, the interfering voltage at the detector input is

$$e_i(t_0) = e(u)R(u) \cos [(\omega_0 + u)t_0 + \varphi_1] \tag{221}$$

$$\begin{aligned} &= e(u)R(u) \cos(\omega_0 t_0 + \varphi_0) \cos(ut_0 + \varphi) \\ &\quad - e(u)R(u) \sin(\omega_0 t_0 + \varphi_0) \sin(ut_0 + \varphi), \end{aligned} \tag{222}$$

where $\varphi = \varphi_1 - \varphi_0$.

In this case,

$$\begin{aligned} r_i(t_0) &= e(u)R(u) \cos(ut_0 + \varphi), \\ q_i(t_0) &= -e(u)R(u) \sin(ut_0 + \varphi). \end{aligned} \tag{223}$$

B.2.1 *Frequency Modulation*

In the case of FM, (212) becomes

$$\eta_i \cong -\frac{e(u)R(u)}{2ER(-\bar{\omega})} \left(1 + \frac{u}{\bar{\omega}}\right) \cos(ut_0 + \varphi). \tag{224}$$

The rms interference in FM with all phases, φ, equally probable is

$$\begin{aligned}\bar{\eta}_i &\cong \frac{e(u)R(u)}{2ER(-\bar{\omega})}\left(1+\frac{u}{\bar{\omega}}\right)\left[\frac{1}{\pi}\int_0^\pi \cos^2(ut_0+\varphi)\,d\varphi\right]^{1/2} \\ &= \frac{\bar{e}(u)R(u)}{2ER(-\bar{\omega})}\left(1+\frac{u}{\bar{\omega}}\right),\end{aligned} \tag{225}$$

where

$$\bar{e}(u) = \frac{e(u)}{2^{1/2}}. \tag{226}$$

B.2.2 *Bipolar AM*

For bipolar AM, (215) gives

$$\eta_i = \frac{e(u)R(u)\cos(ut_0+\varphi)}{2ER(0)}, \tag{227}$$

and the rms value with all phases φ equally probable is

$$\bar{\eta}_i = \frac{\bar{e}(u)R(u)}{2ER(0)}. \tag{228}$$

B.2.3 *Unipolar AM with Envelope Detection*

For unipolar AM with envelope detection, (223) in (217) yields, for the interference during a space,

$$\eta_i^{(s)} = \frac{e(u)R(u)}{ER(0)} = 2^{1/2}\frac{\bar{e}(u)}{ER(0)} \tag{229}$$

and (220) gives, for interference during a mark,

$$\eta_i^{(m)} = \frac{e(u)R(u)}{ER(0)}\cos(ut_0+\varphi), \tag{230}$$

with an rms value

$$\bar{\eta}_i^{(m)} = \frac{\bar{e}(u)R(u)}{ER(0)}. \tag{231}$$

B.3 *Impulse Noise*

In the case of idealized impulse noise, the interfering voltage is of the general form

$$\begin{aligned}e_i(t_0) &= p_i(t_0)\cos(\omega_0 t_0+\varphi_i) \\ &= p_i(t_0)\cos\varphi\cos(\omega_0 t_0+\varphi_0) - p_i(t_0)\sin\varphi\sin(\omega_0 t_0+\varphi_0),\end{aligned} \tag{232}$$

where $\varphi = \varphi_i - \varphi_0$ and $p_i(t_0)$ is the envelope of the noise pulse at the detector input. The shape of the latter will depend on the receiving filter characteristic, $R(u)$. The phase φ represents the difference between the phase of the carrier within the noise pulse envelope and the phase of the carrier within the signal pulse envelope. This phase difference depends on the instant at which the impulse occurs and can have any value.

B.3.1 *Frequency Modulation*

Comparison of (232) with (202) shows that, in this case,

$$\begin{aligned} r_i(t_0) &= p_i(t_0) \cos \varphi, \\ q_i(t_0) &= -p_i(t_0) \sin \varphi \end{aligned} \tag{233}$$

In (208) the various quantities become:

$$r_i^2 + q_i^2 = p_i^2(t_0), \tag{234}$$

$$r_i \cos \bar{\omega} t_0 + q_i \sin \bar{\omega} t_0 = p_i(t_0) \cos (\bar{\omega} t_0 + \varphi), \tag{235}$$

$$r_i' \sin \bar{\omega} t_0 - q_i' \cos \bar{\omega} t_0 = p_i'(t_0) \sin (\bar{\omega} t_0 + \varphi), \tag{236}$$

$$q_i' r_i - r_i' q_i = 0. \tag{237}$$

With these values in (208), the amplitude of a noise pulse after demodulation becomes

$$\eta_i(t_0) = \frac{\mu_i}{2D} \left[\mu_i p_i^2(t_0) - p_i(t_0) \cos (\bar{\omega} t_0 + \varphi) - \frac{p_i'(t_0)}{\bar{\omega}} \sin (\bar{\omega} t_0 + \varphi) \right], \tag{238}$$

$$D = 1 + \mu_i^2 p_i^2(t_0) - 2\mu_i p_i(t_0) \cos (\bar{\omega} t_0 + \varphi). \tag{239}$$

An error will occur if, at the sampling instant, $\eta_i(t_0) \geqq \frac{1}{2}$, which gives the following relation for determining the peak amplitude of a noise pulse at the detector input that will produce an error in the demodulated signal:

$$\begin{aligned} \tfrac{1}{2}\mu_i \left[\mu_i p_i^2(t_0) - p_i(t_0) \cos (\bar{\omega} t_0 + \varphi) - \frac{p_i'(t_0)}{\bar{\omega} \sin (\bar{\omega} t_0 + \varphi)} \right] \\ \geqq \tfrac{1}{2} [1 + \mu_i p_i(t_0) - 2\mu_i p_i(t_0) \cos (\bar{\omega} t_0 + \varphi)] \end{aligned} \tag{240}$$

With $\mu_i = 1/ER(-\bar{\omega})$, this can be written

$$\frac{p_i(t_0) \cos (\bar{\omega} t_0 + \varphi) - \frac{1}{\bar{\omega}}\, p_i'(t_0) \sin (\bar{\omega} t_0 + \varphi)}{2ER(-\bar{\omega})} = \tfrac{1}{2}. \tag{241}$$

B.3.2 *FM vs. Bipolar AM*

In the case of bipolar AM, the corresponding relation is

$$\frac{p_i(t_0)\cos(\omega t_0 + \varphi)}{2ER(0)} = \tfrac{1}{2}. \tag{242}$$

If it is now assumed that the impulses occur at sampling instants $t_0 = 0$, then $p_i'(t_0) = 0$ and the following relations apply.

For FM:

$$\frac{p_i(0)\cos(\bar{\omega} t_0 + \varphi)}{2ER(-\bar{\omega})} = \tfrac{1}{2}. \tag{243}$$

For bipolar AM:

$$\frac{p_i(0)\cos(\bar{\omega} t_0 + \varphi)}{2ER(0)} = \tfrac{1}{2}. \tag{244}$$

The peak amplitudes of the impulses that will cause errors are thus smaller in FM than in AM by the factor

$$\frac{\hat{p}(0)_{\text{FM}}}{\hat{p}(0)_{\text{AM}}} = \frac{R(-\bar{\omega})}{R(0)}. \tag{245}$$

In accordance with (104) and (105), the following relation applies between the average signal powers in FM and AM:

$$\frac{P_{\text{FM}}}{P_{\text{AM}}} = 2\left[\frac{R(-\bar{\omega})}{R(0)}\right]^2. \tag{246}$$

From (245) and (246) it follows that

$$\frac{\hat{p}(0)_{\text{FM}}}{\hat{p}(0)_{\text{AM}}} = \left(\frac{P_{\text{FM}}}{P_{\text{AM}}}\right)^{1/2} \tfrac{1}{2}^{1/2}. \tag{247}$$

Thus, for equal probability of errors when the impulse occurs at sampling instants and equal average signal power $P_{\text{FM}} = P_{\text{AM}}$, the impulses at the detector input can be greater in AM than in FM by a factor $2^{1/2}$, corresponding to 3 db.

As another limiting case, assume that the impulses occur midways between sampling points corresponding to $\bar{\omega} t_0 = \pi/2$ or $t_0 = \pi/2\bar{\omega}$. In this case (241) and (242) become

for FM and $t_0 = \pi/2\bar{\omega}$:

$$\frac{-p_i(t_0)\sin\varphi + (1/\bar{\omega})p_i'(t_0)\cos\varphi}{2ER(-\bar{\omega})} = \tfrac{1}{2}; \tag{248}$$

for AM and $t_0 = \pi/2\bar{\omega}$:

$$\frac{-p_i(t_0)\sin\varphi}{2ER(0)} = \tfrac{1}{2}. \tag{249}$$

Since φ may have any value, it is permissible in (249) to substitute $\varphi_1 = (\varphi - \pi/4) = \cos\pi/4 \sin\varphi + \sin\pi/4 \cos\varphi$, in which case, for AM

$$\frac{-p_i(t_0)(\sin\varphi + \cos\varphi)}{2ER(0)} = \tfrac{1}{2}{}^{1/2}. \tag{250}$$

The following approximation applies for a representative pulse shape* and $t_0 = \pi/2\bar{\omega}$:

$$-\frac{1}{\bar{\omega}}\, p_i'(t_0) \cong p_i(t_0) \cong \tfrac{1}{2}p_i(0)\,, \tag{251}$$

where $p_i(0)$ is the peak pulse amplitude at the detector input, which will occur at a time $t_0 = \pi/2\bar{\omega}$ from a sampling point.

Hence (248) and (250) can be written

for FM:

$$-p_i(0)\;\frac{\sin\varphi + \cos\varphi}{ER(-\bar{\omega})} \cong 1\,; \tag{252}$$

for AM:

$$-p_i(0)\,\frac{\sin\varphi + \cos\varphi}{ER(0)} \cong 2^{1/2}. \tag{253}$$

The peak amplitudes that will cause errors are thus smaller in FM than in AM by the factor

$$\frac{\hat{p}(0)_{\mathrm{FM}}}{\hat{p}(0)_{\mathrm{AM}}} \cong \frac{R(-\bar{\omega})}{2^{1/2}R(0)}. \tag{254}$$

In view of (246),

$$\frac{\hat{p}(0)_{\mathrm{FM}}}{\hat{p}(0)_{\mathrm{AM}}} = \frac{1}{2}\left(\frac{P_{\mathrm{FM}}}{P_{\mathrm{AM}}}\right)^{1/2}. \tag{255}$$

Thus, for equal probability of errors when the impulses occur midways between sampling points and equal average signal power $P_{\mathrm{FM}} = P_{\mathrm{AM}}$, the peak amplitude of the impulses at the detector input can be greater in AM than FM by a factor of 2, corresponding to 6 db.

* In the case of a pulse shape obtained with a raised cosine spectrum, $p_i(t_0) = \frac{1}{2}\, p_i(0)$ and $p_i'(t_0)\bar{\omega} = 0.475\; p_i(0)$.

The peak amplitudes, $\hat{p}(0)$, required to produce an error when an impulse occurs midways between sampling points is greater than when they occur at a sampling point by a factor of $2^{1/2}$ in FM and a factor of 2 in AM. With a Gaussian amplitude distribution of the pulses, the probability of an error from a pulse midway between two sampling points is in the order of 1 per cent of the probability of an error from a pulse at a sampling point in the case of FM, and is substantially smaller for AM. Hence, virtually all the errors will be caused by pulses that occur near sampling points. The AM advantage over FM for equal error probability is 3 db for impulses that occur at sampling points, and would be expected to be only slightly greater, about 4 db when impulses occurring at all instances with respect to a sampling point are considered.

The above comparisons apply without a postdetection low-pass filter in FM. With an optimum bandpass receiving filter characteristic in FM, the reduction in peak impulse noise afforded by low-pass filter would be expected to be about the same as the reduction in average random noise.

APPENDIX C

Optimum Receiving Filter Characteristic

The optimum receiving filter characteristic in AM and in FM without a postdetection low-pass filter can be determined from the solution of the more general case considered here, of FM with a postdetection filter.

In the latter case, the optimum $R(u)$ is obtained when the product of the two integrals in (125) is a minimum, or for the minimum value of the product:

$$J = J_1 J_2 , \tag{256}$$

where J_1 and J_2 are functions of $R(u)$ given by

$$J_1 = \int_{-\infty}^{\infty} L^2(u) R^2(u) (1 + u/\bar{\omega})^2 \, du , \tag{257}$$

$$J_2 = \int_{-\infty}^{\infty} \frac{S^2(u)}{R^2(u)} \, du = 2 \int_0^{\infty} \frac{S^2(u)}{R^2(u)} \, du . \tag{258}$$

In (257), $L(-u) \neq L(u)$, so that it is convenient to resolve the integrand into one component with even symmetry with respect to u and one with odd symmetry. The integral of the latter component vanishes and that of the component with even symmetry becomes

$$J_1 = \int_0^{\infty} H^2(u) R^2(u) \, du , \tag{259}$$

where

$$H^2(u) = L^2(u)(1 + u/\bar{\omega})^2 + L^2(-u)(1 - u/\bar{\omega})^2. \tag{260}$$

When a small variation, $\delta R(u)$, is made in $R(u)$, the resultant variation in J is

$$\begin{aligned} \delta J &= J_2 \delta J_1 + J_1 \delta J_2 \\ &= J_2 \int_0^\infty 2R(u)H^2(u)\delta R(u)\, du \qquad (261) \\ &\qquad -J_1\, 2 \int_0^\infty 2\left[\frac{R^2(u)}{R^3(u)}\right] \delta R(u)\, du. \qquad (262) \end{aligned}$$

The optimum $R(u)$ is obtained when $\delta J = 0$, or

$$\int_0^\infty \left[J_2 R(u)H^2(u) - \frac{2J_1 S^2(u)}{R^3(u)}\right] \delta R(u)\, du = 0, \tag{263}$$

which is the case when

$$J_2 R(u)H^2(u) - \frac{2J_1 S^2(u)}{R^3(u)} = 0 \tag{264}$$

or

$$R(u) = R^o(u) = \frac{c2^{1/4} S^{1/2}(u)}{H^{1/2}(u)}, \tag{265}$$

where $c = (J_1/J_2)^{1/4}$ is a constant.

With (265) in (258) and (259),

$$J_1 = c^2 2^{1/2} \int_0^\infty S(u)H(u)\, du, \tag{266}$$

$$J_2 = \frac{1}{c^2} 2^{1/2} \int_0^\infty S(u)H(u)\, du, \tag{267}$$

$$\begin{aligned} J_1 J_2 &= 2\left[\int_0^\infty S(u)H(u)\, du\right]^2, \\ &= \frac{1}{2}\left[\int_{-\infty}^\infty S(u)H(u)\, du\right]^2. \end{aligned} \tag{268}$$

The optimum ratio, N_0/S_0, obtained with (268) in (125) becomes

$$(N_0/S_0)^o{}_{\mathrm{FM}} = \frac{n}{4\pi T P_{\mathrm{FM}}} \left[\int_{-\infty}^\infty S(u)H(u)\, du\right]^2. \tag{269}$$

In the case of FM without a postdetection filter, $L(u) = 1$ and, in this case,

$$H^2(u) = 2(1 + u^2/\bar{\omega}^2), \tag{270}$$

in which case (265) gives (138).

In the case of AM, the term $u^2/\bar{\omega}^2$ is absent in (270) and

$$H^2(u) = 2, \tag{271}$$

so that (265) gives (127).

REFERENCES

1. Nyquist, H., Certain Topics in Telegraph Transmission Theory, Trans. A.I.E.E., **47,** April 1928, p. 617.
2. Sunde, E. D., Theoretical Fundamentals of Pulse Transmission, B.S.T.J., **33,** May 1954, p. 721; July 1954, p. 987.
3. Oliver, B. M., Pierce, J. R. and Shannon, C. E., The Philosophy of PCM, Proc. I.R.E., **36,** November 1948, p. 1324.
4. Bennett, W. R., Methods of Solving Noise Problems, Proc. I.R.E., **44,** May 1956, p. 609.
5. Montgomery, G. F., A Comparison of Amplitude and Angle Modulation for Narrow-Band Communication of Binary-Coded Messages in Fluctuation Noise, Proc. I.R.E., **42,** February 1954, p. 447.
6. Cahn, C. R., Performance of Digital Phase Modulation Communication Systems, I.R.E. Trans., **CS-7,** May 1959, p. 3.
7. Turin, G. L., Error Probabilities for Binary Symmetric Ideal Reception Through Nonselective Slow Fading and Noise, Proc. I.R.E., **46,** September 1958, p. 1603.
8. Watson, G. N., *A Treatise on the Theory of Bessel Functions,* Cambridge Univ. Press, London, 1944.
9. Rice, S. O., Statistical Properties of a Sine Wave Plus Random Noise, B.S.T.J., **27,** January 1948, p. 109.

Gyromagnetic Modes in Waveguide Partially Loaded with Ferrite

By H. SEIDEL and R. C. FLETCHER

(Manuscript received June 26, 1959)

Analysis is made of all the propagating modes of a vanishingly small rectangular waveguide partially filled with transversely magnetized ferrite. Each of these modes is shown to propagate in only one direction and to tend to be lossy. Use of these properties can be made in the design of a novel nonresonance isolator. All but one of the propagating modes vary in amplitude along the dc magnetic field. Yet they can apparently be excited experimentally at a boundary by an incident mode, with none of the modes having any variation along the dc field. Theoretical considerations indicate that finite conductivity in the waveguide walls may be responsible for this coupling.

The unidirectional properties of these modes suggest the possibility of building purely reactive isolators, but these can be shown nonexistent from general energy considerations. Experiments are described that show that nature resolves this "paradox" by absorbing power, even in low-loss ferrite, rather than reflecting it. Some possible explanations of this behavior are set forth.

I. INTRODUCTION

It has been shown that, for certain ranges of transverse magnetic field, there are an infinite number of propagating modes in a waveguide completely filled with ferrite,[1,2,3] no matter how small the guide. These modes we will call gyromagnetic modes, since they have no analog in waveguides filled with isotropic material. For symmetrical structures, the modes of these completely filled waveguides show no nonreciprocal behavior. It is the intention of the present paper to study a similar set of gyromagnetic modes for a waveguide only partially filled with ferrite. Here there are displayed some interesting nonreciprocal effects, which we will describe.

The procedure used will be first to derive explicit expressions for the propagating modes for the partially filled waveguide for a given range of magnetic fields. We will then show how the nonreciprocal modes obtained can be used in a straightforward fashion to construct a novel isolator. Experimental evidence will be presented that indicates that these higher order modes may be excited from the dominant TE mode, even with boundaries that have no variation in the direction of the applied magnetic field. Some theoretical considerations will indicate that finite conductivity in the waveguide walls may be responsible for this coupling.

Finally, we will consider the possibility of building purely reactive isolators. These will be shown to be nonexistent from general energy considerations. But, as Button and Lax[4] have pointed out, there are modes which propagate in one direction but are cut off in the reverse, suggesting the possibility of reactive isolation. Some experiments will be described which show that nature resolves this "paradox" by absorbing power rather than by reflecting it. Some possible explanations of this behavior will be set forth.

II. GYROMAGNETIC MODES IN PARTIALLY FILLED RECTANGULAR WAVEGUIDE

We wish to find all of the propagating modes of a rectangular waveguide partially filled with ferrite (Fig. 1). In order to simplify the analysis we will find it convenient to consider the waveguide's transverse dimensions to be small compared to a free-space wavelength. Although this assumption will cut off all of the conventional TE modes, the TE "ferrite dielectric" mode may still propagate, as well as other gyromagnetic modes.

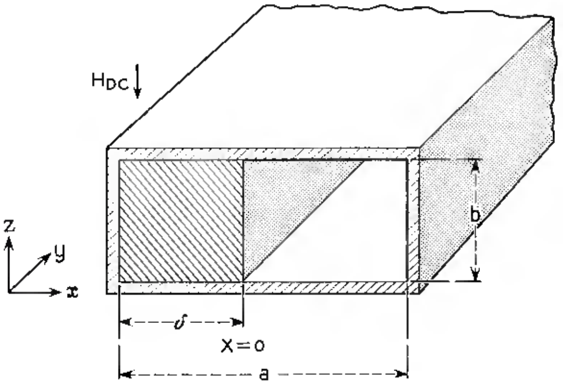


Fig. 1 — Rectangular waveguide partially filled by ferrite slab.

In a loss-free feromagnetic medium, Maxwell's equations are given by

$$\begin{aligned} \text{curl } \mathbf{H} &= i\omega\epsilon\mathbf{E}, \\ \text{curl } \mathbf{E} &= -i\omega\mu_0 T\cdot\mathbf{H}, \end{aligned} \tag{1}$$

where **E** and **H** are the usual field vectors, the time dependence is assumed to be $e^{i\omega t}$ and the tensor T can be written in the Cartesian frame (x,y,z) as

$$T = \begin{pmatrix} \mu & i\kappa & 0 \\ -i\kappa & \mu & 0 \\ 0 & 0 & 1 \end{pmatrix}. \tag{2}$$

Combining the two equations (1), we obtain the wave equation

$$\nabla \times \nabla \times \mathbf{H} - \omega^2 \epsilon\mu_0 T\cdot\mathbf{H} = 0. \tag{3}$$

This has two plane wave solutions, $e^{-i\mathbf{k}_1\cdot\mathbf{R}}$ and $e^{-i\mathbf{k}_2\cdot\mathbf{R}}$. In the limit for which k_x^2, k_y^2 and $k_z^2 \gg \omega^2\mu_0\epsilon$, these solutions are governed[3] by the equations

$$k_{x1}^2 + k_{y1}^2 + \frac{1}{\mu} k_{z1}^2 = 0\,, \tag{4}$$

$$k_{x2}^2 + k_{y2}^2 + k_{z2}^2 = 0\,. \tag{5}$$

Equations (4) and (5) determine individual parallel plane-type modes. We will solve these equations for the situation shown in Fig. 1. The fields in the ferrite which satisfy the boundary conditions in the z direction, $E_x = E_y = 0$ at $y = 0$ and $z = b$, can be derived from (1) through (5) as in Ref. 3. For the ferrite mode[1] corresponding to (4):

$$\mathbf{E}_{1f}^{\pm} = F_1^{\pm} \left\{ \begin{matrix} \left[\left(\dfrac{1-\mu}{\kappa}\right)\cos\varphi_1 \pm i\sin\varphi_1\right]\sin\dfrac{\pi m z}{b} \\ \left[\left(\dfrac{1-\mu}{\kappa}\right)\sin\varphi_1 + i\cos\varphi_1\right]\sin\dfrac{\pi m z}{b} \\ i\mu^{-1/2}\cos\dfrac{\pi m z}{b} \end{matrix} \right\} \tag{6}$$

$$\cdot\exp\left[\frac{\pi m}{b}\,\mu^{-1/2}(\pm x\sin\varphi_1 + y\cos\varphi_1)\right],$$

$$\mathbf{H}_{1f}^{\pm} = F_1^{\pm} \frac{i\pi m}{\omega\mu_0\mu b}\left(\frac{1-\mu}{\kappa}\right)\begin{pmatrix} \pm \sin\varphi_1 \cos\frac{\pi m z}{b} \\ \cos\varphi_1 \cos\frac{\pi m z}{b} \\ -\mu^{+1/2}\sin\frac{\pi m z}{b}\end{pmatrix} \cdot \exp\left[\frac{m\pi}{b}\mu^{-1/2}(\pm x\sin\varphi_1 + y\cos\varphi_1)\right], \tag{7}$$

where

$$k_{z1} = \frac{m\pi}{b} \qquad (m = 1, 2, 3 \cdots),$$

$$k_{y1} = i\frac{m\pi}{b}\mu^{-1/2}\cos\varphi_1,$$

$$k_{x1} = i\frac{m\pi}{b}\mu^{-1/2}\sin\varphi_1,$$

and the variable φ_1 is introduced for convenience, as in Ref. 3, in place of the propagation constant to be determined by the boundary conditions in the x direction. The superscript + or − on field quantities refers to the two solutions $e^{-ik_x x}$ and $e^{+ik_x x}$ respectively, describing the x variation for the same y variation $e^{-ik_y y}$. The constants F^+ and F^- are the corresponding amplitude constants. Note that $m = 0$ is excluded. The assumption of k_z being very large does not apply to $m = 0$ and hence this will be treated separately. For mode 2 corresponding to (5):

$$\mathbf{E}_{2f}^{\pm} = -F_2^{\pm}\frac{i\pi m}{\omega\epsilon b}\begin{pmatrix} \pm \sin\varphi_2 \sin\frac{\pi m z}{b} \\ \cos\varphi_2 \sin\frac{\pi m z}{b} \\ \cos\frac{\pi m z}{b}\end{pmatrix} \cdot \exp\left[\frac{m\pi}{b}(\pm x\sin\varphi_2 + y\cos\varphi_2)\right], \tag{8}$$

$$\mathbf{H}_{2f}{}^{\pm} = F_2{}^{\pm} \begin{bmatrix} \left(\cos\varphi_2 \pm \dfrac{i\kappa}{\mu - 1}\sin\varphi_2\right)\cos\dfrac{\pi m z}{b} \\ \left(\pm\sin\varphi_2 + \dfrac{i\kappa}{\mu - 1}\cos\varphi_2\right)\cos\dfrac{\pi m z}{b} \\ -\dfrac{i\kappa}{\mu - 1}\sin\dfrac{\pi m z}{b} \end{bmatrix} \cdot\exp\left[\frac{m\pi}{b}\,(\pm x\sin\varphi_2 + y\cos\varphi_2)\right], \tag{9}$$

where

$$k_{z2} = \frac{m\pi}{b},$$

$$k_{y2} = i\,\frac{m\pi}{b}\cos\varphi_2,$$

$$k_{x2} = i\,\frac{m\pi}{b}\sin\varphi_2,$$

and φ_2 is the convenient dependent variable for mode 2. Note that, in order for modes 1 and 2 to have the same y variation,

$$\cos\varphi_2 = \mu^{-1/2}\cos\varphi_1\,. \tag{10}$$

For the fields in the air region, the two independent plane-wave solutions are both governed by the same equation, which, in the small waveguide approximation, is also the same as the ferrite mode 2 given by (5). For convenience in satisfying the air-ferrite interface boundary conditions, we will not use the usual resolution of these two modes into transverse electric and transverse magnetic. Instead, we will choose one mode so that its tangential electric field can be made continuous with ferrite mode 1 across the interface and the second made so that its tangential magnetic field can be made continuous with ferrite mode 2 across the interface. Thus, for air mode 1 we obtain

$$\mathbf{E}_{1a}{}^{\pm} = A_1{}^{\pm} \begin{Bmatrix} \left[\mu^{-1/2}\dfrac{\sin\varphi_1}{\sin\varphi_2}\left(-\dfrac{1-\mu}{\kappa}\cos\varphi_1 \pm i\sin\varphi_2\right)\right]\sin\dfrac{\pi m z}{b} \\ \left(\mp\dfrac{1-\mu}{\kappa}\sin\varphi_1 + i\cos\varphi_2\right)\sin\dfrac{\pi m z}{b} \\ \mu i^{-1/2}\cos\dfrac{\pi m z}{b} \end{Bmatrix} \cdot\exp\left[\frac{\pi m}{b}\,(\pm x\sin\varphi_2 + y\cos\varphi_2)\right], \tag{11}$$

$$\mathbf{H}_{1a}^{\pm} = A_1^{\pm} \frac{i\pi m}{\omega\mu_0\mu b}\left(\frac{1-\mu}{\kappa}\right)\begin{pmatrix} \cos\dfrac{\pi m z}{b} \\ \pm\dfrac{\cos\varphi_2}{\sin\varphi_2}\cos\dfrac{\pi m z}{b} \\ \mp\dfrac{1}{\sin\varphi_2}\sin\dfrac{\pi m z}{b} \end{pmatrix} \cdot\exp\left[\frac{\pi m}{b}\left(\pm x\sin\varphi_2 + y\cos\varphi_2\right)\right]. \tag{12}$$

where we have used div $\mathbf{E} = 0$ and div $\mathbf{H} = 0$ to evaluate E_x and H_x and (10) is employed to simplify the expression. For air mode 2

$$\mathbf{E}_{2a}^{\pm} = -A_2^{\pm}\frac{i\pi m}{\omega\epsilon_0 b}\begin{pmatrix} -\sin\varphi_2\sin\dfrac{\pi m z}{b} \\ \mp\cos\varphi_2\sin\dfrac{\pi m z}{b} \\ \mp\cos\dfrac{\pi m z}{b} \end{pmatrix} \cdot\exp\left[\frac{\pi m}{b}\left(\pm x\sin\varphi_2 + y\cos\varphi_2\right)\right], \tag{13}$$

$$\mathbf{H}_{2a}^{\pm} = A_2^{\pm}\begin{bmatrix} \left(\mp\cos\varphi_2 + \dfrac{i\kappa}{\mu-1}\sin\varphi_2\right)\cos\dfrac{\pi m z}{b} \\ \left(\sin\varphi_2 \pm \dfrac{i\kappa}{\mu-1}\cos\varphi_2\right)\cos\dfrac{\pi m z}{b} \\ \mp\dfrac{i\kappa}{\mu-1}\sin\dfrac{\pi m z}{b} \end{bmatrix} \cdot\exp\left[\frac{\pi m}{b}\left(\pm x\sin\varphi_2 + y\cos\varphi_2\right)\right]. \tag{14}$$

It is easy to verify that these modes satisfy the Maxwell equations in the air region.

The boundary conditions in the x direction now require E_y and E_z to vanish at both metal walls and E_y, E_z, H_y and H_z to be continuous across the ferrite-air interface. This will give us eight linear homogeneous equations to determine the eight unknown constants $A_{1,2}^{\pm}$ and $F_{1,2}^{\pm}$, leading to the secular equation determining the propagation constant.

We wish to concentrate only on those modes that are propagating, i.e., those for which $\cos\varphi_2$ is imaginary. We will also restrict our attention to those modes for which $\mu > 0$, so that $\cos\varphi_1$ will also be imaginary.

This will cause real values for both $\sin \varphi_1$ and $\sin \varphi_2$, so that all the modes will have real exponential decay in the x direction. If we further assume that the waveguide width, a, is much larger than the height, b, we need only consider those modes which decay away from the boundaries.

Thus, at the metal-ferrite wall we need only consider $\mathbf{E}_{1f}^-$ and $\mathbf{E}_{2f}^-$ (taking the sign of $\sin \varphi_1$ and $\sin \varphi_2$ as positive). This leads to exactly the same mode found in Ref. 3 for a completely filled waveguide,

$$\cot \varphi_1 = -i \frac{\mu}{\kappa}. \tag{15}$$

With the use of (9a) and (10), (15) can be solved for k_y :

$$\text{FM:} \qquad k_y = \frac{1}{\kappa} \frac{m\pi}{b} \sqrt{\frac{\mu\kappa^2}{\kappa^2 - \mu^2}}. \tag{15a}$$

We can call this the ferrite-metal (FM) mode since it has a maximum amplitude near the ferrite-metal wall. Notice that there is a solution for this partially filled waveguide only for one direction of propagation for a given value of κ.

At the air-metal wall we need consider only $\mathbf{E}_{1a}^+$ and $\mathbf{E}_{2a}^+$. Since these have exactly the same x dependence there is no nontrivial propagating solution for this case.

At the ferrite-air (FA) interface we need consider only the plus modes in the ferrite and the minus modes in the air. The requirements of continuous E_y, E_z, H_y and H_z lead to the relations

$$\begin{aligned}
&(F_1^+ + A_1^-)\left(\frac{1-\mu}{\kappa} \sin \varphi_1 + i \cos \varphi_1\right) \\
&\qquad\qquad - \left(\frac{F_2^+}{\epsilon} - \frac{A_2^-}{\epsilon_0}\right)\frac{i\pi m}{\omega b} \cos \varphi_2 = 0, \\
&(F_1^+ + A_1^-)(i\mu^{-1/2}) \\
&\qquad\qquad - \left(\frac{F_2^+}{\epsilon} - \frac{A_2^-}{\epsilon_0}\right)\frac{i\pi m}{\omega b} = 0, \\
&\frac{i\pi m}{\omega\mu_0\mu b} \frac{1-\mu}{\kappa}\left(\cos \varphi_1 F_1^+ + \frac{\cos \varphi_2}{\sin \varphi_2} A_1^-\right) \\
&\qquad\qquad + (F_2^+ + A_2^-)\left(-\sin \varphi_2 + \frac{i\kappa}{\mu - 1} \cos \varphi_2\right) = 0, \\
&\frac{i\pi m}{\omega\mu_0\mu b} \frac{1-\mu}{\kappa}\left(-\mu^{-1/2} F_1^+ - \frac{1}{\sin \varphi_2} A_1^-\right) \\
&\qquad\qquad + (F_2^+ + A_2^-)\left(-\frac{i\kappa}{\mu - 1}\right) = 0.
\end{aligned} \tag{16}$$

For solution, the determinant of the coefficients multiplying F_1^+, A_1^-, F_2^+ and A_2^- must vanish. The factoring of this determinant leads to two values of φ_1 :

$$\tan \varphi_1 = -i\frac{\kappa}{\mu}, \tag{17}$$

$$\tan \varphi_2 + \mu \tan \varphi_1 = -i\kappa. \tag{18}$$

These can be solved for the propagation constant with the aid of (9a) and (10):

$$\text{FAI:} \qquad k_y = -\frac{1}{\kappa}\frac{m\pi}{b}\sqrt{\frac{\mu\kappa^2}{\kappa^2 - \mu^2}}, \tag{17a}$$

$$\text{FAII:} \qquad \sqrt{\left(\frac{m\pi}{b}\right)^2 + k_y^2} + \mu\sqrt{\frac{1}{\mu}\left(\frac{m\pi}{b}\right)^2 + k_y^2} = -\kappa k_y . \tag{18a}$$

We will call these the ferrite-air modes (FAI and FAII), since the fields fall off exponentially from the ferrite-air interface.

The modes FM, FAI and FAII represent all the propagating modes for $\mu > 0$ except for the case $m = 0$, for which the approximations used above are not valid. However, $m = 0$ represents a TE mode such as was treated by Button and Lax.[4] In the limit of small waveguide $[b^2 \ll 1/(\omega^2\mu_0\epsilon)]$ the only TE mode that is not cut off is the "ferrite dielectric" mode. Its propagation constant is given by

$$(\mu^2 - \kappa^2)k_a \coth k_a(a - \delta) = \kappa k_y - \mu k_m \coth k_m\delta, \tag{19}$$

where

$$k_a = k_y\sqrt{1 - \frac{\omega^2\mu_0\epsilon}{k_y^2}}, \tag{20}$$

and

$$k_m = k_y\sqrt{1 - \frac{\omega^2\mu_0\epsilon}{k_y^2}\frac{\mu^2 - \kappa^2}{\mu}}, \tag{21}$$

and where δ is the ferrite thickness and a the guidewidth.

A sketch of the propagation constant as a function of magnetic field is shown in Fig. 2 for the various propagation modes. It can be shown from (15a), (17a) and (18a) that the FM mode propagates in the plus y direction between $\mu = \kappa$ and $\mu = 0$, that the FAI mode propagates in the minus y direction between $\mu = \kappa$ and $\mu = 0$ and that the FAII modes propagate in the minus y direction between $\mu = \kappa - 1$ and $\mu = 0$. The FD mode has more complex behavior. For small values of $(a - \delta)/\delta$

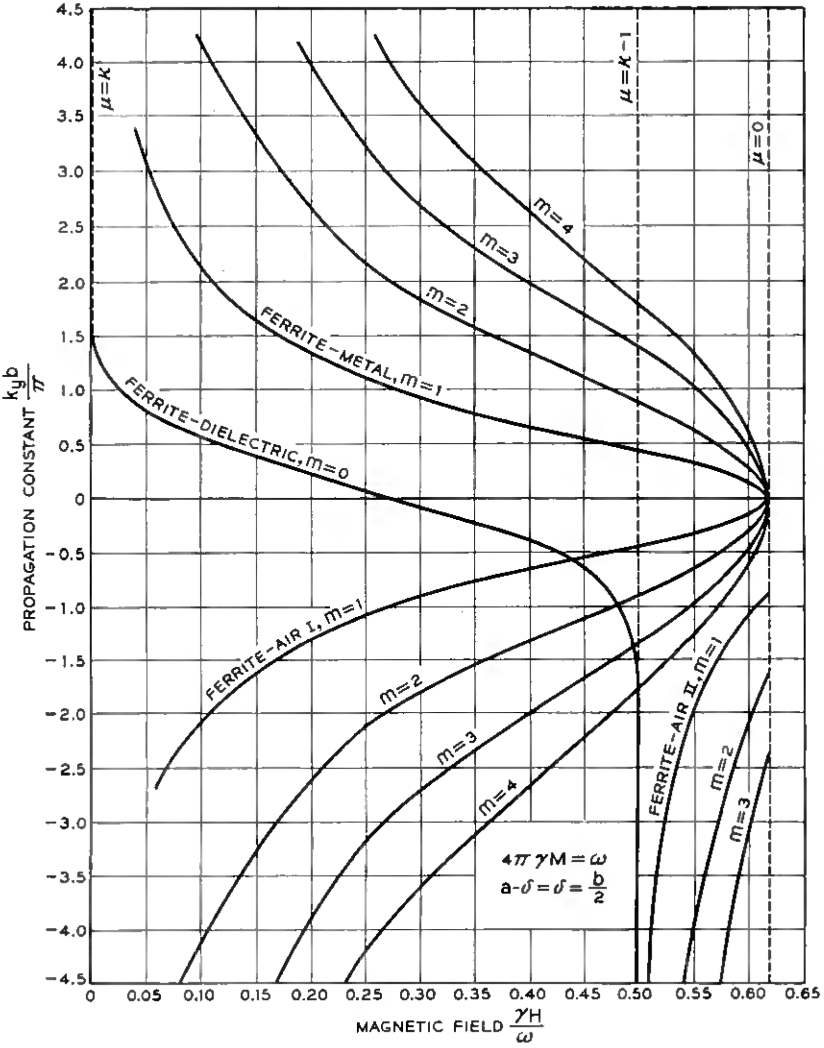


Fig. 2 — Typical mode spectrum of partially filled small guide as a function of magnetic field.

[less than $1 + \omega/(\gamma\pi M)$ for material obeying Polder's relations,[5]], the FD mode has a group velocity in the $+y$ direction between $\mu = \kappa$ and $\mu = \kappa - 1$, but the propagation constant is either plus or minus depending on whether

$$\left(\frac{a - \delta}{\delta} - \frac{\kappa^2 - \mu^2}{\mu}\right)$$

is positive or negative. For large $(a - \delta)/\delta$ [greater than $1 + \omega/(\gamma\pi M)$], the FD mode has both group velocity and phase velocity positive between $\mu = \kappa$ and $\mu = \kappa - 1$, but it is double-valued between $\mu = \kappa - 1$ and $\mu = 0$, having a positive group velocity at whatever magnetic fields it has a negative group velocity.

These gyromagnetic modes tend to be lossy, particularly for the higher orders. This can be demonstrated by allowing μ and κ to be slightly complex: $\mu = \mu' - j\mu''$, $\kappa = \kappa' - j\kappa''$. Then the expressions for the propagation constant can be expanded to give

$$k_y(\mu, \kappa) = k_y(\mu', \kappa') - j\frac{\partial k_y}{\partial \mu}\mu'' - j\frac{\partial k_y}{\partial \kappa}\kappa''. \tag{22}$$

In all the expressions for k_y, FAI, FAII, and FM, $\partial k_y/\partial\mu$ and $\partial k_y/\partial\kappa$ can be seen to be proportional to m. That is, the attenuation increases linearly with the order number, m.

III. GYROMAGNETIC MODE STRIP LINE ISOLATOR

We can use these gyromagnetic modes to make a novel isolator. As a design objective we will try to excite a high-order gyromagnetic mode for one direction of propagation, thus obtaining loss, but will try not to excite any in the opposite direction. To do this we will use a strip line TEM mode incident on a ferrite section, as shown in Fig. 3. The TEM mode has magnetic field components that are symmetric in the z direction and will not couple to the TE mode of the ferrite. However, they are appropriate to couple to the gyromagnetic modes.

To get maximum coupling we need spatial harmonics of the TEM mode to have appreciable amplitudes at the gyromagnetic propagation constant. One way to accomplish this is to break up the ferrite along the y direction.

Now, since the fields fall off in the x direction away from the strip line, we would expect the ferrite-air modes to be excited more than the ferrite-metal modes. Since the FA modes exist for only one direction of propagation and the FM modes for the other, we should get appreciable

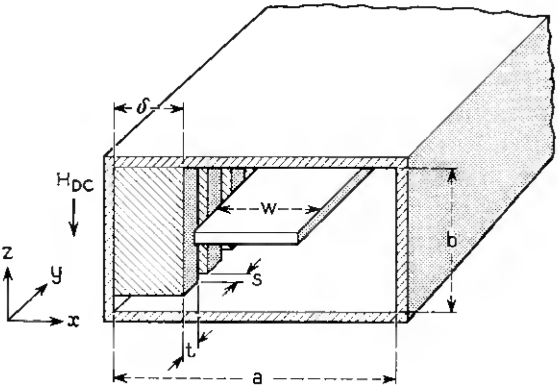


Fig. 3 — Strip line isolator employing periodically broken ferrite slab.

attenuation for the direction of propagation of the FA modes, but little attenuation in the reverse direction.

A physical embodiment of this idea has been built, and a plot of the forward and reverse loss as a function of frequency is shown in Fig. 4. Ratios of reverse to forward loss of greater than 10 can be obtained over a 30 per cent bandwidth. It should be emphasized that this is not a resonance-type isolator. The absorption does not depend on circular polarization, nor does it occur at the ferromagnetic resonance (approxi-

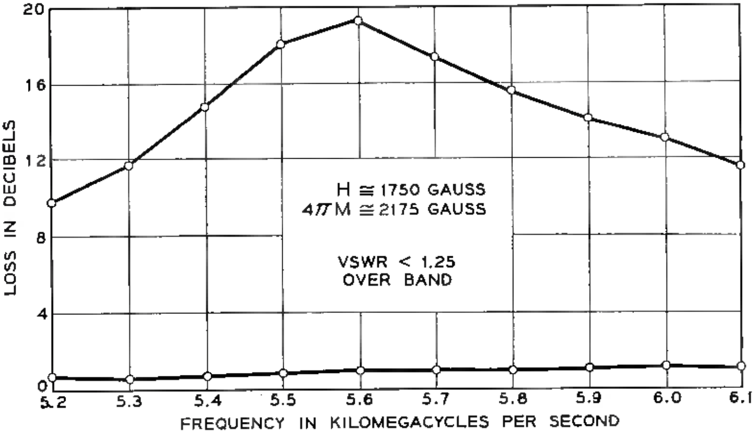


Fig. 4 — Typical characteristic of the gyromagnetic mode strip line isolator of Fig. 3: a = 0.900 inch; b = 0.400 inch; δ = 0.170 inch; S = 0.155 inch; W = 0.588 inch; t = 0.155 inch.

mately 4.4 kmc for the case indicated). On the other hand, the maximum absorption does occur between the frequencies for which $\mu = 0$ (6.2 kmc) and $\mu = \kappa - 1$ (4.5 kmc), as would be expected for coupling to the FAI mode. The small forward loss is believed mainly due to the excitation of the FM mode.

IV. EXCITATION OF GYROMAGNETIC MODES FROM A UNIFORM* TE MODE THROUGH WALL LOSS

We have been surprised to discover experimentally that uniform* TE modes can couple to modes with different symmetry, even when the boundary is uniform.* For instance, let the dominant TE mode of a rectangular guide be allowed to impinge on another rectangular guide containing a slab of ferrite as shown in Fig. 5. The slab completely fills the waveguide in the z direction and ends abruptly on an xz plane. Thus, neither the original mode nor the boundary has any quantity which varies along the magnetic field (z direction). We then insert a probe in the ferrite to probe for nonuniform modes. This probe consists of a metallic plate inserted in the middle of the ferrite slab in the xz plane with metal leads running out in the x direction. Energy will be coupled to this probe only if the average of H_z along the z direction is nonvanishing or if a component E_x appears at $z = b/2$; i.e., only if nonuniform field components appear.

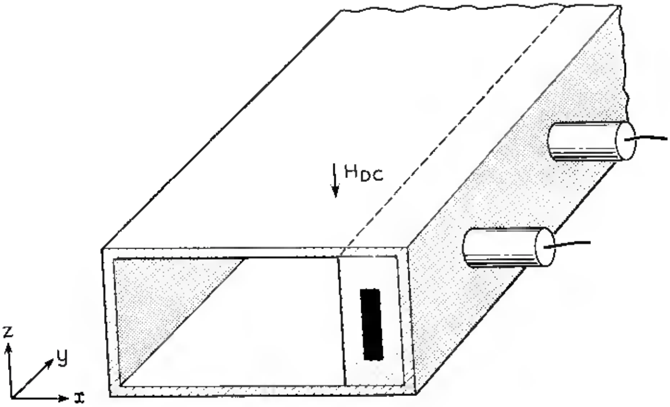


Fig. 5 — Ferrite slab geometry with embedded strip line terminating in coaxial lines.

* By "uniform" we mean to indicate that there are no variations parallel to the applied field.

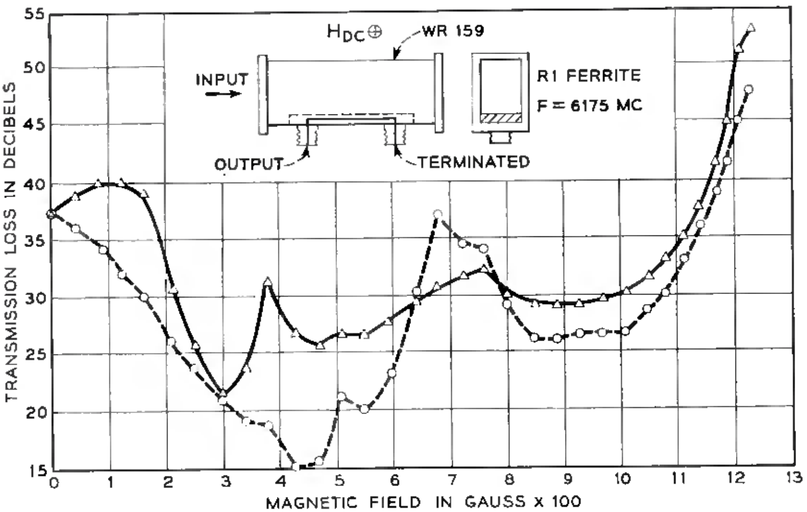


Fig. 6 — Power transmitted to coaxial line with TE incidence on embedded strip line geometry.

In Fig. 6 the fraction of the power coupled to the probe is plotted as a function of magnetic field. In the absence of the field, this fraction is below −37 db, the residual presumably being due to small errors in alignment. However, for a particular applied magnetic field this fraction can increase to −15 db, an increase of 22 db. The fact that there are irregularities in the coupling as a function of field suggests an interference between some of the excited modes.

A similar experiment was performed with a full height slab in the center of a square waveguide, away from both side walls (Fig. 7). A dominant mode of a rectangular guide is made incident on this square guide, exciting one polarization. The transverse electric field is probed by examining the transmitted power into a rectangular waveguide at right angles to the first one. Again, if only uniform modes were excited, there should be no transmission. Yet, as shown in Fig. 8, the transmission increases from −47 db at $H = 0$ to as high as −16 db with an applied magnetic field, an increase of 31 db. That this transmission was not Faraday rotation in a small axial magnetic field component could be assured by observing the transmission to be relatively insensitive to a slight tilting of the magnetic field with respect to the waveguide. Both of these experiments indicate the possibility of coupling appreciable power into nonuniform modes.

The only mechanism we have been able to discover which leads to a compatible model for such a coupling has been the finite conductivity of the walls. Let us consider then what would happen to a TEM wave that was started through a two-dimensional ferrite media contained between metal plates of finite conductivity (Fig. 9). The TEM mode has initially the components E_z and H_x , whose amplitudes are independent of z. As the mode moves through the medium, a magnetic field is developed at right angles to M and H_x ; i.e., a field H_y is developed. This field H_y induces currents, $\mathbf{n} \times \mathbf{y}_0 H_y$, in the metal walls. If the walls have a finite conductivity, σ, an electric field will therefore appear equal to

$$\frac{1}{\sigma} \mathbf{n} \times \mathbf{y}_0 H_y .$$

Since at the top face this is opposite in direction from the bottom, the induced electric field in the ferrite medium, E_x , has a z variation which is antisymmetric about the middle of the waveguide, as indicated in Fig. 9.

As the wall conductivity is made ever larger, this antisymmetric component tends to disappear. The limiting processes as this field disappears are, however, very unclear. For instance, we have considered (in Appendix A) the infinite spectrum of modes in the ferrite medium that would be excited by an incident TEM mode perturbed by the finite conductivity of the walls. Under the simple perturbation assumed, we find an unlimitedly large amount of scattered energy is predicted. Since this is impossible, we conclude that the simple perturbation picture is

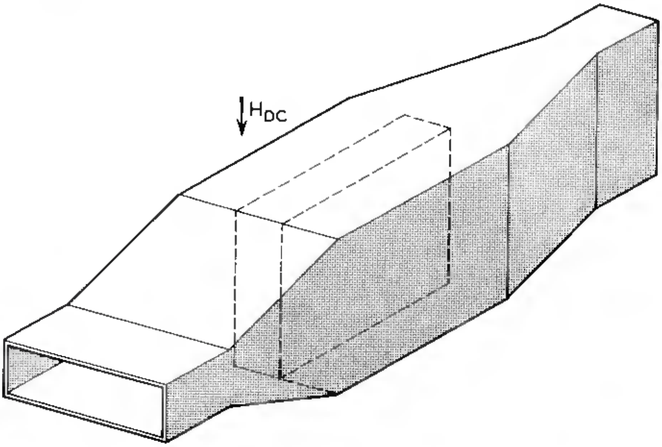


Fig. 7 — Square guide with uniform ferrite slab terminating in orthogonal output.

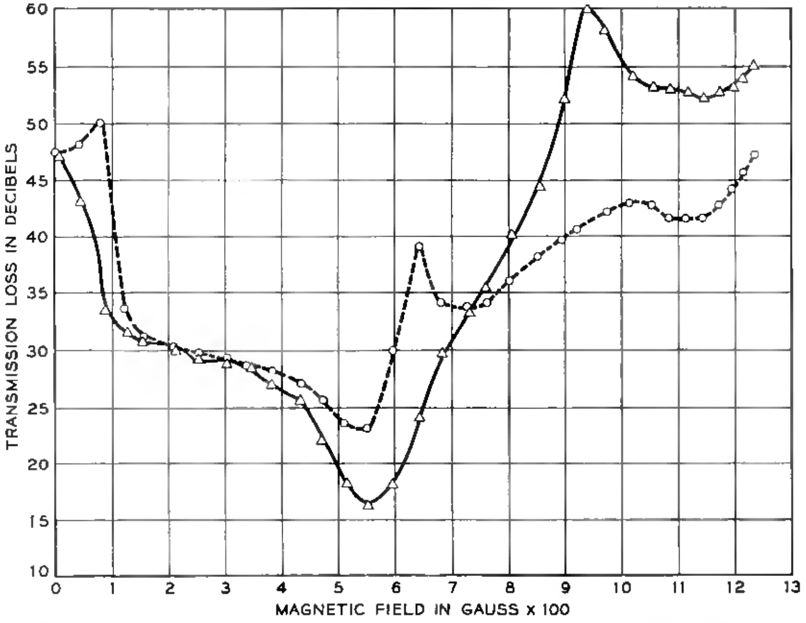


Fig. 8 — Transmission characteristic to orthogonal output by means of a uniform ferrite slab. Solid curve is forward direction; dashed curve is reverse.

incorrect. This leads to the suggestion that modes with variations along the magnetic field are excited at the boundary even in the limit of infinite wall conductivity.

