Frequency Modulation

L. B. ARGUIMBAU
McIntosh Laboratory, Binghampton, New York

AND

R. D. STUART
Lately of Research Laboratory of Electronics
Massachusetts Institute of Technology

LONDON: METHUEN & CO. LTD
NEW YORK: JOHN WILEY & SONS, INC.

METHUEN'S MONOGRAPHS
ON PHYSICAL SUBJECTS

First published in 1956

CATALOGUE NO. 4065/U (METHUEN)
PRINTED AND BOUND IN GREAT BRITAIN AT THE PITMAN PRESS, BATH

PREFACE

THE frequency-modulation system as developed by Armstrong has found its greatest usefulness in reducing the effect of extraneous disturbances introduced during transmission. For this reason any discussion of the system should place primary emphasis on this feature.

The improvement is achieved by using changes of instantaneous frequency that are large in comparison with those occurring naturally. Thus the concept of changing instantaneous frequency is of great importance. Much detailed circuitry has been discussed in the light of this concept.

Attempts to make rapid changes in frequency give rise to circuit response problems that restrict the naïve concepts of instantaneous frequency. These problems must be studied in terms of spectra.

This monograph is largely based on experience gained by the writers in the Research Laboratory of Electronics of the Massachusetts Institute of Technology. Except for the fact that we have rather accidentally been the ones to gather the material in connected form, a dozen names might just as well have been included on the title page. We hope others will recognize their work whether or not specific credit is given.

Free use has been made of material from the chapter on frequency modulation in *Vacuum Tube Circuits and Transistors* by L. B. Arguimbau, John Wiley and Sons, 1956.

L. B. ARGUIMBAU AND R. D. STUART

CONTENTS

CHAPTER I

INTRODUCTION

COMMUNICATIONS is concerned with the transmission of information over large areas or between fixed points. Many different types of information are concerned giving rise to many specialized branches of communications such as telegraphy, telephony, broadcasting, television, facsimile, teletype, and telemetering. In some of these systems a continuous record of the information, as in sound broadcasting and in others, only a finite number of samples, as in television, is transmitted. However, all systems have certain things in common. In each the information is converted into an electrical signal which varies with time in either a continuous manner as in telephony or in a discontinuous manner as in telegraphy.

In some cases these electrical signals may be sent directly along wires, and many such systems are in use today. In some cases the signal to noise ratio of this simple system may be improved by the use of volume compression and frequency pre-emphasis. This direct wire method has not found widespread use for broadcast services. High frequency electrical signals can be radiated through space for great distances and the varying electric current representing the information can be used to modify the high frequency electric current in some way.

There are at present many different ways in which the signal currents can be used to modulate the carrier and several of them are in use.

The method most widely used so far is amplitude modulation in which the amplitude of the high frequency carrier is varied in accordance with the signal to be transmitted. We shall discuss frequency modulation where the frequency instead of the amplitude of the carrier is varied in accordance with the signal to be transmitted. We shall be concerned

mainly with its use for transmitting speech, music, and pictures. Amplitude modulation will be used as a standard of comparison.

In any practical system there will always be a certain amount of noise arising from such sources as static, thermal agitation, and random emission. There is also the question of interference resulting from many unwanted signals arriving at the receiver both from other transmitters and from multiple transmission paths. The desirable features of any communication system are high signal to noise ratio and freedom from interference. Both of these factors are influenced by the type of modulation employed. If the receiver is near to the transmitter then there is little to choose between the various systems. In providing a broadcast service we wish to obtain the maximum possible coverage with a minimum number of transmitters and so we need the best signal to noise ratio obtainable. In providing a communications service where the primary object is the transmission of intelligence, quality being of secondary importance, the signal to noise ratio will limit the range at which the service is useful. The ability to discriminate against interference governs the maximum number of channels it is possible to provide in a given area within the limited bandwidth available for the service.

It has long been known that systems using amplitude modulation suffer from what has been called static interference. Static consists of sharp electrical impulses generated by electrical machinery and by natural atmospheric discharges. Very sharp pulses have a fairly uniform frequency distribution provided the individual pulses constituting the noise last for shorter times than the period of the frequency considered. The spectrum of atmospheric noise energy is more or less constant over the first few megacycles but is definitely smaller at a hundred megacycles and is very small at thousands of megacycles. In either case, the energy distribution is uniform enough so that we can say the received noise energy is proportional to the bandwidth of the receiver. It behaves very much like thermal agitation noise.

One method proposed to overcome noise consisted of

varying the frequency instead of the amplitude of the carrier at an audio rate by very small amounts of the order of 100 c.p.s. This would "eliminate the sidebands" and so reduce the bandwidth required with a consequent reduction in noise. The signals were to be received by a very sharply resonant receiver tuned to a frequency slightly higher than the carrier, and so converted to amplitude variations which were then to be treated in the conventional manner. The fallacy in this system is now well known. The sharply tuned resonant circuit can follow the frequency changes only if they are made very slowly. When the frequency of the variations becomes greater than the bandwidth of the tuned circuits the changes can no longer be followed faithfully.

It will be shown later that a frequency-modulated wave can be regarded as consisting of a carrier with associated sidebands. However, unlike the amplitude modulation case there is an infinite number of sidebands instead of one pair for each modulating frequency. This suggests that frequency modulation rather than reducing the bandwidth would tend to increase it. CARSON (1922, 1928), in a mathematical investigation of the problem, corrected the narrow band fallacy and pointed out that the sideband amplitude depended upon modulation frequency (cf. amplitude modulation).

The problem was attacked in a very different way by ARMSTRONG (1936). It was argued that the way to avoid static was to produce artificially a type of signal differing as radically as possible from those occurring in nature. A receiver could then be built which is insensitive to static interference. It was argued further that random noise constituted an amplitude variation but had no orderly variation of frequency. Thus a signal having large changes in frequency should be separable from random noise. A receiver could then be built which is sensitive to such frequency shifts but not to amplitude variations. The possibilities of an improved signal to noise ratio from a frequency-modulation system were demonstrated.

The difference in the effect of noise in the case of amplitude modulation and the case of frequency modulation may be

seen by considering a single noise component. Suppose we have two unmodulated signals, one wanted and the other unwanted. These signals can be represented by two vectors.

Assume the magnitude of the wanted signal is unity and that of the unwanted signal a. In general, the wanted and unwanted signals will have different frequencies and so will move in and out of phase. This causes the amplitude of the resultant to vary from $1 + a$ to $1 - a$ and the wanted signal is in effect amplitude modulated at the difference frequency.

In order to make the effects of the noise small it would be desirable for the amplitude modulation applied to the wanted carrier to be large compared with the amplitude modulation due to the presence of unwanted signals. However, there is a definite upper limit to what can be accomplished here, since the amplitude modulation on the original wanted signal cannot exceed 100 per cent. In addition to producing amplitude modulation of the wanted carrier the interference will also produce frequency modulation. It is seen from a vector diagram that when the unwanted signal is in quadrature with the wanted signal then the phase of the latter is either advanced or retarded. The frequency modulation of the wanted signal is produced by the unwanted signal moving in and out of phase with the wanted carrier. However, provided the unwanted signal is not larger than the wanted signal then this phase shift can never exceed $\pm 90°$. There is no limitation theoretically on the phase shifts which can be used to represent the wanted modulation and so we have the possibility of a great improvement in the signal to noise ratio. A more precise treatment of this matter will be given in Chapter V.

When the information consists of speech or music there is another method of reducing the effects of noise which is applicable to any system of modulation, and also in other instances, such as disc recording (CROSBY, 1940). The main energy of speech and music is concentrated at the low end of the spectrum, whereas the energy of noise is almost uniformly distributed in the part of the spectrum which concerns us. It

has been found that the high audio frequency components in the original signal can be greatly increased in magnitude without material increase in the peak amplitude of the signal. We then reduce the amplitude of the high frequencies at the receiver to obtain the original waveform and at the same time the high frequency noise components are reduced, giving an improvement in the overall signal to noise ratio. Experiments on a wide variety of programmes have shown that a weighting factor given by $k = \sqrt{[1 + (\omega/\omega_0)^2]}$ does not greatly alter the peak amplitude if ω_0 is correctly chosen. Actually, its magnitude is not critical and values of $(1/\omega_0)$ from 25 μsec to 100 μsec have been successfully used. Assuming uniform distribution of noise, it can be shown that the r.m.s. improvement in signal to noise ratio is given by

$$\left[\frac{f_0}{f_c}\int_0^{f_c}\frac{df}{1+(f/f_0)^2}\right]^{-1/2} \text{ or } [(f_0/f_c)\tan^{-1}(f_c/f_0)]^{-1/2}$$

where f_c is the upper limit of the audio spectrum considered. Taking $f_c = 21$ kc and $f_0 = 2100$ c.p.s. (corresponding to 75 μsec), the improvement in signal to noise ratio is about 8 db.

This artifice has not come into use in amplitude modulation broadcasting for several reasons. The change would require modification of all existing receivers, increasing the high frequency sidebands would increase interchannel interference, and an increase in percentage modulation at high frequencies would put severe requirements on detector design. These arguments do not apply in the case of frequency modulation, and pre-emphasis has become standard practice. The noise interference in frequency modulation is greatest at the highest audio frequencies and so pre-emphasis is particularly appropriate, and for the limits given above results in an improvement of about 16 db rather than 8 db.

A further improvement can be obtained by the use of volume compression and expansion. In this system the low energy levels are raised in volume before transmission, and then reduced at the receiver.

REFERENCES

ARMSTRONG: *Proc. I.R.E.*, **24**, 689 (1936)
CARSON: *Proc. I.R.E.*, **10**, 57 (1922)
CARSON: *Proc. I.R.E.*, **16**, 966 (1928)
CROSBY: *R.C.A. Rev.*, **4**, 349 (1940)

CHAPTER II

THE NATURE OF FREQUENCY MODULATION

2.1. Introduction. The last chapter discussed a frequency-modulation system without any careful analysis of what is meant by a changing frequency. At first glance no fussy academic analysis would seem to be necessary. The frequency is slowly varied by changing the tuning of an oscillator. This shifts the frequency along a resonance curve, varies the response, and transmits intelligence. Without this fundamental notion it is doubtful if any progress would have been made on the development of a frequency-modulation system. It is certainly the most important idea of what frequency modulation is.

For the most part, circuits are considered from a steady-state alternating-current point of view. As a matter of fact, no alternating-current has existed from prehistoric times and backward for an infinite time in the past. All physical circuits and alternating-voltage sources are brought into existence at some point in time. If we treat them carefully (and there is seldom any reason for doing so) the current is found to be the sum of a transient and a steady-state solution. In all ordinary cases the transient becomes totally negligible after a short time and the steady-state response is all that need be considered. In computing the steady solution we assume that the circuit has always been functioning, and that the applied sinusoidal voltage has been connected to it for an indefinitely long time.

We are already accustomed to taking liberties with this steady-state idea. We permit the amplitude of a "sine wave" to vary and assume that as long as the rate of variation is small in comparison to the bandwidths of any circuits involved then the instantaneous response can be found on a steady-state basis. When the change in amplitude is more rapid we still define an instantaneous amplitude even though we cannot

use steady-state methods in finding the response. It is easy to think of the peak value of any one half cycle of the wave as representing the "instantaneous amplitude." Alternatively we consider the momentary length of the rotating vector as the amplitude.

The same extensions are reasonable in the idea of instantaneous frequency; we permit it to vary slowly from cycle to cycle. As long as the rate of variation is small in comparison to the bandwidth involved we can again apply steady-state solutions. The instantaneous frequency is the rate at which cycles are very nearly repeating themselves. When we prefer we can think of the instantaneous angular frequency as the momentary speed of rotation of the vector representing the wave.

2.2. Concept of Frequency. The primitive notion of frequency is that of a certain number of cycles per second. It is the time rate of change of the angle locating the vector. Thus if we write

$$e(t) = a\cos\phi = a\cos\omega t = a\cos 2\pi ft \qquad (2.1)$$

ϕ is the instantaneous angle and ω the angular frequency. The amplitude, a, is assumed constant. It is convenient to consider the radian frequency as the time derivative of the angle,

$$\omega = \frac{d\phi}{dt} \quad . \quad . \quad . \quad . \quad (2.2)$$

In the particular case where a and ω are constants (independent of time) the meaning of this expression is very definite. We have already permitted the generalization of Eq. 2.1 for time-varying amplitude. Equation 2.1 can be set up as an extended definition of frequency. It is convenient to do this, and the extension leads to useful results.

A little care is necessary at this point to avoid a trivial objection. Although the meaning of Eq. 2.1 is clear for constant amplitude and frequency, it should be admitted that the vector chosen to represent a given disturbance such as

$e(t)$ is not unique. For example, consider the ellipse in Fig. 2.1 given by the equations,

$$\begin{cases} x = c\cos\omega t \\ y = d\sin\omega t \end{cases} \quad . \quad . \quad . \quad (2.3)$$

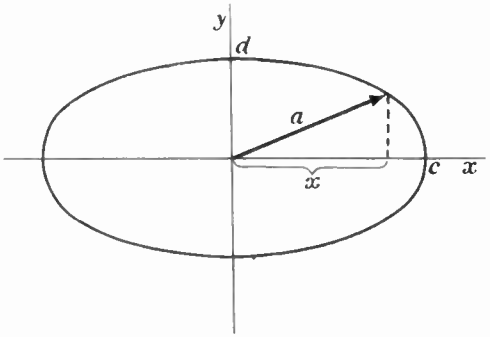

Fig. 2.1. An absurd extension of vector representations

This ellipse can be regarded as generated by a vector of variable length a rotating at a variable speed

$$a = \sqrt{c^2\cos^2\omega t + d^2\sin^2\omega t} \quad . \quad . \quad (2.4)$$

$$\phi = \tan^{-1}\frac{y}{x} = \tan^{-1}\left(\frac{d}{c}\tan\omega t\right) \quad . \quad . \quad (2.5)$$

Here a is by no means constant unless c and d are equal. Furthermore ϕ does not increase linearly with time unless they are equal. In spite of this the projection, x, varies sinusoidally with time. Thus we have the worthless result that we can regard a constant-frequency, constant-amplitude sinusoidal signal as a variable-amplitude, variable-frequency signal.