One might wonder why this process of assuming a finite wall resistivity yields a coupling in the limiting process, whereas a starting assumption

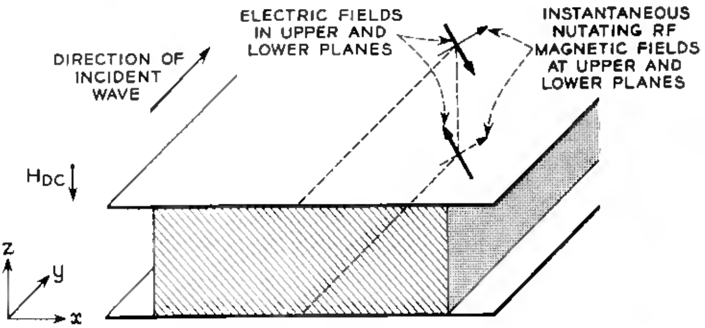


Fig. 9 — Transversely magnetized ferrite slab bounded by finitely conducting parallel planes.

of a loss-free medium does not. It must be remembered that uniqueness of the field representation is obtained only by recognizing the existence of loss terms, however small. Therefore, any reactive physical system has meaning only through this limiting process. Some classical "paradoxes" have owed their existence to the failure to recognize this fact. The above discrepancy in results arrived at by the two processes should therefore not be disturbing.

Thus, although we have not been able to show rigorously that these gyromagnetic modes must be excited from an incident uniform mode in the limit of resistanceless walls, we have demonstrated their excitation experimentally and have produced a plausible explanation of how this might be possible.

V. GENERAL THEOREM ON THE NONEXISTENCE OF PURELY REACTIVE ISOLATORS

We can see from the analyses of Section II and have shown in Fig. 2 that, for $\kappa > \mu > \kappa - 1$, the situation described by Button and Lax[4] is manifested for the ferrite dielectric (FD) mode: it propagates in only one direction. It should be noticed that the same thing is true for any one of the other gyromagnetic modes. This suggests that a lossless device could be made which would be perfectly transmitting in one direction (assuming one could match into one mode, e.g., the ferrite dielectric mode), but be nontransmitting in the opposite direction. However, we can show that such a device is impossible.

Let us consider two reference planes in a waveguide that are so far removed on either side of an arbitrarily loaded section that all modes of the waveguide except a dominant one are vanishingly small. We may then set up a scattering matrix between these reference planes, which we designate 1 and 2, to relate the incident waves, u, and reflected waves, v, at each of these points:

$$\begin{aligned} v_1 &= s_{11}u_1 + s_{12}u_2 , \\ v_2 &= s_{21}u_1 + s_{22}u_2 , \end{aligned} \tag{23}$$

or, operationally,

$$v = su. \tag{24}$$

In a loss-free network under steady-state conditions, s is unitary,[6] so that

$$\begin{aligned} | s_{11} |^2 + | s_{12} |^2 &= 1, \\ | s_{22} |^2 + | s_{21} |^2 &= 1, \\ s_{11}s_{21}^* + s_{12}^*s_{22} &= 0. \end{aligned} \tag{25}$$

By algebraic manipulation of these relations (and assuming s_{22} and s_{21} to be nonvanishing), we find

$$\begin{gathered} |s_{11}|^2 - |s_{22}|^2 = |s_{21}|^2 - |s_{12}|^2, \\ \frac{|s_{11}|}{|s_{22}|} = \frac{|s_{12}|}{|s_{21}|}, \end{gathered} \tag{26}$$

and thus

$$\frac{|s_{21}|^2}{|s_{22}|^2}(|s_{21}|^2 - |s_{12}|^2) = |s_{12}|^2 - |s_{21}|^2. \tag{27}$$

The only way for this last relation to be true is for

$$|s_{12}|^2 = |s_{21}|^2. \tag{28}$$

Hence,

$$|s_{11}|^2 = |s_{22}|^2. \tag{29}$$

These two equations state that the transmission and reflection looking from one direction must equal in magnitude the transmission and reflection, respectively, looking from the other. That is, no isolator action is possible in such a loss-free network.

VI. EXPERIMENTAL BEHAVIOR OF SOME "REACTIVE" ISOLATORS

We have attempted to set up experimentally two situations which were designed to give "reactive" isolation. The first is based on the ferrite dielectric mode described by Button and Lax.[4] A rectangular waveguide partially loaded as shown in Fig. 1 was made small enough so that all the conventional TE modes were cut off. A junction was made between this section of small loaded guide with standard unloaded guide, with suitable tuning screws for matching. The reflection and transmission in both directions are plotted in Fig. 10 as a function of the transverse magnetic field. We observe the predicted "one-way" transmission but note that the loss in the reverse direction is attributable primarily to an absorption, not reflection.

A second type of reactive isolator can be made out of a field-displacement isolator[7] (Fig. 11). In this type of isolator the dominant mode can be made to have an electric field null at one face of ferrite for one direction of propagation. If we were to place a copper sheet at this point (see Fig. 11) it should not affect the propagation in this direction. However, for the reverse direction the field of this mode does not have a null. We would expect, therefore, that if the forward direction were well matched it could be made perfectly transmitting, while one might expect a strong reflection would occur in the reverse direction, contrary to the theorem proved in Section V.*

The experiment was tried on a variant of the geometry shown in Fig. 11, using a partial height slab with the final dimensions shown in Ref. 7. In Fig. 12 the transmission and reflection in the forward and reverse directions are shown as functions of magnetic field. One can see that we can arrange just what we expected in the forward direction with a trans-

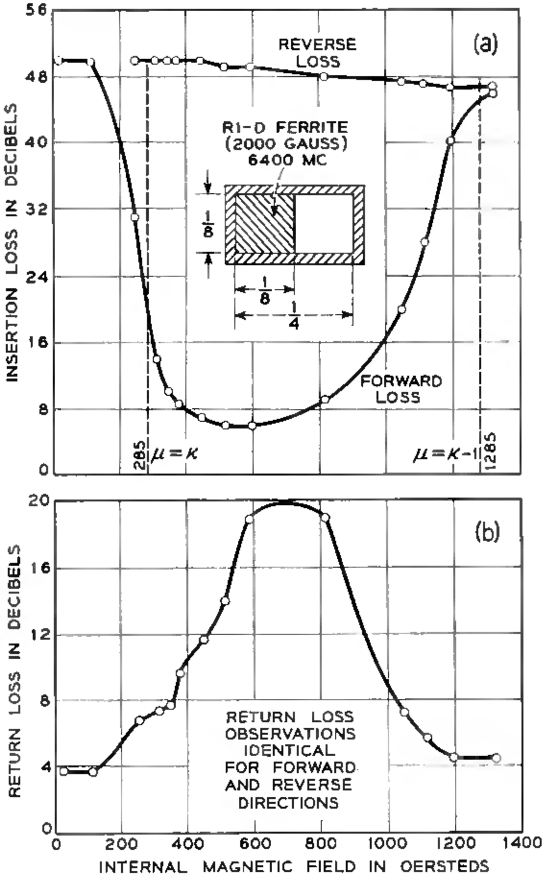


Fig. 10 — Characteristics of a ferrite-dielectric mode "reactive" isolator, showing that loss is caused by absorbtion rather than reflection: (a) insertion loss; (b) return loss.

* It may be shown that the discontinuity in H_x at the ferrite interface permits, to first order, coupling of an electric dipole to the mode having a null **E** field at the interface. This statement in itself might be viewed as a thermodynamic violation.

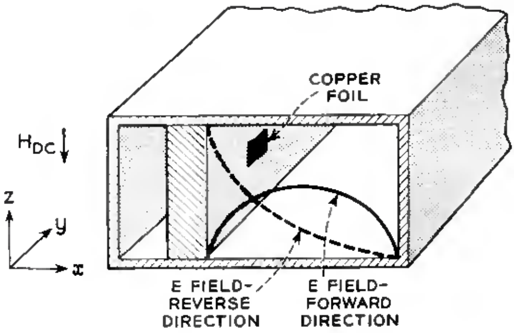


Fig. 11 — Field displacement isolator configuration with "perfectly" conducting scattering element replacing resistive sheet.

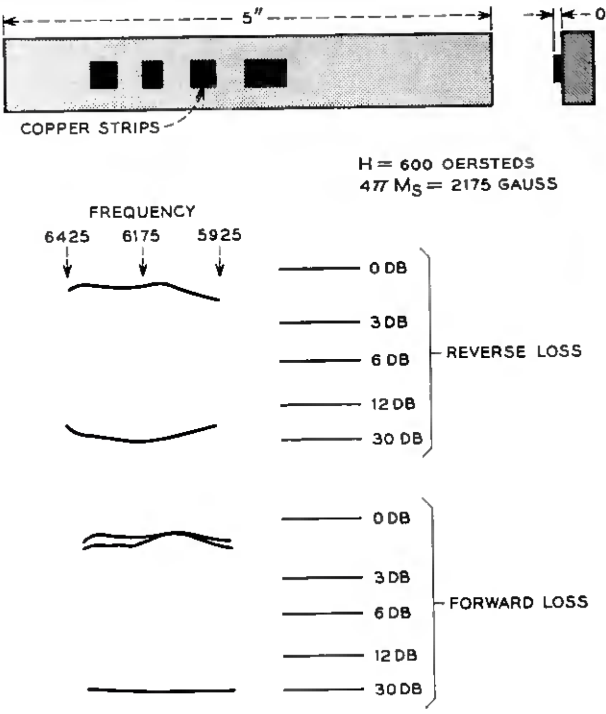


Fig. 12 — Response of copper strip insulator.

mission loss of less than 1 db. But in the reverse direction, instead of reflection, we obtain an absorption of greater than 30 db at some fields.

We thus find that when one attempts to build a theoretically impossible loss-free isolator, nature resolves the paradox, not by equalizing the transmission in both directions as predicted by the theorem in Section V, but by absorbing power in at least one direction, thus violating the assumptions of the theorem.

VII. POSSIBLE EXPLANATIONS OF "REACTIVE" ISOLATOR BEHAVIOR

It seems to us that the most reasonable explanation of the absorption that appears when a "reactive" isolator should be reflecting is caused by the excitation of gyromagnetic modes, which, as shown in Section II, tend to be lossy. It should be noticed that, for the range of magnetic fields investigated ($\mu > 0$) in Section II, there is no magnetic field for which there are propagating modes in one direction and not in the opposite direction.

Thus, it is tempting to suggest that a possible resolution of this paradox of Section VI is that, if we really had a zero loss material, we would get transmission in both directions, one set of modes carrying the power in one direction, but a different set carrying it for the opposite direction when the first set cannot propagate. For example, the FD mode in Fig. 10 could carry the power in one direction, whereas the FA modes would carry it in the other.

It can be appreciated that this suggestion requires a simple boundary to perform some rather startling feats. It must excite one mode for one direction of propagation that has one distribution of fields, say the FD mode with a uniform distribution in the z direction, while for the other direction a different mode must be excited, e.g., the FA modes with a sinusoidal variation in the z direction. In defense of the suggestion, we offer the experimental evidence described in Section IV that a simple boundary apparently can perform startling feats.

An alternate possible explanation has been offered by Walker.[8] He has proposed that power may be transmitted through cutoff modes. Thus, if the FD mode is coupled for one direction of propagation, a set of cutoff TE modes will carry the power in the reverse direction. This similarly requires some extraordinary behavior at the boundaries. In order for a cutoff mode to have appreciable amplitude at the far end, it must have an amplitude at the near end that is exponentially larger, the exponent being proportional to the length of the cutoff section. In the presence of a little loss in the material, such large amplitudes would give rise to large absorption, explaining the observed loss. Our reason for favoring the gyromagnetic mode resolution rather than the cutoff modes is that we have experimental evidence for the coupling to the gyromagnetic modes, but probing has thus far not indicated the existence of excess fields in the vicinity of the boundary.

Another suggestion for resolving the dilemma of the Button-Lax reactive isolator was originally given us by R. L. Martin. This same suggestion is attributed independently to some work of A. D. Bressler in this matter. The viewpoint expressed was that the position of the ferrite slab should be viewed as a limiting process as the ferrite slab approaches contact with the wall. Under this situation, there is an FD mode propagating in both directions that can carry the power. Since we have considered the case mathematically in which the slab exactly contacts the wall, such a limiting procedure does not appear to have any justification. Nevertheless, if we apply this process to the TE mode equations of Lax, Button and Roth,[9] the propagation constant of the returning FD wave approaches

$$k_y = \frac{1}{\delta} \operatorname{arctanh} (\kappa - \mu), \tag{30}$$

where δ is the air separation of the ferrite from the metal wall. This mode has maximum fields at the ferrite surface which fall off exponentially away from it as $e^{-k_y x}$. Thus, most of the energy in this mode is confined within a distance $1/k_y$ of the ferrite surface. As δ goes to zero, we would not be able to excite this mode from an impinging TE mode which has zero transverse field components at the only values of x where the FD mode has any amplitude. Thus, this does not appear to us a valid resolution of this paradox.

VIII. CONCLUSION

We thus see that the consideration of the gyromagnetic modes in a partially filled waveguide has led to some unusual nonreciprocal effects. Nonresonance isolation can be obtained without the deliberate introduction of loss material. Coupling to these modes tends to violate the usual symmetry arguments. Finally, they seem capable of resolving the Button-Lax "paradox" concerning reactive isolation.

We should like to acknowledge, with appreciation, the help of W. A. Dean and J. J. Kostelnick in the experimental studies.

APPENDIX A

Scattering of a TEM Mode from a Uniform Semi-Infinite Ferrite Interface Contained between Walls of Finite Resistivity

Properly, we should find the gyromagnetic modes corresponding to (6) through (9) for walls of finite conductivity. This is an extremely involved process when we attempt to solve the partially filled waveguide problem of Section II. Since we are interested only in showing the breakdown of symmetry arguments for predicting coupling, in what follows we will consider a semi-infinite medium (Fig. 9) bounded by walls of finite conductivity.

We will assume no variation in the fields in the x direction ($k_x = 0$) and confine our attention only to those modes for which $| k_z |^2$ and $| k_y |^2$ in the ferrite are large compared to $\omega^2 \epsilon_0 \mu_0$. The plane wave fields in the ferrite are then given[3] by

$$E_{1F}^{\pm} = \begin{bmatrix} i\left(\dfrac{1-\mu}{\kappa}\right) \\ -1 \\ \pm i\mu^{-1/2} \end{bmatrix} \exp\,(\pm ik_{z_1}z - ik_y y), \tag{31}$$

$$E_{2F}^{\pm} = \begin{pmatrix} 0 \\ -1 \\ \pm i \end{pmatrix} \exp\,(\pm ik_{z_2}z - ik_y y), \tag{32}$$

$$H_{1F}^{\pm} = \frac{ik_{z_1}}{\omega\mu_0}\left(\frac{1-\mu}{\kappa\mu}\right)\begin{pmatrix} 0 \\ \pm 1 \\ i\mu^{+1/2} \end{pmatrix} \exp\,(\pm ik_{z_1}z - ik_y y), \tag{33}$$

$$H_{2F}^{\pm} = +\,\frac{i\omega\epsilon}{k_z}\begin{pmatrix} \pm i \\ \pm\dfrac{\kappa}{\mu-1} \\ \pm\dfrac{i\kappa}{\mu-1} \end{pmatrix} \exp\,(\pm ik_{z_2}z - ik_y y). \tag{34}$$

The relations corresponding to (4) and (5) are

$$k_{z_1} = -i\mu^{+1/2}k_y\,, \tag{35}$$

$$k_{z_2} = -ik_y\,. \tag{36}$$

We will take the origin of the z-axis midway between the metallic boundaries. Note that the plus and minus ($\pm$) now refer to z-directed waves, which is different from the convention used in (6) through (9).

In the metal for $\sigma \gg \omega\epsilon$, Maxwell's equations reduce to

$$\begin{aligned} \text{curl } \mathbf{H}_m &= \sigma \mathbf{E}_m\,, \\ \text{curl } \mathbf{E}_m &= i\omega\mu_0 \mathbf{H}_m\,, \end{aligned} \tag{37}$$

and the wave equation [corresponding to (3)] yields the relation

$$k_y^2 + k_{zm}^2 = -i\omega\mu_0\sigma. \tag{38}$$

Let us first ask under what conditions the modes (6) through (9) are distorted by the resistive walls. This will happen when the electric fields caused by the induced current flowing in the resistive walls become comparable to the maximum electric field of the mode. Now the induced electric field is just $(1/\delta\sigma)(\mathbf{n} \times \mathbf{H}_w)$, where $\mathbf{n}$ is the surface normal, $\mathbf{H}_w$ is the magnetic field at the wall, and δ is the skin depth. Thus, (6) through (9) with $\sin\varphi = 0$ yield

$$\frac{E_1\,(\text{induced})}{E_1\,(\text{max})} = \frac{k_{z_1}}{\omega\mu_0\mu\sigma\delta}, \tag{39}$$

$$\frac{E_2\,(\text{induced})}{E_2\,(\text{max})} = \frac{1}{\sigma\delta}\frac{\omega\epsilon}{k_{z_2}}. \tag{40}$$

We see that the induced fields for mode 2 decrease for large k_z so that this mode is little affected by a finite conductivity. However, for mode 1, when $k_z \gg \omega\mu_0\mu\sigma\delta$, the induced fields are large and the mode 1 is greatly modified.

To get the correct fields for this condition we must combine the plane wave fields of (31) through (34). The problem can be simplified in this large k_z limit by noting that, for mode 1, the ratio of electric fields to magnetic fields varies as $1/k_z$, whereas for mode 2 this ratio varies as k_z. This means that, if the electric fields of the two modes are comparable, we can neglect H_2, while if the magnetic fields are comparable, we can neglect E_1. But in this latter case we have already shown that, if we retain both E_2 and H_2, we can satisfy the boundary conditions for large k_z by the unmodified mode 2 (without including H_1). Therefore we will seek the modified mode 1 from a mixture of the electric fields of mode 1 and mode 2 and neglect H_2.

This considerably simplifies our problem, since we note that $H_{1x} = 0$, By continuity across the metal-ferrite interface this also implies that $H_{mx} = 0$. Therefore, from (35) we see that in the metal we have a TE mode:

$$\mathbf{E}_m = \frac{i}{\sigma}\begin{pmatrix} k_{zm}H_{my} - k_yH_{mz} \\ 0 \\ 0 \end{pmatrix}.$$

Since div H vanishes in the metal,

$$k_{zm}H_{mz} + k_yH_{my} = 0 \tag{41}$$

and the relation (39) with the use of (38) reduces to

$$\mathbf{E}_m = \frac{\omega\mu_0}{k_{zm}} H_{my} \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}. \tag{42}$$

The wave we will select in each of the two metal regions will be that which falls off exponentially away from the metal-ferrite interface. Thus, in the $+z$ region, if we take the root of k_{zm} that has a positive imaginary part, we must select the wave which varies as $e^{+ik_zm_z}$ In the $-z$ region we select e^{-ik_zmz}.

In the ferrite the total field for a particular mode (characterized by a y variation of e^{-ik_yy}) is given by

$$\begin{pmatrix} E_{1F}^{\sigma} \\ H_{1F}^{\sigma} \end{pmatrix} = A_1^+ \begin{pmatrix} E_{1F}^+ \\ H_{1F}^+ \end{pmatrix} + A_1^- \begin{pmatrix} E_{1F}^- \\ H_{1F}^- \end{pmatrix} + A_2^+ \begin{pmatrix} E_{2F}^+ \\ 0 \end{pmatrix} + A_2^- \begin{pmatrix} E_{2F}^- \\ 0 \end{pmatrix}. \tag{43}$$

The boundary conditions require that E_x, E_y and H_y be continuous across the ferrite-metal interface. Let $\theta = (k_zb)/2$. At $z = +b/2$, continuity of E_x and H_y requires that

$$A_1^+e^{i\theta_1} + A_1^-e^{-i\theta_1} = \frac{k_{z1}}{\mu k_{zm}} (A_1^+e^{i\theta_1} - A_1^-e^{-i\theta_1}) \tag{44}$$

and, at $z = -b/2$,

$$A_1^+e^{-i\theta_1} + A_1^-e^{+i\theta_1} = -\frac{k_{z1}}{\mu k_{zm}} (A_1^+e^{-i\theta_1} - A_1^-e^{i\theta_1}). \tag{45}$$

For large $k_z(k_z^2 \gg \omega\mu_0\sigma)$,

$$k_{zM} = -ik_y = k_{z2} = \mu^{-1/2}k_{z1}. \tag{46}$$

With this relation, (44) and (45) have a solution

$$\tan 2\theta_1 = -\frac{2i\mu^{1/2}}{1 + \mu} \tag{47}$$

and

$$\frac{A_1^+}{A_1^-} = -1. \tag{48}$$

In order to evaluate $A_2^{\pm}$ we require E_y to vanish at $z = \pm b/2$:

$$A_1^+e^{+i\theta_1} + A_1^-e^{-i\theta_1} + A_2^+e^{+i\theta_2} + A_2^-e^{-i\theta_2} = 0, \tag{49}$$

$$A_1^+e^{-i\theta_1} + A_1^-e^{+i\theta_1} + A_2^+e^{-i\theta_2} + A_2^-e^{+i\theta_2} = 0. \tag{50}$$

These have the solution

$$A_2^+ = -A_2^- = A_1^+ \frac{\sin\theta_1}{\sin\theta_2}. \tag{51}$$

Collecting (31), (32), (33), (43), (48) and (50), we obtain the resultant fields in the ferrite:

$$\mathbf{E}_{1f}^{\sigma} = F_1 \begin{bmatrix} -i\dfrac{1-\mu}{\kappa}\sin k_{z1}z \\[2ex] \sin k_{z1}z - \dfrac{\sin k_{z1}\dfrac{b}{2}}{\sin k_{z2}\dfrac{b}{2}}\sin k_{z2}z \\[2ex] \mu^{-1/2}\cos k_{z1}z - \dfrac{\sin k_{z1}\dfrac{b}{2}}{\sin k_{z2}\dfrac{b}{2}}\cos k_{z2}z \end{bmatrix} e^{-ik_yy}, \tag{52}$$

$$\mathbf{H}_{1f}^{\sigma} = \frac{k_{z1}}{\omega\mu_0\mu}\,\frac{1-\mu}{\kappa}\,F_1\begin{pmatrix} 0 \\ -\cos k_{z1}z \\ \mu^{1/2}\sin k_{z1}z \end{pmatrix} e^{-ik_yy}, \tag{53}$$

where

$$k_{z1} = \frac{m\pi}{2b} + \frac{i}{2b}\operatorname{arctanh}\frac{2\mu^{1/2}}{1+\mu} \qquad m = 1, 2, \cdots, \tag{54}$$

$$k_{z2} = \mu^{-1/2}k_{z1} = ik_y, \tag{55}$$

and the z-axis is now considered to have its origin halfway between the metal walls.

If the normal modes of the finite conductivity guide are enumerated, $\mathbf{E}_j$ and $\mathbf{H}_j$, we can expand the incident wave, $(\mathbf{E}/\mathbf{H})$, in terms of them according to the relation

$$\begin{pmatrix}\mathbf{E}\\ \mathbf{H}\end{pmatrix} = \sum_j \begin{pmatrix}\mathbf{E}_j\\ \mathbf{H}_j\end{pmatrix} = \sum A_j\begin{pmatrix}\mathbf{e}_j\\ \mathbf{h}_j\end{pmatrix}, \tag{56}$$

where the amplitude A_j is given in terms of the modes of the adjoint set $\mathbf{e}_j^\dagger$ and $\mathbf{h}_j^\dagger$, (Appendix B) by

$$\int_S (\mathbf{E} \times \mathbf{h}_j^\dagger + \mathbf{e}_j^\dagger \times \mathbf{H}) \cdot \mathbf{dS}, \tag{57}$$

and the lower case vectors $\mathbf{e}_j$, $\mathbf{h}_j$, represent a normalized set of modes, which obey the normalization condition:

$$1 = \int_S (\mathbf{e}_j \times \mathbf{h}_j^\dagger + \mathbf{e}_j^\dagger \times \mathbf{h}_j) \cdot dS. \tag{58}$$

From (27) it can be seen that A_j is nonvanishing for the modes E_{1f}^{σ} of large wave number [(22) and (23)], when the TE mode ($\mathbf{E}/\mathbf{H}$) is incident; i.e., these modes will be excited even though there is no variation of the boundary in the direction of the magnetic field to excite them.

It is of interest to evaluate the total amount of power contained in these higher-order modes. By our simplification of the problem to a semi-infinite plane, all of the higher-order modes are cut off [equation (25)]. In order to calculate a power flow, therefore, we need to introduce a little loss into the ferrite medium. We do this by allowing the components, μ and k, of the permeability tensor to take on the complex values, $\mu' - i\mu''$ and $\kappa' - i\kappa''$ and the dielectric constant, ϵ, to become $\epsilon' - i\epsilon''$.

The total power across a cross section, S_y, is given by

$$P = \frac{1}{2} \sum_{j,l} \int (\mathbf{E}_j \times \mathbf{H}_l^* + \mathbf{E}_l^* \times \mathbf{H}_j) \cdot \mathbf{dS}_y = \sum_{j,l} P_{jl}. \tag{59}$$

From Maxwells' equations (1) one can obtain the identity

$$\begin{aligned} \operatorname{div} (\mathbf{E}_j \times \mathbf{H}_l^* + \mathbf{E}_l^* \times \mathbf{H}_j) &= i\omega[\mu_0(\mathbf{H}_j \cdot \mathbf{T}^* \cdot \mathbf{H}_l^* - \mathbf{H}_l^* \cdot \mathbf{T} \cdot \mathbf{H}_j) \\ &\qquad + (\mathbf{E}_j \cdot \epsilon^* \mathbf{E}_l - \mathbf{E}_l^* \cdot \epsilon \mathbf{E}_y)] \\ &= 2\omega[\mu_0 \mathbf{H}_j \cdot \tau \cdot \mathbf{H}_l^* + \epsilon'' \mathbf{E}_j \cdot \mathbf{E}_l], \end{aligned} \tag{60}$$

where τ is the tensor $(1/2i)(T^* - T^T)$ and is given by

$$\tau = \begin{pmatrix} \mu'' & i\kappa'' & 0 \\ -i\kappa'' & \mu'' & 0 \\ 0 & 0 & 0 \end{pmatrix}. \tag{61}$$

The application of Gauss' theorem to (30) in a small-volume element bounded by $S_y(y)$, and $S_y(y + dy)$ yields

$$\begin{aligned} &\frac{\partial}{\partial y} \int (\mathbf{E}_j \times \mathbf{H}_l^* + \mathbf{E}_l^* \times \mathbf{H}_j) \cdot \mathbf{dS}_y \\ &\qquad = 2\omega \int (\mu_0 \mathbf{H}_j \cdot \tau \cdot \mathbf{H}_l^* + \epsilon'' \mathbf{E}_j \cdot \mathbf{E}_l^*) dS_y, \end{aligned} \tag{62}$$

whereupon, from (29),

$$
\begin{aligned}
P_{jl} &= \frac{i\omega \int (\mu_0 \mathbf{H}_j \cdot \boldsymbol{\tau} \cdot \mathbf{H}_l^* + \epsilon''\ \mathbf{E}_j \cdot \mathbf{E}_l^*)\, dS_y}{k_{y_j} - k_{y_l}^*} \\
&= \frac{i\omega A_j A_j^* \int (\mu_0 \mathbf{h}_j \cdot \boldsymbol{\tau} \cdot \mathbf{h}_l^* + \epsilon'' \mathbf{e}_j \cdot \mathbf{e}_l^*)\, dS_y}{k_{y_j} - k_{y_l}^*}.
\end{aligned}
\tag{63}
$$

Let us examine just a part of the power expended in the diagonal power components, P_{jj}, for the resistive wall modes, (22) and (23), which we will call P_{mm}. Here we will be interested in the dependence on m for large m to normalize (22) and (23) according to (28). The amplitudes F_1 for the normal modes must vary as

$$F_1 F_1^{\dagger} \sim m. \tag{64}$$

If we evaluate $A_j A_j^*$ for the TEM mode as modified by the ferrite and resistive wall (as discussed earlier) we find

$$A_j A_j^* \sim \frac{{E_x}^2}{m}.$$

Finally, the component of power, P_{mm}, in the mth mode varies as

$$P_{mm} \sim \frac{\mu'' \mu^{1/2}}{m} {E_x}^2 \sim \frac{\mu''}{m\sigma^2} {H_y}^2. \tag{66}$$

If we sum over all the modes, m, we find that the total power contained in these higher-order modes has a logarithmic singularity. This means that the reflection and transmission coefficients at the boundary must have been such as to reduce the amplitudes of the higher-m modes. This is true no matter how high the conductivity becomes. Thus, a reasonable interpretation of this divergence is that there must be a finite coupling to the gyromagnetic modes, even those of lower m, at the boundary. This argument leads to the suggestion that, if we have the proper explanation of the observed coupling, this coupling might become independent of wall conductivity at high enough conductivities.

APPENDIX B

Orthogonality Relationships

Orthogonality relations in generalized media are well covered in the literature.† Nevertheless, to show explicit forms. we include derivations

† See, typically, Refs. 10 and 11.

specifically directed to gyromagnetic media of the form discussed in this paper.

A real medium supporting electromagnetic propagation is characterized by free space parameters, μ_0 and ϵ_0, and by relative electric and magnetic susceptibilities. Given only a magnetic anisotropy produced by an applied magnetic field in gyromagnetic media, we further characterize the medium by the Polder tensor T, as well as by the ordinary scalar relative dielectric constant, ϵ.

The Polder tensor contains components which are all complex because of losses. Nevertheless, the transverse off-diagonal terms of this tensor are perfectly skew, and the sign associated with either component is prescribed by the direction of precession. The transpose of this tensor simply reverses the signs of the off-diagonal components and corresponds to time reversal in the dynamic classical equation of the spin.

Let us define a medium reciprocal to the real medium of the guide such that the following transformations hold:

$$\mu_0 \rightarrow -\mu_0,$$

$$\epsilon_0 \rightarrow -\epsilon_0,$$

$$T \rightarrow T',$$

where T' is an operator yet to be defined. The reciprocal fields are $\mathbf{E}'$ and $\mathbf{H}'$ and satisfy the Maxwell equations

$$\text{curl } \mathbf{H}' = -i\omega\epsilon_0\epsilon\mathbf{E}', \tag{67}$$

$$\text{curl } \mathbf{E}' = i\omega\mu_0 T'\cdot\mathbf{H}'. \tag{68}$$

We have the identity

$$\begin{aligned} &\text{div } (\mathbf{E} \times \mathbf{H}' + \mathbf{E}' \times \mathbf{H}) \\ &\quad = (\mathbf{E}\cdot\text{curl } \mathbf{H}' + \mathbf{E}'\cdot\text{curl } \mathbf{H}) - (\mathbf{H}\cdot\text{curl } \mathbf{E}' + \mathbf{H}'\cdot\text{curl } \mathbf{E}). \end{aligned} \tag{69}$$

From (67) and (68), the right-hand side of (69) becomes

$$i\omega[\epsilon_0\epsilon(-\mathbf{E}\cdot\mathbf{E}' + \mathbf{E}'\cdot\mathbf{E}) + \mu_0(\mathbf{H}'\cdot T\cdot\mathbf{H} - \mathbf{H}\cdot T'\cdot\mathbf{H}')].$$

If T' is defined such that $T' = T^T$, the right-hand side of (69) vanishes identically, and

$$\text{div } (\mathbf{E} \times \mathbf{H}' + \mathbf{E}' \times \mathbf{H}) = 0. \tag{70}$$

The reciprocal system bears the relation to the real system of creating solutions identical to those of the real system but having a negative time variation. If, then, there exists a solution of form $\epsilon^{i(\beta y + \omega t)}$ in the real system there exists solutions

$$\epsilon^{i[\beta y+\omega(-t)]} = \epsilon^{-i[(-\beta)y+\omega t]}$$

in the reciprocal system, implying, for the reciprocal propagation constant β',

$$\beta' = -\beta. \tag{71}$$

We now perform an integration of (70) over a differential volume of a cylindrical waveguide formed by two infinitesimally separated planes normal to the guide axis along the y direction, and intersecting the guide walls. From Gauss' theorem,

$$\int (\mathbf{E} \times \mathbf{H}' + \mathbf{E}' \times \mathbf{H}) \cdot \mathbf{dS} = 0. \tag{72}$$

Since $\mathbf{E} \times \mathbf{dS}$ vanishes on the guide wall, the surface integral takes on value only over the two transverse planes normal to the axis. The left-hand side of (72) has a value equal to the difference of the surface integrals over these adjacent planes, viz.:

$$dy \frac{\partial}{\partial y} \int (\mathbf{E} \times \mathbf{H}' + \mathbf{E}' \times \mathbf{H}) \cdot \mathbf{dA} = 0, \tag{73}$$

where $\mathbf{A}$ is the transverse cross section of the guide.

Let us assume a mode of order k for $\mathbf{E}$ and order j for $\mathbf{E}'$:

$$\begin{aligned} \mathbf{E} &= \mathbf{E}_k(x,y)\epsilon^{-i\beta_k y} \\ \mathbf{E}' &= \mathbf{E}^{(j)}(x,y)\epsilon^{-i\beta'^{(j)}y}. \end{aligned}$$

Then, from (73),

$$(\beta_k - \beta_j) \int [\mathbf{E}_k \times \mathbf{H}^{(j)} + \mathbf{E}^{(j)} \times \mathbf{H}_k] \cdot \mathbf{dA} = 0, \tag{74}$$

where we have employed (71) to transform $\beta'^{(j)}$ to $-\beta_j$. Equation (74) provides the final result:

$$\frac{\int [\mathbf{E}_k \times \mathbf{H}^{(j)} + \mathbf{E}^{(j)} \times \mathbf{H}_k] \cdot \mathbf{dA}}{\int [\mathbf{E}_k \times \mathbf{H}^{(k)} + \mathbf{E}^{(k)} \times \mathbf{H}_k] \cdot \mathbf{dA}} = \delta_{jk}. \tag{75}$$

The notational change to that employed in (57) is evident.

REFERENCES

1. Suhl, H. and Walker, L. R., B.S.T.J., **33,** May 1954, p. 579.
2. Thompson, G. H. B., Nature, **175,** June 25, 1955, p. 1135.
3. Seidel, H., B.S.T.J., **36,** March 1957, p. 409.
4. Button, K. J. and Lax, B., Trans. I.R.E., **AP-4,** July 1956, p. 531.
5. Polder, D., Phil. Mag., **40,** 1949, p. 99.
6. Montgomery, C. G., Dicke, R. H. and Purcell, E. M., *Principles of Microwave Circuits*, McGraw-Hill Book Co., New York, 1948, pp. 149.
7. Weisbaum, S. and Seidel, H., B.S.T.J., **35,** July 1956, p. 877.
8. Walker, L. R. and Suhl, H., Trans. I.R.E., **AP-4,** July 1956, p. 492.
9. Lax, B., Button, K. J. and Roth, L. M., J. Appl. Phys., **25,** 1954, p. 1413.
10. Walker, L. R., J. Appl. Phys., **28,** March 1957, p. 377.
11. Bressler, A. D., Joshi, G. H. and Marcuvitz, N., J. Appl. Phys., **29,** May 1958, p. 794.

Waveguide Bending Design Analysis

Theory of Bending and Formulae for Determination of Wall Thicknesses

By F. J. FUCHS, JR.

(Manuscript received June 4, 1959)

The art of rectangular tube bending is analyzed, with particular attention being given to tube wall thickness variations. Effects of these variations on tool design are discussed, and methods and formulae for determination of wall distortions are presented.

I. INTRODUCTION

For many years, waveguide bends for most microwave installations have been difficult and expensive to produce. The art of tube bending was not sufficiently advanced to make economically possible the extremely close tolerances required in waveguides. This was true even when waveguides were first introduced into radar equipment. It later became evident, as increased power, higher aircraft speeds and missile applications made waveguide requirements more severe, that the bending technique would have to be improved. First, a faster method of bending had to be found, since the best of existing methods required about 30 minutes to make one bend. Second, to reduce transmission losses, the new method had to produce bends that met closer internal cross-sectional tolerances. Third, bends of much smaller radius, more closely spaced compound bends and bends adjacent to swaged and twisted sections had to be made to meet new design demands. In addition to these specific improvements and innovations in the bending technique, production uniformity was desirable, since it is only through uniformity that statistical quality control can be realized.

At the Western Electric North Carolina Works the development of waveguide bending began in 1951 and continued for the next five years. This article describes the new bending process and indicates how internal cross-sectional accuracy is maintained despite material flow due to the bending action that changes the tubing's external dimensions. This information is of paramount importance to the bending tool designer and can also greatly aid the waveguide component designer who may wish to apply assembly details or machining in the region of the bend where wall thickness changes.

The several methods of bending that existed prior to this development are briefly reviewed. One of these, draw bending, is explained in somewhat more detail, because it was chosen as the basic method on which the improved technique was developed. The tooling used is shown and explained to acquaint the reader with terms used later in the analysis of effects of wall distortions on bend accuracy.

The material flow patterns and cross-sectional distortions are shown qualitatively and related to the individual tool parts. Corrective contouring evolves from these relationships to compensate for such distortions. Then, methods and formulae are advanced to make it possible to calculate accurately the wall thickness changes at any point in the bend region. For several of the more common sizes of waveguide, graphs are presented for reference in designs in which wall thickness must be evaluated.

Most of the formulae used in this discussion contain empirical constants. Therefore, an Appendix is given to explain the derivation of equations and provide supporting data for the constants.

II. TUBE BENDING METHODS

All tube bending methods consist basically of filling a tube with something to prevent its natural tendency to collapse and then bending it around a form, meanwhile constraining the outside of the tube by various methods to keep it from losing shape. There are three common methods of tube bending, all of which have been used to bend waveguides: compression bending, form bending and draw bending.

In compression bending, as shown in Fig. 1, the tube is filled with a close-fitting mandrel, laminated strips, low-melting-point alloy or other material. Then one end of the tube is clamped against a form die and the other end is wrapped around the curved portion of the die.

In form bending a filler is placed in the tube, and the bend is made by use of a punch and die in a manner similar to that employed in a sheet metal press brake. The tube is constrained at its sides by plates mounted on the form die.

These two methods of bending require a filler that is free to bend but will hold the cross section of the waveguide to close tolerance. A soft metal filler (lead, for example) will bend, but it cannot preserve the accuracy of cross section. Laminated strip filler will both bend readily and maintain cross section, but it is extremely difficult to load and unload. Also, edges of the laminations mar the sidewalls of the tube. The link-type mandrel filler works well for large-radius bends in thick-walled tubing, but in thin-walled waveguide the tube wall tends to "oil can" in between the links, as shown in Fig. 2. These are the problems that make form and compression bending impractical for economical, accurate waveguide bends.

The third common tube bending method is draw bending, illustrated in Fig. 3. Fig. 4 is a photograph of a typical draw bending tool.

It can be seen that this process is very similar to compression bending, and the tooling is almost identical. The important difference is that the tube is clamped against the straight portion of the form die and both are rotated, thus pulling the tube through the wiper and pressure dies

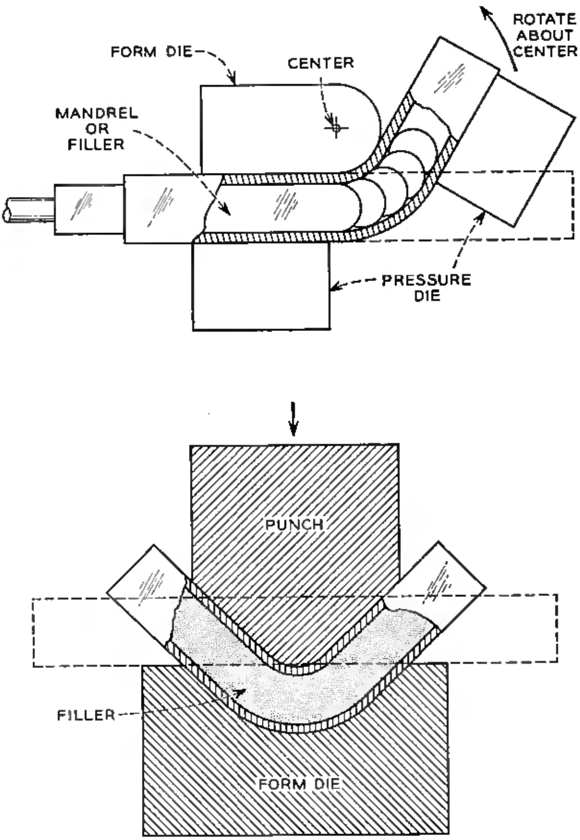


Fig. 1 — (a). Compression bending; (b) form bending.

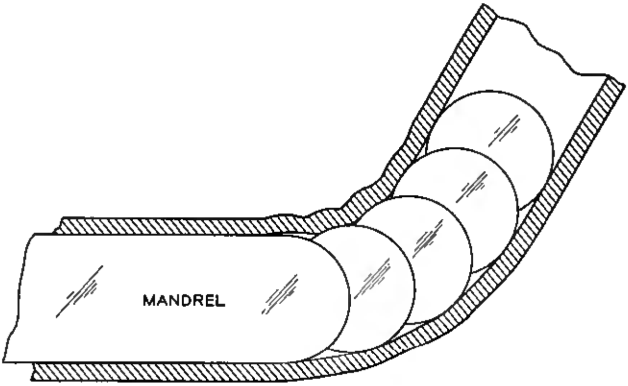


Fig. 2 — Nondrawing bending operation.

as it wraps around the curved form. This makes possible a drawing action of the tube over the mandrel, which prevents the tube wall from buckling because it is being "ironed" by the mandrel links.

When the draw bending process was selected to develop waveguide bending, many improvements had to be made to meet the problems presented. Distortion of the tubing exerted such extreme forces on the mandrel links that breakage was prohibitive on all but very large radius bends. Thin walls of waveguide wrinkled in almost every case, and no mandrels were available to make compound bends. Another fault more pronounced in draw bending is tube breakage on small radius bends, where pulling action subjects the tube to more axial tension. One way

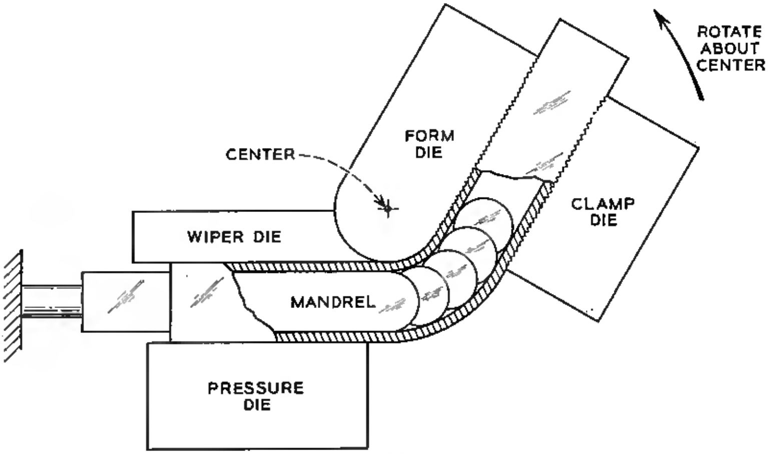


Fig. 3 — Draw bending.

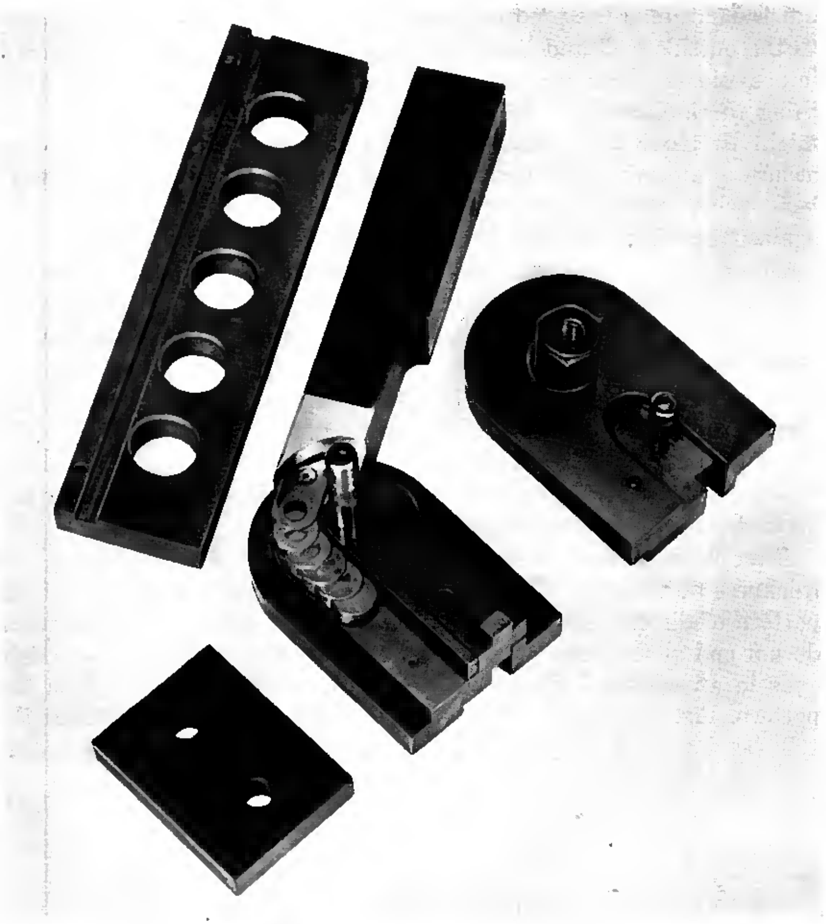


Fig. 4 — A typical bending die.

of preventing this breakage is to employ a "booster" to compress the tube axially as it is bending. However, boosting, while decreasing the tensile forces, increases the wall build-up. Fig. 5 illustrates this effect. These are the specific problems that were the main targets of the development work.

III. REFINING THE DRAW BENDING TECHNIQUE

Now it will be shown how the draw bending technique is refined to give satisfactory performance.

Wall distortions are obviously detrimental to accurate tube bending, but the nature of the process (the necessary application of tensile forces to the outside wall and compressive forces to the inner wall) makes it impossible to eliminate wall distortion. Assuming that the next best thing to elimination is accommodation, it appears that bending tools might be modified to allow for wall build-up in such a way that an accurate internal tube cross section could be preserved. Fig. 6 shows a bend being formed in the tool. The inner wall of the tube, which lies against the wiper and form die, tends to thicken, due to the compressive force involved in the bend. Designs of these parts of the tool are critical, because they must prevent buckling under great pressures. The top and bottom walls of the tube bend thicken toward the inner radius and become thinner at the outer radius. Therefore, the top and bottom plates must prevent the inner parts of these walls from buckling, even though they do not even touch the plates at the outer parts of the walls. The outer portions of the bend are in tension and pull in against the mandrel, which must be strong enough to withstand the forces involved and accurate enough to maintain size.

After investigation of sample bends by cutting and measuring wall thickness at various points in and around the bent portion, a definite pattern of distortion was revealed, as shown in Fig. 7. The distortions do not end at the tangent lines, but extend outward along the straight ends in an elliptical shape, and the wall thickness varies within this pattern. Fig. 8 is a more dramatic illustration of wall distortion. It is a 152° bend made in $\frac{5}{8}$- by $1\frac{1}{4}$-inch waveguide on a die with $\frac{1}{2}$-inch radius.

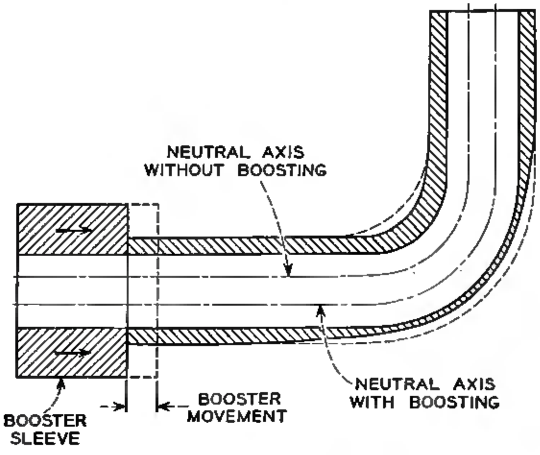


Fig. 5 — Effect of "boosting".

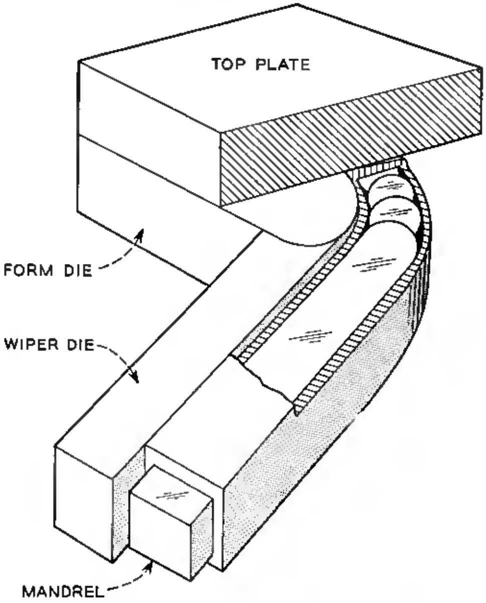


Fig. 6 — Wall distortions in relation to tooling.

These wall changes severely distort the cross section of the tube when an unrefined die is used. Fig. 9 shows what can happen, especially in small-radius bends. The thickening inner wall pushes the mandrel outward, and thus opens up a space behind the mandrel stem and allows the tube to wrinkle. The inner top and bottom walls thicken against the plates, moving them away from the form die and allow the top and bottom walls to bulge away from the mandrel. If the tooling and machinery is made extremely rigid in an attempt to prevent wrinkling and

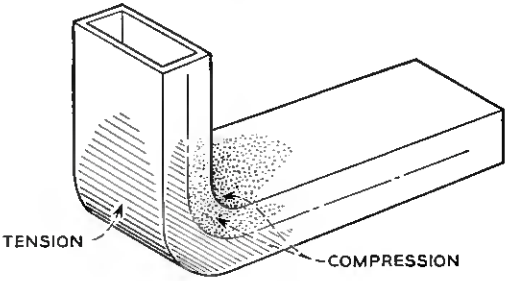


Fig. 7 — Distortion patterns in a bend.

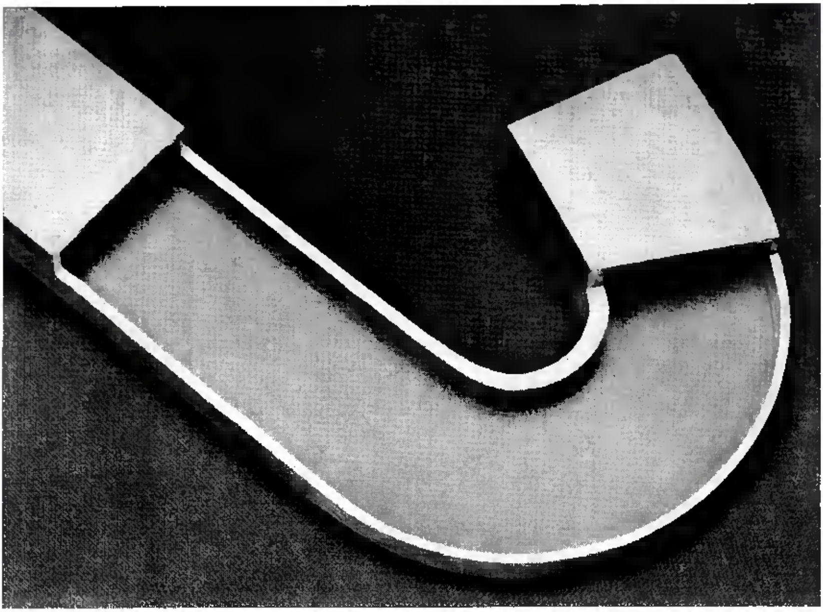


Fig. 8 — Wall distortions in a typical "H" bend.

bulging (this is often done), the mandrel is severely crushed by the thickening walls of the waveguide, while the flexible links, limited in strength, break off. On the other hand, the tools can be carefully contoured to allow for wall build-up, and the tube can maintain its internal cross section and simply "grow" into the recesses provided in the dies.

Fig. 10 shows how this is done. The form die is made smaller in radius by the same amount as the wall thickens. The wiper die is tapered off at the end to match the tapered wall of the tube, and the straight portion of the form die at the clamp end is similarly tapered. Radially tapered recesses are cut into the top and bottom plates to match the wall changes there. This scheme was tried experimentally and proved to be successful when the contouring of the die was accurate in location and amount. The pressures against the tools were greatly relieved, and the internal dimensions of the tube were held accurately. However, in order to contour the dies to sufficient accuracy, a great deal of cut-and-try work was involved. Mathematical evaluation of wall distortions in amount and location thus arises as a practical design necessity, and is undertaken in the following section.

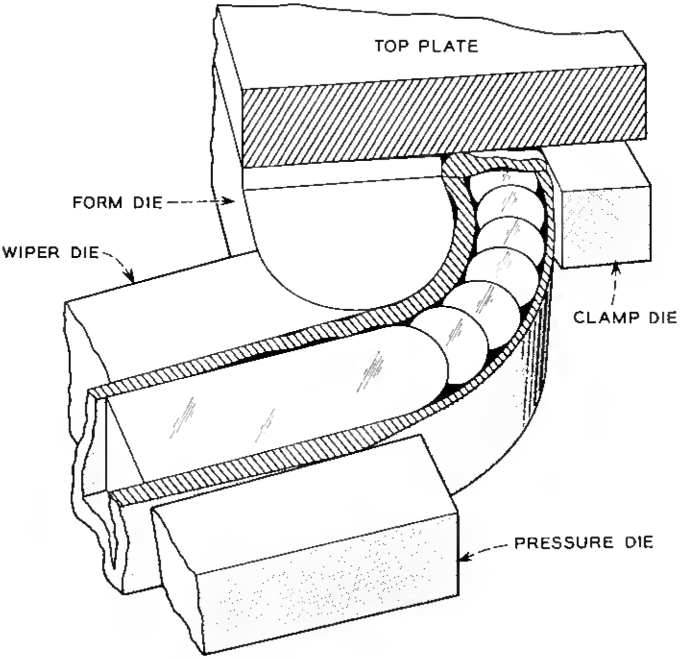


Fig. 9 — Wrinkles caused by wall distortions.

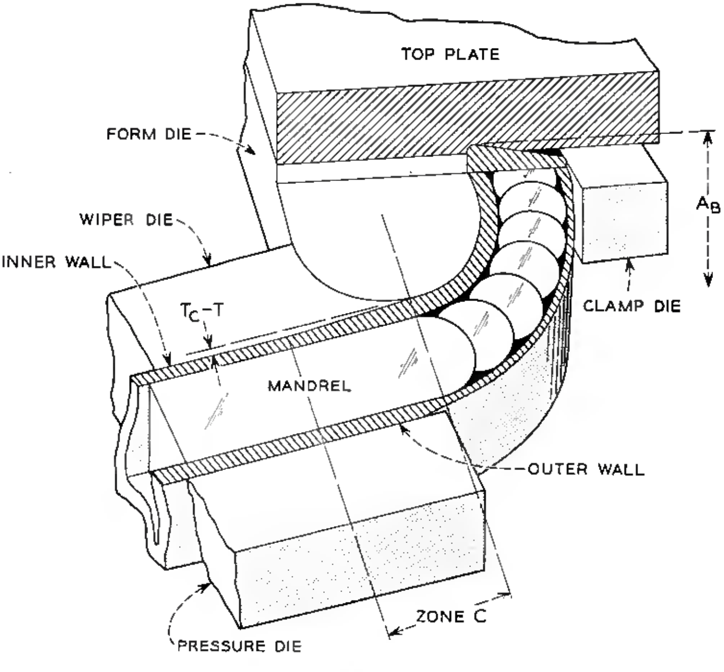


Fig. 10 — Die contoured to prevent wrinkling.

Table I — Equations for Wall Thickness Calculations–Conditions I and II

Condition I $0.0175\,\alpha\,(R_D + S - 0.5T) > S + A$ or $0.0175\alpha\,(R_D + 0.5T) > S + A$ (Special case: $T_{C1} - T = T - T_{T1}$)	Equation Number (See Appendix)
$T_C = -0.263H + \sqrt{0.069H^2 + \dfrac{0.526ATR_C}{R_D + 0.5T + 151S\left(\dfrac{R_D + 0.5T}{R_C}\right)^{0.564}}}$	(7)
$T_T = -0.263H + \sqrt{0.069H^2 + \dfrac{0.526ATR_C}{R_D + S - 0.5T}}$	(8)
$A_B = 0.5H + \sqrt{0.25H^2 + \dfrac{1.9ATR_C}{R_D + 0.5T + 0.151S\left(\dfrac{R_D + 0.5T}{R_C}\right)^{0.564}}}$	(9)
$A_T = 0.5\text{H} + \sqrt{0.25H^2 + \dfrac{1.9ATR_C}{R_D + S - 0.5T}}$	(10)
Condition II $0.0175\,\alpha\,(R_D + 0.5T) < S + A$ or $0.0175\,\alpha\,(R_D + S - 0.5T) < S + A$	
$T_{C1} = T_C\sqrt{1 - \dfrac{[S + A - 0.0175\,\alpha\,(R_D + 0.5T)]^2(T_C{}^2 - T^2)}{T_C{}^2(S + A)^2}}$	(2)
$A_{B1} = H + (A_B - H)$ $\cdot\sqrt{1 - \dfrac{[S + A - 0.0175\,\alpha(R_D + 0.5T)^2][(A_B - H)^2 - (2T)^2]}{(A_B - H)^2(S + A)^2}}$	(3)
$T_{T1} = 2T - (2T - T_T)$ $\cdot\sqrt{1 - \dfrac{[S + A - 0.0175\,\alpha(R_D + S - 0.5T)]^2[(2T - T_T)^2 - T^2]}{(S + A)^2(2T - T_T)^2}}$	(4)
$A_{T1} = 2A - H - (2A - A_T - H)$ $\cdot\sqrt{1 - \dfrac{[S + A - 0.0175\,\alpha(R_D - 0.5T)]^2[(2A - A_T - H)^2 - (A - H)^2]}{(S + A)^2(2A - A_T - H)^2}}$	(5)

IV. MATHEMATICAL EVALUATION OF WALL DISTORTIONS

For explaining the evaluation procedures worked out, the notation of the various quantities shown in Fig. 11 will be used.

In order to evaluate the changes in wall thickness due to cold-flow from the bending stresses, a large number of parts were cut open and measured at several points, and the quantities T_C, T_T, A_B and A_T were recorded. From analysis of these data a mathematical procedure for accurate evaluation was derived. It was found that two basic conditions exist for bends due to variation in angle of bend and radius. Table I defines these two conditions and also provides equations to be used to calculate the dimensions needed. The two conditions are:

i. If the angle of bend, α, and radius of the midpoint of the inner wall thickness, $R_D + 0.5T$, are such that the arc length is larger than the sum of the two nominal dimensions of the waveguide, $S + A$, Condition I exists. In the formulae for this condition the angle of bend is not used, since the changes in wall thickness do not increase with further increase in bend angle.

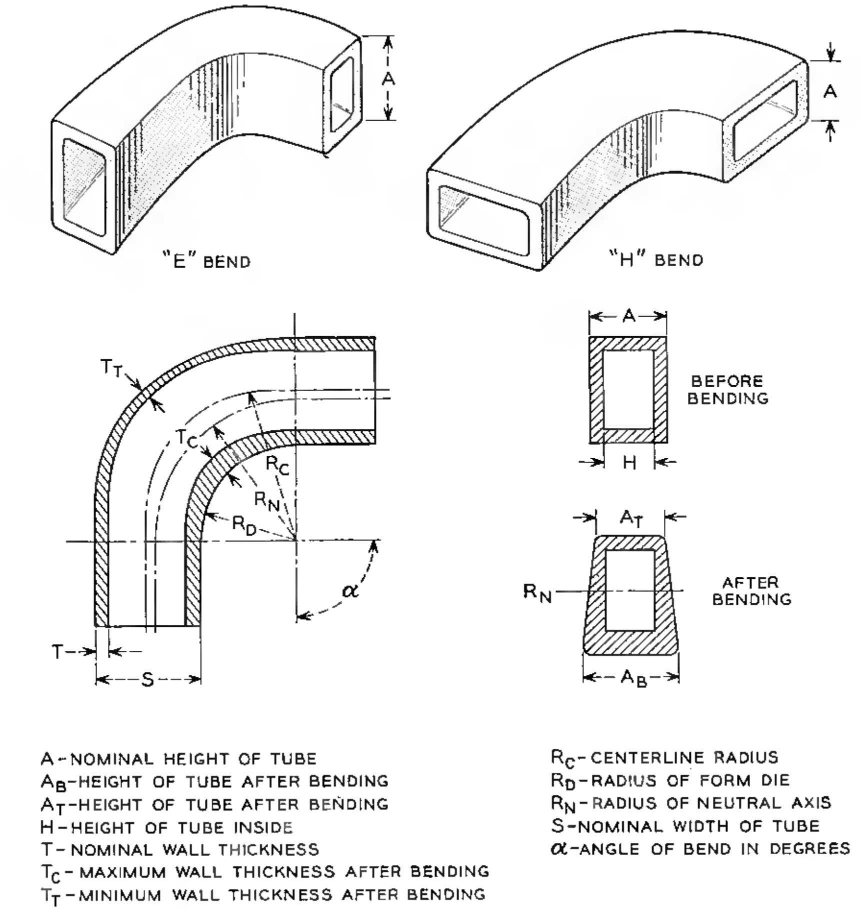

A
"E" BEND
A
"H" BEND
T_T
T_C
R_C
R_N
R_D
α
T
S
A
BEFORE BENDING
H
A_T
R_N
AFTER BENDING
A_B
A - NOMINAL HEIGHT OF TUBE
A_B - HEIGHT OF TUBE AFTER BENDING
A_T - HEIGHT OF TUBE AFTER BENDING
H - HEIGHT OF TUBE INSIDE
T - NOMINAL WALL THICKNESS
T_C - MAXIMUM WALL THICKNESS AFTER BENDING
T_T - MINIMUM WALL THICKNESS AFTER BENDING
R_C - CENTERLINE RADIUS
R_D - RADIUS OF FORM DIE
R_N - RADIUS OF NEUTRAL AXIS
S - NOMINAL WIDTH OF TUBE
α - ANGLE OF BEND IN DEGREES

Fig. 11 — Definitions of terms used in waveguide bend analysis.

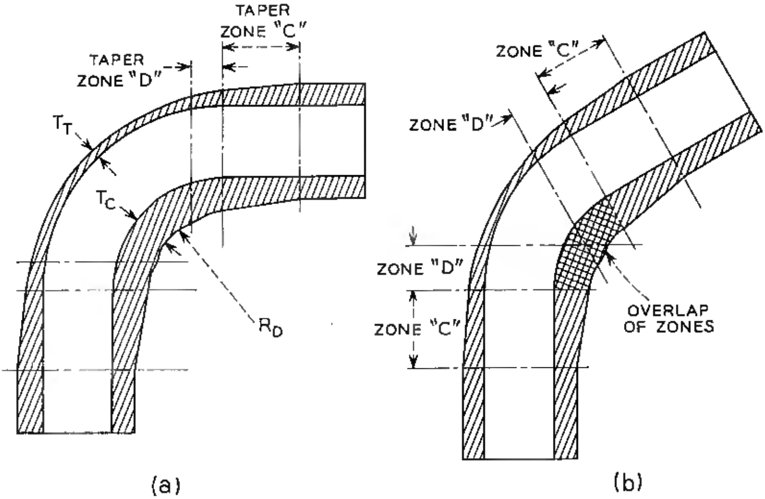


Fig. 12 — Variation of wall build-up with angle of bend.

ii. If the radius of die, R_D, and angle α are such that the arc length $(\pi/180)\alpha(R_D + 0.5T)$ is less than $S + A$, Condition II exists. Here the wall thickness changes are increasing with angle and, therefore, α is used in the calculation. Fig. 12 displays Conditions I and II.

The equations shown are partly analytical and partly empirical; their derivation and verification appears in the Appendix. By use of these calculations a series of graphs has been prepared for quick reference in determining wall thickness values. Figs. 13, 14 and 15 are examples of these graphs. They are based on the assumption that the neutral axis is located such that the outer wall thins down an amount equal to $T_C - T$. This is a condition that normally exists when a standard booster is used or when the form die radius is large enough to make the booster unnecessary.

There are several uses for these charts and evaluation procedures. In the case of new designs of bends, it is desirable to determine whether the bend can be made without breaking the tube, and how much distortion can be expected. The wall-thickness values may be needed to determine the feasibility of assembling details (such as tuning slugs, soldered brackets or clamps) to the guide in the vicinity of the bend. In compound bends, design specifications of the bend spacing can be affected by the wall thickness changes. These points are often overlooked in waveguide designs. In the design of tools for bending, as has already been discussed, these wall thickness dimensions are essential. The following example will serve to illustrate the use of these data.

A new "H"-plane bend is to be designed out of $\frac{1}{2}$- by 1-inch copper waveguide with an 0.050-inch wall thickness. The center line radius is 1.5 inches (nominal 1-inch form die radius), and the angle of bend is to be 180°. To determine if the bend can be made by standard procedures, it is necessary to know how thin the outer wall, T_T, will become, using standard tooling. For annealed copper, the maximum elongation before rupture is 30 to 35 per cent. Therefore, the wall thickness cannot decrease by more than 30 per cent of 0.050, down to 0.035 inch, or the tube will probably split. To determine which condition and thus which formula to use, it is necessary to calculate $(\pi/180)\alpha(R_D + 0.5T)$: in this case, 3.15. Consequently, the arc length is greater than $S + A$,

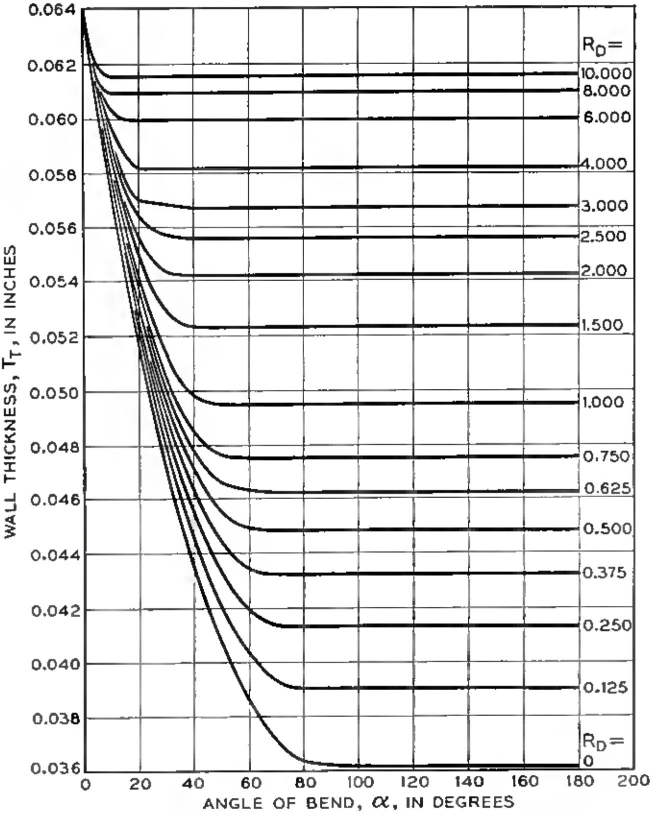


Fig. 13 — Wall thickness, T_T, vs. angle of bend.

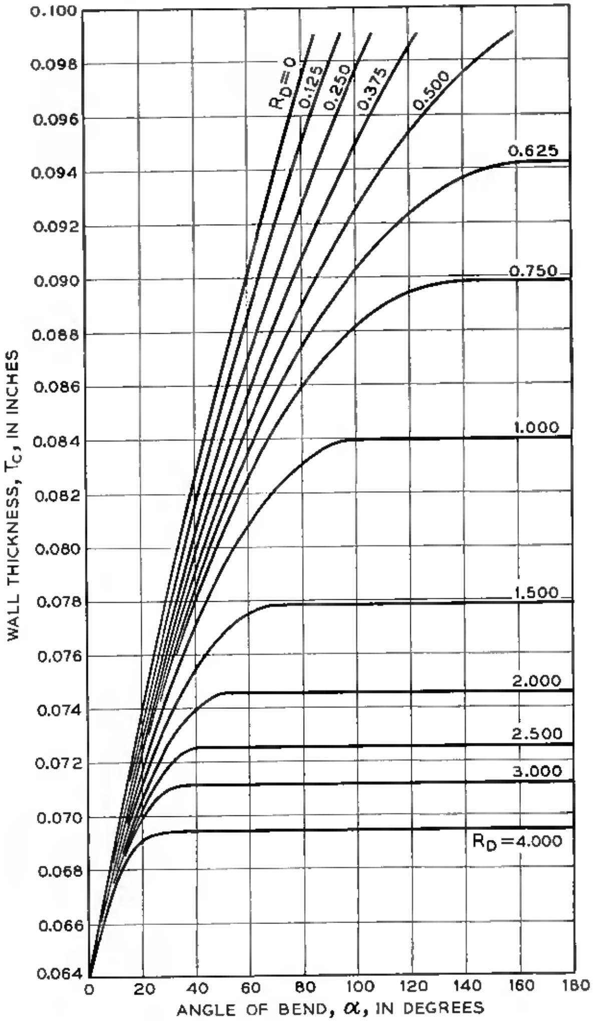


Fig. 14 — Wall thickness, T_C, vs. angle of bend.

(1.5 inches), and Condition I can be used for calculations. By substituting the proper values into the formulae in Table I for the given bend, the three dimensions T_C, T_T and A_B are determined:

$$T_C = 0.063 \text{ inch},$$

$$T_T = 0.037 \text{ inch},$$

$$A_B = 0.521 \text{ inch}.$$

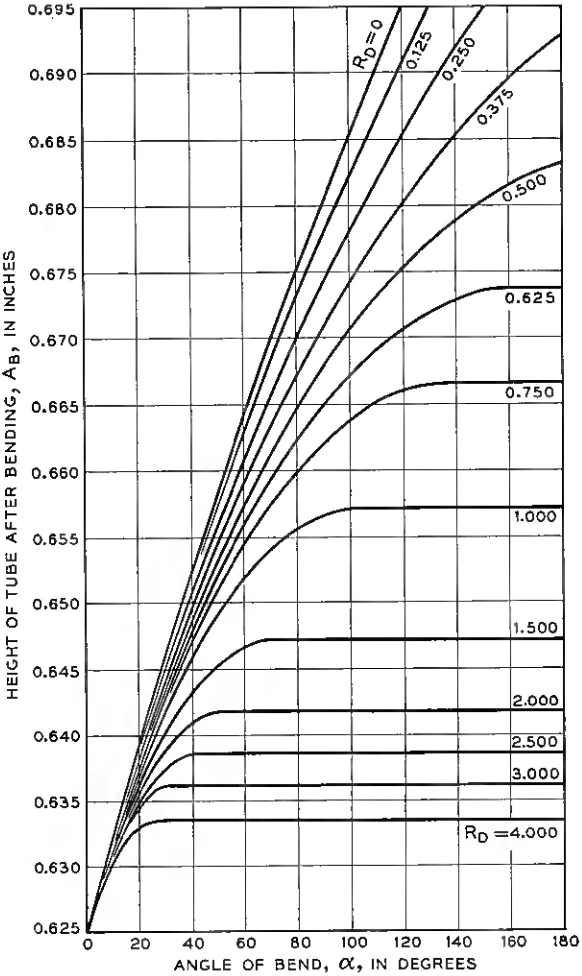


Fig. 15 — Height of tube vs. angle of bend.

Since T_T is greater than 0.035 inch no rupture will occur and the bend can be made economically.

The wall thickness that might be needed for assembly purposes is available from this same calculation. It should be noted, however, that the side wall thickness, $A_B - H/2$, is measured only at the form die radius, and tapers down to the nominal value at the neutral axis and to a proportional amount at the outer radius (refer to Fig. 11). Also, in the vicinity of the tangent lines of the bend the wall thicknesses taper from those calculated (T_T , T_C) down to nominal in zones, as shown in Fig. 12(a). If the exact values are needed in these regions simple proportions can be used, because the variations are, for all practical calculations, straight tapers. The wall thickness at the center line of the bend is increased or decreased close to three-fourths of the maximum distortion and tapers into the straight portion of the tube for a distance that may be evaluated as $S + A$. At the end of this distance, the wall thickness resumes its original value. These same data are used to contour the tools to afford the accuracy required for the particular bend. The nominal 1-inch form die radius is ground undersize an amount equal to $T_C - T$, or 0.013 inch, for the build-up. The straight clamping portion of the form die is tapered from the tangent point for a distance of $S + A$ to reflect the tapered wall at the clamp end, and the wiper die tip is tapered similarly. Now the inner wall of the tube can thicken naturally and exert no undue pressures on the mandrel (see Fig. 10). The top and bottom plates are contoured to fit the side walls of the bend by means of a circular groove whose inside radius coincides with the 0.986-inch radius of the form die and whose outer radius coincides with the radius of the center line of the tube, in this case, 1.5 inches. The depth of this groove is 0.010 inch at the 0.986-inch radius and tapers to zero at the center line of the tube. These circular grooves hold the thickened portion of the sidewalls flat.

It should be noticed that the above contouring design is accurate for the bend only when the angle of bend is great enough so that Condition I exists. If the angle is small, the tube may be in Condition II, where the wall distortions are not fully developed, and, in order to preserve accuracy, a specially contoured die is used. However, in most cases the fully contoured die will produce a small-angle bend whose electrical performance is good.

In very small radius bends, the tools are not only contoured with nominal shapes as described, but they are also refined to reflect the elliptical pattern of distortion shown in Fig. 7. This is done by measuring a tube at several points after bending and tailoring the tools to fit. Since the above example was in Condition I, the angle of bend had little importance in build up, but, as the radius becomes smaller, the angle becomes more and more important. This is quite evident on inspection of Figs. 13, 14 and 15. The slope of each curve increases with decrease in R_D . In any design work, product or tooling, it becomes more important to recognize these wall distortions in smaller radius bends. Tool contouring must be done for smaller ranges of angles. To take the most extreme case — a zero radius bend — the die can be contoured and used

Table II — Equations for Wall Thickness Calculation–General Case

General Case	Equation Number (See Appendix)
$T_C = -0.263H + \sqrt{0.069H^2 + \dfrac{0.526ATR_N}{R_D + 0.5T + 0.132(R_N - R_D - 0.5T)\left(\dfrac{R_D + 0.5T}{R_N}\right)^{0.2}}}$	(11)
$T_T = -0.263H + \sqrt{0.069H^2 + \dfrac{0.526ATR_N}{R_D + S - 0.5T - 0.132(R_D + S - 0.5T - R_N) \cdot \left(\dfrac{R_N}{R_D + S - 0.5T}\right)^{0.2}}}$	(12)
$A_B = 0.5H + \sqrt{0.25H^2 + \dfrac{1.9ATR_N}{R_D + 0.5T - 0.132(R_N - R_D - 0.5T)\left(\dfrac{R_D + 0.5T}{R_N}\right)^{0.2}}}$	(13)
$A_T = 0.5H + \sqrt{0.25H^2 + \dfrac{1.9ATR_N}{R_D + S - 0.5\mathrm{T} - 0.132(R_D + S - 0.5T - R_N) \cdot \left(\dfrac{R_N}{R_D + S - 0.5T}\right)^{0.2}}}$	(14)

to bend no more than a 5° variation in angle to maintain an internal accuracy of 0.004 inch.

The equations shown in Table II are used in the same way as those in Table I. The necessity for two different systems of calculations, along with the method of derivation, is explained in the Appendix.

The charts shown in Figs. 13, 14 and 15 can be used for all bends whose outer walls lose as much in thickness as the inner wall gains, but if a small-radius bend is to be made, and the neutral axis must be boosted outward to prevent splitting, then the wall values must be computed for the specific case.

Perhaps it would be helpful to present such a case here. Assume a design involving quite a small radius — a $\frac{1}{2}$- by 1-inch waveguide with a 90° "H" bend and a 0.25-inch R_D. With this radius, Condition I calculations may be applied for the outer wall, since its length is less than $S + A$. To check the possibility of making the bend by standard procedure the value T_T is calculated using (8). It is found to be 0.033 inch and, since this is less than the breaking point of 0.035 inch, special tooling will probably be necessary. Although the bend could be made by bending 45° and then annealing before finish-bending, this practice is undesirable, because it is more costly and produces a weaker product. A better procedure is to use the boosting principle to move the neutral axis outward until the wall thickness is greater than 0.035 inch. By substituting this value into (12) and solving for R_N, the new neutral axis position is found to be 0.778 inch. The neutral axis found using 0.033 inch for T_T was 0.732 inch. The difference, 0.046, is the amount of adjustment necessary for the booster. Now, by using 0.778 inch, any of the values can be determined from (11) and (7) for T_C and from (13) and (3) for A_B, and the tools can be accurately contoured.

Compound bends are often designed with close spacing between bends. When this spacing becomes less than the $S + A$ dimension of distortion beyond the bend, there is an overlap of stress patterns and the wall thickness at any point in this region must be computed for each bend; the resultant change in wall thickness at any point in that area will be the algebraic sum of the individual bend changes, a minus value for decreased wall and a plus value for a thickened wall. Usually, when the bends are spaced this closely, it is also necessary to hold the spacing quite accurately. This is reflected in centerline-to-centerline dimensional tolerances that often are less than ±0.005 inch. Fig. 16 is an illustration of such a product. In designing dies to make the second bend for parts such as this it is necessary to provide a clamping nest that will support the first bend. This will prevent damage from the ensuing bending action and will accurately position the part to insure meeting of the desired centerline tolerance. The clamping nest must be contoured to

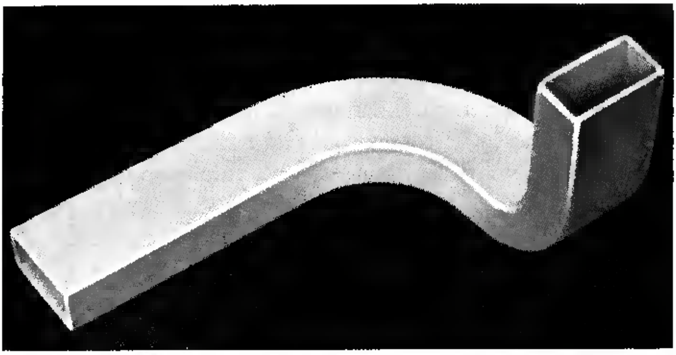


Fig. 16 — A close-coupled compound bend.

fit the first bend's wall distortions and the build-up of both bends must be allowed for in locating the nest. Contouring of the nest can be readily prescribed from the contours of the die used to make the first bend; these values are calculated as previously described. The distortions for the second bend are calculated in the same manner, but the section of the die corresponding to the place where overlapping distortions between the two bends occur is dimensioned accordingly.

This product illustrated in Fig. 16 consists of an "E" and an "H" bend spaced 0.312 inch apart in $\frac{5}{8}$- by $1\frac{1}{4}$-tubing. The "E" bend has a centerline radius of $\frac{3}{4}$ inch, and the "H" bend's centerline radius is 2 inches. The centerline offset dimension required is 3.062 ± 0.005 inch.

Fig. 17 shows how the die for the product in Fig. 16 is contoured from the values taken from Figs. 13, 14 and 15. In addition to the contoured form die and top plate as shown, the clamp die is similarly faced, and a curved clamping block is used to back up the outer wall of the nested bend. The subscripts 1 and 2 on the values in Fig. 17 refer to the nested bend and the second bend, respectively.

V. CONCLUSION

These procedures and techniques developed in the waveguide bending project have had a marked effect on production costs, quality and uniformity.

The cost of waveguide bends had been quite high, because of the

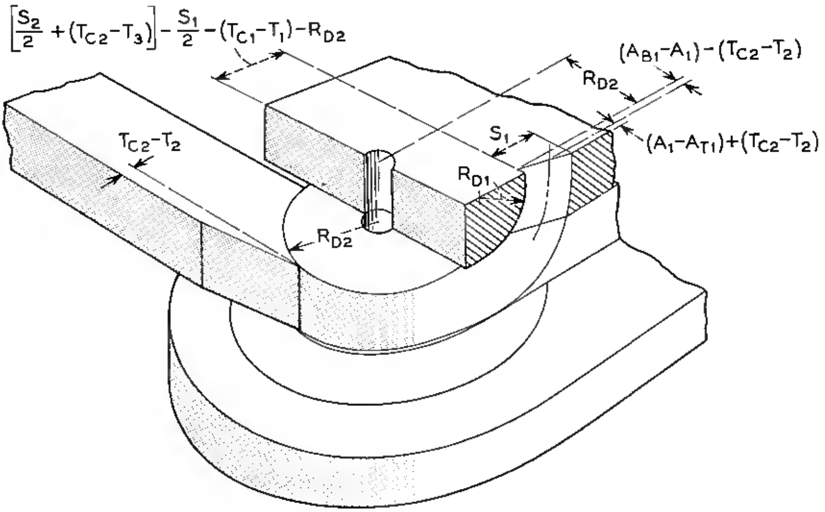

$\left[\frac{S_2}{2} + (T_{C2} - T_3)\right] - \frac{S_1}{2} - (T_{C1} - T_1) - R_{D2}$
$(A_{B1} - A_1) - (T_{C2} - T_2)$
R_{D2}
$(A_1 - A_{T1}) + (T_{C2} - T_2)$
$T_{C2} - T_2$
S_1
R_{D1}
R_{D2}

Fig. 17 — Die contours for a compound bend.

amount of labor involved in the loading and unloading of tube fillers and because of the very high spoilage rate. The new methods of contoured dies, permitting use of draw bending mandrels, have lowered the average cost of bends from over \$7.00 to less than \$2.00 each. Also, the capacity of the bending shop has increased tenfold per man hour.

Electrical transmission and physical appearance of the new bends are markedly improved. The elimination of hand work has considerably reduced surface scars and irregularities that not only were objectionable from a visual standpoint but also impaired electrical performance. It has been found that the uniformity of the process is reflected in an electrical performance from part to part which is exceptionally constant.

APPENDIX

This Appendix is used to explain the derivations of equations and present the experimental data used.

The basic method of computing wall thicknesses of bent tubing is to compare the length of the bent wall to the length of the neutral axis of the bend, and assume that the volume of the wall after bending is the same as it was before. The formula for T_C, Fig. 11, would be derived as follows:

Volume before bending:

$$V = \pi R_N \frac{\alpha}{180} AT.$$

Volume after bending:

$$V = \pi(R_D + 0.5T) \frac{\alpha}{180} A_B T_C.$$

Equating the two:

$$\pi R_N \frac{\alpha}{180} AT = \pi(R_D + .5T) \frac{\alpha}{180} A_B T_C.$$

Solving for T_C:

$$T_C = \frac{ATR_N}{(R_D + 0.5T)A_B}. \tag{1}$$

This basic calculation was made with two assumptions: first, that the distortion of the thickening wall ends abruptly at the two centerlines of bend; second, that wall increase is the same for any angle of bend. Neither of these conditions is found to be true, although the two are interrelated. It is found that the wall thickness changes beyond the tangent point of bend, as shown in Figs. 7 and 12. This distance beyond the centerline, Zone C, is approximately equal to $S + A$ in over-all

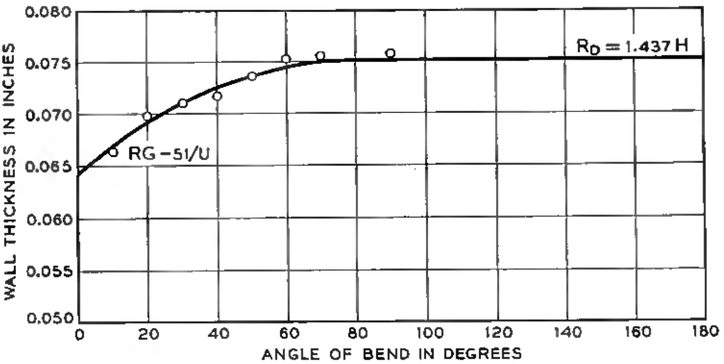


Fig. 18 — Wall thickness, T_C, for a typical bend.

length. Fig. 12 also shows a Zone D, which is in the bent portion of the waveguide. This zone is also reasonably constant in length, approximating $0.5(S + A)$. Since the wall distortion is a maximum at the end of Zone D and something less at the centerline of bend (found to be 0.75 of the maximum by experiment), it is evident that the wall thickness will change from nominal to its maximum value only after the bend has progressed sufficiently to provide an arc length, center to center, greater than $S + A$. Fig. 18, a curve of measured values for a typical bend, shows this effect clearly. The wall increases from nominal at 0° bend to a maximum at about 75° bend. For this inner wall, the angle of bend that gives an arc length of $S + A$, (1.875 inches in this case) is

$$\frac{(S + A)\,180}{\pi(R_D + 0.5T)} = 73^\circ.$$

In order to describe this build-up effect mathematically, use is made of the equation for an ellipse,

$$\frac{x^2}{a^2} + \frac{y^2}{b^2} = 1,$$

that fits the wall distortion curve very well. Referring to Fig. 19, b is equal to T_C at $X = 0$. At the known point of 0° bend, $X = S + A$ and $y = T$. By substituting these values into the general equation for an ellipse, the value a^2 can be found:

$$\frac{(S + A)^2}{a^2} + \frac{T^2}{T_C^2} = 1,$$

$$a^2 = \frac{(S + A)^2\, T_C^2}{T_C^2 - T^2}.$$

At any other point between $X = 0$ and $X = S + A$, X would be equal to the arc length of bend subtracted from $S + A$:

$$x = S + A - (R_D + 0.5T)\frac{\alpha\pi}{180}.$$

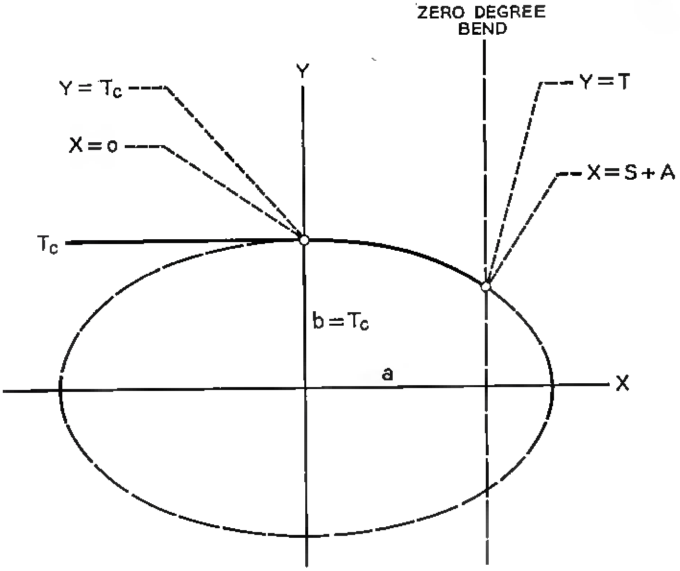
For the purpose of clarity, the value of y at this point is denoted T_{C1}. By substituting these values and the value of a^2 as above, the equation for an ellipse can be again rewritten:

$$\frac{\left[S + A - (R_D + 0.5T)\alpha\,\frac{\pi}{180}\right]^2}{\dfrac{(S+A)^2T_C^{\,2}}{T_C^2 - T^2}} + \frac{T_{C1}^{\,2}}{T_C^2} = 1.$$

After solving for T_{C1}, the equation appears as:

$$T_{C1} = T_C\sqrt{1 - \frac{[S + A - 0.0175\alpha(R_D + 0.5T)]^2[T_C^{\,2} - T^2]}{T_C^2(S+A)^2}}. \qquad (2)$$

This is the equation that was used for T_{C1} in Table II. The accuracy



Y = WALL THICKNESS
X = S + A MINUS ARC LENGTH OF BENT WALL

Fig. 19 — Ellipse used to derive equation (2).

of this calculation is within 0.002 inch on sizes of waveguide ranging from 0.391 by 0.702 inch to $1\frac{1}{2}$ by 3 inches. In Fig. 18 the curve shows the calculated values in relation to the experimental values for a typical bend.

The foregoing derivation was for the inner wall thickness only. The sidewall thickness is calculated in exactly the same way to evaluate the height of tube after bending, A_B. The only variation is that the calculated sidewall thickness is doubled and added to the internal height of the tube, H. This variation of (2) becomes

$$A_{B1} = H + (A_B - H) \cdot \sqrt{1 - \frac{[S + A - 0.0175\alpha(R_D + 0.5T)^2][(A_B - H)^2 - (2T)^2]}{(A_B - H)^2(S + A)^2}}. \quad (3)$$

Another value needed for bend analysis is the outer wall thickness, T_T. The same method of setting up an equation is used, with Zones C and D being the same length for tension as they are for compression. After solving the equation for the ellipse as shown in Fig. 20, the ex-

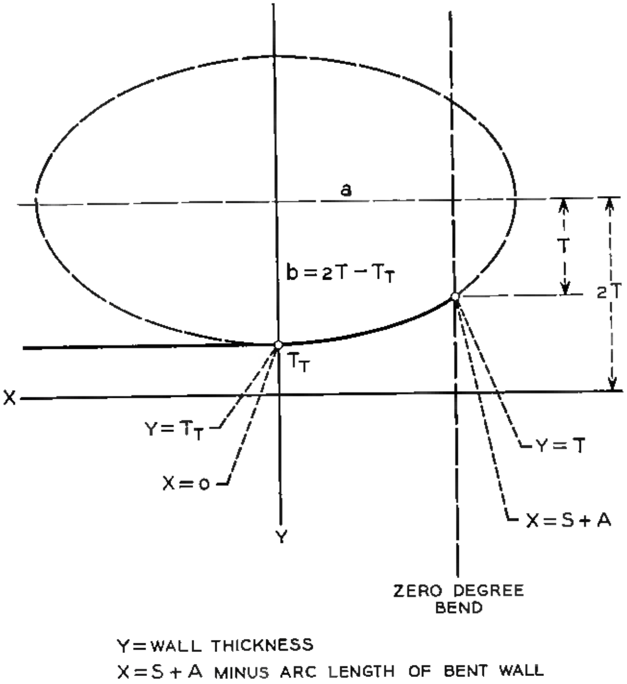


Fig. 20 — Ellipse used to derive equation (4).

pression for T_{T1} is as follows:

$$T_{T1} = 2T - (2T - T_T) \cdot \sqrt{1 - \frac{[S + A - 0.0175\alpha(R_D + S - 0.5T)]^2[(2T - T_T)^2 - T^2]}{(S + A)^2(2T - T_T)^2}} . \tag{4}$$

The only remaining value to determine is the height of tube after bending at the outer wall. This is derived from the same type of elliptical function as the foregoing, and is

$$A_{T1} = 2A - H - (2A - A_T - H) \cdot \sqrt{1 - \frac{[S + A - 0.0175(R_D - 0.5T)]^2 \cdot [(2A - A_T - H)^2 - (A - H)^2]}{(S + A)^2(2A - A_T - H)^2}} . \tag{5}$$

Equations (2), (3), (4) and (5) evaluate the wall distortions of a bend which is still in Condition II. The equations all include the values T_C, A_B, T_T and A_T, which must be determined first. Equation (1) expresses the wall distortions assuming no distortion beyond the bend. By rewriting the equation, the fact that the wall does thicken beyond the centerline is represented by an empirical addition to $(R_D + 0.5T)$. Measurements of samples which were boosted by standard methods were substituted into (6) and the value K was obtained:

$$T_C = - \frac{ATR_N}{A_B[(R_D + 0.5T) + K]} . \tag{6}$$

By assuming R_N equal to R_C and plotting K against various parameters, it was found that K equals

$$0.151S\left(\frac{R_D + 0.5T}{R_C}\right)^{0.564}$$

for T_C and close to zero for T_T. Therefore, equations for T_C and T_T appear as follows:

$$T_C = \frac{ATR_C}{A_B\left[R_D + 0.5T + 0.151S\left(\frac{R_D + 0.5T}{R_C}\right)^{0.564}\right]},$$

$$T_T = \frac{ATR_C}{A_T(R_D + S - 0.5T)} .$$

From Table III, it is evident that the sidewalls after bending assume a value close to 0.95 times their adjacent walls. From this, $A_B = H + 1.9T_C$ and $A_T = H + 1.9T_T$, and final equations for T_C and T_T can be

Table III — Measurements of Wall Thicknesses Taken from Sample Production Bends

Description of Bends	T_C	T_T	A_B	A_T	Ratio $\frac{A_B - H}{T_C}$	Ratio $\frac{A_T - H}{T_T}$
0.391 × 0.702 E, 0.196 R_D 90°	0.050	0.030	0.714	0.678	1.84	1.87
0.391 × 0.702 E, 0.562 R_D 150°	0.047	0.033	0.712	0.687	1.91	1.97
0.391 × 0.702 H, 0.351 R_D 90°	0.052	0.030	0.412	0.367	1.90	1.86
0.5 × 1 E, 0.5 R_D 90°	0.060	0.040	1.014	0.976	1.90	1.90
0.5 × 1 E, 0.348 R_D 90°	0.063	0.037	1.017	0.969	1.86	1.86
0.5 × 1 H, 0.625 R_D 90°	0.066	0.035	0.525	0.467	1.93	1.92
0.5 × 1 H, 0.625 R_D 180°	0.068	0.035	0.530	0.473	1.86	1.97
0.625 × 1.25 E, 1.5 R_D 180°	0.071	0.056	1.260	1.226	1.94	1.86
0.625 × 1.25 H, .5 R_D 90°	0.088	0.043	0.667	0.575	1.96	1.95
0.625 × 1.25 H, 0 R_D 45°	0.083	0.044	0.655	0.592	1.90	1.90
0.625 × 1.25 H, 1.5 R_D 90°	0.076	0.0523	0.640	0.587	1.88	1.91
0.625 × 1.25 H, 2 R_D 90°	0.075	0.053	0.645	0.593	1.95	1.86
1.273 × 2.418 E, 6.86 R_D 90°	0.069	0.059	2.423	2.399	1.93	1.85
1.273 × 2.418 H, 6.29 R_D 90°	0.073	0.056	1.285	1.250	1.92	1.88
Average					1.90	1.89

found by substituting $H + 1.9T_C$ and $H + 1.9T_T$ into the above and solving for the values T_C and T_T:

$$T_C = -0.263H + \sqrt{0.069H^2 + \frac{0.526ATR_C}{R_D + 0.5T + 0.151S\left(\frac{R_D + 0.5T}{R_C}\right)^{0.524}}}, \tag{7}$$

$$T_T = -0.263H + \sqrt{0.069H^2 + \frac{0.526ATR_C}{R_D + S - 0.5T}}. \tag{8}$$

The sidewall thicknesses A_B and A_T can be evaluated similarly:

$$A_B = H + \frac{1.9ATR_C}{A_B\left[R_D + 0.5T + 0.151S\left(\frac{R_D + 0.5T}{R_C}\right)^{0.564}\right]}.$$

Solving for A_B:

$$A_B = 0.5H + \sqrt{0.25H^2 + \frac{1.9ATR_C}{R_D + 0.5T + 0.151S\left(\frac{R_D + 0.5T}{R_C}\right)^{0.564}}}. \tag{9}$$

Similarly,

$$A_T = 0.5H + \sqrt{0.25H^2 + \frac{1.9ATR_C}{R_D + S - 0.5T}}. \tag{10}$$

These equations can be used to determine wall thicknesses within 0.002 inch for a range of waveguide sizes of from 0.391 to 2.418 inches and for any radius of bend.

Equations (7), (8), (9) and (10) can be used only where the tube wall distortions of outer and inner walls are almost equal. Fortunately, this is the normal situation, and these equations are very useful because they cover the great majority of cases. However, in cases where it is desirable to determine the neutral axis, R_N, or where R_N is known, it would also be desirable to have expressions for T_C, T_T, A_B and A_T in terms of R_N.

These expressions are written in general form as:

$$T_C = \frac{ATR_N}{A_B(R_D + 0.5T + K)},$$

$$T_T = \frac{ATR_N}{A_T(R_D + S - 0.5T - K)},$$

$$A_B = H + \frac{2ATR_N}{A_B(R_D + S - 0.5T - K)},$$

$$A_T = H + \frac{2ATR_N}{A_T(R_D + S - 0.5T - K)},$$

where K has the same significance as previously explained. By substitution of T_C and T_T taken from experimental samples in Table III and solving for R_N and K simultaneously, K is found to be equal to

$$0.132(R_N - R_D - 0.5T)\left(\frac{R_D + 0.5T}{R_N}\right)^{0.2}$$

and

$$0.132(R_D + S - 0.5T)\left(\frac{R_N}{R_D + S - 0.5T}\right)^{0.2}$$

for the respective cases.

Finally, T_C and T_T are evaluated by:

$$T_C = \frac{ATR_N}{A_B\left[R_D + 0.5T + 0.132(R_N - R_D - 0.5T)\left(\frac{R_D + 0.5T}{R_N}\right)^{0.2}\right]}.$$

Substituting $A_B = H + 1.9T_C$ and solving for T_C:

$$T_C = -0.263H + \sqrt{0.069H^2 + \frac{0.526ATR_N}{R_D + 0.5T + 0.132(R_N - R_D - 0.5T) \cdot \left(\frac{R_D + 0.5T}{R_N}\right)^{0.2}}}, \qquad (11)$$

$$T_T = \frac{ATR_N}{A_T\left[R_D + S - 0.5T - 0.132(R_D + S - 0.5T - R_N) \cdot \left(\frac{R_N}{R_D + S - 0.5T}\right)^{0.2}\right]}.$$

Substituting $A_T = H + 1.9T_T$, and solving for T_T:

$$T_T = -0.263H + \sqrt{0.069H^2 + \frac{0.526ATR_N}{R_D + S - 0.5T - 0.132(R_D + S - 0.5T - R_N)\left(\frac{R_N}{R_D + S - 0.5T}\right)^{0.2}}}. \qquad (12)$$

The sidewall thicknesses are found to be less than the outer and inner wall thicknesses. Table III shows this relationship to be a constant ratio: the sidewalls are 0.95 times the outer and inner walls. By use of this ratio, formulae for A_B and A_T are written as:

$$A_B = H + \frac{1.9ATR_N}{A_B\left[R_D + 0.5T + 0.132(R_N - R_D - 0.5T)\quad\left(\frac{R_D + 0.5T}{R_N}\right)^{0.2}\right]}.$$

Solving for A_B:

$$A_B = 0.5H + \sqrt{0.25H^2 + \frac{1.9ATR_N}{R_D + 0.5T + 0.132(R_N - R_D - 0.5T) \cdot \left(\frac{R_D + 0.5T}{R_N}\right)^{0.2}}}. \qquad (13)$$

Similarly,

$$A_T = H + \frac{1.9ATR_N}{A_T\left[R_D + S - 0.5T - 132(R_D + S - 0.5T - R_N)\left(\frac{R_N}{R_D + S - 0.5T}\right)^{0.2}\right]}.$$

Solving for A_T:

$$A_T = 0.5H + \sqrt{0.25H^2 + \frac{1.9ATR_N}{R_D + S - 0.5T - 0.132(R_D + S - 0.5T - R_N)} \cdot \left(\frac{R_N}{R_D + S - 0.5T}\right)^{0.2}}. \quad (14)$$

Here, A_T and A_B have been shown as separate calculations, but where T_T and T_C are already known or computed A_T and A_B can be evaluated by:

$$A_T = H + 1.9T_T, \quad (15)$$

$$A_B = H + 1.9T_C. \quad (16)$$

These relations are found to be true for both Conditions I and II.

Error-Correcting Codes — A Linear Programming Approach

By E. J. McCLUSKEY, JR.

(Manuscript received June 5, 1959)

Two theorems are proved that characterize the matrices used to construct systematic error-correcting codes. A lower bound on the number of required check bits is derived, and it is shown that, in certain cases, this bound for systematic codes is identical with Plotkin's bound on the size of any error-correcting code. A linear program whose solutions correspond directly to a minimum-redundancy error-correcting code is derived. This linear program can be solved by an algorithm that is essentially the simplex method modified to produce integer solutions. Explicit solutions in closed form that specify the codes directly are derived for the cases when the specified code parameters satisfy certain restrictions. Several theorems are proved about minimum redundancy codes with related parameters.

I. INTRODUCTION

This paper is concerned with the problem of transmitting binary signals over a noisy channel. Some situations in which this problem occurs are: when telephone lines are being used to transmit data in binary form; when an imperfect medium such as magnetic tape or a photographic emulsion is used to store binary data; or when operations on binary signals are being carried out by means of circuits constructed of devices such as relays, diodes or transistors, which have a probability of error. It has been shown by Shannon[1] that it is possible to add redundant bits to the transmitted messages so as to reduce the probability of error in the received messages to an arbitrarily small quantity. Since Shannon did not exhibit efficient codes for achieving this reduction in error probability, considerable attention has been devoted to the search for useful coding schemes. The usefulness of a coding scheme is determined by the number of redundant bits that must be added, by the complexity of the equipment required for inserting the redundant bits before transmission and for removing the redundant bits and correcting errors after transmission, and by the error-correcting capabilities.

In 1950, Hamming[2] published schemes for constructing codes for (a) detecting the presence of an error in one out of n bits, (b) correcting an error in one out of n bits or (c) correcting an error in one out of n bits *and* detecting errors in two out of n bits. In all these codes, it is possible to separate the transmitted message into information or message bits and redundant or check bits. Hamming defined codes that have this property as *systematic* codes, and proved that all systematic codes can be constructed by means of parity constraints on the transmitted bits. He also proved that the codes that he constructed contained the minimum number of check or redundant bits. While Hamming did not obtain any codes for correcting more than one error, he did show that a code for correcting e errors can always be changed into a code for correcting e errors *and* detecting $e + 1$ errors by adding one extra check bit that makes the over-all parity of the transmitted message always even.

A procedure for constructing codes for multiple errors was obtained by Reed[3] and Muller.[4] The resulting codes are commonly called Reed-Muller codes, since they were obtained independently by both Reed and Muller. A Reed-Muller code can be constructed for detecting e errors whenever e is a power of two ($e = 2^x$). The number of bits in the resulting code will also be a power of two. This paper presents a method for constructing minimum-redundancy codes for correcting or detecting any specified number of errors.

II. THE HAMMING MATRIX

A *binary word* is defined as a sequence of n binary digits, $x = x_1x_2 \cdots x_n$; and the *distance* between two binary words is defined as $d(x, y) = (x_1 \oplus y_1) + (x_2 \oplus y_2) + \cdots + (x_n \oplus y_n)$,* which is equal to the number of bit locations in which the two words differ. An *e-error-correcting code*[2] is a collection of binary words for which the distance between any two words is greater than or equal to $2e + 1$. If an error-correcting code consists of all binary words whose digits satisfy certain parity-check requirements, the code is called a *systematic error-correcting code*. For example, the collection of six-bit binary words that satisfy $x_3 \oplus x_4 \oplus x_5 = 0$, $x_2 \oplus x_4 \oplus x_6 = 0$ and $x_1 \oplus x_5 \oplus x_6 = 0$ forms a one-error-correcting code. The problem of obtaining a systematic error-correcting code is equivalent to that of finding a set of parity-check requirements that will generate a set of words with the required distance property.

The parity-check requirements can be specified by a matrix of zeros and ones in which the jth column corresponds to the jth bit of the

* The symbol $\oplus$ represents addition modulo two.

binary code words and the ith row corresponds to the ith parity check. The entry in the ith row and jth column is one if the jth bit is involved in the ith parity check and is zero otherwise. This matrix will be called a *Hamming matrix*, and its elements will be represented by the symbol h_{ij}. This is the Hamming matrix for parity rules $x_3 \oplus x_4 \oplus x_5 = 0$, $x_2 \oplus x_4 \oplus x_6 = 0$, $x_1 \oplus x_5 \oplus x_6 = 0$:

$$\begin{array}{c} \begin{array}{cccccc} x_1 & x_2 & x_3 & x_4 & x_5 & x_6 \end{array} \\ \begin{bmatrix} 0 & 0 & 1 & 1 & 1 & 0 \\ 0 & 1 & 0 & 1 & 0 & 1 \\ 1 & 0 & 0 & 0 & 1 & 1 \end{bmatrix} \end{array} = [h_{ij}].$$

The first problem considered in this paper is that of characterizing Hamming matrices by determining the necessary and sufficient conditions that a matrix of zeros and ones be a Hamming matrix for a code with minimum distance d (between any two code words).