If the choice of amplitude and frequency variation is properly made there is no practical ambiguity in using

$\omega = d\phi/dt$ to define frequency. This procedure requires that the percentage change in a or in ω be kept small during a cycle, a condition that is not met by the example of Fig. 2.1.

A constant-amplitude, variable-frequency signal can be regarded as generated by a vector of constant length rotating with slowly-varying speed. For amplitude-modulated waves we write

$$e(t) = a(t) \cos pt$$

or

$$e(t) = a_0(1 + m \cos qt) \cos pt$$

We are tempted to argue by analogy and write

$$e(t) = a_0 \cos [\omega(t)]t \quad . \quad . \quad . \quad (2.6)$$

or

$$e(t) = a_0 \cos [p + \Delta\omega \cos qt]t \quad . \quad (2.7)$$

The expressions are crossed out because it happens to be one of the instances of snap judgment failing us. Differentiation of the angles in the two cases with respect to time gives

$$\omega = [\omega(t)] + t \frac{d\omega(t)}{dt}$$

and

$$\omega = [p + \Delta\omega \cos qt] - tq\Delta\omega \sin qt \quad . \quad (2.8)$$

The portions in the brackets are what we hoped the frequencies would be. For large values of t the second, unbracketed terms become very important.

This trouble is analogous to the difficulty we might have had when first studying accelerated motion. We might have written $s = vt$ for variable velocities. Clearly we should write $v = ds/dt$ and $s = \int v\, dt$. In the present case we should write $\omega = d\phi/dt$ and $\phi = \int \omega\, dt$.

This gives

$$e(t) = a \cos \{\int \omega(t)\, dt\} \quad . \quad . \quad . \quad . \quad (2.9)$$

or for the special case we were trying to treat in Eq. 2.8

$$\omega(t) = p + \Delta\omega \cos qt \quad . \quad . \quad . \quad . \quad (2.10)$$

$$\phi(t) = \int \omega(t)\, dt = \int (p + \Delta\omega \cos qt)\, dt \quad . \quad (2.11)$$

$$\phi(t) = pt + (\Delta\omega/q) \sin qt \quad . \quad . \quad . \quad (2.12)$$

and $$e(t) = a \cos [pt + (\Delta\omega/q) \sin qt] \quad . \quad . \quad (2.13)$$

For convenience the constants of integration have been chosen at zero.

The most important application of these ideas about instantaneous frequency comes in determining the effect of interference in a frequency-modulated system. We add two vectors, determine the resulting angle, and then differentiate this to find the effect on frequency. An interfering signal as we saw in the last chapter introduces an uncertainty (equal to its own magnitude) in our knowledge of amplitudes. It also brings an uncertainty in our knowledge of the instantaneous angle and hence of the frequency.

2.3. Spectrum of a Frequency-modulated Signal. Up to this point we have discussed the methods of transmitting and receiving signals whose frequency varies rather slowly with time, the message being carried by frequency variations. In amplitude modulation, attempts to signal too rapidly by abrupt changes of amplitude lead to trouble. The tuned circuits act sluggishly and it is necessary for them to respond not only to the carrier but also to frequencies separated from the carrier by the audio frequency. The circuits must have more or less constant response over a frequency range of twice the audio bandwidth.

In the present chapter we have carefully avoided any discussion of tuned circuit dynamics because we have wanted to emphasize the concept of a slowly varying instantaneous frequency. We have assumed frequency changes large in comparison with the audio frequencies. In the typical frequency-modulation broadcast systems we use bandwidths of at least 150 kc and with such bandwidths the audio speeds of frequency variation are so low that the dynamics of the tuned circuits are of little consequence.

It is fortunate that the dynamics of a tuned circuit are often of very little importance because it is *most difficult* to obtain any accurate knowledge of the response of circuits to a signal whose frequency changes at a rate comparable to the bandwidth. Perhaps some day a neat, simple answer will be found but that day has not yet arrived.

In just the same way that an amplitude-modulated wave can be expanded into a trigonometric series the spectrum of a frequency-modulated signal can be found. Such a wave is also composed of sine waves. We can find the response of our circuits to each of these sine waves, add the results, and thereby find the response to the frequency-modulated wave. Unfortunately, however, a sinusoidally frequency-modulated wave, unlike its amplitude-modulated counterpart, contains not three terms, but tens, hundreds, or even thousands of significant components. There is no *theoretical* difficulty in finding the responses to all of these waves, adding them and thereby getting our answer. However, the answer when obtained (hours or years later) would not throw much light on our problem, as it would cover pages of paper and would be most difficult to visualize and interpret.

The *idea* of a variable instantaneous frequency has led to the important results that frequency modulation has achieved and we shall return to it later in studying interference. A signal of unit amplitude having the instantaneous angular frequency

$$\omega(t) = p + \Delta\omega \cos qt \quad . \quad . \quad (2.14)$$

has the instantaneous voltage value

$$e(t) = \cos [pt + (\Delta\omega/q) \sin qt] \quad . \quad (2.15)$$

Here p is the average carrier frequency, $\Delta\omega$ is the peak variation of this frequency, and q is the audio-frequency rate of variation. It is convenient to define the frequency-deviation ratio as

$$m_f = \Delta\omega/q \quad . \quad . \quad . \quad (2.16)$$

It is the ratio of the peak deviation to the audio rate at which it occurs. Thus Eq. 2.15 becomes

$$e(t) = \cos (pt + m_f \sin qt) \quad . \quad . \quad (2.17)$$

This is readily expanded to give

$$e(t) = \cos pt \cos (m_f \sin qt) - \sin pt \sin (m_f \sin qt). \quad (2.18)$$

We must now try to find the meaning of these expressions, $\cos (m_f \sin qt)$ and $\sin (m_f \sin qt)$. Whatever other properties they may have, one thing is obvious: they are periodic functions with the period $2\pi/q$. The argument $m_f \sin qt$, repeats itself as t changes by $2\pi/q$. Thus, every single-valued function of $m_f \sin qt$ must repeat itself with the same period. This suggests an attempt to represent the functions as Fourier series,

$$\cos (m_f \sin qt) = a_0(m_f) + a_1(m_f) \cos qt + a_2(m_f) \cos 2qt + \ldots \quad . \quad (2.19)$$

$$\sin (m_f \sin qt) = b_1(m_f) \sin qt + b_2(m_f) \sin 2qt + \ldots \quad . \quad (2.20)$$

The sine terms in the first expression and the cosine terms in the second are missing because they are even and odd functions respectively.

Following the ordinary rules, we have

$$a_0(m_f) = \frac{1}{T}\int_0^T \cos (m_f \sin qt)\, dt \quad . \quad (2.21)$$

$$a_n(m_f) = \frac{2}{T}\int_0^T \cos (m_f \sin qt) \cos nqt\, dt \quad . \quad (2.22)$$

Similarly, $b_n(m_f) = \frac{2}{T}\int_0^T \sin (m_f \sin qt) \sin nqt\, dt \quad . \quad (2.23)$

As we shall see shortly, these expressions can be evaluated without much difficulty by series methods, but we should recognize that the integrals are merely functions of t integrated

over fixed intervals and must result, by graphical integration if necessary, in constants dependent on m_f.

Expanding the integral of Eq. 2.21 in a power series gives

$$a_0(m_f) = \frac{1}{T}\int_0^T \left[1 - \frac{1}{2!}(m_f \sin qt)^2 + \frac{1}{4!}(m_f \sin qt)^4 - \ldots\right] dt \quad . \quad . \quad (2.24)$$

But

$$\int_0^{2\pi} \sin^k x\, dx = \frac{1 \cdot 3 \cdot 5 \ldots (k-1)}{2 \cdot 4 \cdot 6 \ldots k} \cdot 2\pi, \quad k \text{ even.} \quad . \quad (2.25)$$

This leads to

$$a_0(m_f) = 1 - \frac{1}{2^2} m_f^2 + \frac{1}{2^2 4^2} m_f^4 - \frac{1}{2^2 4^2 6^2} m_f^6 + \ldots \quad (2.26)$$

In a similar way it can be shown that for even values of n,

$$a_n(m_f) = 2\left\{\frac{m_f{}^n}{2^n n!}\left[1 - \frac{1}{2(2n+2)} m_f^2 + \frac{1}{2 \cdot 4 \cdot (2n+2)(2n+4)} m_f^4 - \ldots\right]\right\} \quad . \quad (2.27)$$

For odd values of n, $a_n(m_f) = 0$. The sine coefficients (the b's of Eq. 2.23) have the same form as the a's except that odd values of n must be used instead of even values.

The various coefficients have fortunately been computed and tabulated. It turns out that they are Bessel functions,

$$J_0(m_f) = a_0(m_f) = \frac{1}{T}\int_0^T \cos(m_f \sin qt)\, dt \quad . \quad . \quad . \quad (2.28)$$

$$J_n(m_f) = \tfrac{1}{2} a_n(m_f) = \frac{1}{T}\int_0^T \cos(m_f \sin qt) \cos nqt\, dt, \; n \text{ even} \quad . \quad (2.29)$$

$$J_n(m_f) = \tfrac{1}{2}b_n(m_f)$$

$$= \frac{1}{T}\int_0^T \sin\,(m_f \sin qt) \sin nqt\, dt,\ n \text{ odd} \quad . \quad (2.30)$$

Thus,

$$\cos\,(m_f \sin qt) = J_0(m_f) + 2\sum_{n=1}^{\infty} J_{2n}(m_f) \cos 2nqt \quad . \quad . \quad (2.31)$$

$$\sin\,(m_f \sin qt) = 2\sum_{n=0}^{\infty} J_{2n+1}(m_f) \cos\,(2n+1)qt \quad . \quad . \quad (2.32)$$

Direct substitution shows that these functions satisfy Bessel's equation,

$$\frac{d^2y}{dx^2} + \frac{1}{x}\frac{dy}{dx} + \left(1 - \frac{n^2}{x^2}\right)y = 0 \quad . \quad (2.33)$$

which arises from attempts to solve Laplace's equation or the wave equation in cylindrical coordinates.

It remains to substitute the results back in Eq. 2.15. This substitution gives

$$\begin{aligned} e(t) = {} & J_0(m_f)\cos pt + J_1(m_f)\cos\,(p+q)t - J_1(m_f)\cos\,(p-q)t \\ & + J_2(m_f)\cos\,(p+2q)t + J_2(m_f)\cos\,(p-2q)t \\ & + J_3(m_f)\cos\,(p+3q)t - J_3(m_f)\cos\,(p-3q)t \\ & + J_4(m_f)\cos\,(p+4q)t + J_4(m_f)\cos\,(p-4q)t + \ldots \end{aligned} \quad . \quad . \quad (2.34)$$

Eq. 2.34 shows that the spectrum consists of a carrier and an infinite number of sidebands, all of whose amplitudes are various-order Bessel functions of m_f. Graphs of these functions are given in Fig. 2.2. It will be noticed that $J_0(m_f)$ decreases slowly until it vanishes altogether for $m_f \cong 2{\cdot}4$. The variable m_f is the ratio of the frequency deviation to the audio-frequency rate. Thus, if the audio frequency is left constant and the magnitude of the frequency deviation is

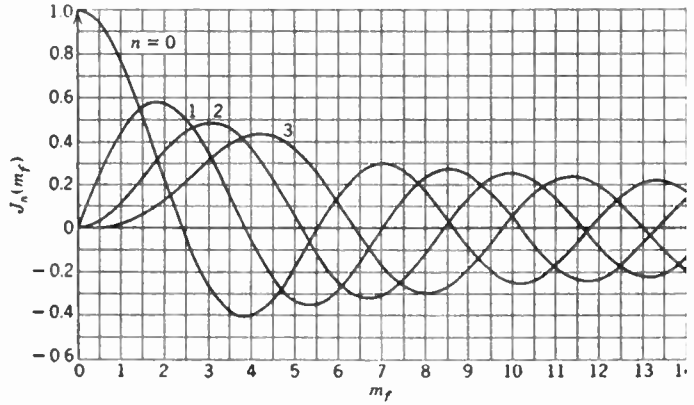

Fig. 2.2. Bessel functions of the first kind $J_n(m_f)$

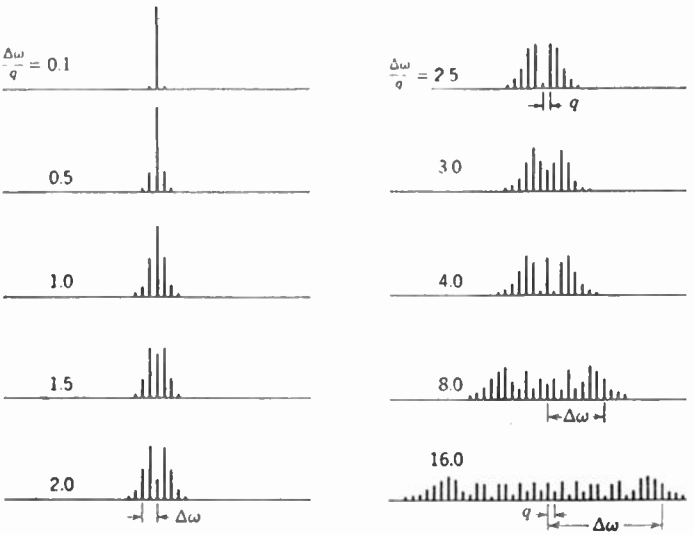

Fig. 2.3. Spectra of signals modulated sinusoidally in frequency. The audio frequency is held constant as the deviation is increased. (Adapted from Balth. van der Pol 'Frequency Modulation,' *Proc. I.R.E.*, 18 [1930]. Courtesy Institute of Radio Engineers)

increased, the carrier decreases in magnitude until it disappears when the frequency deviation is a little less than two and a half times the audio frequency. The carrier then increases again in opposite phase. As the frequency swing is

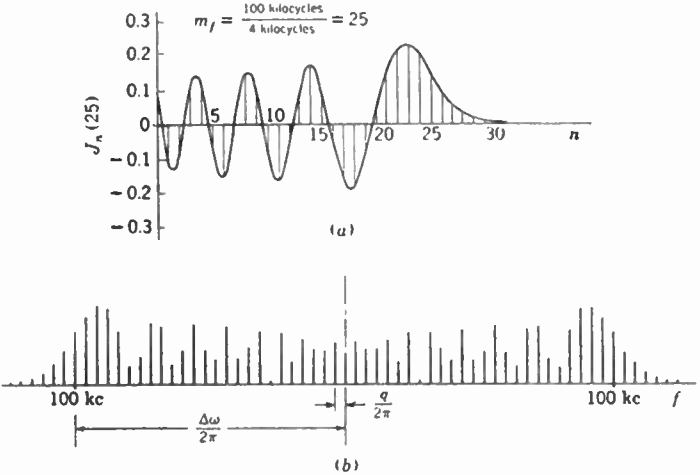

Fig. 2.4. Evolution of a frequency-modulation spectrum. (Reprinted by permission from M.I.T. Staff, *Applied Electronics* (Wiley, 1943))

increased continually, the carrier disappears again and again, reversing phase at each null.