In the following, the binary code words will be represented by column matrices,

$$x] = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix},$$

and the *Boolean product* of two matrices with elements a_{ij} and b_{ij} will be defined as a matrix $[c_{ij}] = [a_{ij}] \circ [b_{ij}]$ with elements $c_{ij} = \sum_k a_{ik} b_{kj}$ (modulo 2).

Example 1:

$$\begin{bmatrix} 0 & 1 & 0 & 1 \\ 0 & 1 & 1 & 0 \\ 1 & 0 & 1 & 1 \end{bmatrix} \circ \begin{bmatrix} 1 & 1 \\ 0 & 1 \\ 1 & 0 \\ 0 & 0 \end{bmatrix} = \begin{bmatrix} 0 & 1 \\ 1 & 1 \\ 0 & 1 \end{bmatrix}.$$

Definition: A matrix of zeros and ones with k rows and n columns is the *Hamming matrix for a code of minimum distance* d if and only if $d(x, y) \geqq d$ for all x and y $(x \neq y)$ for which $[H]\circ x] = 0]$ and $[H]\circ y] = 0]$, where $0]$ represents a column matrix of k zeros.

Definition: The *weight of a matrix*, $w[a_{ij}]$, is equal to the number of entries of the matrix which are equal to one (for matrices of only zeros and ones).

Lemma 1: $d(0, x) = w[x]$, where 0 represents a sequence of n zeros.

Definition: The *sum (modulo 2) of two or more columns of a matrix* is the column matrix with each element equal to the sum modulo 2 of the elements in the same row of the columns being summed.

Lemma 2: If $d(x, y) = d_1$, then $y] = x \oplus z]$, where $w[z] = d_1$.

Theorem 1: H is the Hamming matrix for a code of minimum distance, d, if and only if $[H]\circ z] \neq 0]$ for all $z]$ $(z] \neq 0])$ for which $w[z] < d$.

Proof: First suppose the H is a Hamming matrix for a code of minimum distance d and that $w[z] < d$; then $[H]\circ 0] = 0]$ and $d(0, z) < d$, so that $[H]\circ z]$ cannot equal $0]$, by the definition of a Hamming matrix. Next suppose that $[H]\circ z] \neq 0]$ for all $z]$ $(z] \neq 0])$ for which $w[z] < d$, and $d(x, y) < d$. Then $y]$ can be expressed as $x \oplus z]$, where $w[z] < d$; and $[H]\circ y] = [H]\circ x \oplus z] = [H]\circ x] + [H]\circ z]$. Thus, if x is a code word, $[H]\circ x] = 0]$, y cannot be a code word, since $[H]\circ y] = [H]\circ x] + [H]\circ z] = 0] + [H]\circ z] \neq 0]$. This shows that, if $[H]\circ z] \neq 0$ for all $z]$ with $w[z] < d$, then $d(x, y)$ must be equal to or greater than d for all x and y with $[H]\circ x] = 0$ and $[H]\circ y] = 0$.

Corollary 1: H is the Hamming matrix for a code of minimum distance d if and only if no set of $d - 1$ or fewer columns sums to the all-zero column.

Proof: If $d - 1$ or fewer columns sum to zero there is a corresponding z with $w[z] < d$ such that $[H]\circ z] = 0]$.

This theorem makes it possible to attack the problem of finding a systematic code with the specified d by constructing a matrix satisfying the given conditions. However, no satisfactory procedure for constructing such a matrix directly is known, and the construction procedure to be developed here is based on Theorem 2, which characterizes the parity-check matrix, a submatrix of the Hamming matrix.

III. THE PARITY-CHECK MATRIX

Hamming showed that a Hamming matrix can always be put in the form of a $k \times k$ unit matrix (matrix with ones on the main diagonal and zeros elsewhere) and a $k \times n - k$ arbitrary matrix called the *parity-check matrix* (see Ref. 2, Section 7). This form of the Hamming matrix will be called the *standard form.* In the following it will be assumed that the Hamming matrices are always in standard form.

It is customary to use the term redundant bits or *check bits* for the bits of the code words which correspond to columns of the unit matrix part of the Hamming matrix. The remaining $n - k = m$ bits are called *information* or *message* bits. This usage derives from the fact that each of the check bits occurs in only one of the parity checks, and therefore the values of each check bit can be calculated directly from the values of the information bits, independent of the values of the other check bits. If the elements of the parity-check matrix are denoted by p_{ij}, the check bits (u_i) are obtained from the message bits (x_j) according to the following expression:

$$u_i = \sum_{j=1}^{m} p_{ij}\, x_j \quad (\text{modulo } 2). \tag{1}$$

A systematic error-correcting code is thus completely specified by the parity-check matrix. The main object of this paper is to present methods for obtaining parity-check matrices corresponding to systematic codes that have a specified minimum distance d between any pair of code words and requiring the minimum number of check bits.

The following is an example of a systematic code of minimum distance 3 (one-error-correcting code) having two message bits and three check bits.

Example 2:

Matrix: $$P = \begin{bmatrix} 1 & 1 \\ 1 & 0 \\ 0 & 1 \end{bmatrix};$$

equations: $$\begin{aligned} u_1 &= x_1 \oplus x_2\,, \\ u_2 &= x_1\,, \\ u_3 &= x_2\,; \end{aligned}$$

code: $$\begin{array}{ccccc} u_1 & u_2 & u_3 & x_1 & x_2 \\ \end{array}$$
$$\begin{bmatrix} 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 1 \\ 1 & 1 & 0 & 1 & 0 \\ 0 & 1 & 1 & 1 & 1 \end{bmatrix}.$$

The method to be used for constructing parity-check matrices depends on the following theorem.

Theorem 2: P is a parity-check matrix for a code of minimum distance d if and only if:

i. the weight of each column of P is greater than or equal to $d - 1$;

ii. the weight of the sum (modulo 2) of J columns is greater than or equal to $d - J$.

Proof: First, suppose that the conditions of the theorem are not satisfied, and consider the Hamming matrix made up of a unit matrix and the given P matrix. If there is a column of P with weight $w_1 < d - 1$, then the sum of this column and w_1 of the unit columns (one unit column for each one entry of the column of P) will be equal to zero. Since the total number of columns involved in this sum is $w_1 + 1 < d$, the conditions of the Corollary 1 are violated and P cannot correspond to a code of minimum distance d. Similarly, if the sum of J columns has weight $w_J < d - J$, these J columns of P plus w_J unit columns will sum to the all-zero column. The total number of columns summed is $J + w_J < J + d - J = d$, again violating the conditions of the Corollary 1. Thus, unless conditions i and ii are satisfied by P it cannot be the parity-check matrix for a code of minimum distance d.

Next, suppose that conditions i and ii of this theorem are satisfied, and consider which combinations of columns will sum to the all-zero column. No combination involving only unit columns can sum to zero, since these are linearly independent. As discussed in the preceding paragraph, any combination involving only one of the columns of the P matrix will contain $w_1 + 1 \geqq d$ columns, and any combination involving more than one of the columns of the P matrix will contain $J + w_J \geqq J + d - J = d$ columns. Thus, any combination of columns that sums to the all-zero column must involve at least d columns. The conditions of Corollary 1 are satisfied and P corresponds to a code of minimum distance d.

In this paper the construction of error-correcting codes will be based on finding matrices which satisfy the conditions of Theorem 2. The matrices will be obtained directly from the solutions to a set of linear inequalities.

IV. FORMATION OF LINEAR PROGRAM

In order to check that a given matrix P satisfies the conditions of Theorem 2, it is necessary to form the sums modulo 2 of all pairs of columns of P, compute the weights, and compare the weights with $d - 2$; then this must be repeated for all triples of columns, comparing with $d - 3$; all quadruples of columns, comparing with $d - 4$, etc. A systematic procedure for doing this can be given in terms of the following definition.

Definition: P_J $(J = 1, 2, \cdots m)$ is the matrix formed from P by taking, as the columns of P_J, the sums of all possible combinations of J columns of P (P_1 is identical with P).

Example 3: P is the parity check matrix for a code of minimum distance 3 since the weight of each column of P is at least $3 - 1 = 2$, and the weight of each column of P_2 is at least $3 - 2 = 1$:

$$\begin{array}{cccc}
\begin{bmatrix} 0 & 1 & 1 & 1 \\ 1 & 0 & 1 & 1 \\ 1 & 1 & 0 & 1 \end{bmatrix} &
\begin{bmatrix} 1 & 1 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 & 1 & 0 \\ 0 & 1 & 0 & 1 & 0 & 1 \end{bmatrix} &
\begin{bmatrix} 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \end{bmatrix} &
\begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}. \\
(\mathrm{a})P & (\mathrm{b})P_2 & (\mathrm{c})P_3 & (\mathrm{d})P_4
\end{array}$$

A method for checking a matrix P is to form P_2, P_3, $\cdots$ and then to verify that the weight of each column of P_J is at least $d - J$. While this method is quite satisfactory for verifying that a given matrix satisfies the conditions for a parity-check matrix of a code of minimum distance d, it is of little use for the more important problem of constructing a matrix satisfying these conditions. For this reason, a modified method for testing a matrix will be presented as a preliminary to the discussion of methods for constructing parity-check matrices.

In any $k \times m$ parity-check matrix there are only $2^m - 1$ different rows that can occur, since the all-zero row never appears in such a matrix. This does not mean that the total number of rows cannot be larger than $2^m - 1$, since the same row may appear more than once. For any given m, it is possible to compute the rows of $P_2, P_3, \cdots, P_m$ that correspond to each possible row of P. Any specific P matrix with m columns can then be tested by selecting the appropriate P_J rows, taking into account any multiple occurrences of rows in the P matrix being tested. This procedure can be stated more precisely in terms of the following definitions.

Definition: P^m is the matrix having m columns (and $2^m - 1$ rows) in which each possible m-bit binary word, except the all-zero word, appears exactly once. The rows are ordered in the following fashion:

First, all the rows containing a single one are written down. These rows are ordered so that, when the rows are interpreted as binary numbers, they occur in decreasing arithmetic order (this means that the first m rows form a unit matrix). Next, the rows containing exactly two ones are written down, with these rows arranged so that they occur in decreasing arithmetic order. This procedure is repeated by writing down the rows with three ones, four ones, etc. until finally the row with m ones is written down. Within each set of rows that all contain the same number of ones, the rows are arranged in decreasing arithmetic order.

Example 4:

$$P^3 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \\ 0 & 1 & 1 \\ 1 & 1 & 1 \end{bmatrix}, \qquad P^4 = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 \\ 0 & 1 & 1 & 0 \\ 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 1 \\ 1 & 1 & 1 & 0 \\ 1 & 1 & 0 & 1 \\ 1 & 0 & 1 & 1 \\ 0 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \end{bmatrix}.$$

By making use of this definition of P^m, a concise specification of any P matrix having m columns can be given by listing which rows of P^m occur in P.

Definition: If P is a parity-check matrix having m columns, then $z_i(P)$, $i = 1, 2, \cdots 2^m - 1$, is equal to the number of times that the ith row of P^m occurs in P. Usually $z_i(P)$ will be written simply as z_i when the appropriate P is clear from the context.

Example 5:

$$P = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 1 & 1 & 1 \\ 1 & 0 & 1 \\ 0 & 1 & 0 \end{bmatrix}, \qquad \begin{aligned} z_1(P) &= 1 \\ z_2(P) &= 2 \\ z_3(P) &= 0\,. \\ z_4(P) &= 0 \\ z_5(P) &= 1 \\ z_6(P) &= 0 \\ z_7(P) &= 1 \end{aligned}$$

It is now possible to state the requirements for parity-check matrices in terms of $z_i(P)$, P^m and P_J^m, where P_J^m is the matrix formed of all sums of J columns of P^m.

Theorem 3: A matrix P with each entry equal to zero or one is a parity-check matrix for a code of minimum distance d if and only if

$$[z_1(P)\ z_2(P)\ \cdots\ z_{2^m-1}(P)]\ [P_J^m] \geqq [d - J, d - J, \cdots, d - J]^* \qquad (2)$$

for $1 \leqq J \leqq m$.

Proof: For $J = 1$,

$$[z_1(P)\ z_2(P)\ \cdots\ z_{2^m-1}(P)]\ [P_1^m]$$

is just equal to the weights of the columns of P, since each row of P^m is multiplied by the number of times it occurs in P $[z_i(P)]$ and then a sum for each column is formed. Similarly, for $J \neq 1$,

$$[z_i(P)\ z_2(P)\ \cdots\ z_{2^m-1}(P)]\ [P_J^m]$$

is equal to the weights of the columns of P_J. By Theorem 1, these weights must be greater than or equal to $d - J$.

* The multiplication here is ordinary matrix multiplication. The inequality is satisfied if, and only if, each element of the row matrix obtained by the multiplication is at least as large as the corresponding element of the row matrix given on the right side of the inequality.

Example 6:

$$P = \begin{bmatrix} 1 & 1 & 0 \\ 1 & 0 & 1 \\ 0 & 1 & 1 \\ 1 & 1 & 1 \end{bmatrix}, \qquad \begin{aligned} z_1 &= 0 \\ z_2 &= 0 \\ z_3 &= 0 \\ z_4 &= 1 \\ z_5 &= 1 \\ z_6 &= 1 \\ z_7 &= 1 \end{aligned},$$

$$[z_1z_2 \cdots z_7]\cdot[P_1^3] = [0001111]\cdot\begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \\ 0 & 1 & 1 \\ 1 & 1 & 1 \end{bmatrix} = [333] \geqq [d-1, d-1, d-1],$$

$$[z_1z_2 \cdots z_7]\cdot[P_2^3] = [0001111]\cdot\begin{bmatrix} 1 & 1 & 0 \\ 1 & 0 & 1 \\ 0 & 1 & 1 \\ 0 & 1 & 1 \\ 1 & 0 & 1 \\ 1 & 1 & 0 \\ 0 & 0 & 0 \end{bmatrix} = [222] \geqq [d-2, d-2, d-2],$$

$$[z_1z_2 \cdots z_7]\cdot[P_3^3] = [0001111]\begin{bmatrix} 1 \\ 1 \\ 1 \\ 0 \\ 0 \\ 0 \\ 1 \end{bmatrix} = [1] \geqq [d - 3].$$

Thus, P is the parity-check matrix for a code of minimum distance 4.

Theorem 3 is merely a restatement of Theorem 2 using different notation. The reason for introducing this new notation is that, by means of Theorem 3, the problem of constructing minimum redundancy codes can be formulated as an integer linear programming problem.

Lemma 3:

$$k = \sum_{i=1}^{2^m-1} z_i .$$

Proof: By definition, k equals the number of rows in the matrix P and z_i equals the number of times that row i of P^m occurs in P. Since each row of P is identical with some row of P^m, the total number of rows of P just equals

$$\sum_{i=1}^{2^m-1} z_i .$$

Theorem 4: The problem of finding a minimum-redundancy, systematic, error-collecting code for specified values of m (the number of message bits) and d (the minimum distance between any pair of code words) is equivalent to the problem of solving the following linear program:

minimize:

$$k = \sum_{i=1}^{2^m-1} z_i$$

subject to:

$$\text{(LP)}$$

(1) z_i integers,

(2) $z_i \geqq 0$,

(3) $[z_1 \, z_2 \cdots z_{2^m-1}] \cdot [P_J{}^m] \geqq [d - J, d - J, \cdots, d - J]$

for $J = 1, 2, \cdots m$.

Proof: The solution to (LP) will be a set of values for $z_1, z_2, \cdots, z_{2^m-1}$. These values can be used to construct a matrix P by interpreting them as $z_i(P)$. By Theorem 3, these values for $z_i(P)$ must satisfy (LP-3) and by the definition of $z_i(P)$ they must satisfy (LP-1, 2). Since a minimum-redundancy code is desired, it is necessary to minimize k. Lemma 3 establishes the expression for k in terms of $z_i(P)$.

The remainder of this paper will consist mainly of obtaining solutions to this linear program.

V. BOUNDS ON REDUNDANCY

In a certain sense, the formulation of (LP) solves the problem of constructing the desired codes, since a numerical procedure exists for solving this type of integer linear program.[5] Practically, this procedure is of limited usefulness, since the size of the program to be solved soon exceeds the capability of the largest electronic computer. Also, numerical solutions do not provide information about the interrelations among various codes with different parameters. A much more desirable solution would be a closed solution of (LP) in which the values of $z_1, z_2, \cdots, z_{2^m-1}$ and k are expressed as functions of m and d. The derivation of such closed solutions and of various properties relating solutions for different parameters will constitute the remainder of this paper.

The first step in obtaining solutions of (LP) will be to remove the matrix notation and express (LP-3) as a set of simultaneous inequalities. This is done to simplify the proofs of the theorems to follow.

The inequalities represented by (LP-3) can be expressed in terms of a single matrix by defining a matrix A^m in which all of the columns of $P_1^m, P_2^m, \cdots, P_m^m$ appear.[6]

Definition: The matrix A^m is formed as follows:

(1) The first m columns of A^m are identical with P^m.

(2) The jth column of A^m $(j > m)$ is formed by taking the sum modulo 2 of the columns of P^m that have one entries in the jth *row* of P^m. When the value of m is clear from the context, A^m will be written as A.

Example 7: For $m = 3$,

$$P^3 = \left[\begin{array}{ccc} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \\ 0 & 1 & 1 \\ 1 & 1 & 1 \end{array}\right]\begin{array}{c} 1 \\ 2 \\ 3 \\ 4 \\ 5 \\ 6 \\ 7 \end{array}, \qquad A^3 = \begin{array}{c} \begin{array}{ccccccc} 1 & 2 & 3 & 4 & 5 & 6 & 7 \end{array} \\ \left[\begin{array}{ccccccc} 1 & 0 & 0 & 1 & 1 & 0 & 1 \\ 0 & 1 & 0 & 1 & 0 & 1 & 1 \\ 0 & 0 & 1 & 0 & 1 & 1 & 1 \\ 1 & 1 & 0 & 0 & 1 & 1 & 0 \\ 1 & 0 & 1 & 1 & 0 & 1 & 0 \\ 0 & 1 & 1 & 1 & 1 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 & 0 & 1 \end{array}\right] \end{array}.$$

The fifth column of A^3 is formed from the sum modulo 2 of the first and third columns of P^3 since the fifth row of P^3 has ones in the first and third columns, etc.

Definition: Let $\epsilon_j(m)$ be defined as follows:

$$\begin{aligned} \epsilon_j(m) &= 1 \qquad \text{for} \quad 1 \leqq j \leqq m, \\ \epsilon_j(m) &= 2 \qquad \text{for} \quad 1 + m \leqq j \leqq m + \binom{m}{2}, \\ \epsilon_j(m) &= s \qquad \text{for} \quad \sum_{\nu=0}^{s-1} \binom{m}{\nu} \leqq j \leqq \sum_{\mu=1}^{s} \binom{m}{\mu}, \\ \epsilon_j(m) &= m \qquad \text{for} \quad \sum_{\nu=0}^{m-1} \binom{m}{\nu} \leqq j \leqq 2^m - 1. \end{aligned}$$

Theorem 3 can be stated in terms of A^m as follows:

Theorem 3': A matrix P is a parity-check matrix for a code of minimum distance d if and only if:

$$[z_1(P)\ z_2(P)\ \cdots\ z_{2^m-1}(P)] \cdot [A^m] \geqq [w_1\ w_2\ \cdots\ w_{2^m-1}], \qquad (2')$$

where $w_j = d - \epsilon_j(m)$.

Some properties of A^m are stated in the following lemmas, in which $a_{ij}{}^m$ represents the element of the ith row and jth column of A^m. When there is no ambiguity possible, $a_{ij}{}^m$ is written simply as a_{ij}. Proofs will be given in the Appendix.

Lemma 4: The matrix A^m can be partitioned into submatrices as follows:

$$A^m = [P_1{}^m \vdots P_2{}^m \vdots P_3{}^m \vdots \cdots \vdots P_m{}^m].$$

Lemma 5:

$$a_{ij} = a_{i1}a_{j1} \oplus a_{i2}a_{j2} \oplus \cdots \oplus a_{im}a_{jm} \qquad \text{for} \quad j > m.$$

Lemma 6: The transpose of A is identical with A: $A^T = A$, or $a_{ji} = a_{ij}$.

Lemma 7:

$$\sum_{i=1}^{2^m-1} a_{ij}{}^m = \sum_{j=1}^{2^m-1} a_{ij}{}^m = 2^{m-1}.$$

Lemma 8:

$$\sum_{i=1}^{m} a_{ij}{}^m = \epsilon_j(m).$$

Lemma 9: The inverse of A, A^{-1}, is obtained from A by replacing each one entry of A by 2^{1-m} and replacing each zero entry by -2^{1-m}:

$$a_{ij}{}^{-1} = \quad 2^{1-m} \quad \text{if} \quad a_{ij} = 1, \qquad \text{and}$$
$$a_{ij}{}^{-1} = -2^{1-m} \quad \text{if} \quad a_{ij} = 0.$$

Theorem 3 can be stated directly in terms of the $a_{ij}{}^m$ as follows:

Theorem 3″: A matrix P is a parity-check matrix for a code of minimum distance d if and only if:

$$\sum_{i=1}^{2^m-1} a_{ij}z_i \geqq w_j\,, \qquad (2'')$$

where $w_j = d - \epsilon_j(m)$.

The corresponding formulation for the program (LP) is

minimize:

$$k = \sum_{i=1}^{2^m-1} z_i$$

subject to:

$$\text{(1) } z_i \text{ integers,} \qquad \text{(LP}'\text{)}$$

$$\text{(2) } z_i \geqq 0,$$

$$\text{(3) } \sum_{i=1}^{2^m-1} a_{ij}z_i \geqq w_j\,,$$

where $w_j = d - \epsilon_j(m)$.

The following theorem presents a lower bound on k, the number of check bits that are required for specified values of m, the number of information bits and d, the minimum distance between code words.

Theorem 5: For an error-correcting code having minimum distance d and m message bits, the number of check bits, k, must satisfy

$$k \geqq \left(\frac{2^m - 1}{2^{m-1}}\right) d - m. \tag{3}$$

Proof: By $(2'')$,

$$\sum_{i=1}^{2^m-1} a_{ij}z_i \geqq w_j\,,$$

$$\sum_{j=1}^{2^m-1} \sum_{i=1}^{2^m-1} a_{ij}z_i \geqq \sum_{j=1}^{2^m-1} w_j\,,$$

$$\sum_{i=1}^{2^m-1} \sum_{j=1}^{2^m-1} a_{ij}z_i \geqq \sum_{j=1}^{2^m-1} w_j\,,$$

$$\sum_{i=1}^{2^m-1} z_i \sum_{j=1}^{2^m-1} a_{ij} \geqq \sum_{j=1}^{2^m-1} w_j\,.$$

But, by Lemma 7,

$$\sum_{j=1}^{2^m-1} a_{ij} = 2^{m-1},$$

so

$$2^{m-1} \sum_{i=1}^{2^m-1} z_i \geqq \sum_{j=1}^{2^m-1} w_j\,,$$

and

$$k = \sum_{i=1}^{2^m-1} z_i \geqq 2^{1-m} \sum_{j=1}^{2^m-1} w_j\,, \quad \text{by Lemma 3.} \tag{i}$$

By the definition of w_j,

$$\sum_{j=1}^{2^m-1} w_j = \sum_{s=1}^{m} \binom{m}{s} d - \binom{m}{s} s = (2^m - 1)\, d - \sum_{s=1}^{m} \binom{m}{s} s \tag{ii}$$

but

$$\sum_{s=1}^{m} \binom{m}{s} s = m2^{m-1},$$

so

$$\sum_{j=1}^{2^m-1} w_j = (2^m - 1)d - m2^{m-1}.$$

Substituting this in (i) yields

$$k \geqq 2^{1-m}[(2^m - 1)\, d - m2^{m-1}]$$

or

$$k \geqq \left(\frac{2^m - 1}{2^{m-1}}\right) d - m.$$

Since the total number of bits in each code word, n, is just equal to $m + k$, this bound on k yields a bound on n.

Corollary 2: For an error-correcting code having minimum distance d and m message bits, the total number of bits in each code word must satisfy:

$$n \geqq \left(\frac{2^m - 1}{2^{m-1}}\right) d. \tag{4}$$

If $d < m$, the bounds given in (3) and (4) can be improved, since some of the w_j in (2″) will be negative and should be replaced by zeros.

Corollary 3: When $d < m$, k and n must satisfy the following inequalities:

$$k \geqq \left(\frac{2^m - 1}{2^{m-1}}\right) d - m + 2^{1-m} \sum_{s=d+1}^{m} \binom{m}{s} (s - d), \tag{3′}$$

$$n \geqq \left(\frac{2^m - 1}{2^{m-1}}\right) d + 2^{1-m} \sum_{s=d+1}^{m} \binom{m}{s} (s - d) \tag{4′}$$

Proof: From (ii) of Theorem 5,

$$\sum_{j=1}^{2^m-1} w_j = \sum_{s=1}^{m} \binom{m}{s} (d - s),$$

but when $w_j < 0$ it can be replaced by 0. This is equivalent to defining $w_j' = d - \epsilon_j(m)$ for $d \geqq \epsilon_j(m)$ and $w_j' = 0$ for $d < \epsilon_j(m)$. Then,

$$\sum_{j=1}^{2^m-1} w_j' = \sum_{s=1}^{d} \binom{m}{s} (d - s) \qquad \text{for } d < m,$$

or

$$\sum_{j=1}^{2^m-1} w_j' = \sum_{s=1}^{m} \binom{m}{s} (d - s) + \sum_{s=d+1}^{m} \binom{m}{s} (s - d) \qquad \text{for } d < m$$

$$= (2^m - 1)\, d - m2^{m-1} + \sum_{s=d+1}^{m} \binom{m}{s} (s - d).$$

By (i) of Theorem 5,

$$\begin{aligned} k &\geqq 2^{1-m} \sum_{j=1}^{2^m-1} w_j \\ &\geqq 2^{1-m} \sum_{j=1}^{2^m-1} w_j' \\ &\geqq 2^{1-m} \left[(2^m - 1)\, d - m2^{m-1} + \sum_{s=d+1}^{m} \binom{m}{s} (s - d) \right] \\ &\geqq \left(\frac{2^m - 1}{2^{m-1}} \right) d - m + 2^{1-m} \sum_{s=d+1}^{m} \binom{m}{s} (s - d). \end{aligned}$$

Whenever $d \neq h2^{m-1}$, the bounds (3) and (4) are not integers and therefore cannot be met exactly.

Definition: Let $\{N\}$ equal N if N is an integer and equal the smallest integer larger than N if N is not an integer.

Definition: Let

$$k^*(m,d) = \left\{ \left(\frac{2^m - 1}{2^{m-1}} \right) d - m \right\}$$

and

$$n^*(m,d) = \left\{ \left(\frac{2^m - 1}{2^{m-1}} \right) d \right\}.$$

Since the total number of bits per code word and the number of check bits must both be integers, the following inequalities follow directly from Theorem 5 and Corollary 2.

Corollary 4: For an error-correcting code of minimum distance d and having m message bits:

$$k \geqq k^*$$

and

$$n \geqq n^*$$

VI. MINIMUM-REDUNDANCY CODES

Since n and k must be integers, the bounds given by (2) and (3) can be met exactly only when d is divisible by 2^{m-1}, that is, when d can be written as $d = h2^{m-1}$, where h is a positive integer. The following theorem shows that the bounds can always be achieved in these cases. The appropriate P matrix is formed by including each possible distinct row h times, except for rows of weight one, which are included only $h - 1$ times.

Theorem 6: Whenever $d = h2^{m-1}$, where h is any positive integer, a minimum-redundancy code exists with

$$k = h(2^m - 1) - m,$$

$$n = h(2^m - 1)$$

and

$$\begin{aligned} z_i &= h - 1 \qquad && \text{for } 1 \leqq i \leqq m \\ &= h && \text{for } m + 1 \leqq i \leqq 2^m - 1. \end{aligned}$$

Proof: Let $z_i = h - 1$ for $1 \leqq i \leqq m$ and $z_i = h$ for $m + 1 \leqq i \leqq 2^m - 1$. Then,

$$\sum_{i=1}^{2^m-1} a_{ij}z_i = \sum_{i=1}^{2^m-1} ha_{ij} - \sum_{i=1}^{m} a_{ij},$$

$$\sum_{i=1}^{2^m-1} ha_{ij} = h \sum_{i=1}^{2^m-1} a_{ij} = h2^{m-1} \quad \text{by Lemma 7},$$

$$\sum_{i=1}^{m} a_{ij} = \epsilon_j(m) \quad \text{by Lemma 8}.$$

Thus,

$$\sum_{i=1}^{2^m-1} a_{ij}z_i = h2^{m-1} - \epsilon_j(m).$$

But, by Theorem 3″, this is exactly the condition for a code of minimum distance $h2^{m-1}$.

The number of check bits, k, is given by

$$k = \sum_{i=1}^{2^m-1} z_i = \sum_{i=1}^{2^m-1} h - \sum_{i=1}^{m} (1)$$

$$= h(2^m - 1) - m.$$

By Theorem 5,

$$k \geqq \left(\frac{2^m - 1}{2^{m-1}}\right)(h2^{m-1}) - m$$

$$\geqq (2^m - 1)\,h - m.$$

Therefore, the code is a minimum-redundancy code.

Example 8: For $m = 3$, $d = 8 = h \cdot 2^{m-1} = 2 \cdot 2^{3-1}$:

$$k = h(2^m - 1) - m = 2(2^3 - 1) - 3 = 11,$$

$$P = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \\ 0 & 1 & 1 \\ 0 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \end{bmatrix}$$

and

$$z_1 = z_2 = z_3 = 1,$$

$$z_4 = z_5 = z_6 = z_7 = 2.$$

Theorem 5 can be extended to the case when $d = h2^{m-1} - 1$ by means of the following theorem, which has originally proved by Hamming.[2]

Theorem 7 (Hamming): From any minimum-redundancy code* containing n bits per code word and having minimum distance d, with d an even number, it is possible to obtain a minimum-redundancy code containing $n - 1$ bits per code word and having minimum distance $d - 1$ by removing one of the bits from each of the code words (the same bit must be removed from each word). If the original code was a systematic code, the bit removed should be one of the check bits.

Conversely, from any minimum-redundancy code† containing n bits per code word and having minimum distance d, with d an odd number, it is possible to obtain a minimum-redundancy code with $n + 1$ bits per code word and having minimum distance $d + 1$. This is done by

* Not necessarily a systematic code.

adding a check bit that is a parity check over all of the bits of each code word.

Corollary 5: If $z_1, z_2, \cdots, z_{2^m-1}$ specify a minimum-redundancy systematic code for d, where d is an even number, it is possible to obtain a minimum-redundancy code for $d - 1$ by decreasing any nonzero z_i by one.

Corollary 6: Whenever $d = h2^{m-1} - 1$, where h is any positive integer, a minimum-redundancy code exists with

$$k = h(2^m - 1) - (m - 1),$$

$$n = h(2^m - 1) - 1.$$

This follows directly from Theorem 7 and Corollary 5.

There is a large class of codes for which

$$\left(\frac{2^m - 1}{2^{m-1}}\right) d$$

is not an integer, but for which minimum-redundancy codes with $k = k^*$ can be derived.

Theorem 8: Whenever $d = h_1 2^{m-1} - 2^{h_2}$, where h_1 is a positive integer and h_2 is a positive integer with $h_2 < m - 1$, there exists a minimum-redundancy code with

$$k = h_1(2^m - 1) - 2^{h_2+1} - m + 1,$$
$$n = h_1(2^m - 1) - 2^{h_2+1} + 1$$

and

$$z_i = z_i' - z_i'',$$

where

$$\begin{aligned} z_i' &= h_1 - 1 && \text{for } 1 \leqq i \leqq m \\ &= h_1 && \text{for } m + 1 \leqq i \leqq 2^m - 1, \end{aligned}$$

and

$$\begin{aligned} z_i'' &= 1 \text{ if the corresponding row of } A^m \text{ has all zeros in its} \\ &\qquad \text{first } m - h_2 - 1 \text{ columns} \\ &= 0 \text{ if the corresponding row of } A^m \text{ does not have} \\ &\qquad \text{all zeros in its first } m - h_2 - 1 \text{ columns.} \end{aligned}$$

Proof: Let z_i, z_i' and z_i'' be defined as in the statement of the theorem.

By the methods used in proving Theorem 6, it follows that

$$\sum_{i=1}^{2^m-1} a_{ij}z_i' = h_1 2^{m-1} - \epsilon_j(m).$$

Consider

$$\sum_{i=1}^{2^m-1} a_{ij}z_i''.$$

This is equal to

$$\sum_{i=1}^{2^{h_2+1}-1} b_{ij},$$

where the b_{ij} are the entries of a matrix B, which is made up of those rows of A^m for which the corresponding z_i'' are equal to one. The first $m - h_2 - 1$ columns of B contain all zeros. Therefore,

$$\sum_{i=1}^{2^{h_2+1}-1} b_{ij} = 0 \qquad \text{for } 1 \leqq j \leqq m - h_2 - 1.$$

The next $h_2 + 1$ columns of B are identical with P^{h_2+1}, since each combination of zeros and ones (except the all-zero combination) occurs exactly once. Thus, since the weight of each column of P^{h_2+1} is 2^{h_2},

$$\sum_{i=1}^{2^{h_2+1}-1} b_{ij} = 2^{h_2} \qquad \text{for } m - h_2 \leqq j \leqq m.$$

Every other column of B is formed from the sum modulo 2 of some of the first m columns. Since the all-zero columns do not have any effect on the sum modulo 2 operation, every other column of B is equal either to one of the columns for $m - h_2 \leqq j \leqq m$ or to the sum modulo 2 of several of these columns. Thus, every remaining column of B is identical with some column of A^{h_2+1}. Therefore, for $m + 1 \leqq j \leqq 2^m - 1$,

$$\sum_{i=1}^{2^{h_2+1}-1} b_{ij} = \sum_{i=1}^{2^{h_2+1}-1} a_{ij}^{h_2+1} = 2^{h_2}.$$

Thus,

$$\begin{aligned}\sum_{i=1}^{2^m-1} a_{ij}z_i'' = \sum_{i=1}^{2^{h_2+1}-1} b_{ij} &= 0 \quad \text{for } 1 \leqq j \leqq m - h_2 - 1 \\ &= 2^{h_2} \text{ for } m - h_2 \leqq j \leqq 2^m - 1,\end{aligned}$$

and

$$\begin{aligned}\sum_{i=1}^{2^m-1} a_{ij}z_i &= h_1 2^{m-1} - \epsilon_j(m) - 0 \quad \text{for } 1 \leqq j \leqq m - h_2 - 1 \\ &= h_1 2^{m-1} - \epsilon_j(m) - 2^{h_2} \text{ for } m - h_2 \leqq j \leqq 2^m - 1.\end{aligned}$$

Thus,

$$\sum_{i=1}^{2^m-1} a_{ij}z_i \geqq h_1 2^{m-1} - 2^{h_2} - \epsilon_j(m),$$

or the given z_i satisfy the requirements for a code with $d = h_1 2^{m-1} - 2^{h_2}$. This proves that a code constructed from the given z_i will have $d = h_1 2^{m-1} - 2^{h_2}$. A proof must now be given for the fact that the resulting code is minimum-redundancy

$$k = \sum_{i=1}^{2^m-1} z_i = \sum_{i=1}^{2^m-1} z_i' - \sum_{i=1}^{2^m-1} z_i''.$$

The z_i' are the same as the z_i of Theorem 6; therefore,

$$\sum_{i=1}^{2^m-1} z_i' = h_1(2^m - 1) - m.$$

Since there are $2^{h_2+1} - 1$ rows of A^m that have all zeros in the first $m - h_2 - 1$ columns,

$$\sum_{i=1}^{2^m-1} z_i'' = 2^{h_2+1} - 1$$

and

$$k = h_1(2^m - 1) - m - 2^{h_2+1} + 1.$$

Now,

$$k^* (m,h_1 2^{m-1} - 2^{h_2}) = \left\{\left(\frac{2^m - 1}{2^{m-1}}\right)(h_1 2^{m-1} - 2^{h_2}) - m\right\},$$

which can be rewritten as

$$k^* = \{(2^m - 1)h_1 - 2^{h_2+1} + 2^{1+h_2-m} - m\},$$

but, since $m > 1 + h_2$,

$$2^{1+h_2-m} < 1,$$

so that

$$k^* = (2^m - 1)h_1 - 2^{h_2+1} - m + 1,$$

and therefore $k = k^*$ and the code is minimum-redundancy.

Corollary 7: Whenever $d = h_1 2^{m-1} - 2^{h_2} - 1$, where h_1 is a positive integer and h_2 is a positive integer with $h_2 < m - 1$, a minimum-redundancy code can be obtained from the code of Theorem 8 by the method given in Corollary 5.

Minimum-redundancy codes for $d = 2$, 3 and 4 were given in a paper by Hamming.[2] A code for $d = 2$ can be obtained by using all n-bit words that contain an even number of ones, since Theorem 2 just requires each column of P to contain at least one one. Thus, the minimum-redundancy codes for $d = 2$ all have $k = 1$.

For codes of distance 3, Theorem 2 requires that each column of P contain at least two ones and that no two columns be identical. In this case, m is equal to the number of different columns having k rows, and at least two entries equal to 1; or $m = 2^k - k - 1$.

VII. RELATIONSHIPS AMONG CODES

For values of d which are greater than 4 and do not satisfy the conditions of either Theorem 6 or Theorem 8, it has not been possible to obtain closed solutions of the linear program (LP). Computation using Gomory's algorithm[5] is necessary to obtain minimum-redundancy codes for these values of d. The following theorems present various general properties of minimum-redundancy codes that are useful in obtaining codes for new values of d from the codes obtained by use of the algorithm.

Definition: Let $K(m,d)$ be the minimum value of k that is possible for a code having m message bits and minimum distance d.

Definition: Let $N(m,d)$ be the minimum value of n that is possible for a code having m message bits and minimum distance d.

Lemma 10:

$$K(m-1,d) \leqq K(m,d).$$

Proof: A parity-check matrix for $m - 1$ can be obtained from the matrix for m by simply removing one column. Since the conditions of Theorem 2 must be satisfied by the reduced matrix if they were satisfied by the original matrix, the reduced matrix corresponds to a code of distance d if the original matrix corresponded to a code of distance d.

Theorem 9: For $m \leqq d < 2^{m-2}$,

$$N(m,d) > n^*(m,d),$$

$$K(m,d) > k^*(m,d).$$

Proof:

$$k^*(m,d) = \left\{\left(\frac{2^m - 1}{2^{m-1}}\right) d - m\right\} \quad \text{for } m \leqq d$$

$$= \{(2 - 2^{1-m})\, d - m\}$$

$$= \left\{2d - \frac{d}{2^{m-1}} - m\right\}$$

$$k^*(m,d) = 2d - m \qquad \text{for } m \leqq d < 2^{m-1},$$

$$k^*(m-1,d) = 2d - m + 1 \qquad \text{for } m - 1 \leqq d < 2^{m-2},$$

but

$$K(m,d) \geqq K(m-1,d),$$
$$K(m-1,d) \geqq k^*(m-1,d),$$
$$k^*(m-1,d) > k^*(m,d) \qquad \text{for } m \leqq d < 2^{m-2},$$
$$K(m,d) \geqq K(m-1,d) \geqq k^*(m-1,d) > k^*(m,d),$$

so that

$$K(m,d) > k^*(m,d) \qquad \text{for } m \leqq d < 2^{m-2}.$$

Since $n = k + m$, it follows that

$$N(m,d) > n^*(m,d) \qquad \text{for } m \leqq d < 2^{m-2}.$$

Theorem 10:

$$N(m,d_1 + d_2) \leqq N(m,d_1) + N(m,d_2),$$
$$K(m,d_1 + d_2) \leqq K(m,d_1) + K(m,d_2) + m.$$

Proof: Let $z' = [z_1' \; z_2' \; \cdots \; z_{2^m-1}']$ be the values of z_i corresponding to $N(m,d_1)$ and $z'' = [z_1'' \; z_2'' \; \cdots \; z_{2^m-1}'']$ be the values of z_i corresponding to $N(m,d_2)$. Thus,

$$\sum_{i=1}^{2^m-1} a_{ij}^m z_i' \geqq d_1 - \epsilon_j(m)$$

and

$$\sum_{i=1}^{2^m-1} a_{ij}^m z_i'' \geqq d_2 - \epsilon_j(m).$$

Let

$$\hat{z}_i = z_i' + z_i'' + 1 \qquad \text{for } 1 \leqq i \leqq m$$

and

$$\hat{z}_i = z_i' + z_i'' \qquad \text{for } m + 1 \leqq i \leqq 2^m - 1.$$

Then,

$$\begin{aligned} \sum_{i=1}^{2^m-1} a_{ij}^m \hat{z}_i &= \sum_{i=1}^{2^m-1} a_{ij}^m z_i' + \sum_{i=1}^{2^m-1} a_{ij}^m z_i'' + \sum_{i=1}^{m} a_{ij}^m \\ &\geqq d_1 - \epsilon_j(m) + d_2 - \epsilon_j(m) + \epsilon_j(m) \\ &\geqq d_1 + d_2 - \epsilon_j(m). \end{aligned}$$

Thus, $\hat{z}_i$ satisfy the conditions for a code of distance $d_1 + d_2$. Furthermore,

$$\sum_{i=1}^{2^m-1} \hat{z}_i = \sum_{i=1}^{2^m-1} z_i' + \sum_{i=1}^{2^m-1} z_i'' + \sum_{i=1}^{m} (1),$$

$$k(m,d_1+d_2) = \sum_{i=1}^{2^m-1} \hat{z}_i = K(m,d_1) + K(m,d_2) + m,$$

so that

$$K(m,d_1 + d_2) \leqq K(m,d_1) + K(m,d_2) + m.$$

Corollary 8:

$$N(m,d_1 + h2^{m-1}) \leqq N(m,d_1) + h(2^m - 1),$$

$$K(m,d_1 + h2^{m-1}) \leqq K(m,d_1) + h(2^m - 1).$$

Definition: Let $\max(z_1, z_2, \cdots, z_m)$ equal the largest of the values of the z_i.

Theorem 11: Let $z_1, z_2, \cdots, z_{2^m-1}$ correspond to a code for which $k = K(m,d)$. Then, if $\max(z_1, \cdots, z_m) = M$,

$$K(m - 1,d) \leqq K(m,d) - M.$$

Proof. Let z_I, $(1 \leqq I \leqq m)$ be one of the z_i such that $z_I = M$. Then, if the Ith column is removed from the matrix P specified by $(z_1, z_2, \cdots, z_{2^m-1})$, the resulting matrix must still correspond to a code of minimum distance d (Lemma 10). However, M of the rows of the reduced matrix consist of all zeros, since there are M rows which contain a one only in column I. Thus, these M rows can be removed without affecting the minimum distance d. Removal of M rows decreases k by M, giving

$$K(m - 1,d) \leqq K(m,d) - M.$$

VIII. COMPARISON WITH PLOTKIN'S BOUND

The approach of this paper has been to search for codes which require the minimum number of check bits, k, for specified values of m and d. Another common approach to the study of error-correcting codes is to specify the total number of bits per code word, n, and the minimum distance, d, and then to try to construct codes which contain the largest number of messages. A bound on this maximum number of messages has been proved by Plotkin.[7]

Theorem 12 (Plotkin): Let $A(n,d)$ equal the maximum number of binary n-bit words in an error-correcting code (not necessarily a systematic code), for which the distance between any two code words is at least equal to d. Then

$$A(n,d) \leqq \frac{2d}{2d - n} \qquad \text{for} \quad 2d > n.$$

For a systematic code, the number of messages must be a power of two, and therefore this bound can be met exactly only when $2d/(2d - n)$ is a power of two. The following theorem shows that, whenever this is true, a systematic code exists which does meet the bound.

Theorem 13: For values of n and d such that $2d/(2d - n) = 2^m$, for some m, a systematic code exists with m message bits and therefore 2^m code words. For such values of n and d, no code of any type is possible with more code words.

Proof: The equation $2d/(2d - n) = 2^m$ can be written as

$$d = \left(\frac{n}{2^m - 1}\right) 2^{m-1}.$$

Since d and 2^{m-1} are integers, and $2^m - 1$ does not divide 2^{m-1}, $n/(2^m - 1)$ must be an integer. Let $h = n/(2^m - 1)$, then $d = h2^{m-1}$. By Theorem 6, a code exists with $d = h2^{m-1}$ and

$$n = h(2^m - 1) = \left(\frac{n}{2^m - 1}\right)(2^m - 1) = n$$

and m message bits.

By Plotkin's theorem, no code with more code words is possible.

IX. ACKNOWLEDGMENT

The author wishes to acknowledge the many helpful comments contributed by D. H. Evans, S. H. Unger and M. C. Paull. Thanks are due also to W. Ulrich and C. Y. Lee for many stimulating discussions on the general subject of error-correcting codes.

APPENDIX

Various proofs have been omitted in order not to disturb the continuity of the paper. These proofs will be presented here.

Lemma 4: The matrix A^m can be partititioned into submatrices as follows:

$$A^m = [P_1^m \vdots P_2^m \vdots P_3^m \vdots \cdots \vdots P_m^m].$$

Proof: By definition, the first m columns of A^m are identical with P_1^m.

The rows of P^m are ordered so that all rows with one one come first, then all rows with two ones, etc. The columns of P^m which must be summed to form the jth column of A^m are determined by the jth row of P^m. Because of the ordering of the rows of P^m, all sums of pairs of columns of P^m will occur as columns of A^m, then all sums of three columns of P^m, etc. Since P_J^m contains all sums of J columns of P^m, A^m can be partitioned as shown.

Lemma 5: $a_{ij} = a_{i1}a_{j1} \oplus a_{i2}a_{j2} \oplus \cdots \oplus a_{im}a_{jm}$, for $j > m$.
Proof: The jth column of $A^m (j > m)$ is formed by taking the sum modulo 2 of the first m columns of A^m which have a one entry in the jth *row* of A^m. Thus, if the jth column of A^m is denoted by A_j^m,

$$A_j^m = a_{j1}A_1^m \oplus a_{j2}A_2^m \oplus \cdots \oplus a_{jm}A_m^m \qquad \text{for } j > m,$$

since the column A_1^m is to enter into the sum only if $a_{j1} = 1$, etc. It follows from this that the ith element of the jth column $(j > m)$ is given by

$$a_{ij} = a_{i1}a_{j1} \oplus a_{i2}a_{j2} \oplus \cdots \oplus a_{im}a_{jm} .$$

Lemma 6:

$$a_{ij} = a_{ji} .$$

Proof: It follows directly from Lemma 5 that $a_{ij} = a_{ji}$ for $j > m$ or $i > m$. For $j < m$ and $i < m$, the definition of A requires that $a_{ij} = 0$ unless $i = j$, so that $a_{ij} = a_{ji} = 0$ for $i \neq j$ and, for $i = j$, a_{ij} is identical with a_{ji}.

Lemma 7:

$$\sum_{i=1}^{2m-1} a_{ij}^m = \sum_{j=1}^{2m-1} a_{ij}^m = 2^{m-1}.$$

Proof: The first m columns of A^m contain each m-bit binary number, except the all-zero number, exactly once. Consider these rows which have a one entry in the first column. There must be 2^{m-1} such rows, since there are 2^{m-1} different $(m - 1)$-bit binary numbers, and each of these must occur once in the remaining $m - 1$ columns. Thus,

$$\sum_{i=1}^{2m-1} a_{i1}^m = 2^{m-1}.$$

A similar argument shows that

$$\sum_{i=1}^{2m-1} a_{ij}^m = 2^{m-1} \qquad \text{for } 1 \leqq j \leqq m.$$

For $j > m$, each entry a_{ij}^m is the sum modulo 2 of the entries in s of the first m columns. Consider these columns. Each s-bit binary number must occur exactly 2^{m-s} times (except the all-zero number, which occurs $2^{m-s} - 1$ times), since there are exactly 2^{m-s} ways to choose the entries in the remaining $m - s$ columns. There are 2^s different s-bit numbers and 2^{s-1} of these contain an odd number of one entries. Thus, there are $(2^{m-s})\ (2^{s-1}) = 2^{m-1}$ rows containing an odd number of one entries, and hence 2^{m-1} of the a_{ij}^m are equal to one. Thus,

$$\sum_{i=1}^{2m-1} a_{ij} = 2^{m-1} \qquad \text{for } j > m.$$

Since $a_{ij} = a_{ji}$, it follows that

$$\sum_{j=1}^{2m-1} a_{ij} = 2^{m-1}.$$

Lemma 8:

$$\sum_{i=1}^{m} a_{ij}^m = \epsilon_j(m).$$

Proof: The definition of $\epsilon_j(m)$ is:

$$\begin{aligned} \epsilon_j(m) &= 1 && \text{for} \quad 1 \leqq j \leqq m, \\ \epsilon_j(m) &= 2 && \text{for} \quad 1 + m \leqq j \leqq m + \binom{m}{2}, \\ &\vdots \\ \epsilon_j(m) &= s && \text{for} \quad \sum_{\nu=0}^{s-1} \binom{m}{\nu} \leqq j \leqq \sum_{\mu=1}^{s} \binom{m}{\mu}. \end{aligned}$$

Consider the first m columns of A^m. The first m rows contain a single one, since they are all the m-bit numbers containing one one. The next $\binom{m}{2}$ rows contain two ones, the next $\binom{m}{3}$ rows contain three ones, etc. Thus,

$$\sum_{j=1}^{m} a_{ij}^m = \epsilon_i(m).$$

Since $a_{ij} = a_{ji}$, it follows from this that

$$\sum_{i=1}^{m} a_{ij}^m = \epsilon_j(m).$$

Lemma 9: If the elements of A^{-1} are represented by a_{ij}^{-1}, then

$$a_{ij}^{-1} = 2^{1-m} \qquad \text{if } a_{ij} = 1$$

and

$$a_{ij}^{-1} = -2^{1-m} \qquad \text{if } a_{ij} = 0.$$

Proof: The a_{ij}^{-1} given above are the elements of the inverse of A if and only if

$$b_{ij} = \sum_{s=1}^{2m-1} a_{is}a_{sj}^{-1} = 1 \qquad \text{if } i = j$$

and

$$b_{ij} = \sum_{s=1}^{2m-1} a_{is}a_{sj}^{-1} = 0 \qquad \text{if } i \neq j.$$

If $i = j$, then

$$\sum_{s=1}^{2m-1} a_{is}a_{sj}^{-1} = \sum_{s=1}^{2m-1} a_{is}a_{si}^{-1} = \sum_{s=1}^{2m-1} a_{si}a_{si}^{-1},$$

which equals

$$2^{1-m} \sum_{s=1}^{2m-1} a_{si}.$$

But

$$\sum_{s=1}^{2m-1} a_{si} = 2^{m-1} \quad \text{by Lemma 7,}$$

so that

$$\sum_{s=1}^{2m-1} a_{is}a_{si}^{-1} = 1.$$

If $i \neq j$,

$$b_{ij} = \sum_{s=1}^{2m-1} a_{is}a_{sj}^{-1} = \sum_{s=1}^{2m-1} a_{si}a_{sj}^{-1}.$$

Three cases will be considered:

Case 1: $i < m$ and $j < m$.

In this case, $b_{ij} = 2^{1-m}$ (number of 11 entries in columns i and j) minus 2^{1-m} (number of 10 entries in columns i and j). By the argument used in proving Lemma 7, there are 2^{m-2} 11 entries and 2^{m-2} 10 entries, so that $b_{ij} = (2^{1-m})(2^{m-2}) - (2^{1-m})(2^{m-2}) = 0$.

Case 2: $i < m$ and $j > m$.

In this case, the elements a_{sj} are formed as the sum modulo 2 of the entries in ν of the first m columns of A, so that $b_{ij} = 2^{1-m} N_0 - 2^{1-m} N_e$, where

$N_0 =$ the number of rows of A which have a 1 in the ith column and an odd number of 1's in the ν columns from which a_{sj} is formed.

$N_e =$ The number of rows of A which have a 1 in the ith column and an even number of 1's in the ν columns from which a_{sj} is formed.

Again, by the argument of Lemma 7, in the $\nu + 1$ columns consisting of column i and the ν columns used to form a_{sj}, each different binary number (except the all-zero number) must occur exactly $2^{m-\nu-1}$ times so that $N_e = N_0$ and $b_{ij} = 0$.

Case 3:

A similar argument shows that $b_{ij} = 0$ for the case when $1 > m$ and $j > m$.

REFERENCES

1. Shannon, C. E., A Mathematical Theory of Communication, B.S.T.J., **27,** July 1948, p. 279; October 1948, p. 623.
2. Hamming, R. W., Error Detecting and Error Correcting Codes, B.S.T.J., **29,** January 1950, p. 147.
3. Reed, I. S., A Class of Multiple-Error-Correcting Codes and the Decoding Scheme, Trans. I.R.E., **PGIT-4,** 1954, p. 38.
4. Muller, D. E., Metric Properties of Boolean Algebra and Their Application to Switching Circuits, Report No. 46, Digital Computer Lab., Univ. of Illinois, April 1953.
5. Gomory, R. E., Outline of an Algorithm for Integer Solutions to Linear Programs, Office of Ordnance Research Symposium on Combinational Problems, New York, April 1958.
6. Slepian, D., A Class of Binary Signalling Alphabets, B.S.T.J., **35,** January 1956, p. 203.
7. Plotkin, M., Binary Codes with Specified Minimum Distance, Research Division Report 51-20, Moore School of Electrical Engineering, Univ. of Pennsylvania, January 1951.

The Analysis of Valve-Controlled Hydraulic Servomechanisms

By R. G. RAUSCH

(Manuscript received July 10, 1959)

The nonlinear equations that represent the behavior of valve-controlled hydraulic servomechanisms are derived, and the assumptions necessary for their linearization are discussed. Solutions of the nonlinear equations obtained by analog computation are compared with solutions of the linear equations. Attention is directed to the influence of the hydraulic parameters on the nonlinear closed-loop system behavior.

I. INTRODUCTION

Since the development of hydraulic control valves such as that employed in the Nike missile,[1] emphasis has been given to the analysis of hydraulic phenomena in valved systems,[2,3,4] with much of the literature having been devoted to hydraulic component design. In this paper, the nonlinear closed-loop performance is given major emphasis; the effect of hydraulic parameter variations on the closed-loop frequency and transient responses is examined by linear and nonlinear methods.

The basic servomechanism under consideration in this study, as shown in Fig. 1, consists of a summing device, an amplifier, a flow source and control system, a hydraulic actuator (or motor) and a load. In the following sections, the nonlinear differential equations which represent the behavior of this closed-loop system are derived, a linear and an incremental-linear representation are discussed and solutions of the nonlinear equations obtained by analog computation are compared with linear solutions.

II. MOTOR AND LOAD ANALYSIS

In this section, emphasis is on the derivation and validity of the equations used to represent the behavior of the actuator and load; the mechanization of the flow source and its method of control will be discussed in detail in the following section.

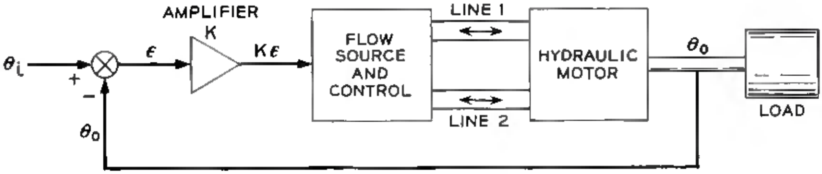


Fig. 1 — Basic hydraulic positional servomechanism.

In Fig. 2, a piston-type actuator is shown connected hydraulically to a flow source and mechanically to a load. The initial step in the analysis is to relate the flows Q_1 and Q_2 to the dependent variable x, the piston displacement from the center position.

The instantaneous volumes between the piston and two arbitrary sections in the lines leading to the cylinder are designated V_{T1} and V_{T2} (in cubic inches), the numerical subscripts indicating a particular side of the piston. Q_1 and Q_2 (in cubic inches per second) represent the flows from the source and depend upon the source mechanization; they are considered positive in the directions indicated in the diagram. Q_L, the leakage flow past the piston, is also shown in the assumed positive direction.

A control volume is chosen so that it coincides with the volume V_{T1} (where V_{T1} is a function of time) and the equation for the conservation of mass flow is written for this volume. This relationship states that the rate of mass accumulation in the control volume is equal to the net rate of mass flow into the volume. The net rate of mass flow into V_{T1} is given by

$$\text{net rate of mass flow into } V_{T1} = \rho(Q_1 - Q_L), \tag{1}$$

where ρ is the mass density (in pound-seconds^2 per inch^4) of the fluid and Q_L is the flow out of V_{T1}.

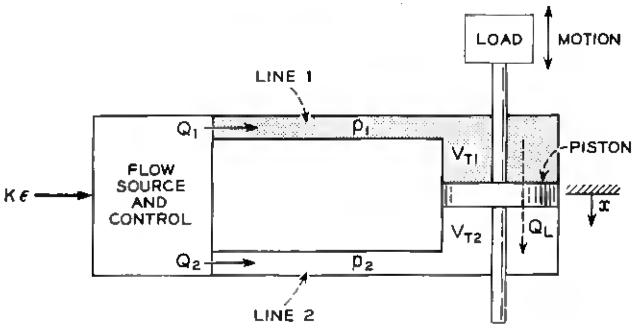


Fig. 2 — Variable displacement actuator.

In writing (1), it is assumed that the mass density, ρ, is uniform throughout the control volume; i.e., it is assumed that ρ is a function of time only, and not a function of position in the volume. Since ρ is dependent upon the instantaneous pressure, p_1, existing in V_{T1}, this assumes that p_1 is uniform throughout V_{T1}. The justification for this assumption is based on the calculation of the velocity of propagation of a longitudinal compression wave in the fluid. In general, the velocity of propagation, for adiabatic conditions, is given by

$$v_p = \sqrt{\frac{\beta}{\rho}}, \tag{2}$$

where β is the adiabatic bulk modulus of compression of the fluid. For a typical oil, v_p is approximately 50,000 inches per second. If the largest linear dimension of the volume V_{T1} is small, the pressure wave will complete many cycles in a short time, and the perturbation will be rapidly attenuated. For valve-controlled high-performance systems, the volumes are small; under the assumption that the largest dimension is one inch, the time of travel is 0.02 millisecond. Since this value is small compared to system time constants of the order of five milliseconds or more, the pressure can be assumed uniform throughout the volume. A calculation of the frequency of oscillation at which nonuniform pressure distribution becomes important shows that it is much higher than the frequencies of interest: 10,000 cps versus 200 cps.

The rate of mass accumulation in the volume V_{T1} is given by

$$\text{rate of mass accumulation in } V_{T1} = \frac{d(\rho V_{T1})}{dt} = \rho \dot{V}_{T1} + V_{T1}\dot{\rho}. \tag{3}$$

Equating (1) to (3) and solving for Q_1 yields

$$Q_1 = Q_L + \dot{V}_{T1} + \frac{V_{T1}}{\rho}\dot{\rho}. \tag{4}$$

The adiabatic bulk modulus of compression, β, of the fluid is defined as

$$\beta = \frac{dp}{\left(\frac{d\rho}{\rho}\right)}, \tag{5}$$

where β is assumed constant. The elimination of ρ from (4) by use of (5) results in

$$Q_1 = Q_L + \dot{V}_{T1} + \frac{V_{T1}}{\beta}\dot{p}_1. \tag{6}$$

In the same manner, application of the equation for the conservation of mass flow to the volume V_{T2} yields

$$Q_2 = -Q_L + \dot{V}_{T2} + \frac{V_{T2}}{\beta}\dot{p}_2. \tag{7}$$

Experimental tests show that the leakage flow is proportional to the pressure differential across the motor (laminar flow), so that

$$Q_L = L_m(p_1 - p_2), \tag{8}$$

where L_m is the leakage coefficient (in inches^5 per pound-second) of the actuator. In addition, define

$$V_T = \frac{V_{T1} + V_{T2}}{2}, \tag{9}$$

where V_T is a constant, so that

$$\begin{aligned} V_{T1} &= V_T + Ax, \\ V_{T2} &= V_T - Ax, \end{aligned} \tag{10}$$

where A is the cross-sectional area (in square inches) of the piston and x (in inches) is the piston displacement measured from the center position. Equations (6) and (7) are thus

$$Q_1 = L_m(p_1 - p_2) + A\dot{x} + \frac{(V_T + Ax)}{\beta}\dot{p}_1 \tag{11}$$

and

$$Q_2 = -L_m(p_1 - p_2) - A\dot{x} + \frac{(V_T - Ax)}{\beta}\dot{p}_2\,. \tag{12}$$

These flow equations have been developed for the linear piston-type actuator, but the same equations are valid for vane motors.

For a fixed-stroke axial-piston rotary motor, the control volume V_{T1} is a discontinuous function of time, since, as the cylinder block rotates, the individual cylinders transfer from one side of the motor to the other. Since the volume of one cylinder is small compared to the total volume on one side of the motor, the volume variation due to this discontinuity may be neglected without serious error. Thus, the control volume V_{T1} is essentially constant, so that

$$V_{T1} = V_{T2} = V_T\,, \tag{13}$$

where V_T is constant and is equal to one-half of the volume in the system. With this assumption, the flow equations for the fixed-stroke axial-piston motor can be written

$$Q_1 = L_m(p_1 - p_2) + V_m\dot{\theta}_0 + \frac{V_T}{\beta}\dot{p}_1 \tag{14}$$

and

$$Q_2 = -L_m(p_1 - p_2) - V_m\dot{\theta}_0 + \frac{V_T}{\beta}\dot{p}_2, \tag{15}$$

where $\dot{\theta}_0$ is the motor shaft angular rate and V_m (in cubic inches per radian) is the fluid displacement per unit rotation of the motor shaft. Since V_T and β do not occur separately in (14) and (15), it will be convenient to define a "compliance coefficient" K_c as

$$K_c = \frac{V_T}{\beta}, \tag{16}$$

where K_c has the units of inches^5 per pound.

The remaining discussion will be concerned with a rotational system, for which (14) and (15) have been developed; the same relationships will be valid for a translational system having small displacement x, with appropriate changes in the definitions of the parameters.

In addition to the flow equations, two torque equations can be derived. The first is an energy relationship that equates the work done by the forces on the motor during a rotational of θ_0 radians to the work output from the motor shaft. The work input is $(p_1 - p_2)\ V_m\theta_0$ and is the flow work commonly encountered in the Bernoulli equation. The corresponding work output is $T\theta_0$, where T is the opposing torque. Since these two expressions for work must be equal, there results

$$T = (p_1 - p_2)\ V_m . \tag{17}$$

Another torque equation is obtained from Newton's Second Law of Motion. In general, the load may consist of inertia, damping and friction torques, and disturbing torques. Thus, the following equation may be written:

$$T = J\ddot{\theta}_0 + T_0, \tag{18}$$

where J (in pound-inch-seconds^2) is the total inertia (including that of the motor and fluid) referred to the motor shaft, and T_0 (in pound-inches) is the total friction and disturbing torque acting on the motor shaft.

The elimination of T from (17) and (18) results in

$$(p_1 - p_2)\, V_m = J\ddot{\theta}_0 + T_0\,. \tag{19}$$

A restriction must be placed on the values allowable for the pressures p_1 and p_2. If absolute pressure units are employed, these pressures must always be equal to or greater than zero. A more accurate representation would be obtained if the vaporization pressure were considered as the limiting value, but, in view of the fact that the differential pressures in the system are normally very large, this degree of refinement is not warranted. A good approximation is

$$p_1 \geqq 0 \qquad \text{and} \qquad p_2 \geqq 0, \tag{20}$$

where it is understood that absolute pressure units are employed.

In addition to the equations derived, the usual expression for the position error ϵ in terms of the input angle θ_i and the output angle θ_0 for a servomechanism having unity feedback (as in Fig. 1) is given by

$$\epsilon = \theta_i - \theta_0\,. \tag{21}$$

The equations which have been derived in this section and which apply to the axial-piston rotary motor are summarized below:

$$\begin{aligned} Q_1 &= L_m(p_1 - p_2) + V_m\dot{\theta}_0 + K_c\dot{p}_1\,, \\ Q_2 &= -L_m(p_1 - p_2) - V_m\dot{\theta}_0 + K_c\dot{p}_2\,, \\ (p_1 - p_2)\, V_m &= J\ddot{\theta}_0 + T_0\,, \\ p_1 &\geqq 0 \qquad \text{and} \qquad p_2 \geqq 0, \\ \epsilon &= \theta_i - \theta_0\,. \end{aligned} \tag{22}$$

The units employed in these equations are given in Table I.

The expressions for Q_1 and Q_2, the flows from the controlled source, are discussed in the next section.

III. CONTROL-VALVE ANALYSIS

In Section II, the equations relating the flows (into the motor) to the dynamic state of the system were derived; the expressions for the flows Q_1 and Q_2 were not specified. In this section, these quantities are discussed for the particular case of a valve-controlled system and analytical expressions relating flow to error signal and pressure are obtained.

The schematic diagram of a typical valve configuration is given in

TABLE I — DEFINITIONS OF SYMBOLS; UNITS

Symbol	Definition	Units
Q	flow	in^3/sec
L_m	motor leakage coefficient	in^5/lb-sec
p	pressure	lb/in^2
V_m	motor displacement	in^3/radian
θ_i	position angle input	radians
θ_0	position angle output	radians
ϵ	position angular error	radians
V_T	one-half total trapped volume	in^3
β	bulk modulus of compression	lb/in^2
J	total inertia	lb-in-sec^2
K_c	compliance coefficient	in^5/lb
T_0	viscous, friction and disturbing torque	lb-in
ρ	fluid mass density	lb-sec^2/in^4

Fig. 3. The type of actuator is not important in this discussion and, for simplicity, it is pictured as a translational piston.

A source having a pressure p_s supplies fluid to the valve as shown and the main spool controls the direction and magnitude of this flow to the motor. Fluid is returned to a sump at pressure p_d. As in Section II, the pressure on each side of the actuator piston is designated as p with the appropriate numerical subscript.

The main spool controls the flow by means of four orifices O_1, O_2, E_1 and E_2, and the spool position, in turn, is controlled by a transducer. The configuration of the transducer varies considerably in current

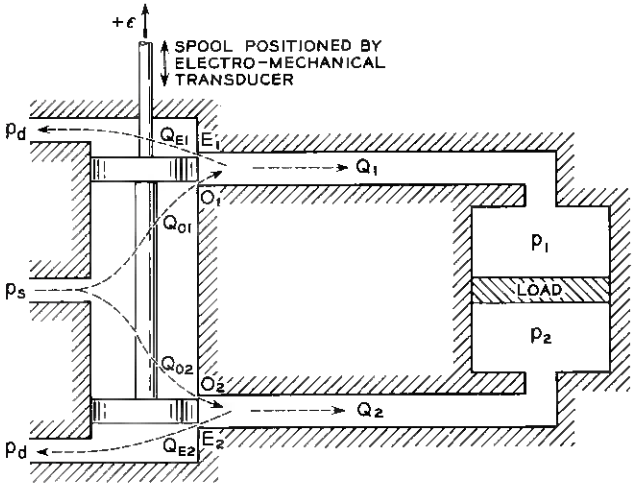


Fig. 3 — Orifice flow conventions.

production models and may have the form of either a torque motor or a hydraulic preamplifier preceded by a torque motor or solenoid. If a hydraulic preamplifier is employed, the entire valve is designated as a "two-stage" valve. In the analysis an ideal spool-positioning mechanism will be assumed; i.e., there will be only one position of the valve spool corresponding to a given error signal, ϵ.

The following nodal equations are obtained from Fig. 3:

$$\begin{aligned} Q_1 &= Q_{O1} - Q_{E1}, \\ Q_2 &= Q_{O2} - Q_{E2}. \end{aligned} \tag{23}$$

Here, as in Section II, the flows are chosen positive in the directions indicated in the figure.

It is necessary to express the flows through the orifices in terms of the pressures and the orifice openings. Consider first the general orifice equation; application of Bernoulli's equation to the case of flow through an orifice of area A results in the relation

$$Q = \frac{AC_cC_v}{\sqrt{1 - C_c^2\left(\frac{A}{A_1}\right)^2}} \sqrt{\frac{2\Delta p}{\rho}}, \tag{24}$$

where:

Q = flow,
A = orifice area,
C_c = orifice contraction coefficient,
C_v = velocity coefficient,
A_1 = upstream line area,
Δp = pressure differential across the orifice,
ρ = mass density of the fluid.

For application to a valve orifice, note that A/A_1 is small compared to unity and that C_c is less than unity. It follows, therefore, that a good approximation is obtained by

$$Q = AC_cC_v \sqrt{\frac{2\Delta p}{\rho}}. \tag{25}$$

The usual procedure in hydraulics is to define a discharge coefficient, C, as

$$C = C_cC_v, \tag{26}$$

so that (25) becomes

$$Q = AC \sqrt{\frac{2\Delta p}{\rho}}. \tag{27}$$

This equation has been derived for the steady-state flow through an orifice; in the analysis, it will be assumed that the same relationship is valid for the dynamic state. It is not necessary to assume a value for the discharge coefficient C, since it will be included in an over-all gain constant.

In the present application, the orifice area, A, is proportional to the displacement of the valve spool (assuming rectangular orifices) and the displacement is proportional to the electrical activating signal, $K\epsilon$, received by the positioning mechanism (see Fig. 3). It follows that A is proportional to $K\epsilon$, and (27) may be written as

$$Q = (K\epsilon)C_0\sqrt{\frac{\Delta p}{\rho}}, \tag{28}$$

where C_0 is a new coefficient that is proportional to the area of the orifice per radian of input signal. If the mass density, ρ, of the fluid is assumed constant in this expression, the following equation may be written:

$$Q = (K_b\epsilon)\sqrt{\Delta p}, \tag{29}$$

where

$$K_b = \frac{KC_0}{\rho^{1/2}} \tag{30}$$

is a constant for a given system. K_b has the units of inches^4 per pound$^{1/2}$-second-radians.

The relationship of (29) is indicated graphically in Fig. 4. In this figure, the equation has been normalized with respect to the three maximum values $Q_{\max}$, $(K_b\epsilon)_{\max}$ and $\Delta p_{\max}$. Since there is a limit to the magnitude of the spool displacement, a spool-displacement saturation region is indicated. (Spool-displacement saturation occurs when the electrical error signal becomes greater than that corresponding to the maximum valve spool-displacement; i.e., the error signal demands a spool position that is physically impossible.)

Equation (29) implies that the flow through the orifice is zero for zero error ϵ. This is not generally true for the orifices in most valves, and in fact, this condition would be very difficult to obtain. Equation (29) must be specialized for each orifice.

There are three general types of valves, which can be classified according to the flow conditions at zero signal input ϵ; these are: (a) the open-center valve, (b) the critical-center valve and (c) the closed-center

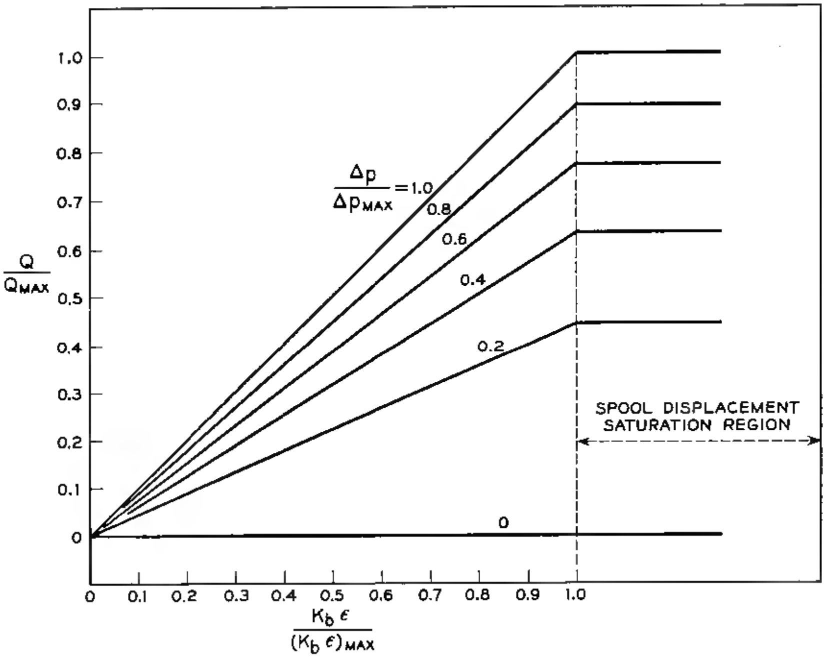


Fig. 4 — Orifice flow characteristics.

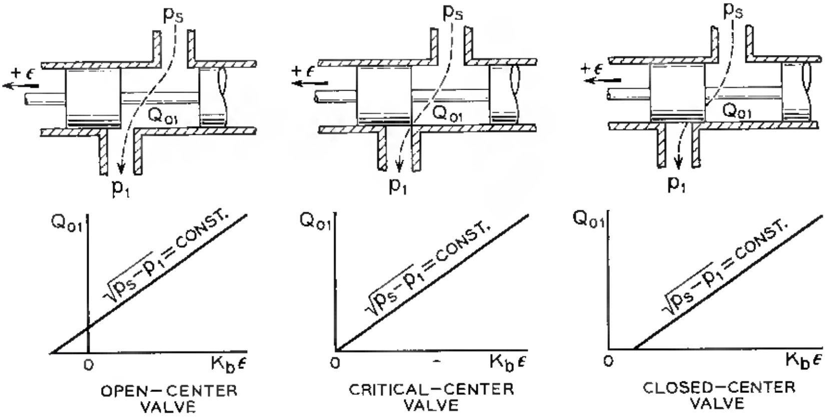


Fig. 5 — (a) Neutral position for the three basic valve types; (b) typical characteristics for the three basic valve types.

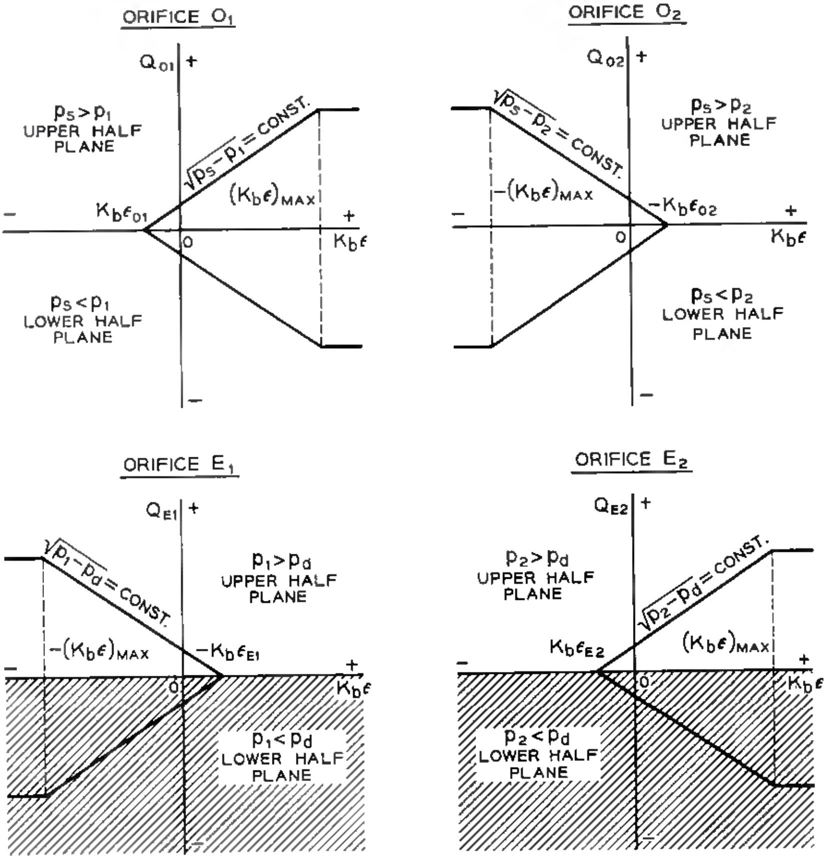


Fig. 6 — Flow characteristics for the orifices in an open-center valve.

valve. Fig. 5(a) indicates the zero error position of the spool for each of these types for the orifice O_1. In the open-center valve, flow passes through the orifice for the condition $\epsilon = 0$, i.e., for the spool in the neutral position. The critical-center valve allows no flow to pass in the neutral position; however, a slight positive displacement opens the orifice O_1. The closed-center valve has a dead zone in that a relatively large displacement is required to open the orifice from the neutral position.

Fig. 5(b) shows a typical differential isobar (corresponding to those of Fig. 4) for each type of valve. The origin is translated according to the neutral spool position.

It is now possible to express the individual orifice flows as shown in Fig. 3 with the aid of Fig. 6 and (29). Fig. 6 shows a typical differential isobar for each of the four orifices O_1, O_2, E_1 and E_2, the case shown representing the open-center type valve. (The equations to be derived will be applicable to all three valve types, depending upon the choice of the neutral position error-flow constants.) It should be noted that the direction of flow is dependent on the sign of Δp for each orifice. In general, the equations for the flows are as given by (31). In these equations ϵ_{O1}, ϵ_{O2}, ϵ_{E1}, ϵ_{E2} are positive for an open-center valve, $\epsilon_{\max}$ is taken as positive and sgn denotes the signum (sign) function:

$$
\begin{aligned}
Q_{O1} &= \begin{cases} K_b \left| \epsilon_{\max} + \epsilon_{O1} \right| \sqrt{\left| p_s - p_1 \right|} \operatorname{sgn} (p_s - p_1) & \epsilon > \epsilon_{\max} \\ K_b \left| \epsilon + \epsilon_{O1} \right| \sqrt{\left| p_s - p_1 \right|} \operatorname{sgn} (p_s - p_1) & -\epsilon_{O1} < \epsilon < \epsilon_{\max} \\ 0 & \epsilon < -\epsilon_{O1}, \end{cases} \\
Q_{O2} &= \begin{cases} 0 & \epsilon > \epsilon_{O2} \\ K_b \left| \epsilon - \epsilon_{O2} \right| \sqrt{\left| p_s - p_2 \right|} \operatorname{sgn} (p_s - p_2) & -\epsilon_{\max} < \epsilon < \epsilon_{O2} \\ K_b \left| \epsilon_{\max} - \epsilon_{O2} \right| \sqrt{\left| p_s - p_2 \right|} \operatorname{sgn} (p_s - p_2) & \epsilon < -\epsilon_{\max}, \end{cases} \\
Q_{E1} &= \begin{cases} 0 & \epsilon > \epsilon_{E1} \\ K_b \left| \epsilon - \epsilon_{E1} \right| \sqrt{\left| p_1 - p_d \right|} \operatorname{sgn} (p_1 - p_d) & -\epsilon_{\max} < \epsilon < \epsilon_{E1} \\ K_b \left| \epsilon_{\max} - \epsilon_{E1} \right| \sqrt{\left| p_1 - p_d \right|} \operatorname{sgn} (p_1 - p_d) & \epsilon < -\epsilon_{\max}, \end{cases} \\
Q_{E2} &= \begin{cases} K_b \left| \epsilon_{\max} + \epsilon_{E2} \right| \sqrt{\left| p_2 - p_d \right|} \operatorname{sgn} (p_2 - p_d) & \epsilon > \epsilon_{\max} \\ K_b \left| \epsilon + \epsilon_{E2} \right| \sqrt{\left| p_2 - p_d \right|} \operatorname{sgn} (p_2 - p_d) & -\epsilon_{E2} < \epsilon < \epsilon_{\max} \\ 0 & \epsilon < -\epsilon_{E2}; \end{cases}
\end{aligned}
\tag{31}
$$

Simplification of (31) is obtained by assuming that the sump pressure

Table II — Definitions of Symbols; Units

Symbol	Definition	Units
Q	flow	in^3/sec
A	orifice area	in^2
A_1	upstream line area	in^2
C_c	orifice contraction coefficient	—
C_v	orifice velocity coefficient	—
C	orifice discharge coefficient	—
C_0	area per radian input	in^2/rad
ϵ	angular actuating signal	radians
K	dimensionless gain constant	—
K_b	over-all gain constant	in^4/lb$^{1/2}$-sec-rad
p	pressure	lb/in^2
Δp	differential pressure	lb/in^2
ρ	fluid mass density	lb-sec^2/in^4

p_d is zero psi absolute rather than atmospheric pressure; this is a good approximation since the pressure differentials are usually several hundred psi. This assumption eliminates consideration of the shaded regions in Fig. 6 for the orifices E_1 and E_2. Further simplification is obtained if the valve is assumed to be perfectly symmetrical; i.e.,

$$\epsilon_{O1} = \epsilon_{O2} = \epsilon_{E1} = \epsilon_{E2} = \epsilon_0 \,,$$

and if it is assumed that the maximum spool position is never attained. With these assumptions, (31) simplify to (32):

$$\begin{aligned}
Q_{O1} &= \left[\begin{array}{ll} K_b\,|\epsilon + \epsilon_0|\sqrt{|p_s - p_1|}\;\mathrm{sgn}\,(p_s - p_1) & \qquad \epsilon > -\epsilon_0 \\ 0 & \qquad \epsilon < -\epsilon_0 \,, \end{array}\right. \\
Q_{O2} &= \left[\begin{array}{ll} 0 & \qquad \epsilon > \epsilon_0 \\ K_b\,|\epsilon - \epsilon_0|\sqrt{|p_s - p_2|}\;\mathrm{sgn}\,(p_s - p_2) & \qquad \epsilon < \epsilon_0 \,, \end{array}\right. \\
Q_{E1} &= \left[\begin{array}{ll} 0 & \qquad \epsilon > \epsilon_0 \\ K_b\,|\epsilon - \epsilon_0|\sqrt{p_1} & \qquad \epsilon < \epsilon_0 \,, \end{array}\right. \\
Q_{E2} &= \left[\begin{array}{ll} K_b\,|\epsilon + \epsilon_0|\sqrt{p_2} & \qquad \epsilon > -\epsilon_0 \\ 0 & \qquad \epsilon < -\epsilon_0 \,. \end{array}\right.
\end{aligned} \tag{32}$$

These equations, together with (23) and (22), complete the preliminary analysis for the valve-controlled servomechanism. Definitions of the symbols, together with a consistent set of units, are given in Table II.

IV. LINEAR AND INCREMENTAL-LINEAR ANALYSIS

The linearization of the equations representing the motor as given by (22) results in (for a pure inertial load):

$$\begin{aligned}
Q_1 &= L_m(p_1 - p_2) + V_m\dot{\theta}_0 + K_c\dot{p}_1 \,, \\
Q_2 &= -L_m(p_1 - p_2) - V_m\dot{\theta}_0 + K_c\dot{p}_2 \,, \\
(p_1 - p_2)\,V_m &= J\ddot{\theta}_0 \,, \\
\epsilon &= \theta_i - \theta_0 \,.
\end{aligned} \tag{33}$$

The restrictions on the values of p_1 and p_2 and their derivatives are not applicable to a linear theory and have been omitted; in addition, the friction and output disturbing torques have been assumed to be zero.

For the case of a symmetrical valve having characteristics as given by (32), the flows through the orifices are functions of the gain constant, K_b, the error signal, ϵ, the open-center constant, ϵ_0, and the respective orifice pressure drops. If the region of applicability is restricted to those cases in which $p_s > p_1$ and $p_s > p_2$ (pressures not limited) and in which $| \epsilon | < | \epsilon_{\max} |$, (32) may be written:

$$\begin{aligned} Q_{O1} &= K_b(\epsilon + \epsilon_0)\sqrt{p_s - p_1} && \text{for} \quad \epsilon > -\epsilon_0 , \\ Q_{O2} &= -K_b(\epsilon - \epsilon_0)\sqrt{p_s - p_2} && \text{for} \quad \epsilon < \epsilon_0 , \\ Q_{E1} &= -K_b(\epsilon - \epsilon_0)\sqrt{p_1} && \text{for} \quad \epsilon < \epsilon_0 , \\ Q_{E2} &= K_b(\epsilon + \epsilon_0)\sqrt{p_2} && \text{for} \quad \epsilon > -\epsilon_0 . \end{aligned} \tag{34}$$

From (23) and (34) it can be seen that there are three distinct ranges of ϵ that must be considered, the magnitude of the range depending upon the value of ϵ_0 :

Region A ($| \epsilon | \leqq \epsilon_0$):

$$\begin{aligned} Q_1 &= K_b(\epsilon + \epsilon_0)\sqrt{p_s - p_1} + K_b(\epsilon - \epsilon_0)\sqrt{p_1} , \\ Q_2 &= -K_b(\epsilon - \epsilon_0)\sqrt{p_s - p_2} - K_b(\epsilon + \epsilon_0)\sqrt{p_2} . \end{aligned} \tag{35}$$

Region B+ ($\epsilon \geqq \epsilon_0$):

$$\begin{aligned} Q_1 &= K_b(\epsilon + \epsilon_0)\sqrt{p_s - p_1} , \\ Q_2 &= -K_b(\epsilon + \epsilon_0)\sqrt{p_2} . \end{aligned} \tag{36}$$

Region B− ($\epsilon \leqq \epsilon_0$):

$$\begin{aligned} Q_1 &= K_b(\epsilon - \epsilon_0)\sqrt{p_1} , \\ Q_2 &= -K_b(\epsilon - \epsilon_0)\sqrt{p_s - p_2} . \end{aligned} \tag{37}$$

Expanding $\sqrt{p_s - p}$ and $\sqrt{p}$ about the steady-state value $p_s/2$ yields

Region A:

$$Q_1 - Q_2 = 2K_b \left(2\epsilon \sqrt{\frac{p_s}{2}} - \epsilon_0 \frac{p_1 - p_2}{\sqrt{2p_s}}\right). \tag{38}$$

Region B+:

$$Q_1 - Q_2 = K_b \, (\epsilon + \epsilon_0) \left(2 \sqrt{\frac{p_s}{2}} - \frac{p_1 - p_2}{\sqrt{2p_s}}\right). \tag{39}$$

Region B−:

$$Q_1 - Q_2 = K_b(\epsilon - \epsilon_0) \left(2 \sqrt{\frac{p_s}{2}} + \frac{p_1 - p_2}{\sqrt{2p_s}}\right), \tag{40}$$

where the higher-order terms in $[(p_s/2) - p]$ have been neglected. This further restricts the region of validity of our analysis to those cases where the pressure differential across the load is not large.

For the region A, combination of (33) and (38) results in

$$4K_b \sqrt{\frac{p_s}{2}}\,\epsilon = \frac{K_c J}{V_m}\dddot{\theta}_0 + \frac{2J}{V_m}\left(L_m + \frac{K_b\epsilon_0}{\sqrt{2p_s}}\right)\ddot{\theta}_0 + 2V_m\dot{\theta}_0\,, \tag{41}$$

and the open-loop transfer function becomes

$$G_A(s) = \frac{\theta_0(s)}{\epsilon(s)} = \frac{\dfrac{4V_mK_b}{K_cJ}\sqrt{\dfrac{p_s}{2}}}{s\left[s^2 + \dfrac{2}{K_c}\left(L_m + \dfrac{K_b\epsilon_0}{\sqrt{2p_s}}\right)s + \dfrac{2V_m^{\,2}}{K_cJ}\right]}, \tag{42}$$

or

$$G_A(s) = \frac{\theta_0(s)}{\epsilon(s)} = \frac{\dfrac{\omega_a}{\omega_n}}{\left(\dfrac{s}{\omega_n}\right)\left[\left(\dfrac{s}{\omega_n}\right)^2 + 2\zeta_a\left(\dfrac{s}{\omega_n}\right) + 1\right]}, \tag{43}$$

where

$$\omega_n = V_m\sqrt{\frac{2\beta}{V_TJ}}, \tag{44}$$

$$\zeta_a = \frac{1}{V_m}\left[L_m + \frac{K_b\epsilon_0}{\sqrt{2p_s}}\right]\sqrt{\frac{J\beta}{2V_T}}, \tag{45}$$

$$\omega_a = \frac{2K_b}{V_m}\sqrt{\frac{p_s}{2}}. \tag{46}$$

Here, ω_n is the undamped natural frequency of the system, ζ_a is the damping ratio and ω_a is the velocity gain constant. (The subscript "a" indicates the region A.) It is interesting to note that ζ_a, the damping ratio, is the sum of the motor damping (motor leakage) and a term related to the steady-state flow through the valve. The term, $K_b\epsilon_0/\sqrt{2p_s}$, contributes the major damping to the system. In the quiescent state, since $\epsilon = 0$ and $p_1 = p_2 = p_s/2$, it follows from (34) that

$$(Q_{O1})_s = (Q_{O2})_s = (Q_{E1})_s = (Q_{E2})_s = K_b\epsilon_0\sqrt{\frac{p_s}{2}}, \tag{47}$$

where the subscript "s" indicates quiescent values. Now designate

$$Q_s = (Q_{O1})_s + (Q_{O2})_s\,, \tag{48}$$

where Q_s is the quiescent (i.e. $\epsilon = 0$ and $\dot{\theta} = 0$) total flow from the source, so that

$$Q_s = 2K_b\epsilon_0 \sqrt{\frac{p_s}{2}}. \tag{49}$$

An effective leakage coefficient is given by

$$L_a = \left(L_m + \frac{K_b\epsilon_0}{\sqrt{2p_s}}\right) = \left(L_m + \frac{Q_s}{2p_s}\right) \tag{50}$$

and therefore (45) becomes

$$\zeta_a = \frac{L_a}{V_m}\sqrt{\frac{J\beta}{2V_T}}. \tag{51}$$

Now consider the region B+, where $\epsilon \geqq \epsilon_0$. The combination of (33) and (39) results in

$$2K_b\sqrt{\frac{p_s}{2}}\,(\epsilon + \epsilon_0) = \frac{K_cJ}{V_m}\dddot{\theta}_0 + \frac{2J}{V_m}\left[L_m + \frac{K_b(\epsilon + \epsilon_0)}{2\sqrt{2p_s}}\right]\ddot{\theta}_0 + 2V_m\dot{\theta}_0. \tag{52}$$

An approximate "incremental linear" transfer function may be obtained for this region by making the substitutions

$$\begin{aligned} \epsilon &= \epsilon^* + \Delta\epsilon, \\ \theta_0 &= \theta_0^* + \Delta\theta_0, \end{aligned} \tag{53}$$

where both the starred and the incremental symbols are considered as functions of time. The incremental variables are assumed small, so that their products may be neglected. In this manner, a linear equation in the incremental quantities can be obtained if the equation defining the starred variables [(52) with the symbols starred] is subtracted from that obtained by the substitutions indicated previously. Then, if the resultant equation, which is linear in the incremental variables, contains any starred quantities, these can be considered to be varying slowly with time — that is, essentially constant when compared with the incremental variations with time. Therefore, a quasilinear incremental transfer function can be obtained. In the present case, the incremental transfer function valid for a small constant-acceleration, α_0, will be derived. This will be used to obtain the incremental transfer function for constant-velocity operation.

Thus, the analysis is initiated by obtaining the linear equation in the incremental quantities as previously outlined; this equation is

$$\begin{aligned} &2K_b\left(\sqrt{\frac{p_s}{2}} - \frac{J\ddot{\theta}_0^*}{2V_m\sqrt{2p_s}}\right)\Delta\epsilon = \\ &\qquad \frac{K_cJ}{V_m}\Delta\dddot{\theta}_0 + \frac{2J}{V_m}\left[L_m + \frac{K_b(\epsilon^* + \epsilon_0)}{2\sqrt{2p_s}}\right]\Delta\ddot{\theta}_0 + 2V_m\Delta\dot{\theta}_0. \end{aligned} \tag{54}$$

Noting that a similar equation may be obtained for the region B−, and denoting both regions by B, the following transfer function is obtained for operation about a constant acceleration α_0 :

$$G_B(s) = \frac{\Delta\theta_0(s)}{\Delta\epsilon(s)} = \frac{\dfrac{\omega_b}{\omega_n}}{\left(\dfrac{s}{\omega_n}\right)\left[\left(\dfrac{s}{\omega_n}\right)^2 + 2\zeta_b\left(\dfrac{s}{\omega_n}\right) + 1\right]}, \tag{55}$$

where

$$\omega_n = V_m \sqrt{\frac{2\beta}{V_T J}}, \tag{56}$$

$$\zeta_b = \frac{1}{V_m}\left[L_m + \frac{K_b(|\epsilon^*| + \epsilon_0)}{2\sqrt{2p_s}}\right]\sqrt{\frac{J\beta}{2V_T}}, \tag{57}$$

$$\omega_b = \frac{K_b}{V_m}\sqrt{\frac{p_s}{2}}\left(1 - \frac{J|\alpha_0|}{2V_m p_s}\right). \tag{58}$$

For an inertia load (no viscous or coulomb friction), (58) can be written:

$$\omega_b = \frac{K_b}{V_m}\sqrt{\frac{p_s}{2}}\left(1 - \frac{|\Delta p^*|}{2p_s}\right), \tag{59}$$

where Δp^* is the constant-differential pressure acting on the motor.

The preceding equations give an approximate solution for the case in which the acceleration α_0 is small; i.e., ϵ^* varies slowly with time. For this case, the incremental damping ratio ζ_b is appreciably increased over that given by (45) for the region A. The incremental gain constant ω_b is less than the gain constant in the A region as defined in (46); for Δp^* equal to p_s , the gain is down 12 db from that of the A region.

For operation about a constant velocity ω_i , ω_n and ζ_b remain as in (56) and (57), but the gain constant becomes

$$\omega_b = \frac{K_b}{V_m}\sqrt{\frac{p_s}{2}}. \tag{60}$$

In this case, the gain is down 6 db from that given for the A region. For constant-speed operation, $\theta_0^* = \omega_i t$, the error is given by

$$\epsilon^* = \sqrt{\frac{2}{p_s}}\left(\frac{V_m\omega_i}{K_b}\right) - \epsilon_0 , \tag{61}$$

so that the effective leakage coefficient becomes:

$$L_b = L_m + \frac{K_b(|\epsilon^*| + \epsilon_0)}{2\sqrt{2p_s}} = L_m + \frac{V_m\omega_i}{2p_s}. \tag{62}$$

Thus, since the "displacement" motor flow is given by

$$Q_m = V_m\dot{\theta}_0 = V_m\omega_i , \tag{63}$$

the effective leakage coefficient is

$$L_b = L_m + \frac{Q_m}{2p_s} . \tag{64}$$

In many cases, L_m is small compared to Q_m so that the incremental damping is primarily a function of the motor speed.

The velocity-lag error, ϵ_v (steady-state position error under the conditions $\dot{\theta}_0 = \omega_i$ where ω_i is a constant rate input), is as follows for the two regions:

Region A:

$$\epsilon_v = \frac{\omega_i}{\omega_a} = \frac{V_m\omega_i}{K_b\sqrt{2p_s}} . \tag{65}$$

Region B:

$$| \epsilon_v | = \frac{2V_m | \omega_i |}{K_b\sqrt{2p_s}} - \epsilon_0 , \tag{66}$$

where it is understood that $| \epsilon_v |$ is greater than ϵ_0. The velocity lag error for the B region is thus approximately twice that predicted by (65) if ϵ_0 is small compared to the lag error.

Some qualitative information on the nature of the system performance can be obtained by comparison of (43) and (55). In the region A (i.e., $| \epsilon | < \epsilon_0$), the system is essentially linear for small pressure differentials. In the region B ($| \epsilon | > \epsilon_0$), the system is nonlinear *even though* the pressure differentials are assumed small. In this region, the behavior is "amplitude sensitive"; as the error amplitude increases, the incremental gain decreases. The incremental damping increases with increasing error in the B region and is considerably greater than the damping in the A region; the incremental damping is proportional to the total flow through the valve. The velocity-lag error is, of course, greater in the B region.

The transfer function $G_A(s)$ as given by (43) was derived for a fully symmetrical valve; the analysis of an unsymmetrical valve shows that the basic form of the transfer function is similar to that for the symmetrical valve. For the unsymmetrical valve, however, the damping ratio and gain expressions differ from those of (45) and (46). The damping ratio is

$$\zeta_a = \frac{1}{V_m}\left[L_m + \frac{Q_s p_s}{8p_q(p_s - p_q)}\right]\sqrt{\frac{J\beta}{2V_T}}, \qquad (67)$$

where p_q is the quiescent value of the pressures p_1 and p_2. For $p_q = p_s/2$, this expression is identical to that given by (45); for p_q greater or less than $p_s/2$, the damping is greater than that given by (45) (assuming the same quiescent flow Q_s). The gain constant is:

$$\omega_a = \frac{K_b}{V_m}\left(\sqrt{p_q} + \sqrt{p_s - p_q}\right), \qquad (68)$$

and comparison with (46) shows that the two give the same solution for $p_q = p_s/2$. For p_q greater or less than $p_s/2$, the gain of the unsymmetrical valve is less than that of the symmetrical valve.

V. SOLUTIONS OF THE LINEARIZED EQUATIONS

The equations representing the valve-controlled servomechanism were linearized in the last section, and it was found that the open-loop transfer function had the following general form:

$$G(s) = \frac{\theta_0(s)}{\epsilon(s)} = \frac{\dfrac{\omega_0}{\omega_n}}{\left(\dfrac{s}{\omega_n}\right)\left[\left(\dfrac{s}{\omega_n}\right)^2 + 2\zeta\left(\dfrac{s}{\omega_n}\right) + 1\right]}, \qquad (69)$$

where ω_0 is the velocity gain constant, ω_n the natural resonant frequency and ζ the dimensionless damping ratio. If this is considered as a frequency function, the resulting open-loop attenuation and phase vary as shown in Figs. 7 and 8. In these figures, the damping ratio, ζ, has been taken as a parameter.

Equation (69), when solved for the closed-loop function, results in

$$\frac{\theta_0(s)}{\theta_i(s)} = \frac{\dfrac{\omega_0}{\omega_n}}{\left(\dfrac{s}{\omega_n}\right)^3 + 2\zeta\left(\dfrac{s}{\omega_n}\right)^2 + \left(\dfrac{s}{\omega_n}\right) + \left(\dfrac{\omega_0}{\omega_n}\right)}. \qquad (70)$$

The relationships for the closed-loop operation are exhibited graphically in Figs. 9 through 12. In Fig. 9, the gain margin is shown as a function of the peak attenuation, M_p (the maximum value of $|\theta_0/\theta_i|$), for the range of values of interest. In general, the gain margin decreases with increasing peak magnitude and is less for the lightly damped cases.

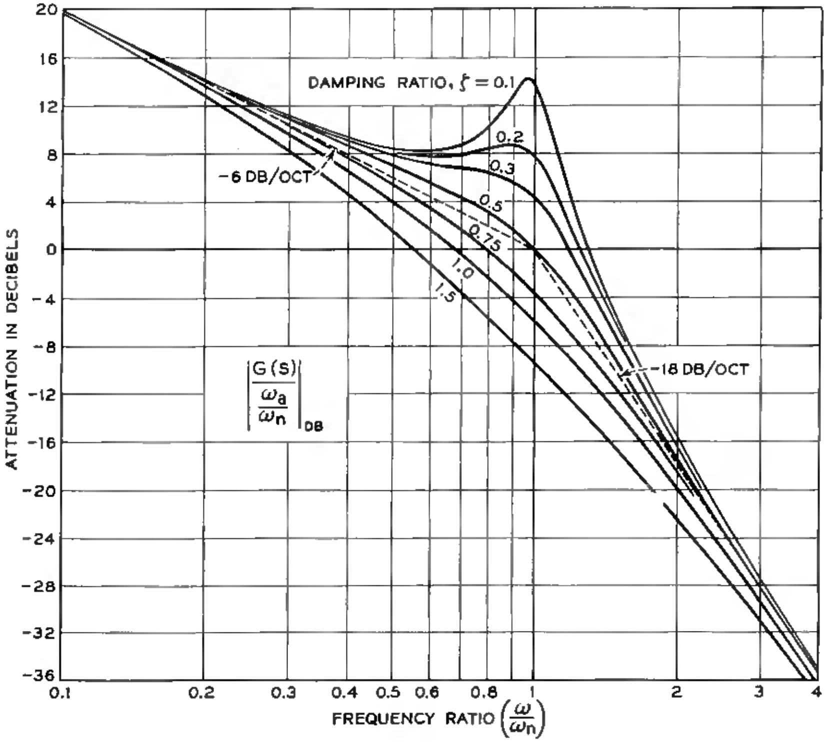


Fig. 7 — Open-loop attenuation of $G(s)$.

The phase margin variation with the peak magnitude is shown in Fig. 10; the smaller values of ζ have the largest phase margins.

Fig. 11 shows the variation of the peak frequency, ω_p, with the peak attenuation, M_p, for the various damping ratios ζ. For $\zeta = 0.1$, the peak frequency is approximately the undamped frequency of the system and is independent of the peak attenuation. As the damping is increased, the peak frequency decreases, and it is lower for the lower peak attenuations. In general, the lower the value of ω_p, the lower will be the bandwidth of the closed-loop system.

The relation between the velocity gain constant ω_0 and the peak magnitude is shown in Fig. 12 as a function of the damping ratio, ζ. As the gain constant is increased, the peak magnitude increases; in most cases (for constant peak magnitude), ω_0 is less for the lightly damped systems. This graph shows that, for $\zeta = 0.1$, a change in ω_0 of approximately 3 db is sufficient to cause M_p to increase 8 db, while,

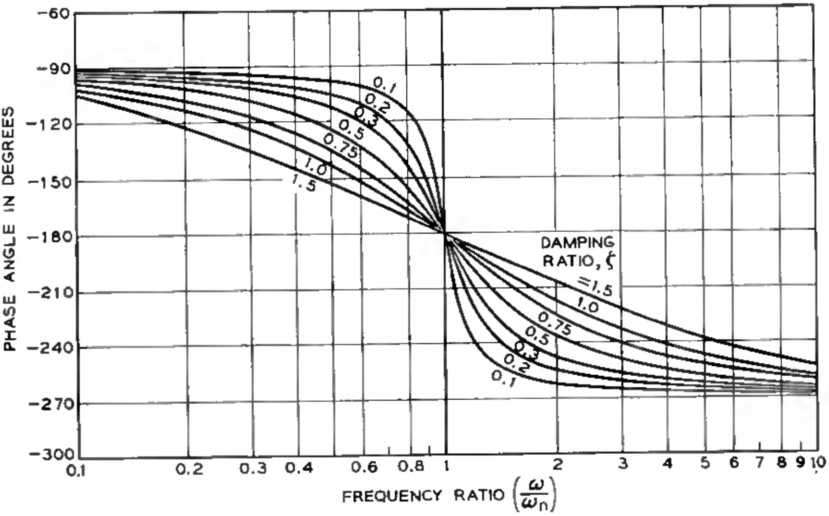


Fig. 8 — Open-loop phase angle of $G(s)$.

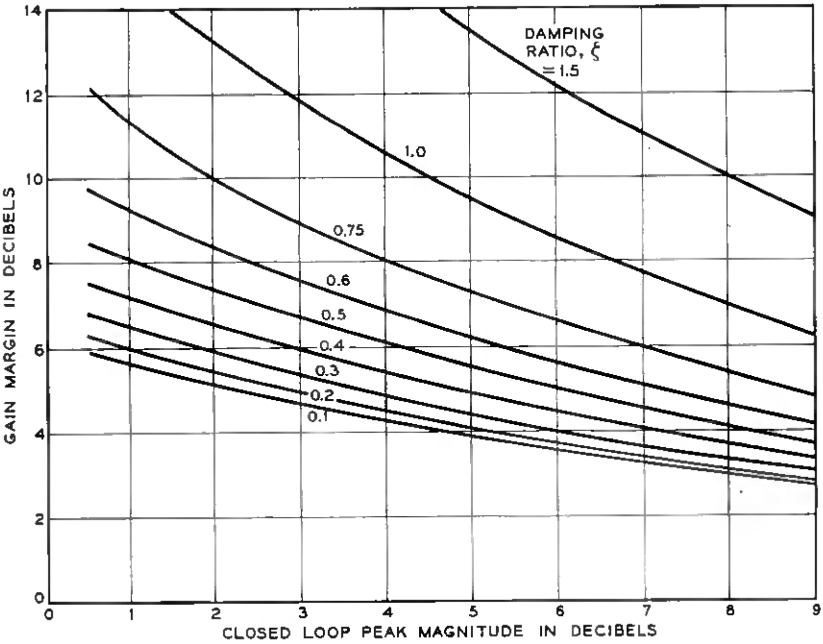


Fig. 9 — Gain margin vs. closed-loop peak magnitude as a function of damping ratio.