As the carrier decreases with increased swing the sidebands begin to appear. This is clearly indicated in Fig. 2.3, which shows the spectra of frequency-modulated waves. The spectra are drawn for a fixed modulating frequency and variable frequency swing. When $m_f = 0{\cdot}1$ a pair of sidebands is visible. For $m_f = 1$ the carrier has decreased slightly and three pairs of side frequencies are in evidence. For $m_f = 2{\cdot}5$ the carrier amplitude has passed through zero. At $m_f = 4$ the carrier has reappeared and the first-order sidebands have passed through zero. It should be noticed that the spacing

between the frequencies is unchanged but the number of significant terms increases when the frequency swing is increased. When the frequency swing reaches sixteen times the audio frequency it will be noted that the spectrum covers the whole range over which the radio frequency is swept. As a matter of fact, several pairs of side frequencies are noticeable beyond this range of frequency variation. The fine structure of one of these spectra is made a little clearer by a careful study of Fig. 2.4*a*, which shows an envelope of

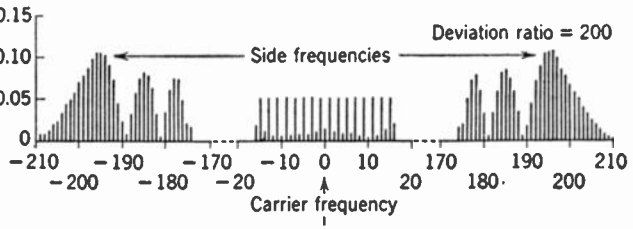

Fig. 2.5. The spectrum of a signal with a high deviation ratio (J. F. Morrison, *Proc. I.R.E.*, 28 [October, 1940]. Courtesy Institute of Radio Engineers)

the individual terms for a deviation ratio of twenty-five. The corresponding spectrum is drawn in Fig. 2.4*b*. It shows the spectrum of a signal which is varied sinusoidally by plus and minus 100 kc four thousand times per second. The spectrum corresponding to a much higher deviation ratio is shown in Fig. 2.5. (See also CORRINGTON, 1947.)

At first sight these spectra are a little surprising. We have become accustomed to the idea of an amplitude-modulated signal which consists of a carrier and two sidebands, but the notion of a carrier and many sidebands is a little confusing. What, if anything, is the inner meaning of the situation? Before a working system was demonstrated, these questions were bothersome. People wondered whether or not frequency modulation might ever prove useful. Although Bessel functions were not so familiar as trigonometric functions, the formal expansion of a frequency-modulated signal into its

series representation was not of any fundamental difficulty. On the other hand, the attention that was focused on this formal side of the problem tended to obscure the direct physical approach later used so successfully by Armstrong.

Having derived an expression for the spectrum of a frequency-modulated wave it is desirable that we should try to understand it. An idea of the meaning of sidebands in amplitude modulation can be gained by a graphical, vector treatment. This approach helps to a certain extent here as well. An amplitude-modulated signal is given by

$$e(t) = \cos pt + \frac{m}{2}\cos(p+q)t + \frac{m}{2}\cos(p-q)t \quad (2.35)$$

which is represented by Fig. 2.6. The components out of phase with the carrier cancel, while the in-phase components of the sidebands add, to produce a vector of sinusoidally varying length which adds and subtracts from the carrier to produce a sinusoidally varying amplitude.

For small deviation ratios all but the carrier and one pair of sidebands can be neglected and a frequency-modulated signal is given by

$$e(t) \cong \cos pt + J_1(m_f)\cos(p+q)t - J_1(m_f)\cos(p-q)t \quad . \quad . \quad . \quad (2.36)$$

For very small values of m_f this is approximately

$$e(t) \cong \cos pt + \frac{m_f}{2}\cos(p+q)t - \frac{m_f}{2}\cos(p-q)t \quad . \quad (2.37)$$

since the initial value of the derivative of the first-order Bessel function is 0·5. It should be noted that the two expressions for the amplitude- and frequency-modulated signals are identical except for a change in the direction of the lower sideband. Fig. 2.7 indicates that the two sidebands can be considered as uniting to form a variable-length vector at right angles to the carrier. This vector varies the phase of the resultant producing the desired frequency modulation.

When the value of m_f increases, the angular deviation gets larger and the resultant vector would vary slightly if this variation were not compensated in some way. At the instants of maximum positive and negative angular frequency the resultant would have its largest value, and for zero deviation the value would be a minimum; thus the resultant vector would vary in length at twice the audio frequency. We know

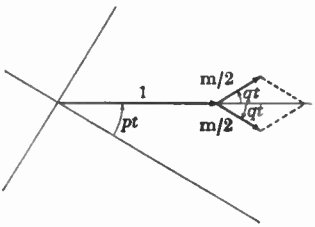

Fig. 2.6. Vector diagram of an amplitude-modulated signal

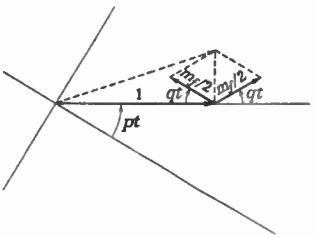

Fig. 2.7. Vector diagram of a frequency-modulated signal

as a matter of fact that no such variation in amplitude occurs in the purely frequency-modulated signal that has been assumed. The variation is suppressed by the second-order sidebands which produce an in-phase vector varying in length at twice the audio-frequency rate. The third-order sidebands can be thought of as producing an out-of-phase vector which compensates for the nonsinusoidal variation of phase contributed by the first-order sidebands.

TABLE 1

Roots of $J_n(m_f)$

$n = 0$	$n = 1$	$n = 2$	$n = 3$	$n = 4$	$n = 5$
2·4048	3·832	5·135	6·379	7·586	8·780
5·520	7·016	8·417	9·760	11·064	12·339
8·654	10·173	11·620	13·017	14·373	15·700
11·792	13·323	14·796	16·224	17·616	18·982
14·931	16·470	17·960	19·410	20·827	22·220

One application of the spectrum idea is found in the measurement of frequency deviation. If the carrier is observed by a heterodyne frequency meter or oscillating receiver it is

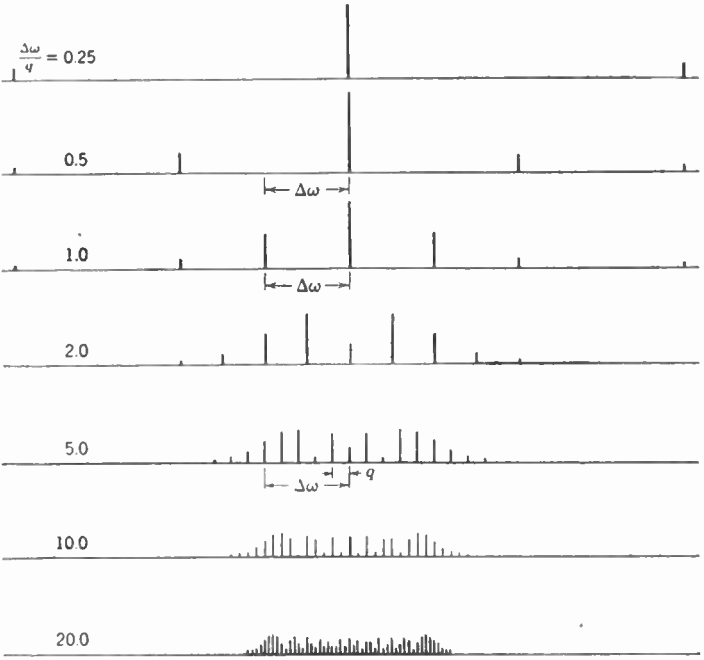

Fig. 2.8. Spectra of signals modulated sinusoidally in frequency. Here the peak frequency deviation is held constant and the audio frequency is varied. (Adapted from Balth. van der Pol, 'Frequency Modulation,' *Proc. I.R.E.*, 18 [1930]. Courtesy Institute of Radio Engineers)

found to pass through a succession of zeros as the frequency deviation is increased. Correlation of these disappearances with the known roots of the Bessel functions gives a measure of m_f. Table 1 may be useful in this respect. It shows the

values of m_f for which the amplitudes of the various components vanish.

A very much more important application of the series can be recognized easily when the spectra are plotted somewhat differently as in Fig. 2.8. Here the frequency deviation is left constant but the audio rate is varied. A careful consideration of these spectra shows the fallacy of attempting to compress the bandwidth by means of frequency modulation. When the audio rate is one-twentieth of the deviation ($m_f = 20$) the components are fairly closely confined to the range over which the frequency is varied. As the audio rate is increased ($m_f = 10$), the spacing between the sidebands increases and their magnitude increases, but the range of the spectrum covered is not greatly influenced. When the audio rate is still further increased to four times the deviation ($m_f = 0{\cdot}25$), the carrier itself is the only component included in the original frequency range. Only one pair of sidebands is in evidence, and they are separated from the carrier by four times the frequency deviation. If the frequency spectrum for a slowly-varying frequency ($m_f = 20$) is impressed on a tuned circuit, the response can be computed by static ideas as implied in the early articles of this chapter. If a tuned circuit is arranged to pass frequencies within the band over which the frequency is swept and suppress those far out of this band, the spectrum for $m_f = 0{\cdot}25$ gives the same response as would a totally unmodulated carrier. No information is conveyed.

This result is important because it gives a quantitative explanation of the fallacy of attempting to compress the spectrum by frequency modulation. We have two alternate ways of considering the problem. The tuned circuit can be thought of as sluggish and suppressing rapid changes, or it can be thought of as clipping the sidebands which lie outside the frequency region that intuition might indicate.

REFERENCE

CORRINGTON: *Proc. I.R.E.*, **35**, 1013 (1947)

CHAPTER III

TRANSMITTERS

3.1. Introduction. The function of the transmitter is to put the information to be transmitted on to the high frequency carrier wave and then to radiate this modulated carrier. Transmitters for frequency modulation are largely conventional; the information in this case causes the frequency of

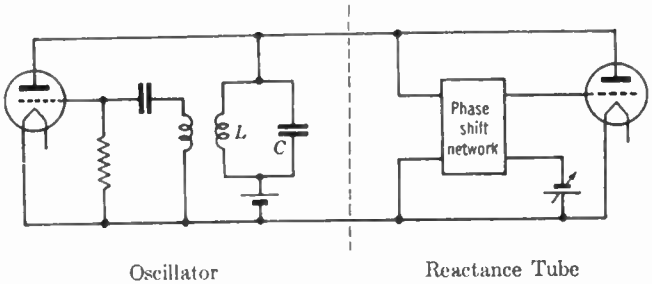

Fig. 3.1. The oscillator frequency is varied by the reactance-tube bias

the carrier to vary and so the differences characteristic of the use of frequency modulation lie mainly in the modulator. The absence of any amplitude variations enables the final stages of the transmitter to be operated in class C, which gives high efficiency.

3.2. Reactance-tube Modulator. Several different types of frequency-modulated transmitters are in common use. The reactance-tube type is the most direct in its operation. The circuit is shown in Fig. 3.1. A tube has its plate circuit connected in parallel to the tuned circuit of the oscillator which is to be modulated. If the tube is to achieve the desired results, it should act like a variable reactance whose value depends

upon the grid bias. As the grid bias varies at an audio rate, this modulates the oscillator frequency.

The reactance-tube circuit conducts an alternating plate current whose vector value depends upon the plate and grid voltages and upon the tube parameters. In the circuit of Fig. 3.1, the plate voltage is merely the applied input voltage developed by the oscillator across its tank circuit. The grid voltage is a constant times the input voltage, the complex

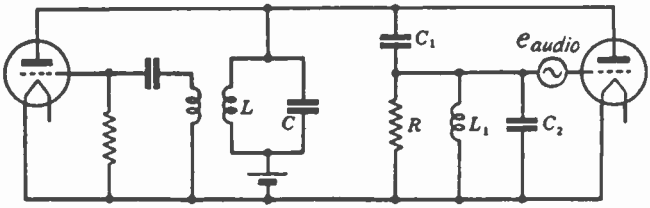

Fig. 3.2. A simple frequency-modulation transmitter

value of the constant depending upon the phase-shifting network.