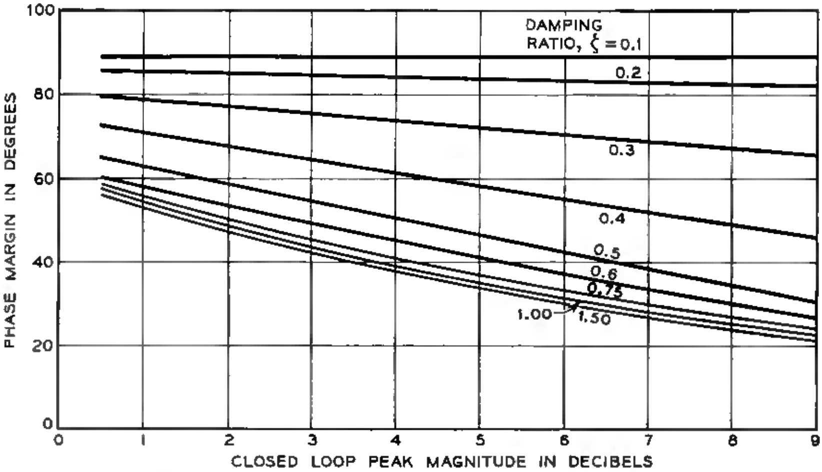


Fig. 10 — Phase margin vs. closed-loop peak magnitude as a function of damping ratio.

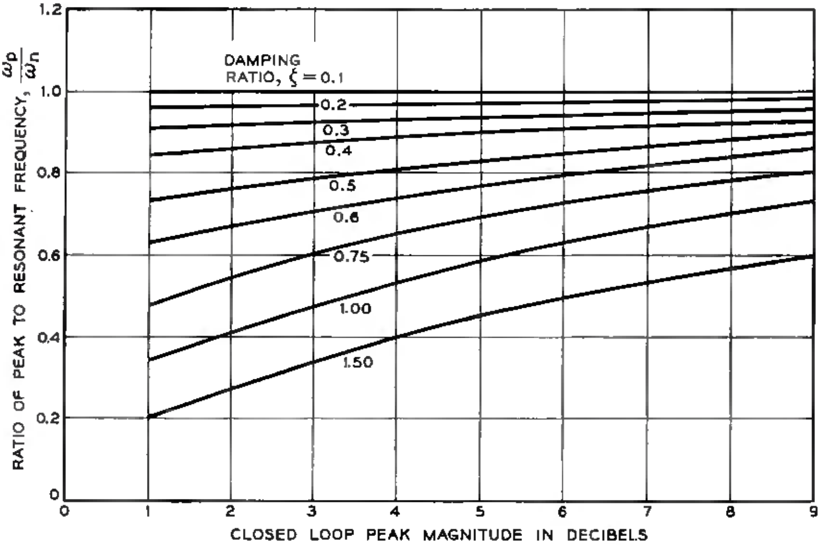


Fig. 11 — Closed-loop peak frequency vs. peak magnitude as a function of damping ratio.

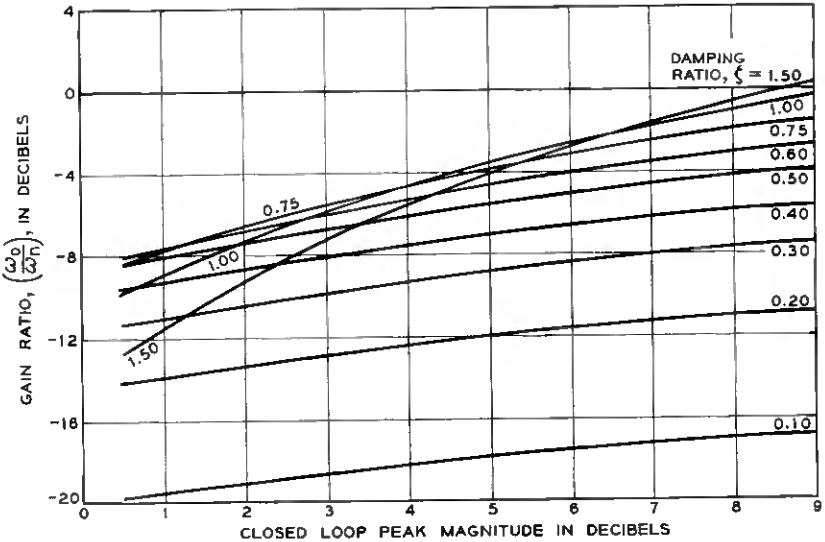


Fig. 12 — Gain ratio vs. closed-loop peak magnitude as a function of damping ratio.

for $\zeta = 1.00$, ω_0 must change 9 db for the same increase in peak magnitude. Since, in a practical system, fluctuations in the value of ω_0 are to be expected, operation of the lightly damped system would be more erratic than that of a system having adequate damping.

Equation (70), when solved for the transient response (θ_i a step function) yields solutions as given in Fig. 13. It is interesting to note that the transient response for a servomechanism having a damping ratio of $\zeta = 0.1$ and a closed-loop peak magnitude $M_p = 3$ db shows little overshoot. Examination of the frequency response shows that this is a result of a large attenuation in the frequency region below resonance. The superimposed oscillation is caused by the gain in the resonant frequency region.

The results of the transient solutions for $\zeta = 0.5$ are summarized in graphical form in Fig. 14, in which the delay time, T_d, rise time, T_r, peak time, T_p, and per cent overshoot are given as functions of the gain, ω_0/ω_n. (The definitions of the various time values are given in Fig. 15.) The response times decrease rapidly with increasing gain for the lower gain values and, as the gain increases, become relatively insensitive to gain variations. The per cent overshoot is a linear function of gain for values of gain above the limiting case in which there is no overshoot ($\omega_0/\omega_n = 0.316$).

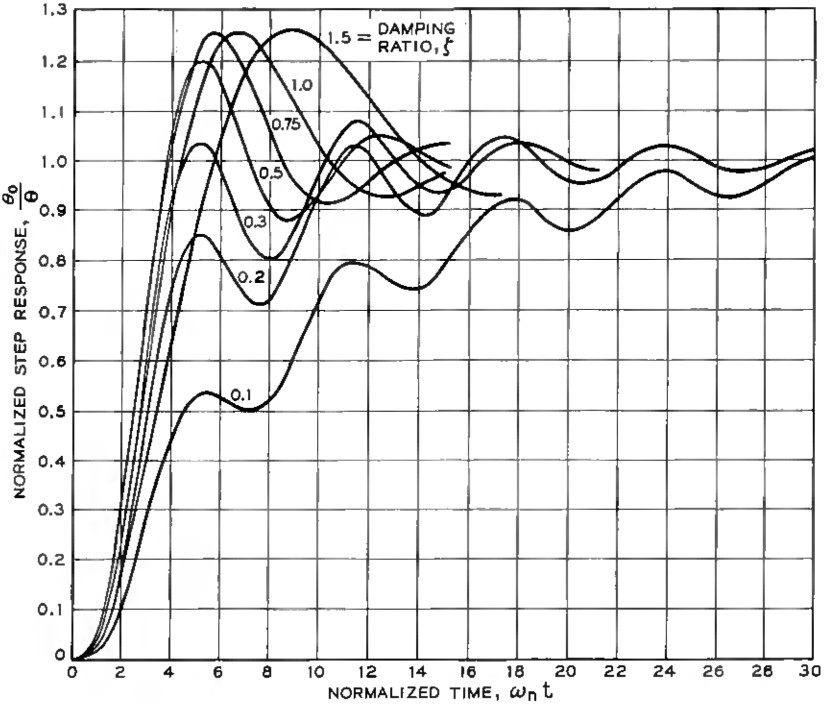


Fig. 13 — Variation of step response of linear system with damping ratio for constant closed-loop peak attenuation $M_p = 3$ db.

The relationships developed in this section can be used to estimate design parameters. For example, from Fig. 9, assume that a gain margin of 6 db is desired for a servomechanism having a damping ratio $\zeta = 0.5$. This fixes the peak magnitude M_p to be 4.15 db and, from Fig. 10, the phase margin is found to be 50 degrees. From Fig. 11, the peak frequency is $\omega_p = 0.8\ \omega_n$ and, from Fig. 12, the gain ratio ω_0/ω_n is -6 db; i.e., $\omega_0 = 0.5\ \omega_n$. Fig. 14 then predicts an overshoot of 24 per cent; a delay time, $T_d = 2.3/\omega_n$; a rise time, $T_r = 2.1/\omega_n$; and a peak time, $T_p = 5.0/\omega_n$.

VI. ANALOG SOLUTIONS OF THE NONLINEAR EQUATIONS

The equations representing the behavior of a valve-controlled servomechanism were derived in Sections II and III, and the approximate linear theory was discussed in Sections IV and V. In this section, representative analog computer solutions of the nonlinear equations (which

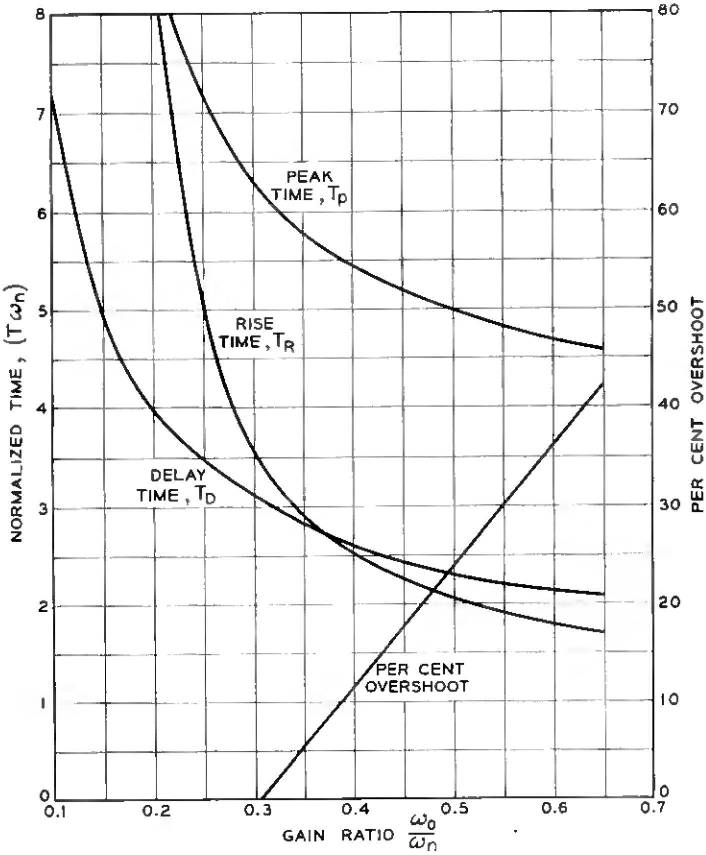


Fig. 14 — Transient response data for a damping ratio of 0.5.

are summarized in Table III) are given for particular values of the parameters, and the results are correlated with the linear solutions.

The valve-controlled servomechanism is assumed to have the numerical constants listed in Table IV. During the course of the discussion, the effects of changes in these parameters will be considered, but, unless otherwise stated, the values will be assumed to be as given in the table. In this manner, a reference system is obtained and the discussion of the effects of parameter variations is facilitated by comparison with the reference behavior.

The first eight constants listed in the Table IV are considered to be the independent variables while the remaining five are dependent. The compliance coefficient, K_c, is given by (16) as the ratio of V_T to β.

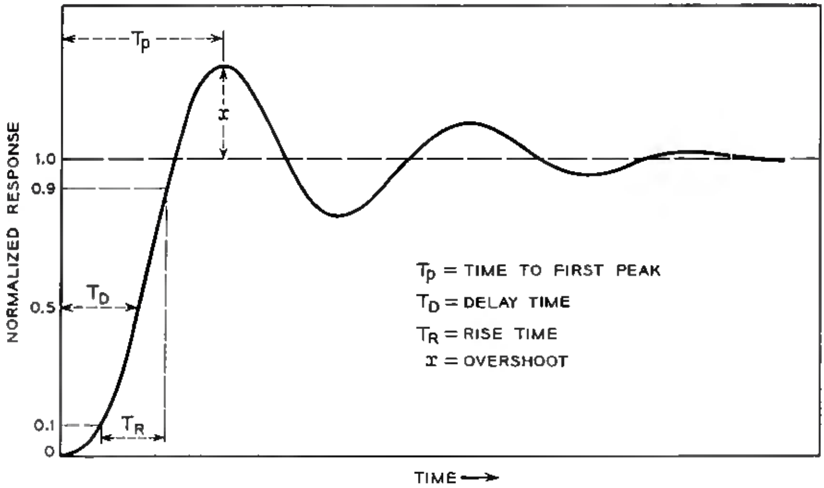


Fig. 15 — Definitions of transient symbols.

The effective leakage coefficient, L_a, is given by (50) as

$$L_a = L_m + \frac{K_b \epsilon_0}{\sqrt{2p_s}} \tag{71}$$

and the undamped natural frequency is given by (44) as

$$\omega_n = V_m \sqrt{\frac{2\beta}{V_T J}}. \tag{72}$$

Table III — Equations Solved by Analog Computation

$$\begin{aligned}
Q_1 &= L_m(p_1 - p_2) + V_m\dot{\theta}_0 + K_c\dot{p}_1 \\
Q_2 &= -L_m(p_1 - p_2) - V_m\dot{\theta}_0 + K_c\dot{p}_2 \\
(p_1 - p_2)V_m &= J\ddot{\theta}_0 \\
\epsilon &= \theta_i - \theta_0 \\
p_1 &\geqq 0 \quad \text{and} \quad p_2 \geqq 0 \\
Q_1 &= Q_{O1} - Q_{E1} \\
Q_2 &= Q_{O2} - Q_{E2} \\
Q_{O1} &= \begin{cases} K_b \mid \epsilon + \epsilon_0 \mid \sqrt{\mid p_s - p_1 \mid}\,\operatorname{sgn}(p_s - p_1) & \epsilon > -\epsilon_0 \\ 0 & \epsilon < -\epsilon_0 \end{cases} \\
Q_{O2} &= \begin{cases} 0 & \epsilon > \epsilon_0 \\ K_b \mid \epsilon - \epsilon_0 \mid \sqrt{\mid p_s - p_2 \mid}\,\operatorname{sgn}(p_s - p_2) & \epsilon < \epsilon_0 \end{cases} \\
Q_{E1} &= \begin{cases} 0 & \epsilon > \epsilon_0 \\ K_b \mid \epsilon - \epsilon_0 \mid \sqrt{p_1} & \epsilon < \epsilon_0 \end{cases} \\
Q_{E2} &= \begin{cases} K_b \mid \epsilon + \epsilon_0 \mid \sqrt{p_2} & \epsilon > -\epsilon_0 \\ 0 & \epsilon < -\epsilon_0 \end{cases}
\end{aligned}$$

Table IV — Reference Values of Parameters

Definition	Symbol	Value	Units
total inertia	J	2.73×10^{-3}	lb-in-sec^2
motor displacement	V_m	0.0151	in^3
trapped volume	V_T	0.125	in^3
bulk modulus	β	2.22×10^5	lb/in^2
motor leakage coefficient	L_m	0.039×10^{-3}	in^5/lb-sec
supply pressure	p_s	1000	lb/in^2
gain constant	K_b	0.0912	in^4/lb$^{1/2}$-sec-rad
open-center constant	ϵ_0	0.0561	radians
		3.21	degrees
compliance coefficient	K_c	0.0563×10^{-5}	in^5/lb
effective leakage coefficient	L_a	0.1533×10^{-3}	in^5/lb-sec
resonant frequency	ω_n	544.7	rad/sec
		86.6	cps
damping ratio	ζ_a	0.5	
gain ratio	$\dfrac{\omega_a}{\omega_n}$	0.496	

The dimensionless damping ratio ζ_a [from (51)] is

$$\zeta_a = \frac{L_a}{V_m} \sqrt{\frac{J\beta}{2V_T}}, \tag{73}$$

and the velocity gain constant is obtained from (46):

$$\omega_a = \frac{2K_b}{V_m} \sqrt{\frac{p_s}{2}}. \tag{74}$$

Fig. 16 shows the theoretical frequency response of the reference servomechanism as a function of the input amplitude. The linear prediction (based on small amplitudes and pressure differentials) is included for comparison. The amplitude sensitivity in the small signal region, as represented for example by a curve of 1° amplitude, is the result of the nonlinear flow characteristics of the valve. The operation is within the region in which $|\epsilon| < \epsilon_0$ (A region) and pressure saturation has not occurred.†

The response for an input of 2° shows more deviation from the linear response, primarily because of pressure saturation; operation is still within the A region. For greater input amplitudes, the response falls off

† A −12 db per octave slope that passes through zero db at the frequency

$$[f_{\text{lim}}]_{0\,\text{db}} = \frac{1}{2\pi}\sqrt{\frac{p_s V_m}{J\theta_i}}$$

divides the graph into regions that represent the saturating and nonsaturating conditions.

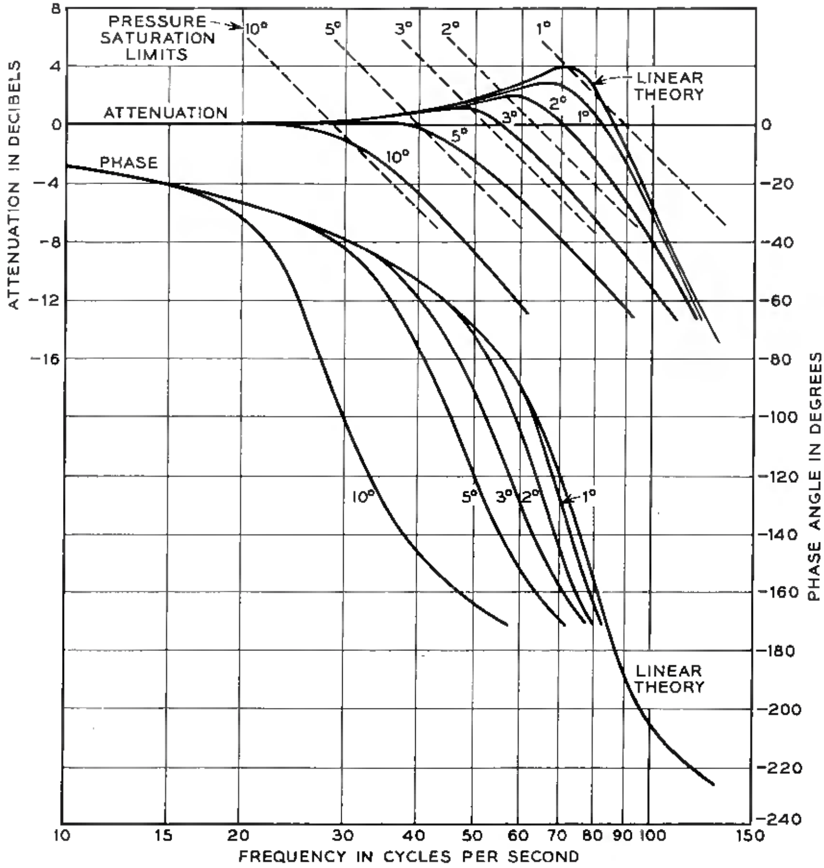


Fig. 16 — Theoretical frequency response of a hydraulic servomechanism as a function of input amplitude.

rapidly and operation is primarily in the B region; pressure saturation effects become more pronounced.

The closed-loop transient response of the reference system as a function of step magnitude is given in Fig. 17, together with the linear solution. For small amplitudes (less than 3°), the linear and nonlinear solutions are essentially identical; as the amplitude is increased, the discrepancy becomes large. The per cent overshoot decreases with amplitude.

It is evident that the transient response of the servomechanism is not as sensitive to amplitude as is the frequency response; for example, comparison of the 3° curves in Figs. 16 and 17 shows that the transient

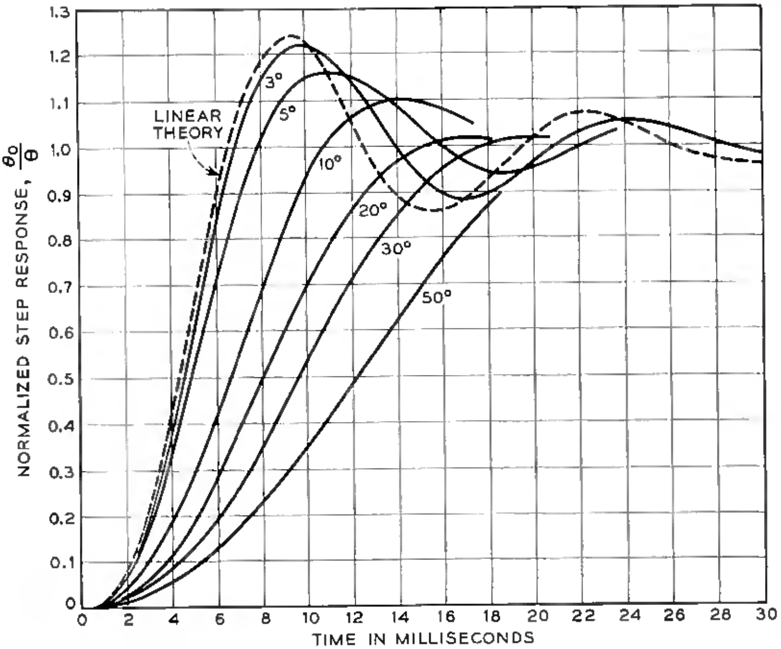


Fig. 17 — Theoretical transient response of hydraulic servomechanism as a function of amplitude.

is much closer to the linear transient solution than the frequency response is to the linear frequency response. This difference is attributable mainly to pressure saturation. Fig. 18 shows the transient response of the system for the 3° step, together with the pressures p_1 and p_2. Since the supply pressure is 1000 psi and the valve is symmetrical, both pressures are initially 500 psi in the quiescent condition. For the case shown, the applied step was in the positive direction, so that θ_0 was also positive; consequently, p_1 has an initial positive slope while p_2 has a negative slope. From 0 to 2 milliseconds, the oil is compressed and very little shaft rotation occurs; maximum acceleration $(p_1 - p_2)$ occurs at 2 milliseconds, at which time the shaft has acquired an appreciable velocity. From 2 to 5 milliseconds, the acceleration decreases from maximum to zero, while the velocity continues to increase to its maximum value. From 5 to 6.8 milliseconds, the acceleration becomes negative, since p_2 is now greater than p_1. The velocity is still positive for this period, and the error signal ϵ decreases from a positive value to zero. The increase in p_2 is due to the compression of the oil in line 2 by the moving inertia, as the orifice area opening to the sump is gradu-

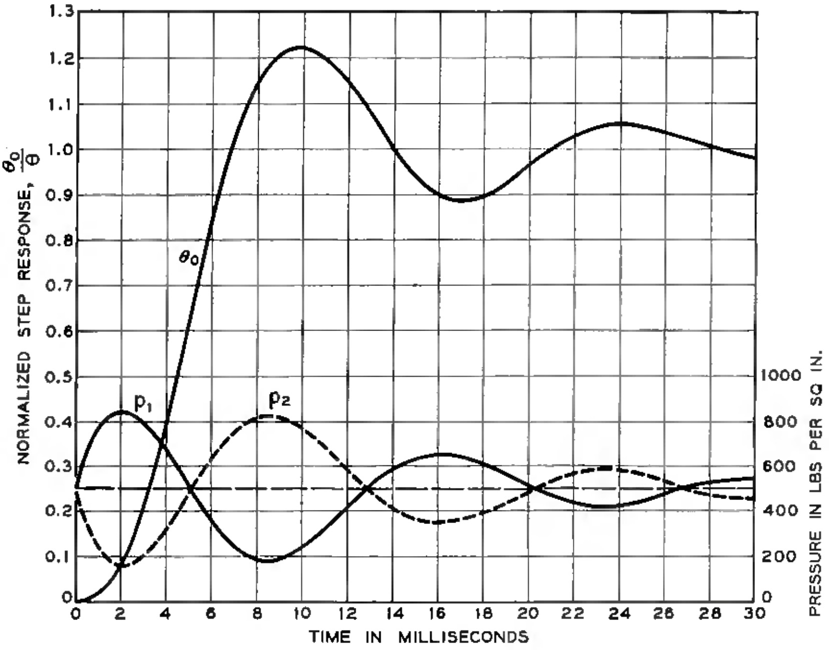


Fig. 18 — Output and pressure variation with time for a 3° step.

ally reduced and the pump pressure is applied to this side of the motor. For the period from 6.8 to 9.5 milliseconds, the error ϵ is negative and the velocity decreases to zero at the peak of the curve.

The nature of the response after 9.5 milliseconds is very similar to that previously described, since the output starts from rest with an initial error. The major difference, aside from the fact that the error is now negative, is that the pressures have appreciable values at 9.5 milliseconds, while at 0 milliseconds the pressures started from the quiescent state.

It should be noticed that the pressures p_1 and p_2 did not limit for the 3° step and that, although the pressure differentials were appreciable, the linear theory still provided a good approximation to the output motion, as indicated in Fig. 17. The value of ϵ_0 as given in Table IV is 3.2°, so that operation was entirely within the A region.

Fig. 19 shows the transient response and pressures for the 20° step. In this case, the pressures just limit at the start, and the pressure differentials are large. In Fig. 20, for the 30° step, the initial pressure saturation is more pronounced and, in addition, p_1 reaches zero and p_2 obtains

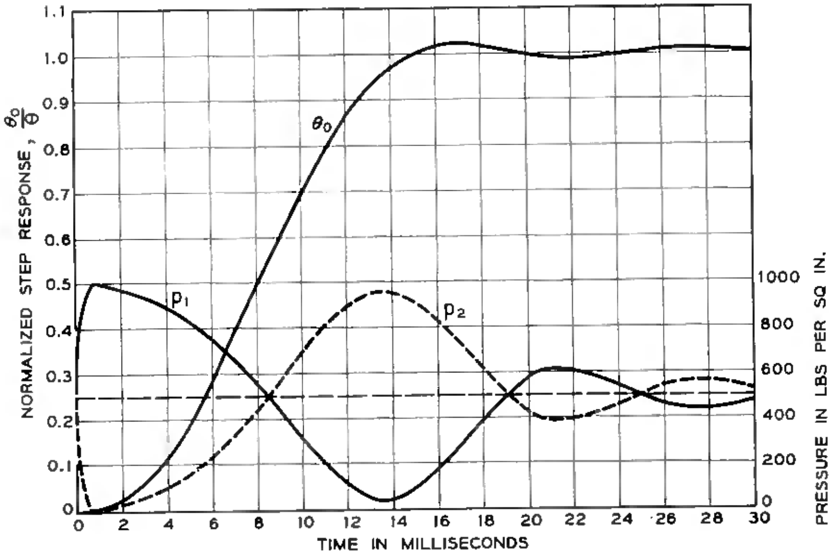


Fig. 19 — Output and pressure variation with time for a 20° step.

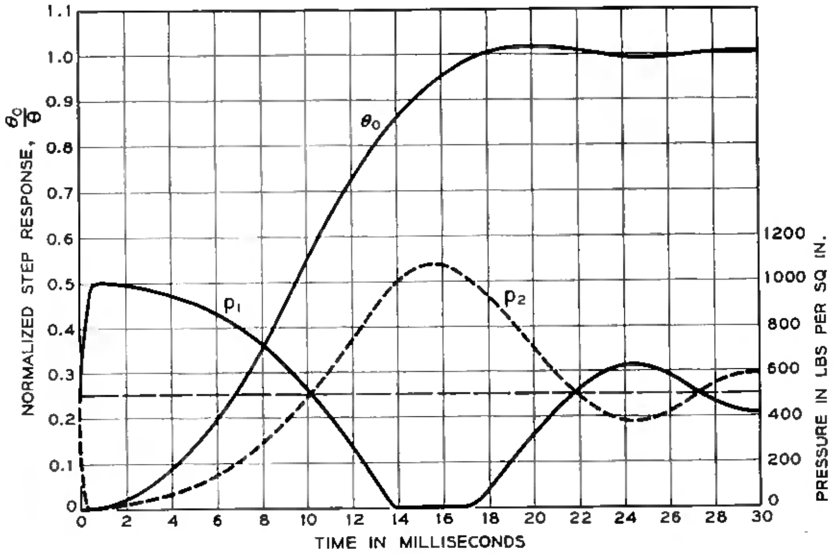


Fig. 20 — Output and pressure variation with time for a 30° step.

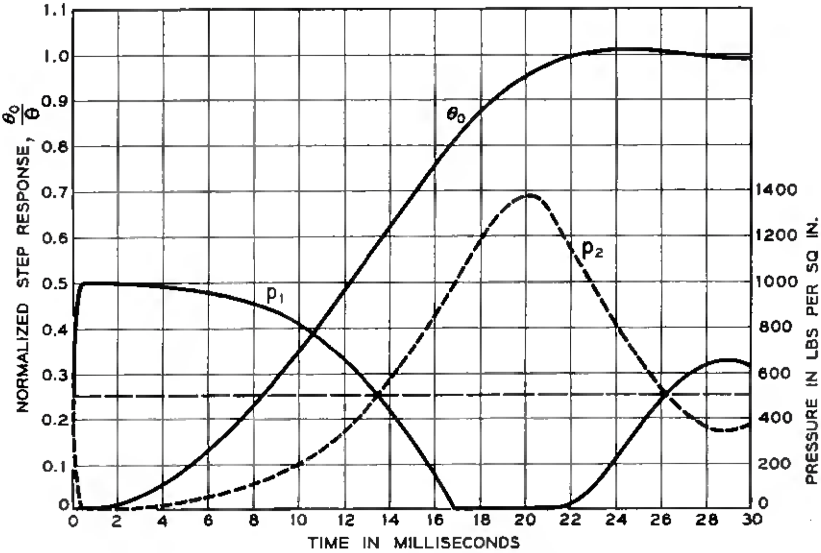


Fig. 21 — Output and pressure variation with time for a 50° step.

a value above that of the supply pressure ($p_s = 1000$ psi). Examination of the curves shows that the maximum value of p_2 of 1080 psi occurs for ϵ positive, the velocity positive and the acceleration negative. This indicates that the inertia of the motor and load combination is forcing oil to flow out of the exhaust orifice and that, as a result, the pressure p_2 achieves a very high value. At the same time, oil is forced through the orifice connecting the pressure supply to line 1, but the velocity is so great that the rate of flow is not sufficient to maintain a pressure in this line. Thus, during the period from 14 to 17 milliseconds cavitation conditions exist on this side of the motor.

In Fig. 21, for the 50° step, the situation is similar to that of Fig. 20, except that the saturation and cavitation periods are of longer duration. Here, the peak pressure is very nearly 1400 psi.

From the preceding discussion, the desirability of including relief valves in each line is apparent. Dangerously high pressures can be generated, especially if a supply pressure of 3000 psi is used. The inclusion of relief valves in effect limits the line pressure, so that the oil cannot be "trapped" by the inertia. This, however, has the disadvantage of increasing the overshoot for large amplitudes and does not decrease the rise or delay times. In addition, the cavitation period is prolonged.

The effect of the variation in gain, K_b, is shown in Figs. 22 and 23.

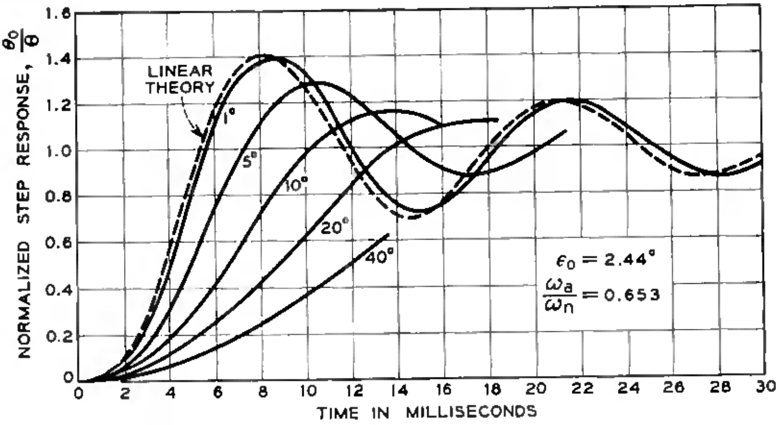


Fig. 22 — Theoretical transient response as a function of amplitude for $K_b = 0.12$.

In Fig. 22, K_b has a value of 0.12, so that the parameters listed on the figure change from those of the reference case; all other parameters remain constant as given in Table IV. Since the open-center constant ϵ_0 is now only 2.44°, it is to be expected that the system exhibit more amplitude sensitivity. Comparison with Fig. 17 shows that this is the case; in the small-signal region, the curves for the higher-gain system differ somewhat more from the linear solution than do those in Fig. 17.

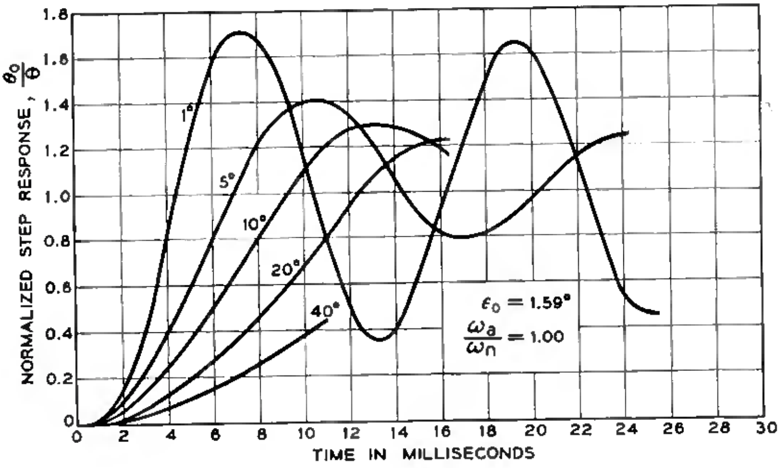


Fig. 23 — Theoretical transient response as a function of amplitude for $K_b = 0.184$.

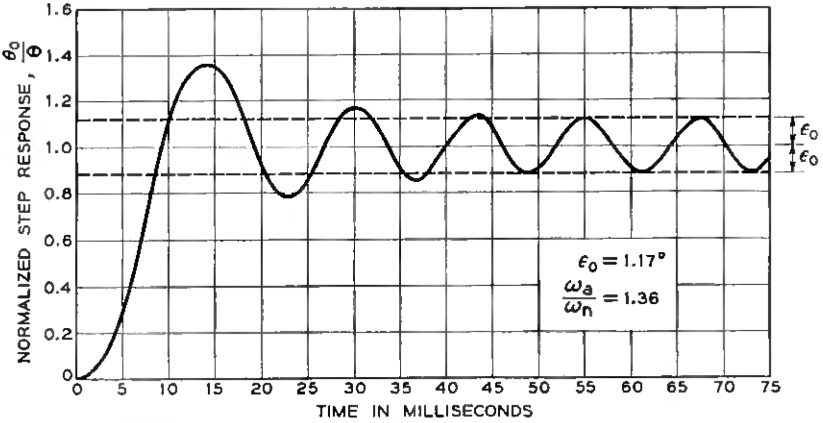


Fig. 24 — Transient response for a 10° step with K_b = 0.25.

As the amplitude increases, this effect is lost, and the major difference is found in the per cent overshoot.

In Fig. 23, the gain constant, K_b, is 0.184. For this value the dimensionless gain ratio is unity and the linear theory predicts zero gain margin. There is a distinct difference in the nature of the response as the amplitude increases. For the 1° step, the operation is entirely within the A region ($\epsilon_0 = 1.59°$) and the oscillations are almost continuous. As the amplitude increases, the per cent overshoot decreases, since pressure saturation occurs.

This effect on the stability is more pronounced at the higher gain values. In Fig. 24, a 10° step is shown for the case in which $K_b = 0.25$ and the gain ratio is 1.36. From the linear theory, this system should be unstable. Examination of the response shows that, in the A region in which ϵ is less than $\epsilon_0 = 1.17°$, the system is unstable, but that it is stable in the B region in which ϵ is greater than ϵ_0. The response, therefore, oscillates indefinitely, but only with the amplitude of ϵ_0. It is evident, therefore, that, if frictional forces are sufficient to overcome the small oscillations of $\pm\epsilon_0$ amplitude, or if these oscillations are not detrimental to the performance in the particular application, the allowable gain is much greater than that predicted by the linear theory. For a given valve, $K_b\epsilon_0$ is a constant, so that, as the gain K_b is increased, the magnitude ϵ_0 of the sustained oscillations decrease. The incremental damping and gain constants for the B region are given by (57) and (59); these equations show that the operation in the B region is inherently more stable than that in the A region, since the incremental damping

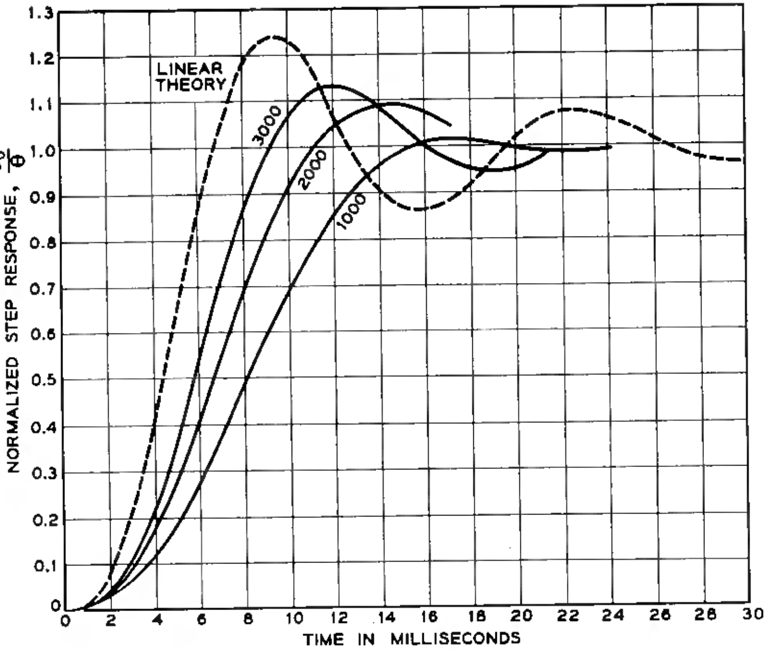


Fig. 25 — Transient response as a function of supply pressure for a 20° step.

is greater than in the A region and the incremental gain is less than the corresponding A region gain. Thus, in the absence of friction or viscous damping, the sustained oscillations are most pronounced in the A region.

The effect of the supply pressure on the transient response is shown in Fig. 25 for a 20° step. The linear theory provides the best approximation for the case in which the supply pressure is greatest. This is the result of two factors: (a) the system with the higher pressure is less susceptible to pressure saturation and (b) the value of ϵ_0 increases with increasing supply pressure, so that operation is more completely in the A region. The solutions show that, whereas the system having a supply pressure of 1000 psi encounters saturation for a 20° step, the 3000-psi servo is not pressure-limited until subjected to a 60° step. The advantage of operating at higher pressures is thus primarily a question of pressure saturation.

In all the previous solutions, the effective leakage coefficient, L_a, has been maintained constant at 0.1533×10^{-3} inches^5 per pound-second, so that the damping ratio, ζ_a, was 0.5. From (71), it is seen that the effective leakage coefficient is the sum of the motor leakage

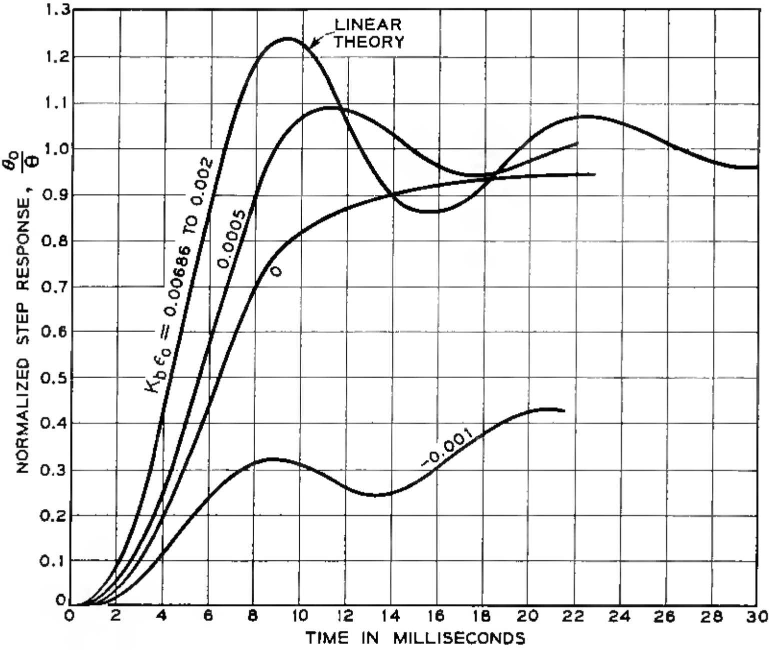


Fig. 26 — Theoretical transient response as a function of $K_b\epsilon_0$ for a 1° step.

coefficient and a term proportional to $K_b\epsilon_0$. It is interesting to observe the response of the system as affected by the nature of the damping. This is given in Fig. 26 for a 1° step. The three upper curves have a damping ratio of $\zeta_a = 0.5$; the difference in the curves results from the method used in obtaining the damping. For cases in which the open-center valve provides appreciable damping, the response does not differ from the linear prediction. This is the result of operation in the A region, where ϵ is less than ϵ_0. As the valve damping decreases and the motor leakage is increased, the response is slower and falls below the linear curve; this occurs for the case in which $K_b\epsilon_0$ is 0.0005. Operation is partly in the B region, since ϵ_0 is equal to 0.31°.

When the damping is contributed entirely by the motor leakage, $K_b\epsilon_0 = 0$ and the valve is of the critical-center type. In this case, Fig. 26 shows that the response does not overshoot and that an increase in gain would be desirable. Computer results show that, for $K_b = 0.20$, the critical-center valve gives a transient response having about 25 per cent overshoot and adequate stability. However, it should be emphasized that the damping ratio for this case was 0.5 and that the damping was obtained by altering the motor leakage. Any attempt to reduce the valve-quiescent flow without compensating the system in some manner to provide additional damping results in an underdamped servomechanism.

The final curve, for which $K_{b\epsilon_0} = -0.001$ in Fig. 26, represents the response to be expected with a closed-center valve. This valve has a dead zone of 0.63° about which the output will wander, and is shown in a particularly poor case, since the step is only 1° and the gain is small. The only damping in this system is contributed by motor leakage, so that the system is underdamped.

VII. CONCLUSIONS

This study shows that the linear approximation to the nonlinear representation of valve-controlled hydraulic servomechanisms can be applied only with the sacrifice of considerable accuracy. However, since the linear theory is readily applied, it can be used in obtaining estimates for preliminary designs if the deviations from the nonlinear solutions are understood.

REFERENCES

1. Schaefer, J. W., An Electrically Operated Hydraulic Control Valve, B.S.T.J., **36,** May 1957, p. 711.
2. Blackburn, J. F. and Lee, S. Y., Contributions to Hydraulic Control — 1. Steady-State Axial Forces on Control-Valve Pistons, Trans. A.S.M.E., **74,** August 1952, p. 1005.
3. Blackburn, J. F., Contributions to Hydraulic Control — 3. Pressure-Flow Relationships for 4-Way Valves, Trans. A.S.M.E., **75,** August 1953, p. 1163.
4. Shearer, J. L., Dynamic Characteristics of Valve-Controlled Hydraulic Servomotors, Trans. A.S.M.E., **76,** August 1954, p. 895.

Some Design Considerations for High-Frequency Transistor Amplifiers

By D. E. THOMAS

(Manuscript received July 10, 1959)

The major problem in the design of high-frequency transistor amplifiers is the interaction between the output and the input of the amplifier caused by the internal feedback of the transistor. This problem is illustrated and the two common design approaches to a solution of the problem are discussed. Nyquist's criterion of stability and Bode's feedback theory are then used to obtain an engineering evaluation of the relative merits of these two design approaches from a stability standpoint. The positive nature of the internal transistor feedback is established in this stability evaluation. Finally, Bode's feedback theory is used to consider the relative merits of some of the broad banding techniques used in transistor video amplifier design. The over-all analysis shows that many of the most practical and stable linear transistor amplifiers are very simple and can be built with a minimum of design effort.

I. INTRODUCTION

A survey of the mass of available literature on high-frequency transistor amplifier design discloses the constantly present problem of amplifier sensitivity and even instability, especially when so-called maximum available gain amplifier designs are attempted. This problem is the result of the internal positive feedback inherent in all known transistors. This paper is particularly directed toward a better understanding of transistor internal feedback and its relationship to transistor amplifier design and performance. A fresh and practical engineering approach to the problem of transistor amplifier sensitivity and stability evaluation is presented. The presentation is largely concerned with basic design principles.* Specific amplifier design discussion is limited to that needed

* The material in the paper covers the basic design principles presented in a talk on "The Design of RF and Video Amplifiers" given by the author as one of a series of six lectures on *Transistors — Their Circuits and Applications*, sponsored by the Dallas, Texas, Section of the Institute of Radio Engineers.

to provide engineering illustrations of these principles. References are then made to published material where more complete details on specific amplifiers can be found.

II. HIGH-FREQUENCY TRANSISTOR CHARACTERIZATION

Before considering amplifier design techniques, we must have some means of characterizing the transistor in terms of its performance as an electrical circuit element.* This paper will rely largely on a small-signal characterization in which the transistor is represented by the generalized equivalent T of Fig. 1 with a single internal generator in the branch corresponding to the collector of the transistor. The details of the impedances, each one of which can be written in terms of lumped constants that are directly relatable to the physical structure of the transistor, will be presented only when needed.

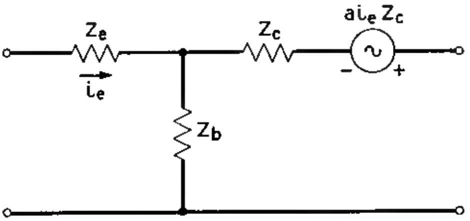


Fig. 1 — Transistor equivalent T.

For simplicity in writing circuit equations, the transistor collector current generator, ai_e, which would normally appear across the collector impedance Z_c, has been replaced by the voltage generator, ai_eZ_c, in series with Z_c, in accordance with Thevenin's theorem. (a is frequency-dependent.) Except for a small phase error, a is closely approximated at frequencies below f_a by the expression

$$a = \frac{a_0}{1 + j\frac{f}{f_a}}, \tag{1}$$

where f_a is the frequency at which the amplitude of a is 3 db below its low frequency value, a_0.[2] For simplicity α, the short-circuit common base current gain will be used interchangeably with a in the discussion to follow.

No parasitic capacitances are shown in Fig. 1, since these will be con-

* For a resume of transistor equivalent circuits, see Pritchard.[1]

sidered as part of the terminating networks, except for the capacitance between output and input — that is, collector-to-emitter capacitance in the common base configuration and collector-to-base capacitance in the common emitter configuration. And, in order to simplify the consideration of the internal feedback effects, these latter capacitances will be neglected except in the discussion of the common emitter neutralized amplifier. However, input-to-output capacitance can be very troublesome, especially in the common base configuration at very high frequencies.

The equivalent T circuit representation illustrated in Fig. 1 is particularly useful in three respects. First, it represents the transistor with sufficient accuracy to be used in generalized circuit evaluation. Secondly, it can be used in any of the three possible transistor connections without change. Finally, since the various components of the circuit are directly relatable to the physical structure of the transistor, the effect of the transistor structure on circuit performance can be better understood, and effects that might otherwise be obscured may be uncovered.

When a precise amplifier circuit design in a particular frequency region is undertaken, a four-pole parameter circuit equivalence may be more accurate and more convenient.* However, this paper will make only limited use of this type of characterization for two reasons. First, the examination of amplifier stability, which is one of the major objectives of the paper, is more easily accomplished with the equivalent T of Fig. 1. In fact, the positive nature of the internal feedback of the transistor is not apparent in the hybrid parameter four-pole analysis of the common emitter transistor configuration. This is because the positive feedback is concealed in the forward transfer current ratio, h_{21e} . Secondly, many electronic circuit engineers are more accustomed to the two-terminal design techniques of vacuum tube circuitry below the UHF region. And this paper shows that those amplifiers that can be built on a two-terminal basis with limited impedance measurements and slide rule computations are often the better transistor amplifiers.

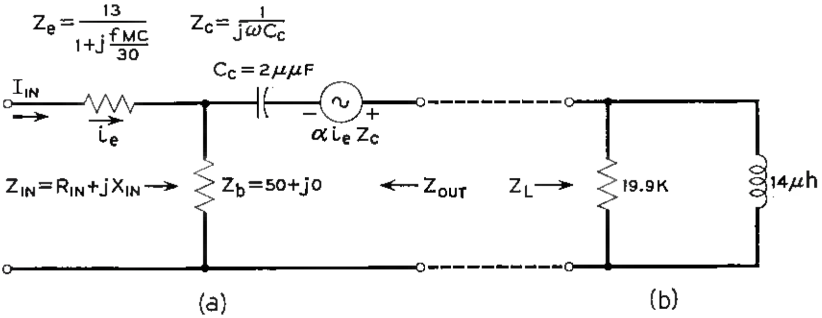
III. RADIO FREQUENCY AMPLIFIER DESIGN

In designing a radio frequency transistor amplifier, the immediate problem is to determine the proper choice of terminal networks for the transistor to obtain the greatest possible gain consistent with the other requirements on the amplifier. The first approach to a solution is given by linear network theory. A conjugate-matched-impedance generator

* For a presentation of the more common four-pole parameter equivalences, see Linvill and Schimpf.[3]

should be connected to the input and a conjugate-matched-impedance load should be connected to the output. However, since the transistor itself is a network of complex impedances ("complex" is used here and hereafter in the sense of having real and imaginary parts) and contains an internal generator that is a complex function of frequency, and also has built-in internal feedback, the required generator and load impedances are themselves complex. To say that the determination of these required impedances is difficult is a gross understatement. Even the computation of the transistor gain between known complex impedances becomes unduly complicated.

An alternate approach to the determination of suitable generator and load impedances is therefore used. The power delivered to the load with either a constant-current or a constant-voltage input generator is determined, and the input-matching problem is then considered separately. This approach will be illustrated by considering an elementary design of a 4-mc wide, 30-mc center-band frequency common base IF amplifier, using a 30-mc alpha-cutoff-frequency germanium transistor. The simplified equivalent T circuit of the transistor is shown in Fig. 2(a). The load impedance should be conjugately matched to the output impedance of the transistor with an open circuit input, since a current generator is being assumed at the input. This impedance is closely approximated by the reactance of the collector junction capacitance, C_c. A positive reactance equal to the negative reactance of C_c at the center-band frequency of 30 mc is therefore chosen as the reactance portion of the load. This is the 14-microhenry inductance of Fig. 2(b). Since the resistance component of the transistor output impedance is very small, bandwidth considerations rather than matching determine the resistance component of the load. A 4-mc bandwidth centered at 30 mc calls for a 19.9K



$Z_e = \frac{13}{1+j\frac{f_{MC}}{30}}$
$Z_c = \frac{1}{j\omega C_c}$
$C_c = 2\mu\mu F$
I_{IN}
i_e
$\alpha i_e Z_c$
$Z_{IN} = R_{IN} + jX_{IN}$
$Z_b = 50 + j0$
Z_{OUT}
Z_L
19.9K
14μh
(a)
(b)

Fig. 2 — Transistor 30-mc single-stage IF amplifier: (a) transistor equivalent circuit; (b) amplifier load impedance.

shunt resistance, as shown in Fig. 2(b). Now, in accordance with the design plan, an input generator conjugately matched to the transistor input impedance with the output terminated in the selected load should be provided to complete the single-stage amplifier design. This input impedance, Z_{IN}, was computed, and its resistance and reactance components are plotted in Fig. 3(a) and 3(b) respectively. The resistance component is seen to vary by a factor of 5 to 1 throughout the desired band, and actually becomes negative at frequencies just below the bottom of the band. The reactance component likewise varies widely throughout the band, going from approximately 200 ohms at the bottom to zero at the top of the band. Anything but a conjugately matched generator at the input would distort the bandpass characteristic designed into the load impedance. Since this generator would have to incorporate the output impedance of the preceding transistor in a multistage amplifier, plus a suitable impedance transformation to obtain gain, its design would be at best very complicated. The design is therefore in serious trouble. An understanding of the source of the trouble is essential to a solution to the problem.