It is usual to use a pentode for the reactance tube. When this is done the plate current depends almost entirely on the grid voltage, becoming merely $g_m E_g$. If the tube is to be a pure reactance the plate current and hence the grid voltage should be in quadrature to the plate voltage.

Suppose the oscillator in Fig. 3.2 is arranged to operate at a frequency of 5 Mc and to be modulated by $\pm$ 10 kc at this frequency. Assuming the plate resistance to be infinite, $C = 200\,\mu\mu$f, $C_1 = 15\,\mu\mu$f, $C_2 = 3\,\mu\mu$f, and $R = 5000\,\Omega$, we shall determine the value L_1 must have if amplitude modulation is to be avoided and the total change in g_m required to produce the desired frequency modulation. In the figure the audio voltage is for convenience drawn in the grid lead, but in practice it would be placed next to ground to avoid stray capacitance. We assume a voltage, E_p, across the tuned circuit and establish conditions for infinite impedance (or current balance) at this point. We can neglect the effect of

oscillator grid-circuit loading and write down the oscillator plate as a negative resistance, including thereby the effect of the oscillator tube's mutual conductance. Were it not for the effect of the reactance tube the whole oscillator could be considered as an infinite impedance, but at equilibrium the oscillator must readjust to give a net infinite impedance including the reactance tube.

$$0 = \frac{I}{E_p} = \frac{1}{-R_{\text{osc}}} + j\omega C + \frac{1}{j\omega L} + Y_{\text{reactance tube}} \quad . \quad (3.1)$$

The reactance-tube current can be considered as having three components, a current through C_1 dependent solely on the passive C_1, R, L_1, C_2 phase network, a plate current for zero e_{audio}, and a changed radio frequency plate current due to the change in tube parameters accompanying e_{audio}. It is perfectly possible to solve for these by brute-force methods, but experience with this particular circuit shows that it is more illuminating to solve for the reactance-tube grid voltage in terms of E_p and proceed from there.

Figs. 3.3 and 3.4 show the grid circuit redrawn. The equivalent current source is found by shorting the C_2 terminals, giving a current of $j\omega C_1 E_p$ as indicated. This gives a 90° phase shift between E_g and E_p when the current can flow through a pure resistance, that is, when L_1 resonates with $C_1 + C_2$. In any case the grid voltage is

$$E_g = \frac{j\omega C_1 E_p}{\frac{1}{R} + j\omega(C_1 + C_2) + \frac{1}{j\omega L_1}} = \frac{j\omega C_1 R E_p}{1 + j\omega(C_1 + C_2)R + \frac{R}{j\omega L_1}}$$

$$\cong j\omega C_1 R E_p \quad . \quad . \quad . \quad (3.2)$$

assuming that the circuit is resonant and varies negligibly over the frequency range concerned. The current drawn by C_1 becomes (Fig. 3.3)

$$I_{C_1} = j\omega C_1(E_p - E_g) = j\omega C_1(1 - j\omega C_1 R)E_p$$

$$= (\omega^2 C_1^2 R + j\omega C_1)E_p \quad . \quad . \quad . \quad . \quad (3.3)$$

Finally then,

$$\frac{I}{E_p} = 0 \cong + \frac{1}{-R_{osc}} + j\omega C + \frac{1}{j\omega L} + \omega^2 C_1^2 R + j\omega C_1 + (g_m + \Delta g_m)(j\omega C_1 R) + \frac{1}{r_p + \Delta r_p} \quad . \qquad (3.4)$$

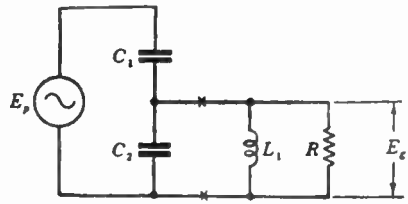

Fig. 3.3. The phase-shift network of Fig. 3.2 is broken between C_2 and L_1

The amplitude of oscillations adjusts itself as usual to make $-1/R_{osc}$ balance out the other real terms. The rectified oscillator grid voltage appearing across the grid resistor is a

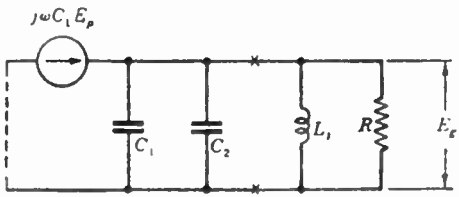

Fig. 3.4. An equivalent current source is substituted for the circuit to the left of the break in Fig. 3.3

fairly good measure of the oscillation envelope. In practice the phase shift network should be adjusted until headphones connected across part of the grid resistor indicate a zero audio voltage corresponding to a constant unmodulated amplitude. This balance occurs very nearly at the point previously mentioned where L_1 resonates $C_1 + C_2$. There may be a

small departure from this if $1/(-R_{osc})$ varies with frequency or if the effect of Δr_p is appreciable.

The frequency adjusts itself as a function of e_{audio} to make the imaginary terms cancel.

$$j\omega[C + C_1(1 + g_m R + \Delta g_m R)] + \frac{1}{j\omega L} = 0 \quad . \qquad (3.5)$$

Thus, we should make L resonant with $C + C_1(1 + g_m R)$ and the change in frequency can then be computed as due to a capacitance, $C_1 R \Delta g_m$, shunted across the circuit.

To avoid amplitude modulation the incremental plate current should be in quadrature to E_g. Assuming r_p infinite this means the grid circuit should be resonant as already assumed, or

$$L_1 = \frac{1}{\omega_0^2(C_1 + C_2)} = \frac{1}{4\pi^2 \times 5^2 \times 10^{12}(200 + 3) \times 10^{-12}}$$
$$= 5{\cdot}0\,\mu\text{h}$$

To compute the change required in g_m we note that

$$\frac{f_0 + \Delta f}{f_0} = \sqrt{\frac{C + C_1(1 + g_m R)}{C + C_1(1 + g_m R_1 + \Delta g_m R)}}$$
$$= \sqrt{\frac{200 + 3(1 + 0{\cdot}002 \times 5000)}{200 + 3(1 + 0{\cdot}002 \times 5000) + 5000\Delta g_m}}$$
$$= \sqrt{\frac{233}{233 + 15{,}000\Delta g_m}}$$
$$1 + \frac{\Delta f}{f_0} \cong 1 - \frac{15{,}000\Delta g_m}{2 \times 233}$$
$$\Delta g_m \cong -\frac{233 \times 2}{15{,}000} \frac{\pm 10}{5000} = 62\,\mu \text{ mhos}$$

On first reading, this calculation may seem a little complicated, but the details should not keep us from the simple physical principles. The admittance term, $j\omega C_1 R \Delta g_m$, is a

grid-bias-controlled capacitance. No matter how queer the reactance-tube circuit may look, it acts like a condenser, pure and simple. Furthermore, it is a condenser whose value can be controlled by the grid bias, and so, as a practical matter, it can be varied dynamically at an audio-frequency rate.

It is useful to pause and think through once more what is meant by impedance. Its meaning here is much like that

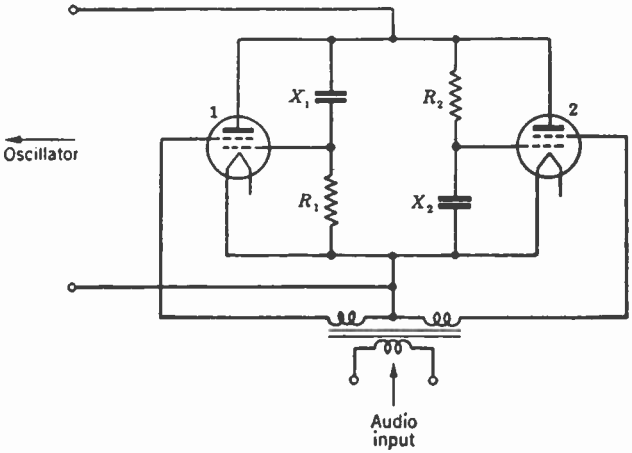

Fig. 3.5. A balanced reactance-tube circuit

obtained under inverse-feedback conditions, and has no new implications. It is the vector ratio of the fundamental components of voltage and current.

A note on the power-handling capacity of the reactance tube is in order. Notice that the tube must be capable of operating at the same plate swing as the oscillator tube. Furthermore, the current drawn may not be at all negligible. The current in the oscillator tank circuit may be very large if the tank condenser is large. This means that the current drawn by the reactance tube, while only a small percentage of the tank current, may be appreciable. As a practical

matter it is sometimes necessary to use a tube of higher power for the reactor than for the oscillator.

The use of a reactance tube has one serious limitation. It must be arranged to vary the frequency dynamically but it must not cause the oscillator frequency to drift from changes in line voltage or similar causes. To avoid this, a balanced circuit such as that of Fig. 3.5 is often used. In commercial transmitters an automatic control circuit must be added. This circuit will be discussed in the next article.

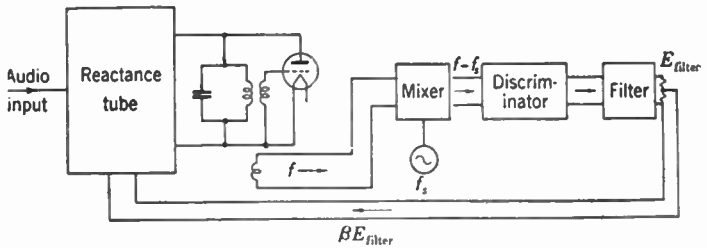

Fig. 3.6. A frequency-modulated transmitter employing automatic control of the mean frequency

As a practical matter, reactance-tube circuits are usually constructed at frequencies below 10 Mc, and the output signal is then fed through a series of frequency multipliers.

3.3. Automatic Control of Mean Transmitter Frequency. The balanced reactance-tube circuit is much less susceptible to slow drift than is a single tube. In spite of this the stability of a transmitter using such a circuit would not be adequate for commercial use. Even if the reactance tube itself could be made perfectly stable, temperature changes in the oscillator tank circuit would be regarded as prohibitive in commercial practice. In a reactance-tube frequency-modulated transmitter use of a crystal oscillator is clearly impractical. For this reason some sort of automatic control is called for.

In Fig. 3.6 a reactance tube is connected across the tank circuit of an oscillator. The resultant frequency, f, is heterodyned in a mixer unit by a standard frequency, f_s. The output

frequency, $f - f_s$, is impressed on a discriminator circuit. This circuit has the property of producing a direct output voltage which varies linearly with the input frequency. The output voltage is put through a filter and a portion is used to bias the reactance tube. If conditions are properly arranged the bias acts in such a manner as to reduce the frequency drift. At the same time the low-pass filter prevents the control circuit from following rapid variations of frequency. Thus, the effect of slow changes in the circuit constants is reduced, but rapid frequency modulation is permitted. The larger the bias that is developed by the discriminator circuit for a given shift in frequency the smaller will be the drift. On the other hand, if too much control is used there is a danger of the frequency variation oscillating.

To determine the improvement effected by the control circuit let us work backward following the method adopted for inverse feedback. We assume as actual shift, $\Delta f_{\text{resultant}}$, occurring sinusoidally. This vector (!) $\Delta f_{\text{resultant}}$, operating first through the discriminator, and then the β divider circuit, places a voltage on the reactance tube that if acting by itself would make the frequency change by $\beta K \Delta f_{\text{resultant}}$. This, added to the factors giving rise to the original shift, gives the resultant shift,

$$\Delta f_{\text{original}} + \beta K \Delta f_{\text{resultant}} = \Delta f_{\text{resultant}}$$

$$\Delta f_{\text{resultant}} = \frac{\Delta f_{\text{original}}}{1 - \beta K} \quad . \quad . \quad . \quad (3.6)$$

3.4. Armstrong Frequency-modulation Transmitter. The reactance-tube type of transmitter has the serious defect of using a relatively unstable oscillator combined with automatic frequency control. It is perfectly possible to derive a frequency-modulated signal from a stable, constant source by means of an amplifier whose phase shift can be varied in some relationship to voice pressure. A circuit for doing this is shown schematically in Fig. 3.7. It serves as an elegant application of the generalized concept of frequency. The primary source is a crystal oscillator having a very stable

output frequency. This is connected to a phase-splitting network which provides two signals differing by 90 degrees in phase but having the same frequency. One of these signals

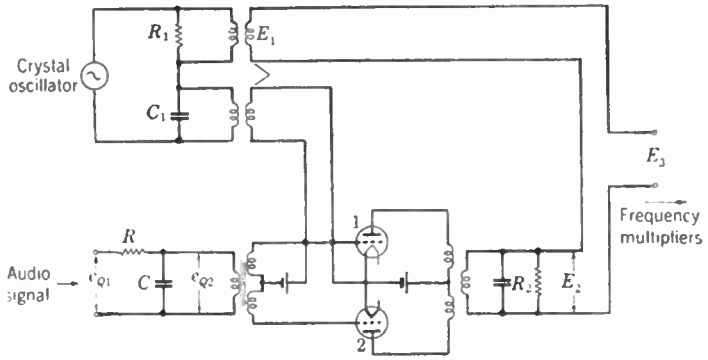

Fig. 3.7. An Armstrong phase modulator

is connected directly to the output, but the other is transmitted through a balanced modulator. The radio-frequency voltage which is fed into the common branch of the modulator gives no output whatsoever when the modulator is in a symmetrical, balanced condition. Thus, if no voltage is connected from

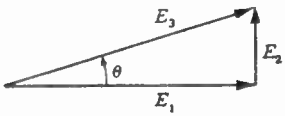

Fig. 3.8. Variation of E_3 results in frequency modulation of E_3

grid to grid of the modulator, the output, E_2, is zero. On the other hand, if the voltage on the grid of tube 1 is made positive relative to the voltage on the grid of tube 2, the alternating plate current in the first tube will exceed that in the second tube and a voltage will appear at E_2. This output voltage will differ in phase from the voltage E_1 by 90 degrees.

Let us say that it leads by this amount. (The direction depends upon the direction of the transformer windings.) As indicated in Fig. 3.8, the phase of the output voltage is given by

$$\theta = \tan^{-1}(E_2/E_1) \quad . \quad . \quad . \quad (3.7)$$

which for small values of E_2/E_1 is essentially

$$\theta \cong E_2/E_1 \quad . \quad . \quad . \quad (3.8)$$

Thus the output phase can be controlled at will by means of the grid-to-grid voltage. The output of a balanced modulator for a small unbalancing voltage varies linearly with that voltage. Thus we can write

$$\theta \cong ke_{Q2} \quad . \quad . \quad . \quad (3.9)$$

where e_{Q2} is the instantaneous voice voltage across the primary of the grid-to-grid transformer. The angle between the output vector voltage E_3 and an arbitrary, fixed time axis is given by

$$\phi = pt + \theta \quad . \quad . \quad . \quad (3.10)$$

The instantaneous value of the output frequency is the time derivative of this angle so that

$$\omega = \frac{d\phi}{dt} = p + \frac{d\theta}{dt} = p + k\frac{de_{Q2}}{dt} \quad . \quad (3.11)$$

The frequency given by Eq. 3.11 is not the desired value. We should like the frequency to be proportional to the voice pressure, not proportional to the derivative of that pressure. As a matter of fact, the output *phase* shift is proportional to e_{Q2}, whereas we should like to have the *frequency* shift proportional to the sound pressure. It is not difficult to bring this about. If a circuit can be made to give an output voltage e_{Q2} which is proportional to the integral of the sound pressure, the result will come out correctly. This is readily done by the integrating circuit shown in Fig. 3.9. If the reactance of the condenser, C, is made small in comparison to the resistance, R, at all frequency components of the input

wave, then the current is very nearly e_{Q1} divided by the resistance. But

$$e_{Q2} = \frac{1}{C}\int i\,dt = \frac{1}{CR}\int e_{Q1}\,dt \quad . \quad . \quad (3.12)$$

and Eq. 3.11 becomes

$$\omega = p + k\frac{d}{dt}\left[\frac{1}{CR}\int e_{Q1}\,dt\right] = p + \frac{k}{CR}e_{Q1} \quad . \quad (3.13)$$

Eq. 3.13 shows that with a proper integrating circuit the frequency can be made to vary linearly with the sound

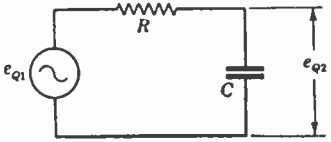

Fig. 3.9. The audio signal is integrated before it is used in the transmitter

pressure. (Of course an additional pre-emphasis circuit of the sort described on p. 5 must be included in an Armstrong transmitter.) If the angular variation is allowed to become too great the tangent is no longer equal to the angle (Eqs. 3.7 and 3.8) and distortion is introduced, limiting the frequency deviation obtainable. The effect is made clearer by studying a numerical case.

Suppose we wish to determine the maximum frequency deviation that can be obtained at an audio-frequency rate of 30 c.p.s. by the phase-variation method if the distortion is to be kept below 2 per cent. Let the instantaneous angular deviation be given by

$$\theta = \tan^{-1} y = \tan^{-1}(y_m \cos qt) \quad . \quad (3.14)$$

where y is the ratio of the instantaneous out-of-phase voltage (E_2 in Fig. 3.8) to the in-phase component (E_1 in Fig. 3.8). Equation 3.14 can be written

$$\theta = y - \tfrac{1}{3}y^3 + \tfrac{1}{5}y^5 - \tfrac{1}{7}y^7 + \ldots \quad . \quad (3.15)$$

But

$$\Delta\omega = 2\pi\Delta f = \frac{d\theta}{dt} = \frac{d\theta}{dy}\frac{dy}{dt}$$

$$= (1 - y^2 + y^4 - y^6 + \ldots)(-y_m q \sin qt)$$

$$\cong -y_m q \sin qt + y_m{}^3 q \cos^2 qt \sin qt$$

$$\cong -y_m q \sin qt + \tfrac{1}{2} y_m{}^3 q(1 + \cos 2qt) \sin qt$$

$$\cong (-y_m q + \tfrac{1}{4} y_m{}^3 q) \sin qt + \tfrac{1}{4} y_m{}^3 q \sin 3qt \quad (3.16)$$

$$\left|\frac{\Delta\omega_{3q}}{\Delta\omega_q}\right| \cong \frac{y_m{}^2}{4} \quad . \quad . \quad . \quad (3.17)$$

Hence

$$\theta_m \cong \tan^{-1} 2\sqrt{\left|\frac{\Delta\omega_{3q}}{\Delta\omega_q}\right|} \quad . \quad . \quad (3.18)$$

In particular if the ratio of third harmonic to fundamental is to be kept below 2 per cent,

$$\theta_m < \tan^{-1} 2\sqrt{0{\cdot}02} = 15{\cdot}8^\circ$$

But

$$\Delta f_m = f_q \tan^{-1} \theta_m \cong f_q \theta_m$$

or

$$\Delta f_m = f_q \frac{15{\cdot}8}{360} \cdot 2\pi \cong 0{\cdot}275 f_q$$

Thus for an audio frequency of 30 c.p.s.

$$\Delta f_m < 0{\cdot}275 \times 30$$

$$\Delta f_m < 8{\cdot}2 \text{ cps}$$

This example shows that to keep the distortion below 2 per cent the input must be limited to a value producing a deviation equal to about one-quarter of the audio frequency. Thus, we must not modulate the frequency by more than plus or minus 8 cycles at a 30-cycle rate, but we can modulate it by 80 cycles at 300 cycles.

The frequency changes just computed are totally inadequate. Even with the improved circuits they must be increased

many-fold by frequency multiplication. It is necessary to digress and consider the relationships obtained during multiplication. When a signal is applied to a nonlinear distorting device, such as a class C amplifier, the output wave contains integral multiples of the applied frequency. If the applied frequency is changed, the harmonics are all changed by the same percentage. Thus, if a signal is varying in frequency the percentage variation of the harmonics is the same, but the numerical variation is multiplied by the order of the harmonic. One further point should be thought through and understood. If the frequency of the fundamental is shifted by a given amount once a second, all the harmonics will be shifted *once* a second, not n times. This is readily seen by direct physical reasoning. At an instant of maximum fundamental frequency all harmonics will have maximum frequency. *The audio-frequency rate at which frequency changes occur is not modified by frequency multiplication.*

The same thing is true of an amplitude-modulated signal. After frequency multiplication the harmonics are maximum when the signal is maximum, and the harmonic amplitudes are minimum when the signal is minimum. This result causes confusion at times if looked at in terms of sidebands. We might think that connecting the modulated signal to a frequency doubler would double all the component frequencies and hence double the sideband separation and with it the effective modulation frequency. That this is not the case is readily seen from the simple physical reasoning that has just been given.

It is illuminating to consider the effect of a square-law device on a frequency-modulated signal. When we impress a frequency-modulated signal given by Eq. 2.17,

$$e(t) = a \cos (pt + m_f \sin qt),\ m_f = \Delta\omega/q,$$

on a tube having the characteristic, $i = e^2$, then,

$$\begin{aligned} i &= a^2 \cos^2 (pt + m_f \sin qt) \\ &= \tfrac{1}{2}a^2[1 + \cos 2(pt + m_f \sin qt)] \quad . \quad (3.19) \end{aligned}$$

This shows without more ado that the instantaneous frequency including the deviations is just doubled.

In order to obtain a clearer picture of the practical transmitter design problem, let us consider the construction of a typical broadcast transmitter. We shall determine the frequency multiplication necessary if a transmitter is to operate at 100 Mc with a frequency deviation of $\pm$ 75 kc. The transmitter is to be capable of full modulation at all audio frequencies above 60 c.p.s., with a distortion of less than 2 per cent. As we saw above, the maximum undistorted frequency deviation obtainable is about one-quarter of the audio frequency. Thus at 60 c.p.s. a deviation of about 15 c.p.s. is permissible and a multiplication of 75,000/15 or 5000 is necessary.

It has been found that modulator circuit components such as the transformer are very readily constructed at some low radio frequency such as 200 kc. If we were to start at this frequency and multiply it by the required amount we would obtain a final frequency of 1000 Mc! This amount is far more than the required 100 Mc. The numerical frequency deviation is of the desired value, but the percentage change is one-tenth of the required amount.

At the present writing it is still considered a little awkward to work with frequencies as high as 1000 Mc. If it were not for this we might heterodyne by 900 Mc and obtain 100 Mc with the required deviation. Notice that numerical frequency deviations are not modified by heterodyning but the percentage deviations can be greatly increased. To avoid the 1000 Mc we can just as well multiply the 200 kc signal by 200, obtaining a 40 Mc signal with a deviation of $\pm$ 200 $\times$ 15 or 3 kc. This signal is then heterodyned to a tenth the frequency (to 4 Mc), which increases the percentage frequency change to the required amount. The 4 Mc signal is then multiplied by 25:1, giving the required 100 Mc + 75 kc.

The example just discussed has two practical difficulties. If the original 200 kc crystal oscillator were to drift by, say, four parts in a million, and the heterodyning oscillator were to drift by an equal amount in the opposite direction, the

resultant signal would shift by twenty times as much, or by eighty parts in a million. This means that as a practical matter the oscillators must be made to have twenty times the final required stability. Furthermore, the five-thousand-to-one frequency multiplication, if conservatively accomplished with doublers, would require approximately twelve stages.

The frequency-stability problem can be solved in a very simple way by deriving the heterodyne oscillator signal from the original 200 kc crystal. Thus, to heterodyne from 40 Mc to 4 Mc in our example, a frequency of 40 − 4, or 36 Mc, is required. This is the 180th harmonic of the 200 kc crystal. A little consideration shows that this expedient makes the net percentage frequency shift the same as that of the original crystal. If desired, both frequency-multiplier chains can be frequency modulated, but in opposite phase, giving double deviation.

A very important advantage of the frequency-modulation system from a transmitter-design viewpoint is the fact that the amplitude is constant and the high-power stages can be operated at full Class C efficiency. It is merely necessary to get the power to the antenna without any regard whatsoever to linearity. The power amplifiers can be operated in a manner that would give atrocious results with an amplitude-modulated transmitter. As a matter of fact, in a certain sense the poorest possible transmitter from an amplitude basis is the best for frequency-modulation use. An amplifier whose output is independent of the input serves as a limiter and reduces any incidental amplitude modulation which may be present in the transmitter.

3.5. Serrasoid Modulator. The extensive development of pulse modulation has given rise to a number of new techniques. We can obtain a phase modulated wave by taking the fundamental component of a pulse-phase modulated wave. A particular form of this type of modulator has been developed by DAY (1948) and has been called a "serrasoid modulator." With this type of modulator much larger phase shifts can be obtained with good linearity and high signal to noise ratios are possible. When used to provide frequency modulation

peak phase shifts of the order of $\pm 150°$ are possible with equivalent FM distortion of a few tenths of 1 per cent, and the noise originating in the modulator can be kept some 80 db below 100 per cent modulation.

REFERENCE

DAY: *Electronics*, **21**, 72 (1948)

CHAPTER IV

RECEIVERS

4.1. Introduction. The function of the receiver is to select the required carrier from the other signals in the spectrum and to recover the original information from the modulation of the wave. Receivers for frequency modulation are largely conventional, the major difference lying in the circuits required for detection of the wave. Since the information is contained in the variation of frequency it has become common practice to remove any amplitude variations either by means of a limiter or by using a detector circuit which is insensitive to variations in input amplitude. It is by no means obvious that this will lead to an optimum system. The variations in amplitude will have been caused by some form of interference, since none exist in the radiated signal. This interference may also disturb the frequency of the carrier. Thus the amplitude variation contains information about the interference and this information may be of use in reducing the interference in the output of the receiver. However, so far, no system of interference reduction based on this principle has been used in practice.

The following discusses methods of obtaining an audio voltage proportional to the instantaneous frequency of the radio signal and independent of its amplitude.

4.2. Limiters. There are many types of amplifier circuits designed to have variable gain, overloading in such a way as to make the output roughly independent of the input once a certain threshold is reached. Three of these are discussed below.

Clipper Circuits. One simple way of thinking about frequency is to regard it as proportional to the number of zero crossings per second executed by the signal. It is pleasant to base the design of limiters on this concept. In Fig. 4.1 we have taken a variable-amplitude wave and chopped off

the tops and bottoms forming a square wave having the same zero-crossing positions as the original wave. Since the second harmonic is absent in a square wave, no frequency below the third harmonic need be considered. We next apply this wave to a broadly-tuned circuit which passes all frequencies in the vicinity of the fundamental component of the square wave but does not pass the other harmonics of the wave. The resultant is approximately a sine wave having

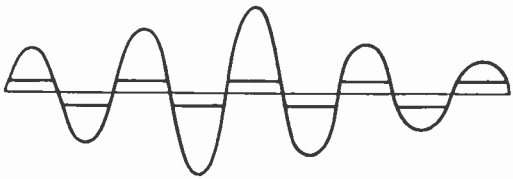

Fig. 4.1. The tops of an amplitude-modulated wave are clipped off to form a square wave with the same axis crossings

nearly the same zero crossings as the original signal but having constant height.

In order to convert a sine wave into a square wave we can use any device that has a saturated characteristic, one in which a negative swing cuts off the current while a positive swing brings the current to its maximum value. A circuit for achieving this is shown in Fig. 4.2. Either a pentode or a special tube (such as the 6BN6) which is more readily saturated than are grid-type tubes can be used. The essential point to be noticed is that this circuit does not depend for its operation on any condenser-resistance time constants and so it is fast acting and able to follow the rapid envelope changes generated by interference. The plate impedance is high so that the tuned circuit acts as a true filter and does not react significantly on the tube action.