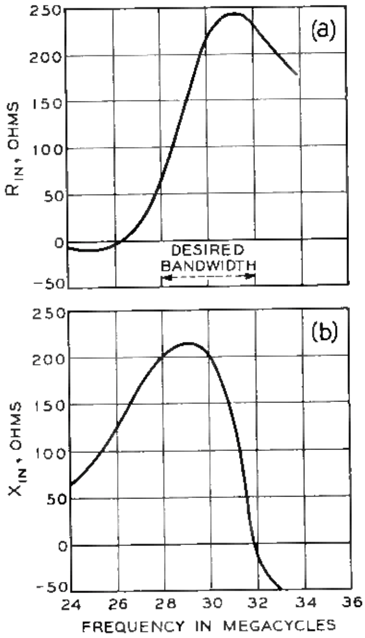


Fig. 3 — Input impedance for amplifier of Fig. 2: (a) resistance component; (b) reactance component.

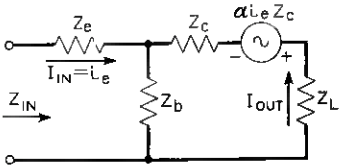


Fig. 4 — Equivalent T for common base connection with load Z_L .

The difficult nature of this input impedance is a direct result of the internal feedback in the transistor. Fig. 4 shows the equivalent circuit of Fig. 1 in the common base connection terminated with a constant-current generator input, I_{IN} , and a load impedance, Z_L . The input impedance Z_{IN} is given by

$$Z_{IN} \doteq Z_e + Z_b\left(1 - \frac{\alpha}{1 + \dfrac{Z_L}{Z_C}}\right). \tag{2}$$

In the common emitter connection, the other of the two more commonly used transistor configurations, the corresponding input impedance is given by

$$Z_{IN} \doteq Z_b + Z_e\left(1 + \frac{\alpha}{(1-\alpha)\left[1 + \dfrac{Z_L}{Z_C\,(1-\alpha)}\right]}\right). \tag{3}$$

Equations (2) and (3) show that, regardless of the common connection, the input impedance to the transistor is a function of the load impedance and the common base short-circuit current gain α, both of which are, as a rule, complex. And so the complicated complex input impedance that was uncovered in the amplifier example above follows naturally. Even if the design problem were not so complicated, and if physically realizable impedances with the proper impedance transformation for interstages of multistage amplifiers could be built, the alignment problem of the multistage amplifier would be an extremely difficult one. This is verified in the large mass of technical literature discussing interstage alignment and band-skewing problems as a result of adjacent interstage interaction. It is therefore apparent that, before practical high-frequency transistor amplifiers can be built, it is necessary to reduce the effect of the load impedance on the input impedance to a point where it is no longer a serious problem. This can be done either by "neutralization"* or by output-to-load-impedance mismatch.

* Neutralization is placed in quotation marks to call attention to the fact that it is quite different from neutralization as we know it in vacuum tubes. The characteristics of transistor neutralization will be discussed in more detail later.

IV. NEUTRALIZED AMPLIFIER DESIGN

The neutralized solution to the input-output impedance interaction will be considered first. The common emitter connection will be used, since this is the more common neutralized configuration. This is because more gain is obtainable in this connection at frequencies below the common base cutoff frequency of the transistor. Fig. 5 shows the transistor equivalent circuit of Fig. 1 in the common emitter connection, with a neutralizing impedance, Z_N, connected between the collector output and the base input and with the base input open. An external generator, V_g, is connected between the collector and the common emitter terminals. The following equations define the voltage and current relations of the circuit of Fig. 5:

$$\begin{aligned} i_1 (Z_e + Z_c - \alpha Z_c) + i_2 Z_c \qquad\qquad &= V_g, \\ i_1 (Z_c - \alpha Z_c) \qquad + i_2(Z_b + Z_c + Z_N) &= 0, \end{aligned} \tag{4}$$

$$i_1 = \frac{V_g \Delta_{11}}{\Delta} \qquad i_2 = \frac{V_g \Delta_{12}}{\Delta}, \tag{5}$$

where Δ is the circuit determinant of (4). Then

$$E_1 = i_2 Z_b - i_1 Z_e = V_g \left[\frac{Z_b \Delta_{12} - Z_e \Delta_{11}}{\Delta} \right], \tag{6}$$

which gives the input voltage, E_1, in terms of the output generator, V_g. If E_1 is made zero regardless of the value of V_g, the input impedance is then independent of the load voltage and therefore of the load impedance when the amplifier is terminated at its output. Solving for the value of Z_N required to make $E_1 = 0$ gives

$$-Z_N = Z_c + \frac{Z_b}{Z_e} Z_c(1 - \alpha) + Z_b. \tag{7}$$

The required neutralization impedance, Z_N, turns out to be negative, which indicates that a phase reversal is needed in the neutralization

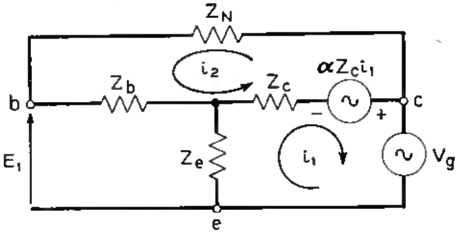


Fig. 5 — Equivalent T for common emitter connection with neutralization.

current feedback path in order to produce a positive neutralization impedance, Z_n', which is the negative of Z_n. A phase-reversing transformer is therefore used, as shown in Fig. 6, which gives a generalized schematic diagram of a common emitter neutralized amplifier. A step-down is used between the collector and neutralizing windings of the phase-reversing transformer in order to distribute the effect of the loading of the neutralization impedance between the collector and the emitter. The transformer is tuned to the desired center-band frequency, and the load is the input impedance to the following identical stage in a multistage amplifier. If the approximation of (1) for α and $-j/\omega C_c$ for Z_c are substituted in (7) for Z_n, and if Z_b and Z_e are assumed to be real and constant, then (7) can be solved for Z_n' in terms of its real and imaginary parts:

$$Z_n' = -Z_n = Z_b \left\{ \frac{1}{\omega_\alpha C_c Z_e} \left[\frac{\alpha_0}{1 + \left(\dfrac{f}{f_\alpha}\right)^2} \right] + 1 \right\} - j \frac{1}{\omega C_c} \left\{ 1 + \frac{Z_b}{Z_e} \left[1 - \frac{\alpha_0}{1 + \left(\dfrac{f}{f_\alpha}\right)^2} \right] \right\} \tag{8}$$

As given by (8), Z_n' can be approximated by two resistances and a capacitance throughout a reasonably broad band of frequency, as shown in the network for Z_n' in Fig. 6. The dotted capacitances, C_{cb} and nC_{cb}, of Fig. 6, show how the input-to-output capacitance in the neutralized common emitter amplifier can be compensated for by a corresponding capacitance in the neutralizing impedance. Only this portion of the neutralization corresponds to the neutralization of the output-to-input capacitance feedback in vacuum tubes. The load impedance, Z_L, of Fig. 6 is given in terms of the input admittance to the following transistor

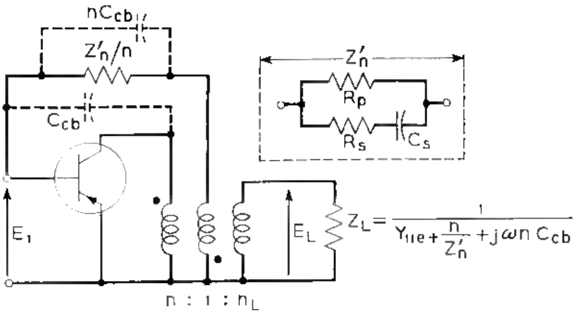


Fig. 6 — Schematic of single-stage common emitter neutralized amplifier.

with its output short-circuited designated as Y_{11e} and the neutralization impedance of the following transistor.

Although (7) gives the neutralization impedance in terms of the transistor equivalent T of Fig. 1 and is useful for qualitative understanding of the neutralization problem, it is not suffiicently accurate to determine the neutralization impedance required for an actual amplifier. A more accurate determination can be made from a four-pole parameter solution to the impedance Z_n' necessary to make Y_{12e}, the reverse transfer admittance of the circuit of Fig. 6, equal to zero. This was the technique used by Webster[4] in determining an expression for Z_n' in connection with the design of one of the best examples of a maximum gain neutralized common emitter amplifier to be found in the literature. However, even the four-pole approach fails to give a satisfactory determination of Z_n' for practical use, and so Z_n' is usually obtained experimentally by adjusting Z_n' until there is no appreciable change in the input impedance to the transistor across the bandwidth of the amplifier when the load is alternately normal and shorted. The input admittance is then given by

$$Y_{IN} = \frac{1}{Z_b + \dfrac{Z_e}{1-\alpha}} + \frac{1}{\dfrac{Z_n'}{n}} + j\omega n C_{cb}\,, \tag{9}$$

which is the common emitter input admittance of the transistor with the collector shorted to the emitter plus the admittance added across the input by the neutralization impedance. The input admittance given by (9) is seen to be independent of the load impedance, and therefore the objective of having input impedance independent of output impedance is achieved.

The load impedance is then conjugately matched to the output impedance of the transistor with the input shorted. The generator is likewise conjugately matched to the input impedance given by (9). The power gain of the transistor can then be easily shown to be given by[4]

$$\text{power gain} = \frac{|\,Y_{21}\,|^2}{4G_{22}G_{11e}}\,, \tag{10}$$

where Y_{21} is the forward transfer admittance of the transistor with the output short circuited, G_{22} is the real part of the output admittance with the input short circuited, and G_{11e} is the real part of the input admittance common emitter with the output short circuited. This is a straightforward computation, since all the parameters are simple functions of the active device only and can be measured on a suitable impedance bridge as discussed by Webster.[4]

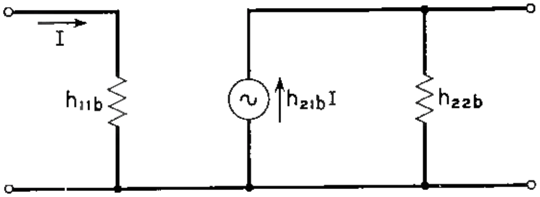


Fig. 7 — Transistor circuit, approximate.

Webster has built a five-stage, 75-db gain, 25-mc center-band frequency IF amplifier using the neutralization technique just discussed. The computed and measured amplifier checked to within 0.5 db in 75 db in gain and to 0.07 mc in 1.6 mc in bandwidth. This is an excellent example of the accuracy of the neutralization amplifier design technique in the prediction of available gain and bandwidth. However, a maximum-gain neutralized amplifier is far from easy to design, since the amplifier represents a delicate balance of feedback effects, which make it difficult to adjust and even more difficult to maintain stable. This will be discussed in greater detail later.

V. MISMATCHED AMPLIFIER DESIGN

The mismatch approach to making the input impedance independent of the output impedance will next be considered. Here we will use the common base connection for our discussion, since this connection has been most frequently used for mismatched RF amplifiers. However, mismatched common emitter RF amplifiers are becoming more frequent, due to the high-cutoff-frequency diffusion transistors currently available. The same principles apply to both types. A reexamination of (2) shows that the common base input impedance can be made substantially independent of load impedance if the load impedance, Z_L, is made small compared to the collector impedance, Z_c. This, of course, involves a loss of gain, but a considerable mismatch can be taken with a relatively small loss of gain. For instance, a 5-to-1 mismatch results in a gain reduction of less than 3 db, and a 10-to-1 mismatch results in a reduction of only 5 db.

With sufficient mismatch to make the input impedance essentially independent of the output impedance, the common base equivalent circuit of the transistor with a constant current generator is given in Fig. 7.* In this circuit, h_{11b} is the impedance looking into the input of the com-

* The equivalent circuit of Fig. 7 has been referred to by Linvill[3] as a "circuit approximate".

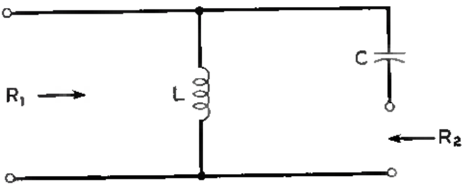


Fig. 8 — Single-tuned reactance-network transformer.

mon base transistor with output short circuited, h_{21b} is the ratio of the common base short-circuit output current to the input current and h_{22b} is the admittance looking into the common base output with the input open.

Even if the degree of mismatch is not great, the circuit of Fig. 7 gives a fair approximation to the true circuit. The approximation is sufficiently good to rough out interstage coupling networks, which can then be adjusted on the bench in the laboratory. Since there is little interaction of the output load circuit on the input impedance, the gain for a given load and generator impedance can be easily computed, in the same manner in which the gain is computed for a neutralized amplifier. It will be somewhat simpler, since no neutralization impedance is present. It must be remembered, however, that the output is no longer matched when the output power is computed.

Since a reversing transformer is not needed for neutralization purposes, simple impedance transformation between the high-impedance collector output of one stage and the low-impedance emitter input of the following stage can be used. The simplest type of impedance transformation corresponding to a single-tuned transformer is shown in Fig. 8. If R_2 is the resistance component of the impedance looking into the emitter of the following stage, and if the reactance component of the impedance is combined with the reactance of C, the load impedance R_1 facing the collector of the preceding stage at center frequency is increased to Q^2R_2, where Q is the ratio of the band center frequency reactance of either C or L to R_2. This simple circuit has the disadvantage that the circuit Q, and consequently the transmission bandwidth, and the transformation ratio are not independent. Therefore, a double-tuned reactance transformation network equivalent to a double-tuned transformer is usually used in the interstage.

Fig. 9 gives the schematic circuit of a single stage of a 70-mc germanium tetrode mismatched amplifier designed by Schimpf[3] using a double-tuned reactance transformation network. The short-circuited input impedance to the transistor was of the order of 75 ohms. This impedance also had a reactance component which varied somewhat throughout the transmitted band, but the variation was not sufficient to seriously complicate the band-width adjustment of the interstage network. The high-side impedance looking into the coupling network was approximately 1500 ohms, which gave sufficient impedance transformation ratio to provide substantial stage gain (approximately 9 db per stage) but presented sufficient mismatch to the collector of the preceding transistor to meet the requirements of making the input impedance substantially equal to the short-circuited input impedance — or, in other words, independent of the load. At the time this amplifier was designed, 70-mc impedance-measuring equipment was not available to the designer. Therefore, judicious extrapolations were made from measurements on a radio frequency bridge at frequencies of 30 mc and below. The interstage transformation network was then designed and built with the adjustable elements shown in Fig. 9. The circuit was then bench-adjusted in the laboratory with a sweep-frequency signal generator and a high-frequency oscilloscope across the load. This is the technique that was referred to earlier when it was stated that excellent amplifiers can be built without complicated impedance measurements and a minimum of slide rule computations.

The relative independence of this circuit design technique on transistor parameters was dramatically demonstrated when Schimpf placed one of the first research models of the germanium diffused-base transistor in a circuit that, except for the omission of the second base of the tetrode, was substantially identical with the circuit of Fig. 9. In spite of the wide difference in electrical characteristics of the diffused-base and tetrode

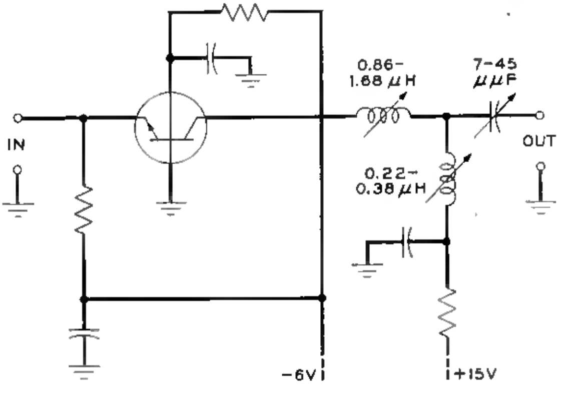


Fig. 9 — Single-stage 70-mc mismatched IF amplifier.

transistors, the circuit was alignable to give a 20-mc IF band centered at 70 mc. Because of the superior high-frequency performance of the diffused-base transistor, the stage gain was 14 db, as compared with 9 db for the tetrode, and the gain was flat to ±0.1 db across the 20-mc band.

VI. TRANSISTOR AMPLIFIER STABILITY

Up to now nothing has been said about the stability of the two types of amplifiers which have been described. And since stability of broadband transistor amplifiers is one of the most important design considerations, the relative stability of the two types of amplifiers that have just been described will now be considered.

Bode[5] has pointed out that the stability of any active network can be determined in terms of the positions of its poles and zeros in the complex frequency plane. However, if we have a known structure whose gain characteristics are satisfactory, it is a long and tedious process in general to determine whether the roots of the structure meet the stability requirement. Furthermore, if the structure is not stable this approach does not necessarily tell us what to do to make it stable. What is needed, therefore, is some means of transferring the restrictions on the poles and zeros into equivalent restrictions on the behavior of the circuit at real frequencies. This we have in the Nyquist criterion of stability,[6] which is used so effectively in the design of negative feedback amplifiers and which it is proposed that we use in the evaluation of the stability of our transistor amplifiers.

The Nyquist criterion of stability is simply stated. The open-feedback loop gain of a feedback amplifier — usually referred to as $\mu\beta$ — is determined in magnitude and phase across a frequency band broad enough to include all frequencies at which the gain is greater than 0.1 in magnitude. The individual values of magnitude and phase are then plotted in polar coordinates and connected to form a closed loop terminating close to the origin. If this loop encloses the point (1,0), the amplifier is unstable; if it does not, it is stable.[6] However, the external gain of the amplifier may be extremely sensitive to changes in amplifier components at frequencies where $\mu\beta$ is in the close vicinity of the (1,0) point, normally called the Nyquist point. In soundly designed negative feedback amplifiers the Nyquist plot approaches the Nyquist point only at frequencies well outside the useful frequency band of the amplifier.

Suppose now that the Nyquist criterion of stability is used to examine critically the stability of our transistor amplifiers. It is generally known that the transistor has built-in feedback, due to its internal base resist-

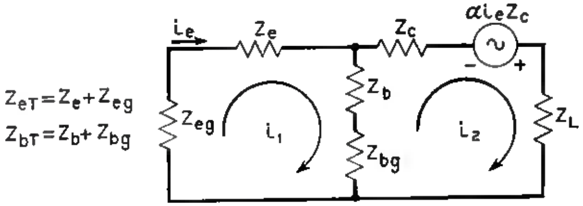


Fig. 10 — Generalized transistor amplifier equivalent circuit.

ance. Furthermore, the behavior of the common emitter transistor amplifier is analogous to that of the common cathode vacuum tube amplifier, since there is a reversal of phase of signal from input to output. Therefore, a reduction in forward gain occurs when a portion of the output signal is fed back to the input through a nonphase-reversing external circuit. As a result, the common emitter configuration is usually considered to be a negative feedback connection.* However, although the incremental feedback due to the external feedback path is negative, the residual or net feedback of the transistor considered as a single-stage amplifier is still positive. In fact, unless external feedback is applied through a suitably phased impedance-matching transformer or other active amplifying devices, a single-stage transistor amplifier—or a single stage of a multistage transistor amplifier—is always of itself a positive feedback amplifier, regardless of which of its elements is made the common connection of the stage.

The positive nature of the feedback is demonstrated in Fig. 10. Here the generalized equivalent circuit of a transistor is shown with generator impedances, Z_{eg} and Z_{bg}, in the emitter and base circuits respectively, and a load impedance, Z_L, in the collector circuit. This load impedance could just as well have been made a generator impedance, thereby making the circuit completely general for any transistor connection. The equations relating the voltages and currents of Fig. 10 are

$$\begin{aligned} i_1(Z_{eT} + Z_{bT}) - i_2 Z_{bT} &= 0, \\ -i_1(\alpha Z_c + Z_{bT}) + i_2(Z_{bT} + Z_c + Z_L) &= 0, \end{aligned} \tag{11}$$

where $Z_{eT} = Z_e + Z_{eg}$, the total impedance in the emitter, and $Z_{bT} = Z_b + Z_{bg}$, the total impedance in the base. The feedback loop gain $\mu\beta$ for this circuit is given by

$$\mu\beta = 1 - \frac{\Delta}{\Delta_0} \doteq \alpha \frac{Z_{bT}}{Z_{eT} + Z_{bT}} \frac{Z_c}{Z_c + Z_L}, \tag{12}$$

* This misconception has been strengthened by hybrid four-pole analysis of the common emitter transistor, since the positive feedback is concealed in the forward transfer parameter or common emitter short-circuit current gain.

where Δ is the circuit determinant of (11) and Δ_0 is the circuit determinant when the active generator, α, is 0. If all the impedances except Z_c are resistive and Z_L is somewhat smaller than Z_c, the feedback loop gain below the common base cutoff frequency will always fall in the right half of the Nyquist polar plot, indicating that the feedback is positive. Note that the choice of the common transistor connection does not influence this result.* In the common emitter iterative amplifier, where the total impedance in the base is much greater than the total impedance in the emitter and the load impedance is small compared to the collector impedance, the feedback loop gain is given by

$$\mu\beta \doteq \alpha = \frac{\alpha_0}{1 + j\dfrac{f}{f_\alpha}}, \tag{13}$$

or the common base short-circuit current gain of the transistor.† Since α has a frequency characteristic which for purposes of discussion can be approximated by an RC cutoff as shown in (13), the Nyquist diagram for this common emitter amplifier becomes a semicircle of diameter α_0, with its center at $\alpha_0/2$ on the zero phase axis and situated below the zero phase axis as shown in the Nyquist plot (a) of Fig. 11.‡ As α_0 approaches unity, a desirable characteristic in a common emitter amplifier, the Nyquist diagram approaches the Nyquist point, (1,0).

The high- and low-frequency cutoff portions of the Nyquist diagram for a soundly designed negative feedback amplifier are also shown in the Nyquist plot (b) of Fig. 11. Note that the negative feedback amplifier stays out of the shaded area bounded by the ±30° axes and gain magnitude greater than 0.5. This shaded area represents the stability margins usually maintained for well-designed negative feedback amplifiers, and corresponds to a loop gain of less than 0.5, or -6 db when the loop phase is between ±30°. This requirement is strictly for stability margins against oscillation in the frequency regions where positive feedback occurs and, these regions are well above or well below the operating amplification band of the amplifier. In contrast, the useful amplification band of our common emitter amplifier falls on that portion of its $\mu\beta$ diagram

* The positive nature of the internal transistor feedback regardless of the common terminal of the transistor has been confirmed by R. B. Blackman of the mathematical research department of Bell Telephone Laboratories.

† Equation (13) neglects a passive component of the feedback loop gain or return ratio which is negligibly small.

‡ The Nyquist plot should include $\mu\beta$ plotted with its imaginary part negative of normal as well as normal. This returns the loop to zero for amplifiers whose gain is not zero at dc as in the present case. However, since this type of plot merely gives a mirror image across its 0–180° axis with the imaginary part of $\mu\beta$ having its normal sign, this half of the plot is not usually shown.

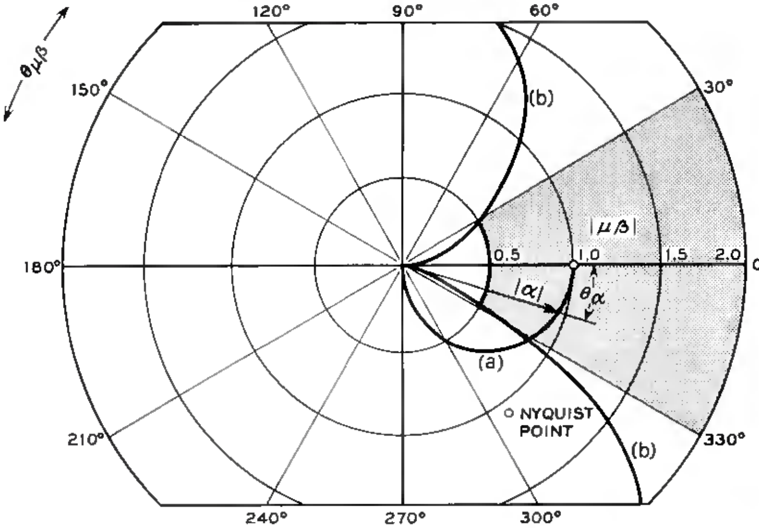


Fig. 11 — Nyquist diagram for amplifier of Fig. 10 with $Z_{bT} \gg Z_{et}$ and $Z_L \ll Z_c$

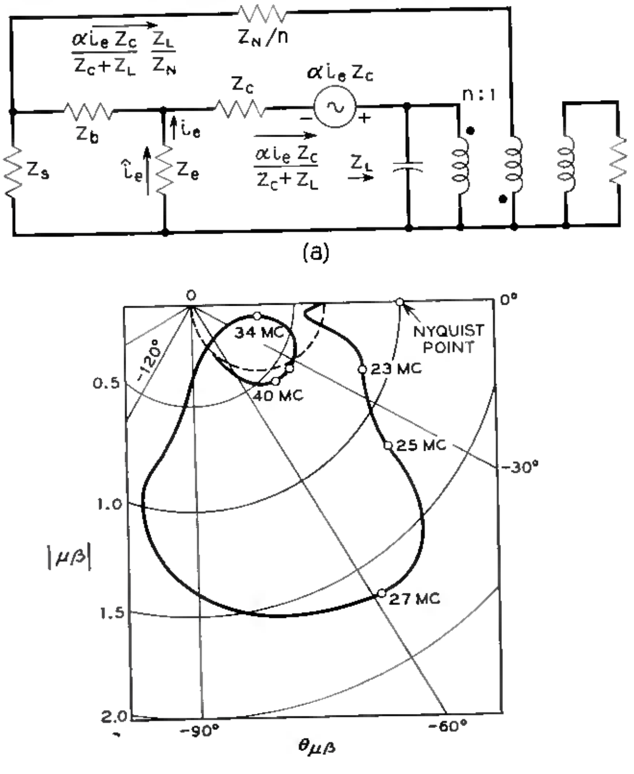
that is in closest proximity to the Nyquist point. The only reason why this is tolerable is because α_0 is a function of the physical structure of the device and because a well-designed and well-behaved junction transistor can be counted upon to stay reasonably constant. In any event, if the dc biases are held reasonably constant, α_0 can be expected to remain less than unity where circuit oscillation will not occur. However, even though there is little danger of oscillation, the positive nature of the inband feedback of the transistor is the basis of the stability problem in high-frequency amplifier design, as will be shown.

Fig. 12(a) shows the schematic diagram of a single-stage neutralized amplifier having a 4-mc bandwidth centered at 25 mc. This amplifier is similar to the amplifier designed by Webster.[4] The feedback loop gain of the amplifier has been computed using a mathematical trick for opening the feedback loop suggested by Blackman.[7] This trick consists of inserting a generator current, i_e, in the emitter and computing the current returned to the emitter, $\hat{i}_e$, through the two feedback paths — the internal feedback of the transistor and the feedback through the neutralization impedance, Z_N/n. The significance of the "^" is that the current so designated is not reamplified by the current gain α of the transistor or, in other words, that the loop transmission is mathematically stopped at a single round trip. The feedback loop gain is then given by $\hat{i}_e/i_e$, and can be written by inspection from the schematic circuit of

Fig. 12(a). This feedback loop gain is given to a close approximation by

$$\begin{aligned}\mu\beta &= \frac{i_e}{\hat{i}_e} = |\,\mu\beta\,|\,\theta_{\mu\beta}\\ &\doteq \frac{\alpha Z_c}{Z_c + Z_L}\,\frac{Z_b + Z_s}{Z_b + Z_s + Z_e} + \frac{\alpha Z_c}{Z_c + Z_L}\,\frac{Z_L}{Z_N}\,\frac{Z_s}{Z_b + Z_s + Z_e} \qquad (14)\\ &\doteq \frac{\alpha Z_c}{Z_c + Z_L}\left[1 + \frac{\dfrac{Z_L Z_s}{Z_N} - Z_e}{Z_b + Z_s + Z_e}\right].\end{aligned}$$

The loop gain given by (14) was computed across a band of frequencies extending well above and well below the pass band of the amplifer. These



(a)

(b)

Fig. 12 — Stability evaluation of neutralized amplifier: (a) circuit schematic; (b) Nyquist diagram.

gains were plotted on a polar diagram to give the Nyquist diagram of Fig. 12(b). It will be noted that the Nyquist diagram does not include the Nyquist point (1,0), and therefore the amplifier would be expected to be stable in the sense of being free from oscillation. However, if the sensitivity of the amplifier to changes in transistor parameters caused by normal drift, battery changes or ambient conditions were examined, large changes in gain would be expected, in view of the fact that the feedback loop gain has real parts in excess of unity at angles of the order of 30° within the transmission band. A complete study of sensitivity can be made in accordance with the techniques described (Ref. 5, Chapters 4 through 6). The fact that the transmission gain of this amplifier is sensitive to environmental and transistor parameter changes due to bias shifts has been confirmed experimentally.

At this point, it is well to stop and consider the nature of the feedback through the neutralization impedance, Z_N. The common emitter connection owes its high current gain to the internal positive feedback in the transistor, which was discussed above. The open input–short-circuited output common emitter amplifier has a Nyquist diagram falling very close to the Nyquist critical point (1,0). [See (12) and (13) and Nyquist diagram (a) of Fig. 11.] When finite impedances are placed in the collector and base circuits, the real part of the positive feedback is reduced, or there is an increment of negative feedback introduced. This moves the Nyquist diagram away from the critical positive feedback area, or in the direction of greater amplifier stability. However, the internal feedback residue is still positive. When the neutralization circuit is added for the purpose of removing the dependence of input impedance on load impedance, the direction of the current through the neutralization circuit is such as to cancel the negative increment of feedback mentioned above. It therefore adds a positive increment to the internal positive feedback residue, and moves the total feedback in the direction of the Nyquist critical point or in the direction of lesser amplifier stability. In the iterative common emitter amplifier, the Nyquist diagram was held within the stability requirement by α_0 holding less than unity. In the neutralized amplifier, a small shift in the critical balance between the positive feedback neutralization current and the negative increment of internal positive feedback caused by the finite impedance terminations can move the Nyquist diagram beyond the (1,0) critical point, and the amplifier will oscillate. Anyone who has built a "maximum available gain" neutralized amplifier is aware of the criticalness of this balance and the tendency toward oscillation.[8] At best, the critical feedback balance results in a gain-sensitive amplifier, as pointed out above in the discussion of the Nyquist diagram of Fig. 12(b).

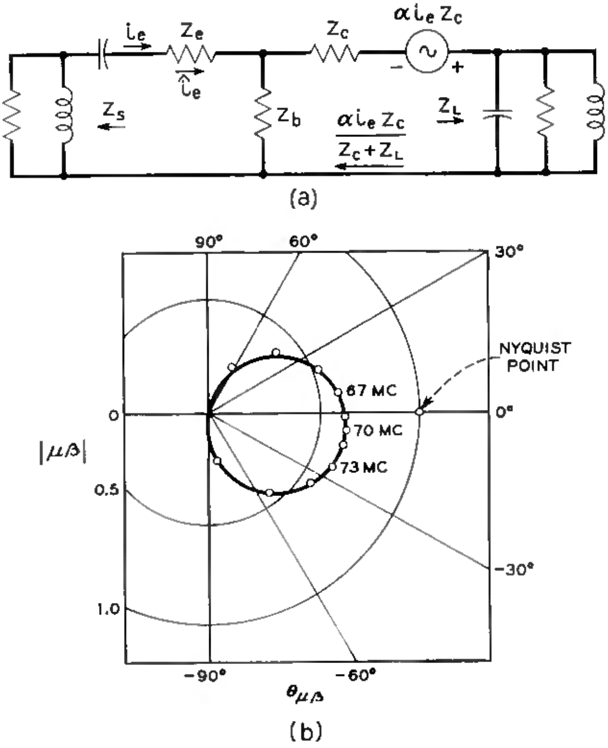


Fig. 13 — Stability evaluation of mismatched amplifier: (a) equivalent circuit; (b) Nyquist diagram.

Now consider the stability of the mismatched type of transistor RF amplifier. Fig. 13(a) shows the schematic circuit of a common base mismatched amplifier using single-tuned reactance coupling of the type referred to above. This amplifier has a transmission bandwidth of 6 mc centered at 70 mc, and a stage gain of approximately 10 db. Looking from the collector into the emitter of the following stage, one sees an effective moderate-Q parallel-resonant circuit whereas, looking from the emitter back toward the collector, one sees a rather high-Q series-resonant tuned circuit, as is shown in the schematic. The feedback loop gain of this amplifier was obtained by the technique used for the amplifier of Fig. 12(a). This gain is given by

$$\mu\beta = \frac{\hat{i}_e}{i_e} = |\, \mu\beta \,|\, \theta$$

$$\doteq \frac{\alpha Z_c}{Z_c + Z_L} \frac{Z_b}{Z_b + Z_e + Z_s}$$

or

$$\mu\beta \doteq \alpha \frac{Z_b}{Z_b + Z_c + Z_s}, \tag{15}$$

when Z_L is appreciably less than Z_c, the condition for mismatch design. Using (15), the loop gain was computed at a sufficient number of points to give the Nyquist diagram shown in Fig. 13(b). The center-band and band-edge frequency points are indicated on the Nyquist diagram. Note the inevitable positive feedback in the transmission band. However, the feedback though positive is fractional (i.e., there is a net loss around the feedback loop) and well below the critical unity value. Furthermore, the loop feedback gain decreases and rapidly goes to zero both above and below the center-band frequencies. An examination of the circuit and the feedback loop gain given by (15) shows the reason for this. Since the load impedance, Z_L, is a moderately high-Q parallel-resonant circuit, it rapidly approaches zero at frequencies outside the transmission band, thereby increasing the degree of mismatch away from the center-band frequency. At the same time, Z_s, the source or generator impedance, is an even greater-Q series-resonant circuit, so that it reaches a high impedance very rapidly away from the center-band frequency. Since Z_s appears only in the denominator of the expression of (15) for feedback loop gain, this means that $\mu\beta$ rapidly goes to zero, due to the high outband impedance of the generator, Z_s. The same behavior would be experienced with a double-tuned interstage circuit, except that the $\mu\beta$ diagram would consist of two loops, due to the added pole and zero in the reactance interstage. These two loops would both pull away from the Nyquist point area towards the origin in the same manner as does the loop gain of Fig. 13(b). Because of the avoidance of positive feedbacks having real parts approaching unity, it would be expected that the mismatch amplifiers would be not only more stable, but also less sensitive to changes than are the neutralized amplifiers, and this is confirmed by experimental results. The price paid for this improved stability and reduction in gain sensitivity is lower stage gain. In return for the gain sacrifice, we also obtain greater ease of design, greater ease of interstage alignment and less complicated circuitry.

How then does one decide on the choice of the neutralized or mismatched techniques? The answer to this question is largely dependent upon economic and system requirement considerations. If a consumer product is being designed where competition demands maximum gain to keep down cost factors and where the failure of an amplifier means only an occasional service call, then the maximum-gain neutralized amplifier might be selected. However, if a system is being designed where amplifier failure would cause malfunctioning of a large and costly system, reliability considerations would favor the more conservative mismatch approach, in spite of the lower stage gains obtained. Intermediate situations might suggest a combination of neutralization and mismatch, with higher gains than could be obtained with the straight mismatched amplifier and with feedback loop diagrams midway between the extremes of Figs. 12 and 13. It is interesting to note that, with the great reduction in the collector capacitances of VHF transistors, the mismatch that automatically occurs from the impracticability of simultaneously obtaining output matching and very broadband interstages results in a compromise mismatch-neutralized circuit of the type just mentioned. Actually, the experience with these circuits has shown that the neutralization is not critical when the degree of mismatch is fairly high, and may be omitted.

VII. VIDEO AMPLIFIER DESIGN

In the design of video amplifiers, the mismatch approach is practically dictated by the broadband requirements and the limitation on the maximum impedance available with a given irreducible circuit capacitance, in accordance with the Bode resistance integral theorem (Ref. 5, Chapters 4 through 6). And so we can use the high common emitter current gain without danger of circuit oscillation. However, the gain sensitivity problem still exists, as will be shown.

With the new high-frequency-cutoff diffusion transistors, common emitter short-circuit current gains of 12 db and higher at 100 mc are now commercially available. These make possible common emitter iterative amplifiers with the collector of one transistor coupled directly into the base of the following transistor — except for a blocking condenser when simple bias circuits are required. Such an amplifier was built by C. E. Paul of Bell Telephone Laboratories with early models of the germanium diffused-base transistor. A picture of this amplifier is shown in Fig. 14. The amplifier has three common emitter iterative stages, a gain of 70 db and a bandwidth of close to 10 mc using the simplest possible resistance-capacitance coupled interstages. The amplifier requires a total power of less than 100 milliwatts and occupies a volume of less than 2 cubic inches. This amplifier demonstrates the great potential of the common emitter transistor connection in video circuits.

With the simple iterative common emitter amplifier, the single-stage bandwidth is determined by $(1 - \alpha_0)f_\alpha$, where f_α is the common base cutoff frequency. This bandwidth will vary widely from transistor to transistor, due to variations in α_0 and f_α. If bandwidths narrower than $(1 - \alpha_0)f_\alpha$ are needed, they can be obtained most easily by choosing transistors with higher α_0 or lower f_α. However, for today's broadband video and baseband amplifiers, bandwidths greater than the normal common emitter bandwidths are frequently required, and some means of trading gain for bandwidth is needed. This can be accomplished by feeding back a portion of the output signal to the input, in accordance with the technique illustrated in Fig. 15. Fig. 15(a) shows a single-stage common emitter amplifier in which the load impedance is small compared to the collector impedance, a situation which exists in the iterative common emitter amplifier. The current gain of the amplifier is given by

$$\frac{i_2}{i_1} \doteq \frac{\alpha}{1 - \alpha}. \tag{16}$$

The current gain given by (16) is plotted as curve A of Fig. 15(b) for a transistor having an α_0 of 0.97 and a common emitter cutoff frequency

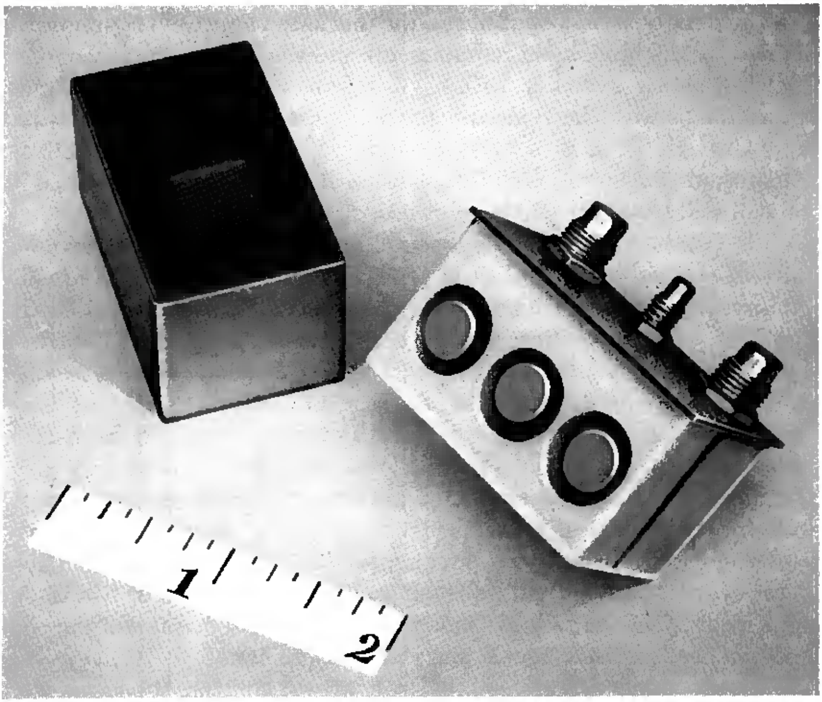


Fig. 14 — Three-stage common emitter video amplifier.

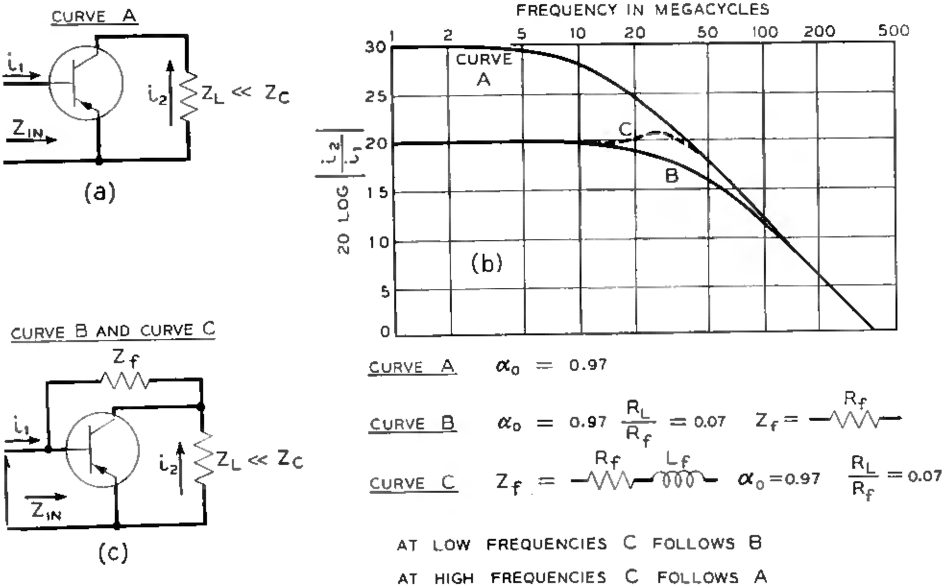


Fig. 15 — Single-stage common emitter amplifier with and without shunt feedback: (a) amplifier without feedback; (b) current gain of amplifier; (c) amplifier with shunt feedback.

of 14 mc — the frequency at which the common emitter current gain is 3 db below its low frequency value of $\alpha_0/(1 - \alpha_0)$. If a broader bandwidth is desired, this can be obtained by feeding back a portion of the output to the input through a feedback impedance, Z_f, connected between the collector and base, as shown in Fig. 15(c). The current gain of this transistor is given by

$$\frac{i_2}{i_1} \doteq \frac{\alpha}{1 - \alpha + \dfrac{Z_L}{Z_f}}. \tag{17}$$

If the simplified expression of (1) for a or α is placed in (17), the current gain as a function of frequency for Z_L/Z_f real is given by

$$\frac{i_2}{i_1} = \frac{\alpha_0}{(1 - \alpha_0)\left[1 + \dfrac{Z_L}{Z_f(1 - \alpha_0)}\right]\left\{1 + j\dfrac{f}{f_\alpha(1 - \alpha_0)\left[1 + \dfrac{\alpha_0 Z_L}{Z_f(1 - \alpha_0)}\right]}\right\}}, \tag{18}$$

so that, except for the ratio

$$\frac{1 + \dfrac{Z_L}{Z_f\,(1 - \alpha_0)}}{1 + \alpha_0 \dfrac{Z_L}{Z_f\,(1 - \alpha_0)}},$$

which is normally close to unity, the low frequency gain is decreased and the cutoff frequency is increased by the same amount, namely

$$1 + \frac{Z_L}{Z_f(1 - \alpha_0)}.$$

This is shown in curve B of Fig. 15(b), where the common emitter current gain is plotted for the transistor assumed for curve A, with a resistance, R_f, connected between its collector and base such that

$$R_L/R_f = 0.07.$$

Note that the low-frequency gain of curve B is down 10 db, or a factor of about one to three in magnitude, from that of curve A, and that the cutoff frequency has been increased by about the same factor. The asymptotic current gains of curves A and B at very high frequencies differ only slightly in magnitude, so they are shown identical in Fig. 15(b). By opening up the feedback path between the collector and base at high frequencies, curve B can be made to move into curve A before the asymptotic region is reached as is illustrated in the dotted curve, curve C. This can be accomplished most simply by making Z_f a resistance and inductance in series. In this way, approximately an extra octave of bandwidth can be obtained with no additional in-band gain sacrifice. However, there will be somewhat greater delay distortion when the amplifier is used for the amplification of narrow pulses than there would be if the cutoff were allowed to proceed in normal RC fashion, as in curve B of Fig. 15(b).

The simplicity of the above technique of trading gain for bandwidth is illustrated in the two-stage diffused-base common emitter video amplifier shown schematically in Fig. 16.* The transistors used in this amplifier have a normal common emitter short-circuit gain given by curve A of Fig. 15, and R_L/R_f is made 0.07 to make $(1 - \alpha_0) + R_L/R_f$ equal to 0.1 and give a low-frequency current gain of magnitude 10 or a

* This amplifier was devised by the author and presented at the June 1955 Semiconductor Device conference in Philadelphia, Pa., to demonstrate the broadband capabilities of the original research models of diffused base germanium transistors. For a description of these transistors see Lee.[9] For more complete information on this type of video amplifier see Ballentine and Blecher.[10]

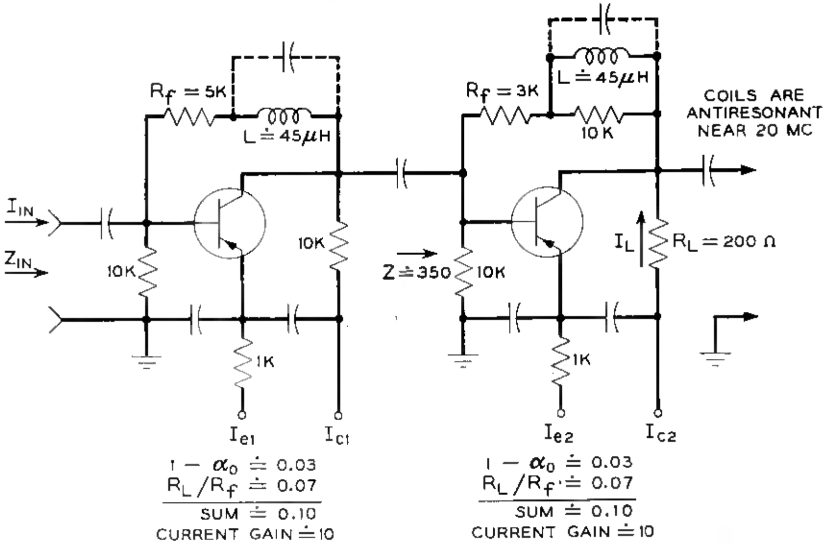


Fig. 16 — Two-stage common emitter amplifier with shunt feedback on each stage.

current gain of 20 db. The feedback path is opened at the high-frequency end of the band by the 45-microhenry coil in series with the feedback resistance of each stage. The dotted capacitances are the distributed capacitances of the coils, which produce a parallel resonance and essentially open-circuit impedance at the top end of the band. Therefore, the feedback path is effectively opened, and the normal common emitter current gain without feedback is obtained.

The current gain of the two-stage amplifier of Fig. 16 is plotted as a function of frequency in Fig. 17. The amplifier has a two-stage gain of 40 db flat to ±0.5 db up to 20 mc.

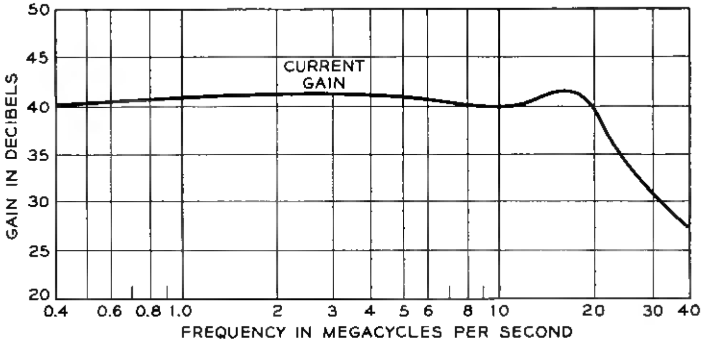


Fig. 17 — Current gain for amplifier of Fig. 16.

Although the technique described above is an easy way to trade gain for bandwidth, it is also an inefficient way. This is a result of the fact that, although a single stage of the amplifier of Fig. 16 behaves like a negative feedback amplifier, in that the forward gain is reduced as the feedback is increased, it is in fact still a positive feedback amplifier in accordance with our earlier analysis of transistor internal feedback. This is shown from Fig. 18, where the generalized T schematic of the transistor is given in the common emitter connection with a load resistance, R_L, and a feedback resistance, R_f. The various currents resulting from an injected emitter current, i_e, are also shown. Using Blackman's technique for determining feedback loop gain, $\mu\beta$, or the return ratio of the amplifier of Fig. 18, can be written by inspection as follows:

$$\hat{i}_e \doteq \alpha i_e - \alpha i_e \frac{R_L}{R_f},$$
$$\mu\beta = \frac{\hat{i}_e}{i_e} \doteq \alpha \left(1 - \frac{R_L}{R_f}\right), \tag{19}$$

or, for the amplifier of Fig. 16,

$$\mu\beta \doteq \alpha\ (1 - 0.07) = 0.93\ \alpha. \tag{20}$$

Equation (19) shows that, even though the magnitude of the feedback has been reduced by the factor $(1 - R_L/R_f)$, $\mu\beta$ is still positive and close to unity. In other words, even though the incremental feedback through the feedback resistance, R_f, is negative, the residual or net feedback is

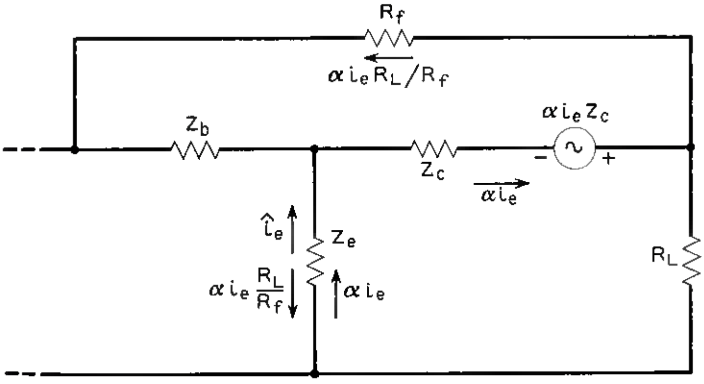


Fig. 18 — Equivalent circuit for common emitter amplifier with shunt feedback.

still positive. This can be practically verified by examining the variation in external gain with change in α.

The current gain of the circuit of Fig. 18 can be obtained from (17) and is given by:

$$\frac{i_2}{i_1} \doteq \frac{\alpha}{(1 - \alpha) + \dfrac{R_L}{R_f}}. \tag{21}$$

If we compute the variation in current gain expressed as a fraction of the initial current gain in terms of the variation in α expressed as a fraction of the original value of α we get,

$$\frac{d\,\dfrac{i_2}{i_1}}{\dfrac{i_2}{i_1}} = \frac{d_\alpha}{\alpha}\left[\frac{1}{1 - \alpha\left(1 - \dfrac{R_L}{R_f}\right)}\right], \tag{22}$$

which, from (19), gives

$$\frac{d\,\dfrac{i_2}{i_1}}{\dfrac{i_2}{i_1}} = \frac{d_\alpha}{\alpha}\left(\frac{1}{1 - \mu\beta}\right). \tag{23}$$

Since $\mu\beta$ from (20) is positive and only slightly less than unity, (23) shows that the terminal current gain changes much more rapidly than does the current generator gain, α, of the active device. This is the reverse of a negative feedback effect and is characteristic of the residual positive feedback which has been shown to exist. The variation in current gain due to a given change in α is less than it was before the R_f feedback path was added, which is in accordance with our statement that R_f represents a negative increment of feedback, but that the feedback loop gain residue is still positive.

In many instances, the decrease in external gain change for a given change in active-element gain obtained by the simple circuit of Fig. 16 is sufficient. However, where lower external gain change is required — and somewhat greater circuit complexity is therefore justified — gain can be more effectively traded for bandwidth by feedback around a minimum of two common emitter stages, as shown in the schematic of Fig. 19. The first stage only is shown in generalized equivalent T form, since it is here that the feedback path is mathematically broken to compute the main feedback gain (i.e. the feedback gain through the R_f feedback path).

Again using the Blackman technique, the feedback loop gain $\mu\beta$ can be obtained by inspection of Fig. 19 as follows:

$$\hat{i}_{e1} = \alpha_1 i_e \frac{Z_b}{Z_b + Z_e + R_{ef}} < 0.5\,\alpha_1 i_e, \tag{24}$$

$$\begin{aligned} \hat{i}_{e2} &\doteq \frac{\alpha_1\alpha_2}{1 - \alpha_2} \frac{R_L}{R_L + R_f} \frac{R_{ef}}{Z_b + Z_e + R_{ef}} i_e \\ &\doteq \frac{\alpha_1\alpha_2}{1 - \alpha_2} \frac{R_L}{R_L + R_f} i_e, \end{aligned} \tag{25}$$

$$\hat{i}_e = \hat{i}_{e1} - \hat{i}_{e2} \doteq -\frac{\alpha_1\alpha_2}{1 - \alpha_2} \frac{R_L}{R_L + R_f} i_e, \tag{26}$$

$$\mu\beta = \frac{\hat{i}_e}{i_e} \doteq -\frac{\alpha_1\alpha_2}{1 - \alpha_2} \frac{R_L}{R_L + R_f}. \tag{27}$$

If $R_L \geqq R_f$ and α_1 and α_2 are close to unity, then $\mu\beta \gg 1$. When the loop gain is much larger than unity, the feedback voltage, V_f, will be approximately equal to the applied generator voltage, V_g, when a steady state signal is applied to the input. Therefore, since

$$V_f = \frac{R_{ef}}{R_f} V_{\text{out}} \qquad \text{and} \qquad V_f \doteq V_g, \tag{28}$$

$$\text{amplifier voltage gain} = \frac{V_{\text{out}}}{V_g} \doteq \frac{R_f}{R_{ef}}, \tag{29}$$

and the voltage gain of the amplifier is substantially independent of change in gain of the active elements — in this case, the transistors. This

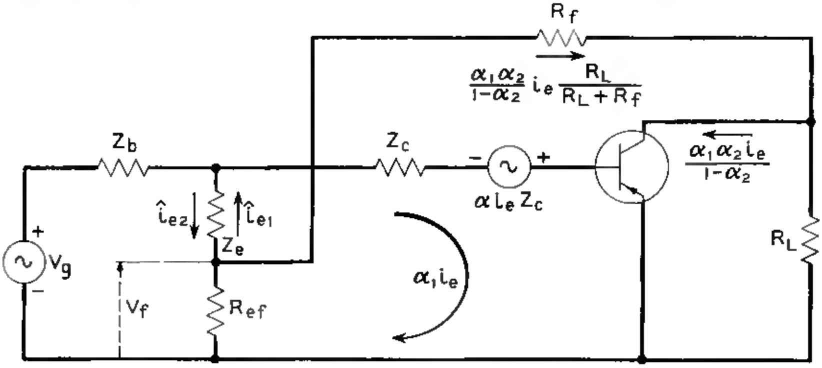


Fig. 19 — Equivalent circuit for two-stage common emitter amplifier with shunt-to-series feedback.

is the anticipated result for a true negative feedback amplifier with feedback loop gain much greater than unity. It is important to notice that the approximation of (29) to the feedback loop gain is good only for $R_L \geqq R_f$, so that this circuit is essentially a voltage feedback amplifier. Since the circuit of Fig. 19 contains basically only two 6-db-per-octave asymptotic cutoffs, it is an intrinsically stable circuit requiring only simple if any feedback loop equalization. The amplifier of Fig. 19 is a voltage amplifier with a high input impedance and a low output impedance, since it has series feedback at the input and shunt feedback at the output. It can be made a current amplifier with low input impedance and high output impedance by feeding back from a resistance in the emitter circuit of the second transistor through a feedback resistance R_f to the base of the first transistor. The approximate design formulae for this configuration can be obtained in the same manner as were those for the voltage amplifier shown schematically in Fig. 19.

If high linearity as well as high stability, or if unusually high stability is required in an amplifier, either of the broadband video or relatively narrowband linear type, then the two-stage amplifier of Fig. 19 is still inefficient from the standpoint of trading gain for bandwidth. In this case, the most efficient circuit is a three-common-emitter-stage single-loop feedback amplifier.[11,12] This, of course, involves the complexity of interstage and feedback network design inherent to the stabilization of a three-stage negative feedback amplifier. This is a consequence of the potential instability associated with the minimum asymptotic cutoff of 18 db per octave associated with three active stages.

In conclusion, it may be stated that the requirements of a large percentage of the radio frequency and video or baseband transistor amplifiers can be met by the circuits of Figs. 9, 16 and 19. These circuits demonstrate the simplicity with which basically sound and stable transistor amplifiers can be built, providing that the basic nature of the internal feedback of the transistor is understood, and the fatal mistake of attempting to obtain so called "maximum available gain" is not made.

Additional material which may be of interest to designers of RF and video amplifiers: neutralization — Cheng;[13] stability — Stern;[14] video amplifiers — Brunn;[15] alignable receivers — Gibbons.[16]

VIII. ACKNOWLEDGMENT

The author wishes to express his appreciation to L. G. Schimpf and C. E. Paul for making available information on the circuits and performance of their amplifiers for use as examples of the various types of amplifiers discussed in this paper. He also wishes to thank R. B. Black-

man for checking the positive feedback theory of single-stage transistor amplification and for his suggestion of the simplified technique for mathematically breaking the feedback loop to compute return ratio.

REFERENCES

1. Pritchard, R. L., Electric Network Representation of Transistors — A Survey, Trans. I.R.E., **CT-3**, March 1956, p. 5.
2. Thomas, D. E. and Moll, J. L., Junction Transistor Short Circuit Current Gain and Phase Determination, Proc. I.R.E., **46**, June 1958, p. 1177.
3. Linvill, J. G. and Schimpf, L. G., The Design of Tetrode Transistor Amplifiers, B.S.T.J., **35**, July 1956, p. 813.
4. Webster, R. R., A Tetrode Transistor Amplifier for 5–40 mc, Elect. Ind. & Tele-Tech., **15**, November 1956, p. 62.
5. Bode, H. W., *Network Analysis and Feedback Amplifier Design*, D. Van Nostrand & Co., New York, 1945.
6. Nyquist, H., Regeneration Theory, B.S.T.J., **11**, January 1932, p. 126.
7. Blackman, R. B., private communication.
8. Hunter, L. P., *Handbook of Semiconductor Electronics*, McGraw-Hill Book Co., New York, 1956, Section 12 by J. B. Angell, p. 12.
9. Lee, C. A., A High-Frequency Diffused-Base Germanium Transistor, B.S.T.J., **35**, January 1956, p. 23.
10. Ballentine, W. E. and Blecher, F. H., Broadband Transistor Video Amplifier, Digest of Tech. Papers, I.R.E.-A.I.E.E. Solid-State Circuits Conf., February 12–13, 1959.
11. Blecher, F. H., Design Principles of Single-Loop Transistor Feedback Amplifiers, Trans. I.R.E., **CT-3**, September 1957, p. 145.
12. Abraham, R. P., A Wide-Band Transistor Feedback Amplifier, I.R.E. Wescon Conv. Rec., 1957, Part 2, p. 10.
13. Cheng, C. C., Neutralization and Unilateralization, Trans. I.R.E., **CT-2**, June 1955, p. 138.
14. Stern, A. P., Stability of Power Gain of Tuned Transistor Amplifiers, Proc. I.R.E., **45**, March 1957, p. 335.
15. Brunn, G., Common Emitter Transistor Video Amplifiers, Proc. I.R.E., **44**, November 1956, p. 1561.
16. Gibbons, J. F., The Design of Alignable Transistor Amplifiers, Tech. Rep. 106, Stanford Elect. Lab., May 7, 1956.

Effects of Tamping and Pavement Breaking on Round Conduit

By G. F. WEISSMANN and DUNCAN M. MITCHEL

(Manuscript received July 20, 1959)

Underground conduits may be subjected to low-frequency dynamic loads caused primarily by the operation of mechanical tamping and pavement-breaking machines. These external loads will produce circumferential bending moments in the conduit wall. The magnitude of the bending moments has been determined by measurement of the circumferential fibre strains in thin-walled metal tubes subjected to the external dynamic forces transmitted through various soil media. Finally, the bending moments are expressed in terms of the the equivalent crushing strength.

I. INTRODUCTION

An extensive investigation to establish the minimum strength requirements for round conduit, based upon the effect of static loads has been reported by one of the authors.[1] It was shown that the minimum required strength depends on the magnitude of the load applied at the surface of the fill, the properties of the backfill material, the height of the backfill over the conduit, the trench width and the bedding condition.

The increasing use of heavy-duty power-activated equipment for tamping backfill in trenches and for breaking pavement has made it necessary to expand this investigation in order to determine the effect of dynamic forces on underground conduit and pipes. This study is intended to show the conditions under which tamping or pavement-breaking equipment may be used without damaging underground conduit that has the minimum strength required to withstand static loads.

External loads acting upon the conduit produce circumferential bending moments in the conduit wall. The magnitude and distribution of these bending moments caused by the operation of tamping and pavement breaking machinery have been determined in tests conducted recently at the Outside Plant Development Laboratory, Chester, New

Jersey, and in Chicago, Illinois. These tests were made with gravel, sand and sandy clay as backfill and with various heights of cover over the conduit. Different energies were applied for both tamping and pavement breaking under conditions simulating as nearly as possible those encountered in the field.

II. TEST APPARATUS AND PROCEDURE

A test method, which was previously developed for the determination of the circumferential bending moments in thin-walled conduits under static loads,[1] was modified for the recording of dynamic loads.

The test device consisted of a thin-walled steel tube one foot in length having an outside diameter of 4 inches and a wall thickness of 0.062 inch. Four SR-4 strain gages (type A-5) were attached, at intervals of 90 degrees, to the inside periphery of the tube at points equidistant from the tube ends. Fig. 1 shows the steel tube with the attached strain gages. The tube ends were sealed with sponge rubber discs to prevent the entry of dirt and moisture. Each strain gage served as the variable arm of a bridge circuit that was connected to an oscillograph (Minneapolis Honeywell Visicorder). This recorder provided a continuous photographic record of the strain readings.

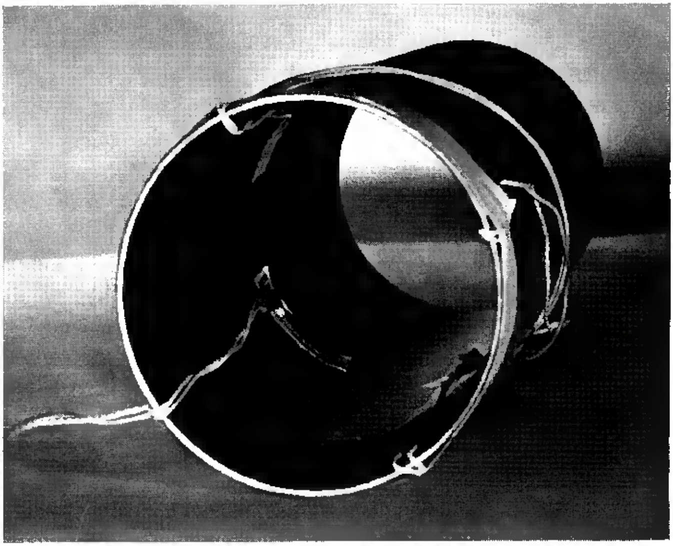


Fig. 1 — Thin-walled tube with SR-4 strain gages.

A Hydrahammer, manufactured by the Ottawa Steel Division of the L. A. Young Spring and Wire Corporation, Ottawa, Kansas, was used to provide the impact loads at the surface above the conduit. The Hydrahammer consists essentially of a weight that is dropped from different heights and is capable of applying energies up to 7500 foot-pounds. Two different weights have been used during this investigation: (a) approximately 1000 pounds using the tamper and (b) approximately 900 pounds using the demolition head.

Fig. 2 shows the three test conditions considered during this investigation.

2.1 *Tamping*

The measuring tube was positioned lengthwise at the bottom of a 24-inch-wide trench between two pieces of plastic conduit, as shown in Fig. 3. It was oriented so that one of the strain gages was at the top of the tube. The backfill material was then placed in the trench to the desired height of cover. Care was taken to insure that no stones were present at the bottom of the trench or in the fill close to the measuring tube. The tests were conducted with 18, 24, 30, 36 and 42 inches of cover, using gravel, sand and sandy clay as backfill materials. The mechanical tamper was positioned so that the tamping head struck the surface of the fill directly over the tube. For consistency in the test it was necessary to restore the fill at the striking point to its original height after each blow.

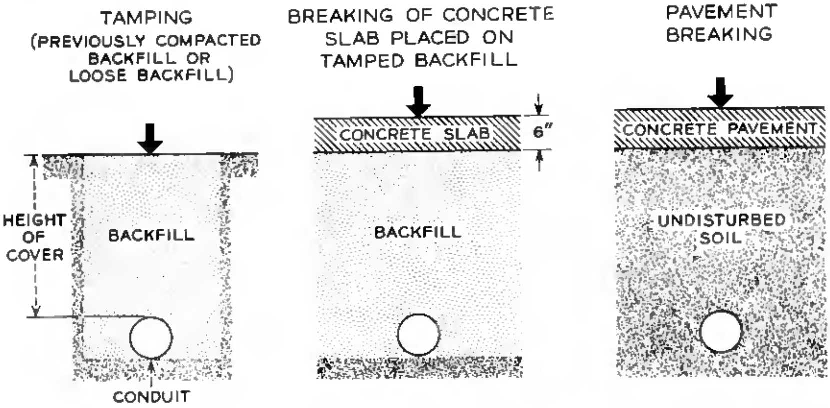


Fig. 2 — Test conditions.

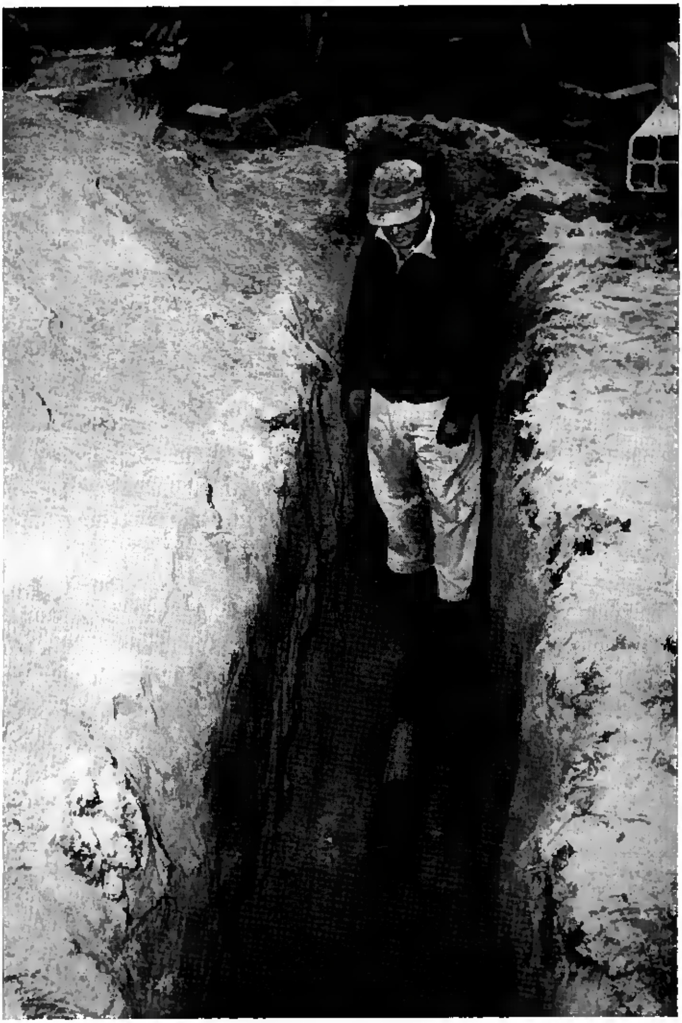


Fig. 3 — Measuring device positioned in trench.

Two procedures were used to apply the impact loads:

i. The energy produced by the Hydrahammer was increased from 1125 to a maximum of 6750 foot-pounds in increments of 1125 foot-pounds. A number of blows were applied at each energy level until the readings did not change significantly.

ii. The maximum energy of 6750 foot-pounds was delivered directly to the loose backfill and repeated until the measurements remained constant.

2.2 *Breaking of Concrete Slab Placed on Tamped Backfill*

To simulate the effect of pavement breaking on round underground conduit, reinforced concrete slabs, three feet square and six inches thick, cast from a 1:2:3 mix and air-cured for 28 days prior to the tests were used. The measuring tube was placed at the bottom of a three-foot-deep pit and covered to a height of 12 or 24 inches with gravel or sandy clay. The fill was lightly compacted and the reinforced concrete slab was positioned so that its center was directly over the measuring tube. Figs. 4 and 5 show the Hydrahammer equipped with the demolition head breaking the concrete slab. An energy of 6750 foot-pounds was used for this part of the investigation.

2.3 *Pavement Breaking*

Actual pavement-breaking tests were conducted at two locations in Chicago, Illinois. Horizontal holes slightly smaller than the outside di-

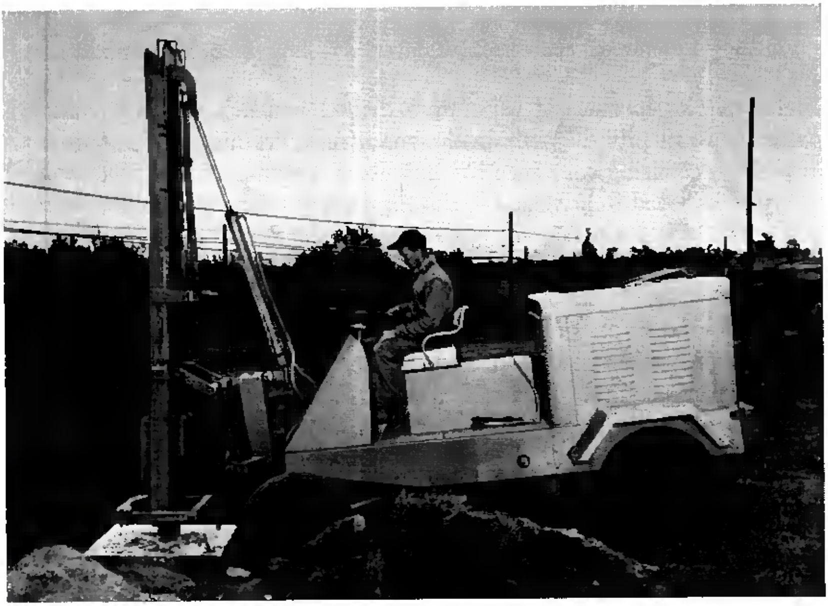


Fig. 4 — Hydrahammer with demolition head breaking reinforced concrete slab.

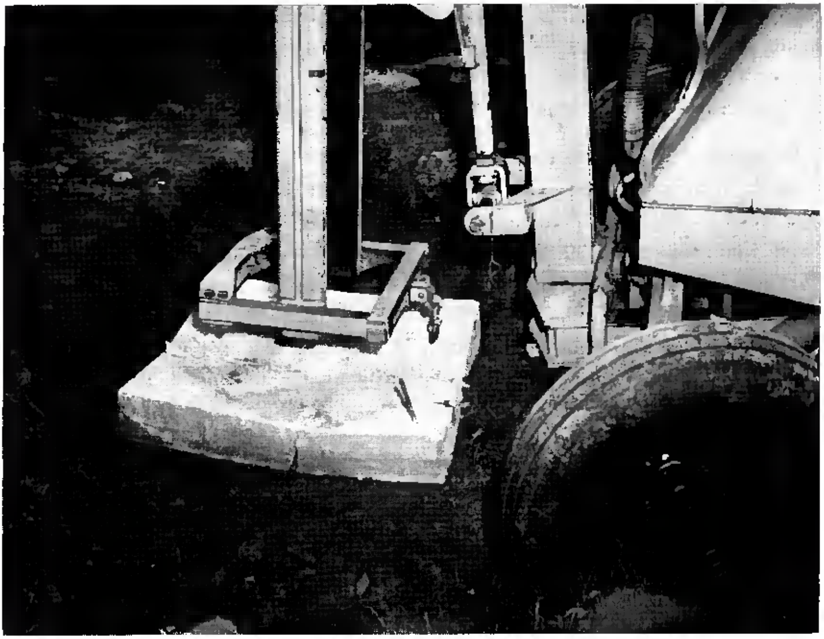


Fig. 5 — Reinforced concrete slab after one blow of 6750 foot-pounds.

ameter of the measuring tube were drilled beneath existing concrete roadways from pits dug beside the road. A hydraulic jack was used to press the measuring tube into the hole for a distance of four feet. A steel tube was inserted in advance of the measuring tube and a piece of four-inch conduit was used to fill the remaining length of the hole. The pit was then backfilled and tamped. Three tests were made, the height of cover consisting of: (a) 3 inches of sandy clay, 21 inches of a mixture of ashes and cinders and 7 inches of concrete; (b) 3 inches of sandy clay, 21 inches of a mixture of ashes and cinders, 5 inches of concrete and 2 inches of asphalt; and (c) 12 inches of crushed stone and 8 inches of concrete. The tamper applied an energy of 6750 foot-pounds directly over the center of the measuring tube. Measurements were taken until the pavement above the conduit was completely broken up.

Table I summarizes the test conditions.

III. TEST RESULTS

The circumferential bending moments in the walls of the test tubes were determined and recorded by means of the test apparatus and procedure described in Section II.[1] A typical example of such a recording is shown in Fig. 6. The duration of the signal caused by the impact is about 0.1 second, and the bending moment at the bottom of the conduit is about double that at the top or at the sides of the conduit. This relation, however, was not observed in all tests; it was more frequent in clay or wet sand than in gravel. The same phenomenon had been observed during an investigation of the effect of static loads on round conduits,[1] when it was concluded that the bedding condition was responsible. The same considerations apply for this investigation. Due to a change in bedding, the moment at the bottom may vary up to 235 per cent, while the moments at the side points change only a maximum of 12 per cent. To compensate for wide variations in the test results attributable to bedding, the maximum bending moment at the bottom of the tube was considered to be double the average of the values measured at the side points.

During mechanical tamping, the maximum bending moment at a given depth increased with the number of blows until it attained a limiting value. This value was generally obtained with the third blow of the tamper. For pavement breaking, the maximum bending moment was obtained immediately after the concrete pavement cracked.

In the remaining sections of this paper, the maximum bending moment

TABLE I — LIST OF TESTS AND TEST CONDITIONS

Test Condition	Type and Height of Cover and Thickness of Pavement	Applied Energy (ft-lbs)
Tamping of previously compacted backfill	24, 30, 36, 42 inches gravel	1125, 2250, 3375, 4500, 5625, 6750
	18, 24, 30, 36 inches sand	1125, 2250, 3375, 4500, 5625, 6200
	18, 24, 30, 36 inches sandy clay	1125, 2250, 3375, 4500, 5625, 6750
Tamping of loose backfill	24, 30, 36, 42 inches gravel	6750
	18, 24, 30, 36 inches sand	6750
	18, 24, 30, 36 inches sandy clay	6750
Breaking of concrete slabs placed on tamped backfill	12 or 24 inches gravel, 6 inches concrete	6750
	18 or 24 inches sandy clay, 6 inches concrete	6750
Pavement breaking	3 inches clay, 21 inches ashes and cinder, 7 inches concrete	6750
	3 inches clay, 21 inches ashes and cinder, 5 inches concrete, 2 inches asphalt	6750
	12 inches crushed stone, 8 inches concrete	6750

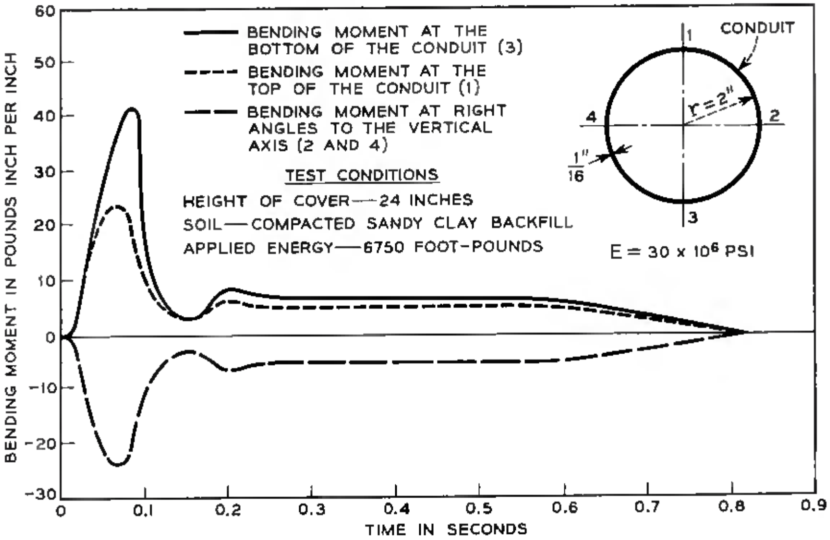


Fig. 6 — Bending moments in round conduit caused by operation of heavy tamping equipment.

will be expressed in terms of the "equivalent two-point load." The equivalent two-point load is the two-point (two-edge bearing) load that, in a compression test on a test tube held between two rigid flat plates, will cause the same maximum bending moment in the tube wall as the maximum bending moment obtained from field measurements such as shown in Fig. 6. The choice of this expression as a measure of the bending moment has been discussed.[1]

3.1 *Tamping*

Figs. 7 and 8 show the equivalent two-point loads for four-inch-diameter round conduit covered with gravel and sandy clay, respectively. In each figure the abscissa represents the energy applied by the tamper and the ordinate the equivalent two-point load, with a logarithmic scale having been used for both coordinates. The data were obtained by tamping the initially loose fill over the conduit, the energy applied by the Hydrahammer being increased from 1125 foot-pounds in steps of 1125 foot-pounds to a maximum of 6750-foot-pounds. A linear relationship between the logarithm of the applied energy and the logarithm of the equivalent two-point load could be observed. For each case, this relationship was derived from the data, using the method of least squares.

The plotted lines were extended beyond 6750 foot-pounds to obtain equivalent two-point load values for energies up to 20,000 foot-pounds.

If a blow were delivered by the tamping equipment upon loose sandy clay or wet sand backfill, the equivalent two-point load values obtained were rather inconsistent: these values were up to three times higher than those obtained when the tamping energy was increased by increments to its maximum. This phenomenon was not observed with gravel or dry sand as backfill.

3.2 *Breaking of Reinforced Concrete Slab Placed on Tamped Backfill*

The equivalent two-point loads obtained when breaking reinforced concrete slabs placed on previously compacted backfill are shown in Table II. The values of Table II should be compared with the results

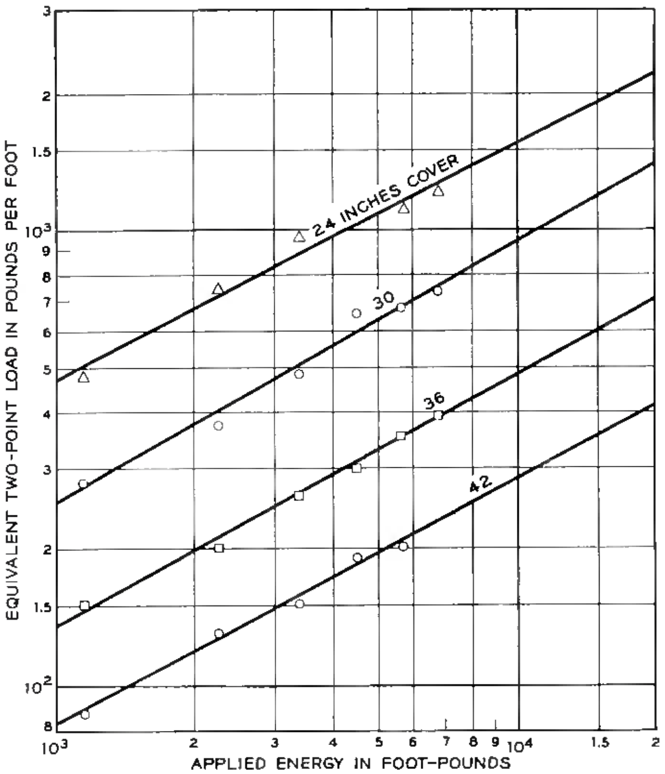


Fig. 7 — Equivalent two-point load vs. applied energy for gravel cover.

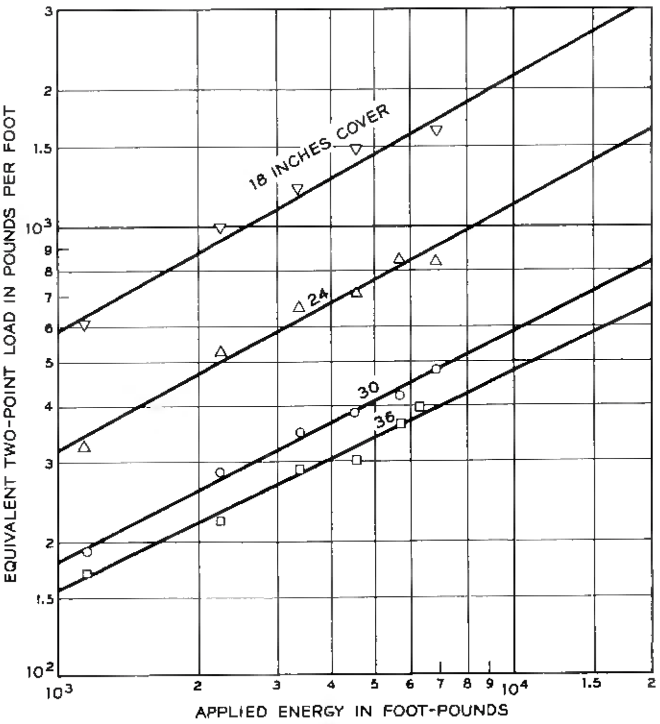


Fig. 8 — Equivalent two-point load vs. applied energy for sandy clay cover.

shown in Figs. 7 and 8, which were obtained by tamping backfill previously compacted at the same energy level. The height of cover is considered to be the distance between the bottom surface of the concrete and the top of the conduit. For the same height of cover and applied energy, the results obtained when breaking the concrete slab are slightly smaller than the values obtained by tamping, since the broken concrete provides some additional protection to the conduit. However, this additional protection is comparatively small and variable, dependent upon the thickness of the concrete, and will be neglected. Energies insufficient to break the slab will produce relatively small forces acting on the conduit.

3.3 *Pavement Breaking*

The equivalent two-point loads obtained by pavement breaking are also shown in Table II. A comparison of the equivalent two-point load values obtained by tamping (Figs. 7 and 8) with those obtained by pave-

Table II — Equivalent Two-Point Load Obtained by Breaking Concrete Slabs and Pavements

Test Condition	Type and Height of Cover and Thickness of Pavement	Applied Energy (ft-lbs)	Equivalent Two-Point Load (lbs/ft)
Breaking of concrete slabs placed on tamped backfill	12 inches gravel, 6 inches concrete	6750	2200
	24 inches gravel, 6 inches concrete	6750	720
	18 inches sandy clay, 6 inches concrete	6750	1300
	24 inches sandy clay, 6 inches concrete	6750	560
Pavement breaking	3 inches clay, 21 inches ashes and cinder, 7 inches concrete	6750	600
	3 inches clay, 21 inches ashes and cinder, 5 inches concrete, 2 inches asphalt	6750	600
	12 inches crushed stone, 8 inches concrete	6750	2000

ment breaking shows the same relationship as that obtained for the breaking of the reinforced concrete slab. For the same height of cover and applied energy, the equivalent two-point loads obtained by pavement breaking are slightly smaller than those obtained by tamping.* As in the case of breaking the slabs, this difference will be neglected.

IV. DISCUSSION OF TEST RESULTS

The test results show that tamping well-compacted fill produces about the same loads acting on the conduit walls as are obtained by pavement breaking, provided the heights of cover, the types of cover and the applied energies are the same.

Figs. 7 and 8 show the effects of the applied energy, the height of cover and the type of cover on the equivalent two-point loads. Results obtained with sand cover were not as consistent as were the values for gravel and sandy clay. This may have been due to a change in moisture content: because of weather conditions, the sand used as backfill material varied from dry to rather wet.

The strength requirements for round underground structures, previously determined on the basis of static loading conditions,[1] showed the highest strength to be required when wet clay was used as backfill material. Fig. 9 shows the required equivalent two-point load as a function of the height of cover for static loads, with the different curves represent-

* The values listed in Table II are maximum values obtained after the pavement failed. Prior to this failure, the measurements were very small.

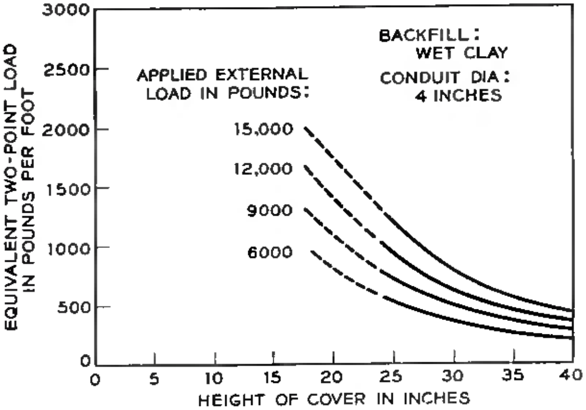


Fig. 9 — Equivalent two-point load caused by static load vs. height of cover.

ing various wheel loads. It is desirable that underground conduit be capable of withstanding a maximum wheel load of 15,000 pounds.

Fig. 10 gives the equivalent two-point load of four-inch-diameter round conduits obtained by tamping well-compacted backfill or by pavement breaking as a function of the height of cover for gravel as the backfill material. Gravel has been chosen because the highest equivalent two-point load values were obtained with this material. Since the type of the

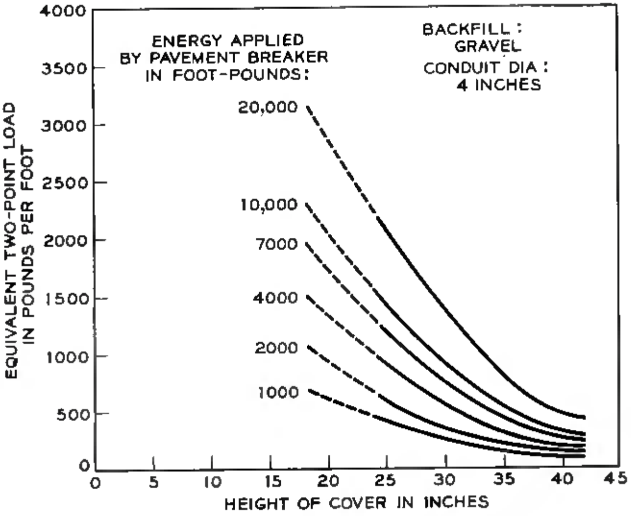


Fig. 10 — Equivalent two-point load caused by pavement breaking vs. height of cover.

subsoil is generally unknown when a pavement breaker is used, the worst conditions should be considered. The different curves in Fig. 10 represent the energies applied by the pavement breaker.

A comparison of Figs. 9 and 10 shows that a conduit having the desired minimum strength requirements based on static considerations should withstand without damage the effects of pavement breaking if equipment of 7500 foot-pounds capacity is employed and the height of cover is at least 24 inches. It appears that the equivalent two-point load due to pavement breaking increases more rapidly with a decrease of the height of cover than does the equivalent two-point load caused by static loads.

Fig. 11 shows the equivalent two-point loads obtained by tamping loose sandy clay backfill. These curves were derived from Fig. 8. To consider the effect of tamping loose fill, the values obtained by tamping compacted sandy clay fill were multiplied by three, in accordance with the experimental data. A comparison of Figs. 9 and 11 shows that tamping loose sandy clay fill with an applied energy of more than 1500 foot-pounds at a height of cover of 24 inches would exceed the assumed minimum crushing strength of the conduit (two-edge bearing load of 1200 pounds per foot) and could cause breakage. The effect of tamping loose backfill is much more severe than that of pavement breaking because the loose backfill in the trench acts like a piston in a cylinder and drives down on the ducts when it is subjected to the blows of the tamper. Furthermore, the tamping head penetrates farther into the loose fill and thus reduces the effective height of cover.

Additional tests were conducted tamping various types of backfill in trenches containing different conduits of known crushing strengths and conduit formations. The results of these tests support the findings of this investigation.

V. SUMMARY AND CONCLUSIONS

Field tests have been conducted to investigate the effect of the use of heavy tamping and pavement breaking equipment on round underground conduits. The results show that conduit may be damaged when heavy-duty equipment with a capacity of 7500 foot-pounds is used as a pavement breaker unless the height of cover over the conduit is at least 24 inches. This is valid for conduits having a crushing strength of 1200 pounds per foot. For stronger conduits the height of cover can be reduced.

The unrestricted use of power-activated machines for tamping loose fill, at a height of cover of 36 inches or less, may cause failure of conduit

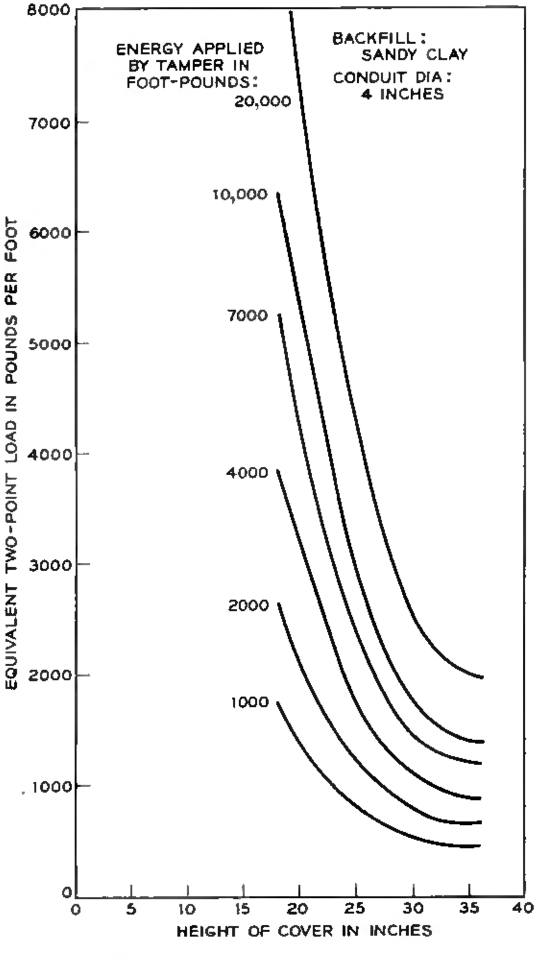


Fig. 11 — Equivalent two-point load caused by tamping on loose fill vs. height of cover.

having a crushing strength of 1200 pounds per foot in two-edge bearing. Compaction of the fill by hand tamping, prior to use of the mechanical tamper, would contribute some improvement. However, this is only of academic interest because, judging from the test results, the height of the hand-tamped cover should be at least 24 inches if a 7500-foot-pound machine is to be used safely at full capacity as a tamper.

REFERENCE

1. Weissmann, G. F., Strength Requirements of Round Conduit, B.S.T.J., **36,** May 1957, p. 737.

Recent Monographs of Bell System Technical Papers Not Published in This Journal*

Amron, I., see Biondi, F. J.

Anderson, E. W., see McCall, D. W.

Arnold, S. M.

Growth of Metal Whiskers on Electrical Components, Monograph 3304.

Ballhausen, C. J. and Liehr, A. D.

Some Comments on Anomalous Magnetic Behavior of Certain Ni(II) Complexes, Monograph 3281.

Becker, E. J., see Biondi, F. J.

Benn, D. R., see Biondi, F. J.

Biondi, F. J., Pondy, P. R., Helmke, G. E., Koontz, D. E., Sullivan, M. V., Amron, I., Feder, D. O., Thomas, C. O., Craft, W. H., Becker, E. J., Elkind, M. J., Benn, D. R., Kern, H. E. and Graney, E. T.

Cleaning of Electronic Devices and Materials, Monograph 3143.

Booth, D. P., see Geller, S.

Bozorth, R. M. and Kramer, V.

Some Ferrimagnetic and Antiferromagnetic Materials at Low Temperatures, Monograph 3308.

Craft, W. H., see Biondi, F. J.

Douglass, D. C., see McCall, D. W.

* Copies of these monographs may be obtained on request to the Publication Department, Bell Telephone Laboratories, Inc., 463 West Street, New York 14, N. Y. The numbers of the monographs should be given in all requests.

Easley, J. W.

Comparison of Neutron Damage in Germanium and Silicon Transistors, Monograph 3294.

Elkind, M. J., see Biondi, F. J.

Feder, D. O., see Biondi, F. J.

Fontana, W. J., see Hovgaard, O. M.

Foote, H. L., Jr., Shair, R. C. and Smith, D. H.

Electrical Storage of Solar Energy, Monograph 3310.

Forster, J. H. and Zuk, P.

Millimicrosecond Diffused Silicon Computer Diodes, Monograph 3290.

Fuller, C. S.

Interactions Between Solutes in Germanium and Silicon, Monograph 3282.

Garn, P. D., see Sharpe, L. H.

Geller, S. and Booth, D. P.

Crystal Structure of Quanidinium Gallium Sulfate Hexahydrate, Monograph 3273.

Geller, S. and Wernick, J. H.

Ternary Semiconducting Compounds with Sodium Chloride-Like Structure, Monograph 3301.

Graham, R. E.

Predictive Quantizing of Television Signals, Monograph 3272.

Graham, R. E. and Kelly, J. L., Jr.

Computer Simulation Chain for Research on Picture Coding, Monograph 3274.

Graney, E. T., see Biondi, F. J.

Green, E. I.

Evolving Technology of Communication, Monograph 3303.

GRUBBS, W. J.

Hall Effect Circulator—A Passive Transmission Device, Monograph 3296.

HAMMING, R. W.

Stable Predictor-Corrector Methods for Ordinary Differential Equations, Monograph 3283.

HAMMING, R. W., see McCall, D. W.

HELMKE, G. E., see Biondi, F. J.

HOUTZ, C. C. and KARLIK, S.

Properties of Tantalum Related to Performance of Solid Electrolytic Capacitors, Monograph 3305.

HOVGAARD, O. M. and FONTANA, W. J.

Versatile Miniature Switching Capsule, Monograph 3306.

IRVIN, H. D.

Sibyl: A Laboratory for Simulation Studies of Man-Machine Systems, Monograph 3292.

ISENBERG, C. R., see Trumbore, F. A.

KARLIK, S., see Houtz, C. C.

KELLY, J. L., JR., see Graham, R. E.

KERN, H. E., see Biondi, F. J.

KOMPFNER, R., see Pierce, J. R.

KOONTZ, D. E., see Biondi, F. J.

KRAMER, H. P.

Symmetrizable Markov Matrices, Monograph 3266.

KRAMER, V., see Bozorth, R. M.

KUNZLER, J. E. and WERNICK, J. H.

Low-Temperature Resistance for Studying Impurities in Zone-Refined Metal Ingots, Monograph 3267.

LAUDISE, R. A.

Kinetics of Hydrothermal Quartz Crystallization, Monograph 3268.

LIEHR, A. D.

Interaction of Vibrational and Electronic Motions in Some Simple Conjugated Hydrocarbons, Monograph 3275.

LIEHR, A. D., see BALLHAUSEN, C. J.

LUNDBERG, J. L., see Nelson, L. S.

MASON, W. P.

Use of Internal Friction Measurements for Causes of Frequency Instabilities, Monograph 3300.

McCALL, D. W., DOUGLASS, D. C. and ANDERSON, E. W.

Self-Diffusion in Liquids: Paraffin Hydrocarbons, Monograph 3298.

McCALL, D. W. and HAMMING, R. W.

Nuclear Magnetic Resonance in Crystals, Monograph 3201.

McCLUSKEY, E. J., JR.

Iterative Combinational Switching Networks—General Design, Monograph 3287.

MILLER, L. E.

Design and Characteristics of a Diffused Silicon Logic Amplifier Transistor, Monograph 3291.

NELSON, L. S. and LUNDBERG, J. L.

Heterogeneous Flash Initiation of Thermal Reactions, Monograph 3269.

OCH, H. G., see Tinus, W. C.

PHILLIPS, J. C.

Vibration Spectra and Specific Heats of Diamond-Type Lattices, Monograph 3270.

PIERCE, J. R. and KOMPFNER, R.

Transoceanic Communication by Means of Satellites, Monograph 3289.

PONDY, P. R., see Biondi, F. J.

PORBANSKY, E. M., see Trumbore, F. A.

REED, E. D.

The Variable-Capacitance Parametric Amplifier, Monograph 3302.

RIDER, D. K., see Schlabach, T. D.

SCHLABACH, T. D. and RIDER, D. K.

Thermal Degradation of Printed Wiring Adhesives, Monograph 3309.

SCHNEIDER, C. and SPAHN, C. F.

Miniature Relay for High Reliability, Monograph 3307.

SHAIR, R. C., see Foote, H. L., Jr.

SHARPE, L. H. and GARN, P. D.

Regenerable Etchant for Copper, Monograph 3288.

SLICHTER, W. P.

Molecular Motion in Polyamides, Monograph 3285.

SMITH, D. H., see Foote, H. L., Jr.

SOBEL, M. and TISCHENDORF, J. A.

Acceptance Sampling With New Life Test Objectives, Monograph 3311.

SPAHN, C. F., see Schneider, C.

SULLIVAN, M. V., see Biondi, F. J.

SYLWESTROWICZ, W. D.

Temperature Dependence of Yield Stress of Copper and Aluminum, Monograph 3286.

TALLEY, H. E.

A Family of Diffused-Base Germanium Transistors, Monograph 3293.

THOMAS, C. O., see Biondi, F. J.

THOMAS, D. G.

Diffusion and Precipitation of Indium in Zinc Oxide, Monograph 3284.

TINUS, W. C. and OCH, H. G.

Systems Engineering for Usefulness and Reliability, Monograph 3278.

TISCHENDORF, J. A., see Sobel, M.

TRUMBORE, F. A., ISENBERG, C. R. and PORBANSKY, E. M.

On Temperature-Dependence of the Distribution Coefficient, Monograph 3279.

VARNERIN, L. J.

Stored Charge Method of Transistor Base Transit Analysis, Monograph 3297.

WARNER, A. W.

Ultra-Precise Quartz Crystal Frequency Standards, Monograph 3299.

WEBER, L. A.

Frequency-Modulation Digital Subset for Data Transmission Over Telephone Lines, Monograph 3280.

WERNICK, J. H., see Geller, S.

WERNICK, J. H., see Kunzler, J. E.

ZUK, P., see Forster, J. H.

Contributors to This Issue

HAROLD L. BARNEY, A.B., 1928, Elon College; M.S., 1929, North Carolina State College; Bell Telephone Laboratories, 1929—. Mr. Barney has been engaged primarily in research in acoustics and speech. He was concerned in work on voice-operated devices for switching and automatic gain control and control terminal equipment for transatlantic radiotelephony. During World War II, he was engaged in studies of speech privacy systems for military use and later made studies of visible speech and speech analysis using the sound spectrograph. In 1948 he did exploratory research on circuit applications of transistors. In 1950 he took charge of a group engaged in design and testing sonar apparatus for the Navy. He is now concerned with studies in psychoacoustics involving speech and hearing processes. Fellow Acoustical Society of America; member I.R.E.

HUGH K. DUNN, A.B., 1918, Miami University; Ph.D., 1925, California Institute of Technology; Bell Telephone Laboratories, 1925—. For a number of years he was engaged in statistical studies of amplitudes and spectra in music and speech, and the characteristics of telephone instruments and circuits in terms of real speech. He took part in the early work on the sound spectrograph, and during World War II he worked on an acoustic torpedo. After the war he returned to speech studies, including development of the transmission-line analog of the vocal tract, and showing how it leads to prediction of vowel formant positions. He has recently been concerned with improvement of the artificial larynx. Fellow Acoustical Society of America, American Association for the Advancement of Science; member American Physical Society, Phi Beta Kappa, Sigma Xi.

ROBERT C. FLETCHER, B.S., 1943, and Ph.D., 1949, Massachusetts Institute of Technology; Bell Telephone Laboratories, 1949—. After joining Bell Laboratories he was engaged in research on traveling wave tubes. In 1952 he turned to studies of semiconductors, including work on irradiation damage and infra-red absorption in germanium and studies of donor impurities in silicon. In 1956 he took charge of a group engaged in development of solid state devices, mainly new magnetic devices such as masers, ferrite sheet memories and twistors. Fellow American Physical Society; senior member I.R.E.; member Sigma Xi.

F. J. Fuchs, Jr., B.S. in M.E., 1947, Duke University; Western Electric Co., 1947—. He has been engaged in manufacturing development engineering and has been particularly concerned with special processes, tooling and machinery for waveguide manufacture. Member American Society of Mechanical Engineers, American Society of Metals.

F. E. Haworth, A.B., 1924, University of Oregon; M.A., 1929, Columbia University; Bell Telephone Laboratories, 1925—. He has been engaged in studies of magnetic materials, dielectrics and crystal analysis by X-ray diffraction and electron diffraction. He has also been engaged in research in physics of electrical contacts and acoustical instruments. Fellow American Physical Society; member Acoustical Society of America, Phi Beta Kappa.

E. J. McCluskey, Jr., A.B., 1953, Bowdoin College; B.S. and M.S., 1953, and Sc.D., 1956, Massachusetts Institute of Technology; Bell Telephone Laboratories, 1955–1959. He is now associate professor of electrical engineering at Princeton University. At Bell Laboratories he was engaged in research in switching theory and logical design, and was a consultant on problems in the design of an electronic telephone central office. He has also been an instructor, lecturer and visiting professor at M.I.T., College of the City of New York and Princeton University. Member I.R.E., Phi Beta Kappa, Tau Beta Pi, Sigma Xi, Eta Kappa Nu.

Duncan M. Mitchel, Bell Telephone Laboratories, 1953—. Since joining Bell Laboratories Mr. Mitchel has been a member of the construction methods group in the Outside Plant Department. He has worked chiefly on underground conduits.

R. G. Rausch, B.S., 1949, M.S., 1950, and Ph.D., 1956, Princeton University; Bell Telephone Laboratories, 1951—. As a member of a military systems studies group he has been engaged in general mathematical investigations of systems, system design, hydraulic servo design and simulation studies. Member Sigma Xi, Phi Beta Kappa.

Harold Seidel, B.E.E., 1943, College of the City of New York; M.E.E., 1947, and D.E.E., 1954, Polytechnic Institute of Brooklyn;

Microwave Research Institute, 1947; Arma Corp, 1947–48; Federal Telecommunications Labs, 1948–53; Bell Telephone Laboratories, 1953—. He has been concerned with general electromagnetic problems, especially regarding waveguide applications and with analysis of microwave ferrite devices. Member I.R.E., Sigma Xi.

ERLING D. SUNDE, E.E., 1926, Technische Hochschule, Darmstadt, Germany; American Telephone and Telegraph Company, 1927–34; Bell Telephone Laboratories, 1934—. He has made theoretical and experimental studies of inductive interference from railway and power systems, lightning protection of the telephone plant and fundamental transmission studies in connection with the use of pulse modulation systems. Author of *Earth Conduction Effects in Transmission Systems*, a Bell Laboratories Series book. Senior member I.R.E.; member A.I.E.E., American Mathematical Society, American Association for the Advancement of Science.

DONALD E. THOMAS, B.S. in E.E., 1929, Pennsylvania State University; M.A., 1932, Columbia University; Bell Telephone Laboratories, 1929—. His early work was in development of telephone submarine cable systems. He later turned to sea and airborne radar development until he left to serve in the Signal Corps and the Air Force as a member of the Joint and Combined Chiefs of Staff Committees on Radio Countermeasures during World War II. When he returned in 1946, he took part in development and installation of the first deep sea repeatered submarine telephone cable system. Since 1950 he has been engaged in development and evaluation of new semiconductor devices. Senior member I.R.E.; member Tau Beta Pi, Phi Kappa Phi.

GERD F. WEISSMANN, Dipl.-Ing., Technical University of Berlin, 1950; M.S., 1953, Pennsylvania State University; Bell Telephone Laboratories, 1953—. His work has been in stress analysis, engineering mechanics, strain measurements and metal properties and testing. He has recently been concerned with soil mechanics studies and with an investigation of the damping properties of materials. Member Society for Experimental Stress Analysis.