Clamping Circuits. The circuit drawn in Fig. 4.3 involves an interesting concept. One or two back-biased diodes are connected across a tuned circuit as indicated. Until the peak value of IR exceeds E_0 the diodes are non-conducting and the quotient of the fundamental components of e and I is

merely R, the tuned-circuit resonant impedance. Now, if the diodes are ideal the voltage across them can never go positive

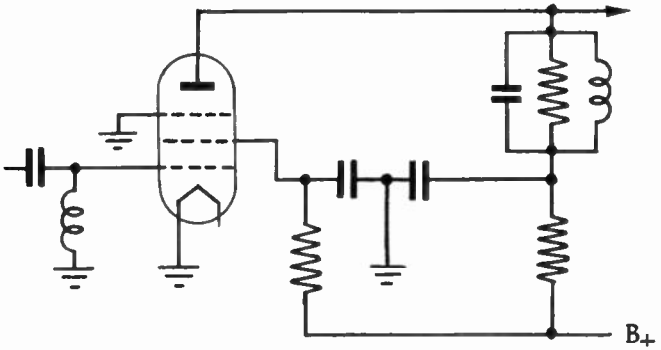

Fig. 4.2. The *Clipper Circuit* produces the result illustrated in Fig. 4.1

without an infinite current flow. Hence, with a good enough tuned circuit to keep e moderately sinusoidal, the peak

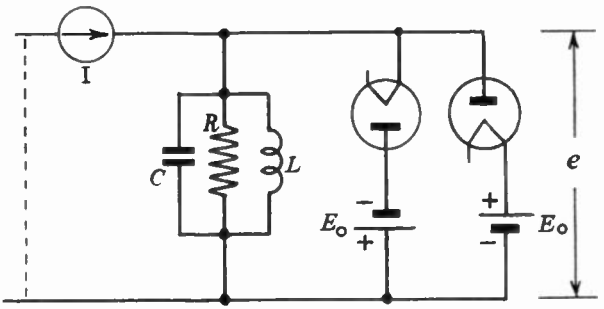

Fig. 4.3. With ideal diodes the voltage is clamped at E_0

fundamental component of e rises until it reaches E_0, and then never goes appreciably beyond this value regardless of how large I may become. The voltage is *clamped* at E_0.

Pentode Limiter. The two limiter circuits we have just discussed are both quick acting. Due to the non-availability of ideal diodes the clamping circuit is relatively ineffective and several stages must be cascaded for good results. At the present writing the pentode limiter is most widely used commercially, even though it is less effective than the clipper type.

A pentode limiter is shown in Fig. 4.4. The grid circuit is biased by grid current flow, the condenser charging to

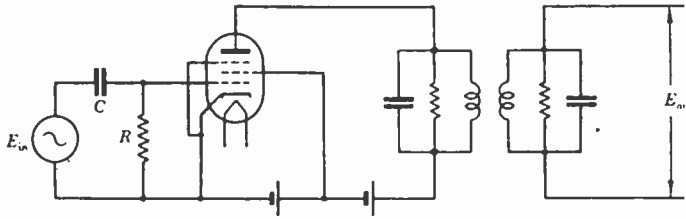

Fig. 4.4. A pentode limiter

approximately the peak value of the input signal. Thus the grid-to-cathode voltage swings between zero and minus twice the peak input voltage. If the screen voltage is adjusted to a rather low value, current flow is limited to comparatively short pulses.

The situation with idealized grid-circuit operation is shown in Fig. 4.5. The grid voltage is assumed to go just to zero. With a large enough grid swing to drive the current to zero the current flows in pulses of constant peak height, the conduction angle depending on the swing. For very low swings the output increases with input, but as soon as the plate current breaks up into pulses the conduction angle decreases with swing and the output decreases.

This account of the operation is oversimplified. Actually the grid is driven slightly positive. The amount it is driven positive depends on the input voltage and upon the grid resistor and tube. When the input voltage is large the grid goes further positive than with a small input voltage. Thus,

when the input is high the conduction angle is decreased but the positive excursion is increased. The two effects tend to balance. This balance is indicated qualitatively in Fig. 4.6.

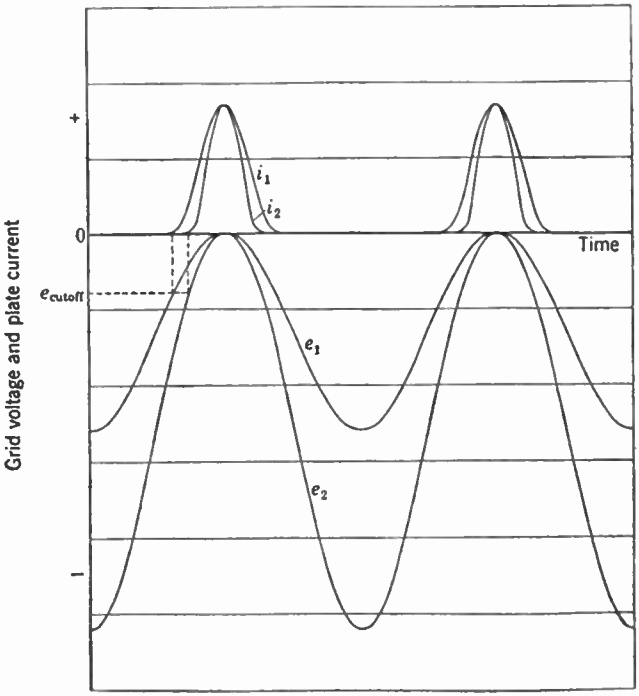

Fig. 4.5. Waveforms in limiter using high-resistance grid leak

With a favourable choice of grid resistor the characteristic is very nearly flat above a certain threshold grid input value of a few volts. The minimum threshold voltage required for effective limiting is fundamentally determined by the curvature of the tube characteristic and hence by the distibution of electron emission velocities.

If a pentode limiter would actually work in the way we have just described, it would be entirely satisfactory. In practice, however, the amplitude variations are not slow but may take

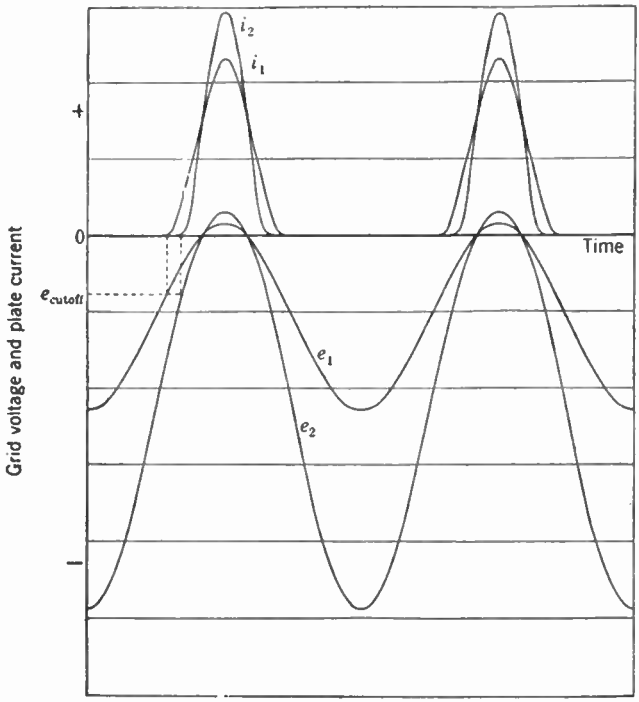

Fig. 4.6. Waveforms in limiter using medium-value grid leak

as little as 6 μsec for a complete envelope cycle. In Fig. 4.7 we have shown the resultant, a, of a desired signal, 1 cos pt, and an interfering signal, b cos $(p + r)t$. The instantaneous angle between them is rt; the instantaneous amplitude is given by

$$a^2 = 1 + b^2 + 2b \cos rt \quad . \quad . \quad (4.1)$$

The grid condenser and grid leak resistor in Fig. 4.4, together with the grid-to-cathode path and the source, act very much like a diode detector circuit. If the grid is to conduct once per r.f. cycle, a condition we have assumed in considering the pentode limiter's operation, then the voltage at the grid must be capable of following the modulation envelope at all times. In other words the circuit must be free from what has been termed "diagonal clipping" in connection with a.m. detectors. Suppose we have a voltage envelope $a(t)$ at the grid. Then the current, averaged over the r.f. cycle, flowing from the CR circuit to the grid is $\frac{a}{R} + C\frac{da}{dt}$. Since the grid-cathode path in the valve can pass current in only one direction this must always be positive if diagonal clipping is to be avoided. Hence

$$C\frac{da}{dt} + \frac{a}{R} \geqslant 0 \quad . \quad . \quad . \quad (4.2)$$

From Eq. 4.1

$$-\frac{br \sin rt}{\sqrt{1 + b^2 + 2b \cos rt}} + \frac{\sqrt{1 + b^2 + 2b \cos rt}}{CR} \geqslant 0$$

$$\frac{1 + b^2}{CR} + \frac{2b}{CR} \cos rt - br \sin rt \geqslant 0 \text{ for all } t.$$

$$. \quad . \quad . \quad (4.3)$$

Since the second two terms are in quadrature their largest negative value is allowed for if

$$\frac{1 + b^2}{CR} - \sqrt{\frac{4b^2}{(CR)^2} + b^2r^2} \geqslant 0 \quad . \quad . \quad (4.4)$$

Simplifying

$$(1 + b^2)^2 \geqslant 4b^2 + C^2R^2b^2r^2$$

$$(1 - b^2)^2 \geqslant C^2R^2b^2r^2$$

or

$$CR \leqslant \frac{1 - b^2}{br} \quad . \quad . \quad . \quad (4.5)$$

$$CR \leqslant \frac{1 - b^2}{2\pi fb} = \frac{1 - 0{\cdot}25}{2\pi \,.\, 150{,}000 \,.\, 0{\cdot}5} = 1{\cdot}6\,\mu\text{sec} \qquad (4.6)$$

The numerical values put in Eq. 4.6 are for a beat frequency of 150 kc and for the modest requirement that the interference be a half the signal. Even if the 1·6 μsec time constant were used the grid-current-pulse areas would be much reduced during part of the beat cycle and the compromise between conduction angle and positive grid drop upon which the

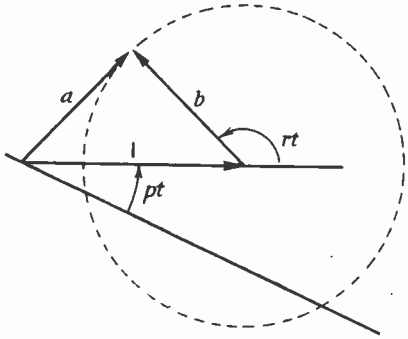

Fig. 4.7. Unit signal and interference *b* combine to give resultant *a*

action of the limiter is based would be destroyed. It is difficult to get the low time constants really needed and at the same time not drop C and R individually below values that would decrease the applied grid voltage. The clipper type that does not depend upon grid current biasing is free from these troubles. Time constants in commercial pentode limiters typically run between 1 and 50 μsec.

In typical broadcast practice multipath conditions all too frequently introduce selective fading and result in amplitude dips twice per audio cycle. An absolute minimum design requirement should be that a limiter be capable of removing these. Ten-to-one amplitude fades with 10 kilocycle modulation require that the product CR not exceed 3 μsec.

4.3. Discriminators. One device for converting frequency changes into variations of a direct voltage is spoken of as a discriminator. In a way it performs two functions: it converts frequency changes into amplitude variations and then rectifies these. Because the two functions are not usually easy to separate in a practical circuit, it is convenient to lump them together.

A simple discriminator is very easily made as indicated in Fig. 4.8. The circuit is tuned at a frequency above that of the signal. As the frequency changes, the voltage amplitude

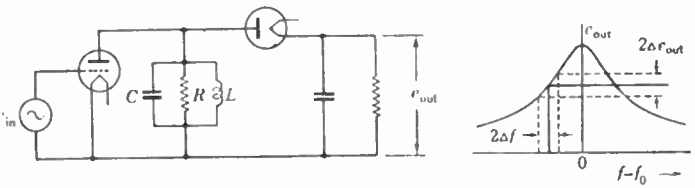

Fig. 4.8. Changes in frequency give rise to changes in direct output voltage

across the circuit varies. The variable amplitude is connected to a detector and converted to a variable direct voltage. If the audio rate of the frequency variation is small in comparison to the bandwidth of the circuit the amplitude follows the frequency variations more or less instantaneously. The amount of the voltage variation is essentially independent of the audio rate. If the frequency variation is limited to the straight portion of the resonance curve the reproduction is faithful.

This elementary discriminator has several drawbacks. Its linearity is not good. A signal above the resonant point would come in equally well—an undesirable selectivity condition. If the limiter were not working properly amplitude modulation would give rise to a variation of detector output.

A slightly more complicated discriminator is indicated in Fig. 4.9. The two tuned circuits are staggered in frequency. Let us assume that the $R_1L_1C_1$ circuit is tuned above the

mean signal frequency and the other below it by an equal amount. Then

$$e = e_1 - e_2 = \sqrt{2}(E_1 - E_2)$$

The normalized output is given by

$$h = \frac{e}{\sqrt{2}E_0} = \frac{1}{\sqrt{1 + (x - a)^2}} - \frac{1}{\sqrt{1 + (x + a)^2}} \qquad (4.7)$$

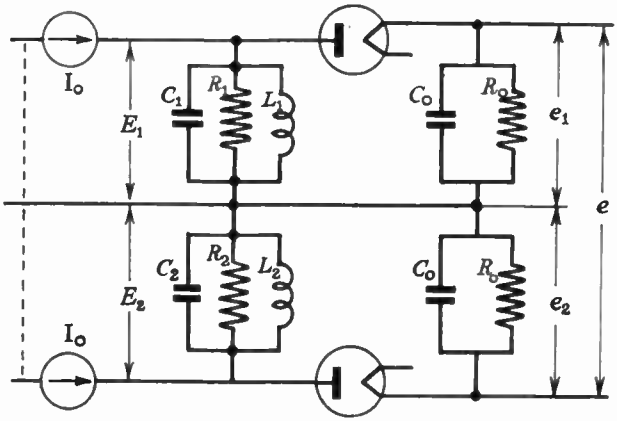

Fig. 4.9. A simple balanced discriminator

where x is the frequency change from resonance and a the separation of the two resonant peaks (both normalized to the half-bandwidth), and E_0 is the r.m.s. signal amplitude. The nature of the curve is clearly seen from Fig. 4.10. The output vanishes midway between the two peaks ($x = 0$) and has a considerable straight region between them.

A symmetrical pair of tuned coupled circuits gives the same response as a pair of isolated staggered circuits. A similar relationship exists for discriminators. The coupled circuit commonly used can be reduced to the simple one we have been discussing. For this reason it is worth while to cover the possibilities of the circuit of Fig. 4.9 carefully.

Eq. 4.7 leads to variously-shaped discriminator curves, depending on the choice of the peak separation, $2a$. If a is

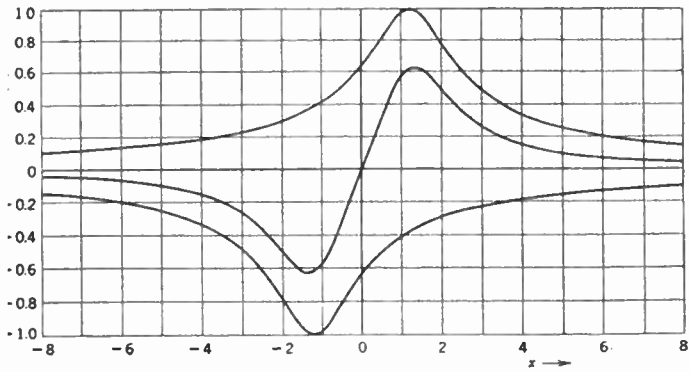

Fig. 4.10. Discrimator characteristic for $a = \sqrt{3/2}$. (Figs. 4.10, 4.12 and 4.13 reprinted from L. B. Arguimbau, 'Discriminator Linearity', *Electronics* [March, 1945], with permission)

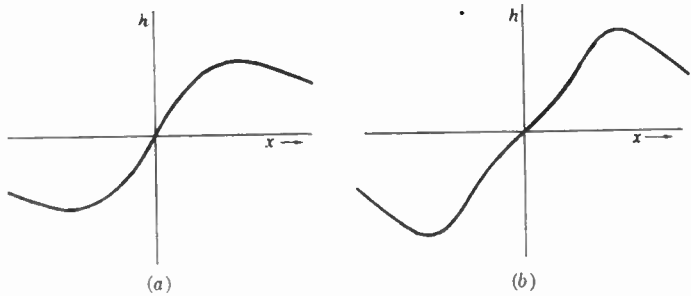

Fig. 4.11. The discriminator characteristic is changed by varying the separation of the peaks

small, the resultant sum has the shape shown in Fig. 4.11*a*. All the curves have points of inflection at the origin, but it is easy to see by inspection that those with large values of a

pass at the origin from being concave downward to being concave upward. The reverse is true for very small values

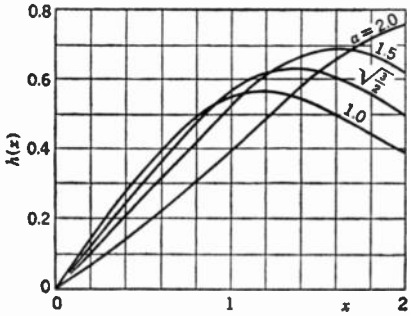

Fig. 4.12. Discriminator characteristic for various values of a

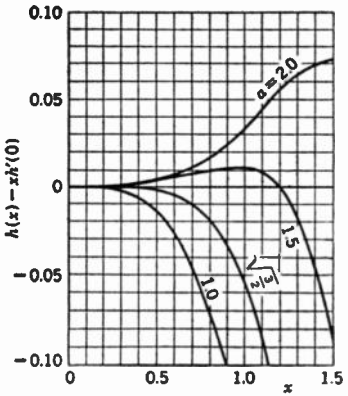

Fig. 4.13. Error curves

of a. Hence we might expect to get a critically straight curve for some intermediate value of a.

One way of attacking the problem consists of expanding h as a power series in x about the origin:

$$h = a_0 + a_1x + a_2x^2 + a_3x^3 + a_4x^4 + a_5x^5 + \ldots \qquad (4.8)$$

It is clear from symmetry that a_0 and all the other even-order coefficients must vanish.

In general,

$$a_n = \frac{1}{n!}\frac{d^n h}{dx^n}\bigg|_{x=0} \qquad (4.9)$$

Differentiating in Eq. 4.7, we have

$$h^{\mathrm{I}}(x) = -[1 + (x-a)^2]^{-3/2}(x-a) + [1 + (x+a)^2]^{-3/2}(x+a)$$

$$h^{\mathrm{I}}(0) = 2a(1 + a^2)^{-3/2}$$

Similarly, differentiating three and five times

$$h^{\mathrm{III}}(0) = 6a(2a^2 - 3)(1 + a^2)^{-7/2}$$

$$h^{\mathrm{V}}(0) = \frac{30a(8a^4 - 40a^2 + 15)}{(1 + a^2)^{11/2}}$$

The third derivative vanishes for $a = \sqrt{3/2}$.

Fig. 4.10 shows a plot of the discriminator curve for this critical value of a. Fig. 4.12 is a family of such curves for various values of a. Fig. 4.13 is a similar family, arranged to show the deviations of the curves from linearity. Instead of plotting the curve itself the difference between the curve and the tangent at the origin is plotted. Notice that the difference curve for the critical case gives the closest approximation to a straight line near the origin. If wider variations of frequency are to be permitted a slightly higher value of a is desirable.

It remains to show that the coupled circuit discriminator (Fig. 4.14) is equivalent to the simple staggered-circuit type. In Fig. 4.15, the vector sum of E_1 and E_2 is merely the sum of the left-hand tuned-circuit current times the impedance of

that tuned circuit and the similar right-hand circuit product. But the two tuned circuits are identical. Hence, the sum is

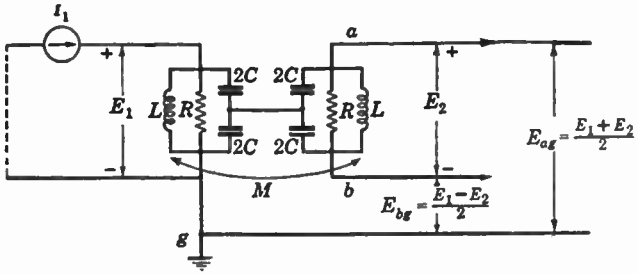

Fig. 4.14. A simple idealized discriminator circuit

merely the product of the circuit impedance and the vector sum of the two currents, namely ZI_0. Thus,

$$\frac{E_1 + E_2}{2} = \tfrac{1}{2} \cdot \frac{R}{1 + jx} \cdot I_0 \quad . \quad . \quad (4.10)$$

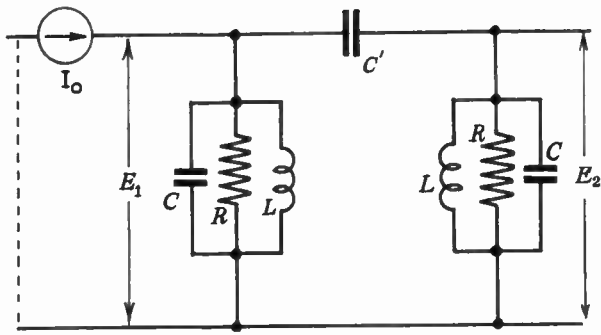

Fig. 4.15. A further idealization

The vector difference between E_1 and E_2 is merely the voltage across C'. We find this by shorting C' and noting from symmetry that I_0 must now divide equally between the

two circuits so that the short-circuit current is merely $I_0/2$. When the short is removed, application of the Thévenin-Norton theorem reduces the circuit to Fig. 4.16. But the load circuit is now merely a single tuned circuit with twice the impedance of a single one but shunted by C'. Thus,

$$\frac{E_1 - E_2}{2} = \tfrac{1}{2} \cdot \frac{R}{1 + j(x + b)} \cdot I_0 \qquad . \quad (4.11)$$

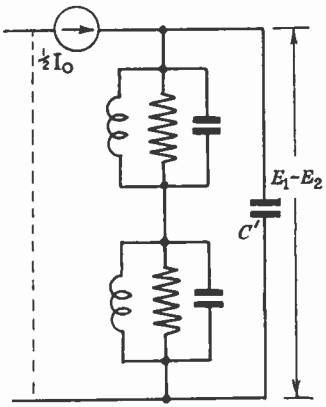

Fig. 4.16. An equivalent circuit is derived from Fig. 4.15 by the application of the Thévenin-Norton Theorem

where b is determined from the relative values of C and C' or the coupling coefficient. Thus our coupled circuit discriminator is reduced to that of the idealized staggered circuits.

4.4. Ratio Detector. The discriminator circuit has the disadvantage that the diodes charge the load condensers in Fig. 4.9 just as in ordinary amplitude-modulation detectors and if the input voltages (E_1 and E_2) decrease too rapidly the rectified output voltages (e_1 and e_2) fail to follow; distortion results. Under conditions of heavy interference very rapid changes are called for and trouble develops. GRANLUND

(1949), while working on an interference problem, felt that the answer to this sluggishness was to send reversed pulses into the condenser to discharge it during rapidly-changing conditions, and thus avoid the time constants of the detector circuit. The circuit of Fig. 4.17 resulted. Later it was found

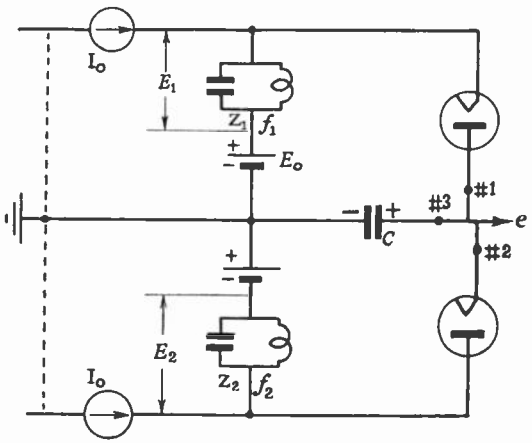

Fig. 4.17. A ratio detector

to be reducible to the original form of the ratio detector (SEELEY and AVINS, 1947; MARKO, 1952).

Notice that in Fig. 4.17 a circuit designed primarily for rapid response, *no* dissipative elements are intentionally added; ideally the damping is accomplished by charging the batteries, using the diodes as loss-free switches.

It should be noted that no direct current can flow through the load condenser, C. Hence, the direct current at point No. 1 is equal to the direct current at No. 2.

We will assume that the diodes are ideal and hence that they have no positive voltage drop and that the currents are in the form of short pulses. We will further assume that the tuned circuits are effective enough to maintain voltages at E_1 and E_2 sinusoidal. It has been shown that these simplifying

assumptions cause little trouble except that for accurate work allowance should be made for internal diode drops and for tuned-circuit losses.

Assuming short conduction angles, it follows that the peak fundamental alternating components are proportional to the direct currents ($= \frac{1}{2}I_{D.C.}$) and hence that the alternating currents at No. 1 and No. 2 are equal.

But the short pulse at No. 1 comes at the instant E_1 is maximum negative and hence the current at No. 1 is in phase with E_1. However, as Z_1 is loss-free the tuned-circuit current is 90° out of phase with E_1 and hence with the current at No. 1. Due to the quadrature current relationships each driving current squared is equal to the sum of the squared tuned circuit current and the squared diode current. Since the two diode currents are equal and since the two driving current sources have been assumed equal, it follows that the two tuned-circuit currents are equal.

We should note that the assumption of ideal diodes and sinusoidal voltages means that the alternating-peak and direct voltage components applied to the diodes must be equal. The load condenser is chosen to be large enough to have negligible alternating drop. Hence,

$$|E_2| = E_0 + e \quad . \quad . \quad . \quad (4.12)$$

$$|E_1| = E_0 - e \quad . \quad . \quad . \quad (4.13)$$

If we subtract Eq. 4.13 from Eq. 4.12, then add the two equations, and finally divide the answers, we get

$$\frac{e}{E_0} = \frac{|E_2| - |E_1|}{|E_2| + |E_1|} \quad . \quad . \quad . \quad (4.14)$$

Since the currents in the two tuned circuits are equal the voltages are proportional to the impedances and we can write,

$$e = \frac{|Z_2| - |Z_1|}{|Z_2| + |Z_1|} \quad . \quad . \quad . \quad (4.15)$$

or dividing by $|Z_1| \cdot |Z_2|$ and writing the admittances for reciprocal impedances,

$$e = \frac{|Y_1| - |Y_2|}{|Y_1| + |Y_2|} \quad . \quad . \quad . \quad (4.16)$$

Now the admittances have particularly simple forms since the circuits are loss-free. In fact, as indicated in Fig. 4.18,

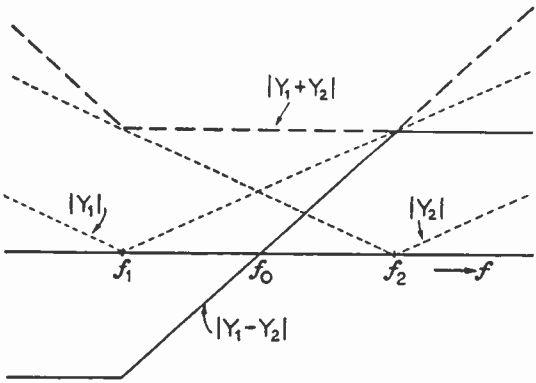

Fig. 4.18. The admittances and their sum and differences

the admittances $|Y_1|$ and $|Y_2|$ are very nearly linearly proportional to the numerical frequency differences from their respective resonant frequencies f_1 and f_2. $|Y_1|$ and $|Y_2|$ are shown as dotted lines in the figure, $|Y_1| - |Y_2|$ as a solid line, and $|Y_1| + |Y_2|$ as a broken line. The ratio is shown in Fig. 4.19. It is merely a broken line consisting of two hyperbolas connected by a straight line.

It should be noted that as long as I_0 is large enough the output voltage is independent of I_0. Thus, a ratio detector combines the *properties* of a limiter and of a discriminator. There is a controversy as to whether or not for patent purposes it actually *contains* a limiter, but this does not concern us (ARMSTRONG, 1948). In any case, experience with heavy

interference has shown that a ratio detector gives better results than a discriminator, but does so only if it is preceded by one or more effective limiter stages.

The derivation we have just carried through may be a little difficult to reproduce. It consists of a long sequence of exceedingly simple individual steps. The only difficulties are

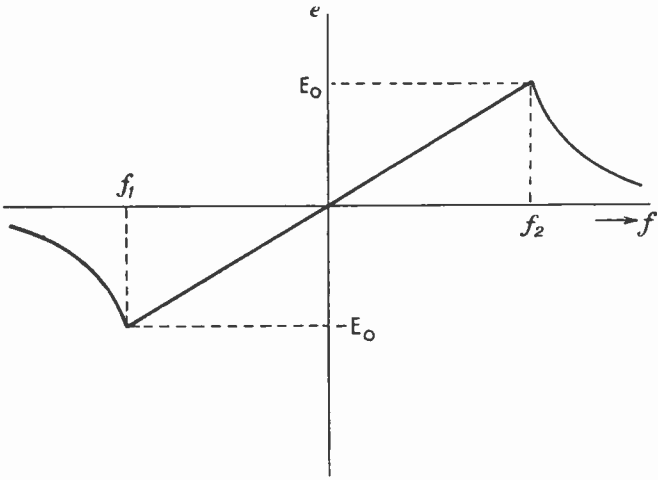

Fig. 4.19. The characteristic of an idealized ratio detector

the initial one of finding the sequence and the final one of seeing through the full physical assemblage. Most troublesome problems behave this way.

It should be noted that unless the current sources are large enough the action will be more complicated, since for some frequencies the diodes will not conduct at all.

4.5. Bradley Detector. One totally different method of detection developed by BRADLEY (1946) is indicated in Fig. 4.20. Here the first three electrodes in a pentagrid tube are connected to self-excite oscillations and break the plate current into a series of short pulses. The plate current pulses

are coupled back to the oscillator circuit to control frequency much as by a reactance tube. The incoming frequency-modulated signal, e_s, modulates the pulse train and causes the oscillations to interlock with the signal. The frequency shift so induced is proportional to the fundamental component of plate current but as the current is in short pulses this is proportional to the direct current. A resistor placed in the plate circuit gives an averaged output proportional to the

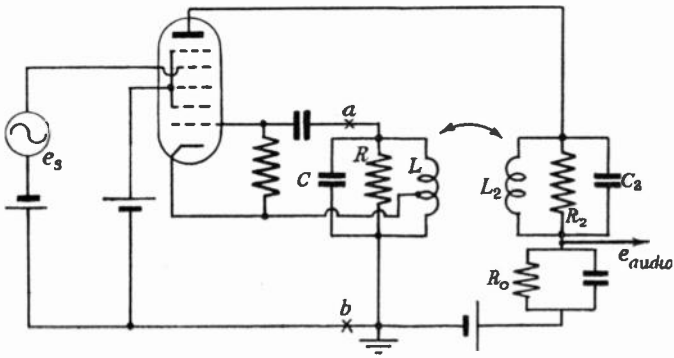

FIG. 4.20. Bradley detector—self-excited

frequency pull and hence varying linearly with the incoming signal frequency.

The circuit of Fig. 4.20 is a little bothersome to analyse because the cathode is not grounded, e_s is hence not measured with respect to the cathode, and further the plate current flows through part of the tuned circuit. These matters make little practical operational difference, but in order to understand the action it is convenient to study the idealized form shown in Fig. 4.21. We have further substituted a fixed battery for e_s and will see how the frequency depends upon its value.

We will assume that the C_1, C_2, L_2, R_2, secondary circuit is adjusted to be in resonance with terminals a and b shorted and to be so broadly tuned as to present an approximately

real impedance, R_2, to I_p for all frequencies of interest. The oscillations are supplied by a dynatron negative resistance device placed across the main tank circuit. In practice a triode with coupled grid circuit would be used, but here again the seldom-used dynatron has been assumed for simplicity of analysis.

It should be recognized that the entire circuit acts as a self-sustaining oscillator. The sum of the currents flowing

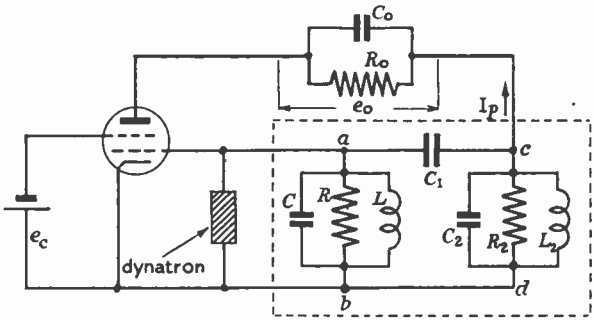

Fig. 4.21. Simplified version of Fig. 4.20

across the terminals a and b must be zero. We have surrounded the main totally linear portions of the circuit by a dotted line. The current flowing into this box from the left is clearly equal and opposite to the current in the shaded dynatron. (Grid current is neglected.) We propose to find the currents flowing into this box by applying the linear superposition theorem to the linear network inside the box. In doing so we shall assume the existence of an approximately sinusoidal voltage, E_g, across a and b. Further, we shall assume the tube current, I_p, little influenced by the voltage across the plate but depending upon the non-linear action of E_g on the tube. The plate current also depends upon the value of e_c.

The application of linear superposition in a circuit containing violently non-linear elements should not cause alarm.

What we are proposing is to piece together a linear circuit and a non-linear one using the properties of the non-linear circuit as boundary conditions for the linear part. If the boundary conditions of a linear network are known it makes no difference whence they come, superposition can be used. We shall apply it by finding the current flowing across a and b due to E_g with I_p removed and then finding the current due to I_p with E_g shorted.

We shall start by finding the linear-network currents due to E_g. Those flowing in C, R, L are by inspection

$$E_g[j\omega C + 1/(j\omega L) + 1/R]$$

To find the current in C_1 due to E_g we first find the voltage due to E_g across R_2. This is done by shorting terminals c and d obtaining a short-circuit current, $j\omega C_1 E_g$. With $C_1 + C_2$ resonant to L_2 we have an impedance across c and d of merely R_2. (E_g being here regarded as a voltage source is of zero impedance.) Thus the voltage developed across c and d is $j\omega C_1 R_2 E_g$ and the voltage across C_1 is $E_g(1 - j\omega C_1 R_2)$. Thus the total current drawn by the network due to E_g is

$$I_{E_g} = E_g \left[j\omega C + \frac{1}{j\omega L} + \frac{1}{R} + j\omega C_1(1 - j\omega C_1 R_2) \right] \quad (4.17)$$

The current I_p makes d positive with respect to c by $R_2 I_p$, when E_g is shorted and the secondary circuit is adjusted to resonance. The current flowing from a toward c due to this voltage is merely $j\omega C_1 R_2 I_p$.

We are now in a position to write down an equation for the total current between a and b:

$$I = E_g[j\omega C + \frac{1}{j\omega L} + \frac{1}{R} + j\omega C_1 + \omega^2 C_1^2 R_2] + f(|E_g|) + j\omega C_1 R_2 I_p(E_g, e_c) = 0. \quad (4.18)$$

In this equation we have represented the dynatron current as $f(|E_g|)$, a current that adjusts itself to balance Eq. 4.18, or rather the real part of that equation.

$$f(|E_g|) + E_g/R + E_g \omega^2 C_1^2 R_2 = 0 \quad (4.19)$$

Now I_p is written $I_p(E_g, e_c)$ to indicate that it is a function of E_g and e_c. It should be noted that if this current flows in short pulses it is equal to twice the direct current, and we may write for the imaginary components of Eq. 4.18,

$$2\omega C_1 R_2 I_{\text{D.C.}} + [\omega(C + C_1) - 1/(\omega L)]E_g = 0 \quad (4.20)$$

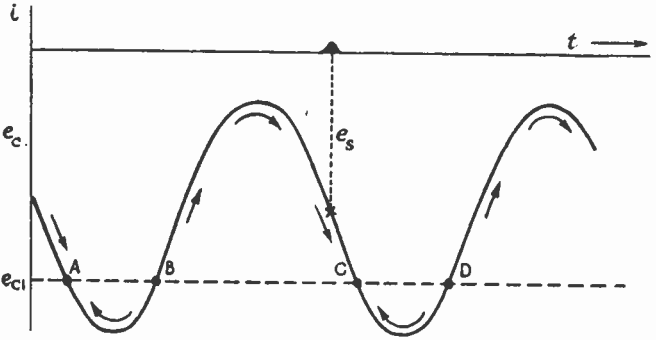

Fig. 4.22. The pulses adjust their position until interlock is achieved

If we write the frequency deviation from the resonant frequency of C, C_1, L, as Δf and note that the change in $1/(\omega L)$ is approximately $\Delta\omega(C + C_1)$, we can solve for $I_{\text{D.C.}}$.

$$I_{\text{D.C.}} = -\frac{(C + C_1)|E_g|}{f_0 C_1 R_2}\Delta f,$$

$$\text{or} \quad e_0 = -\frac{(C + C_1)R_0|E_g|}{f_0 C_1 R_2}\Delta f \quad . \quad (4.21)$$

and the drop in R_0, $e_0 = I_{\text{D.C.}} R_0$.

It remains to show under what conditions the assumed interlocking of the oscillator can really take place. To do this we replace the fixed bias by the incoming signal waveform shown in Fig. 4.22. We assume that some particular bias $-e_{c_1}$, is needed to adjust the oscillator frequency to the signal frequency. If it happens that the pulse (corresponding

to maximum positive control-grid swing) occurs at one of the points A, B, C, D, at which $e_s = e_{c_1}$, then the frequency of the oscillator will be the same as that of e_s. However, let us assume that this is not the case and that the pulse occurs at some point x. This means that the grid action will make the pulse have too large an area, I_{fund} will be too large and in accordance with Eq. 4.21 Δf will have too large a negative value, the oscillator will run more slowly than the incoming signal, and the next pulse instead of occurring at x will move to the right as indicated by the arrows. The pulse position will move to position C. If it overshoots this point it will be brought back as indicated. Notice that points A and C are stable in that any accidental disturbance is self-correcting whereas B and D are unstable, any accidental disturbance builds up. In practice, then, the phase between the oscillator and the incoming signal adjusts itself until interlock is achieved.

The voltage range of e_s must cover the range of biases necessary to achieve interlock over the frequency range travelled by the incoming signal.

It is important to notice that the amount of current the tube can carry in short pulses is limited. If interlock is to be achieved over a fairly wide frequency range the circulating power in the tank circuit must be kept low so that the power in the plate circuit need not be unduly high. Eq. 4.21 indicates that the plate current per cycle shift can be kept low if the ratio $C/C_1R_2f_0$ is kept low.

REFERENCES

ARMSTRONG: *Proc. Radio Club of America.* (11 W. 42nd St., New York), **25**, No. 3 (1948)

BRADLEY: *Electronics*, **19**, 88 (1946)

GRANLUND: Research Lab. of Electronics (M.I.T.) Technical Report, 42 (1949)

MARKO: *Frequenz*, **6**, 1 (1952)

SEELEY and AVINS: *R.C.A. Rev.*, **8**, 201 (1947)

CHAPTER V

INTERFERENCE

5.1. Introduction. In a frequency-modulation receiver we must recover our audio signal from frequency changes rather than from amplitude changes, and this implies new circuitry.

There is a further new element here that is much more important. Our fundamental reason for interest in frequency modulation comes from its potential ability to avoid the effects of interference and noise. In the case of amplitude modulation we saw that interference brings a proportional uncertainty in our knowledge of the signal. A 10 per cent interfering signal causes a 10 per cent change in envelope amplitude and a 10 per cent fluctuation of the detected output. On the other hand we shall see later that a 10 per cent interference in the frequency-modulation broadcast system at present in widespread use gives rise to a frequency change that is only about 0·3 per cent of the frequency change we use for signalling, a potential improvement of 30 : 1 in the signal to noise ratio. In order to realize this improvement, however, it is necessary for the amplitude changes resulting from interference to have little influence on the detected output. This implies that the circuits be arranged to be less sensitive to amplitude than to the corresponding frequency changes by a ratio of about 30:1.

Freedom from the effects of amplitude changes can be achieved in many ways. Most receivers include limiter circuits whose output voltage amplitudes depend little on their inputs. Some of the detector circuits give rectified outputs that are proportional to the product of the frequency deviation and the amplitude; others are intended to be independent of the amplitude or at least independent of rapid amplitude changes. Usually the circuits intended to be free from amplitude effects perform better if they are preceded by one or more limiters. Automatic-volume-control circuits give

further help. For a time certain manufacturers, wishing to produce more economical sets, argued that just as good or better results could be obtained from receivers whose output depended on amplitude in addition to frequency. It is now generally agreed that this is not correct.

In amplitude modulation any antenna will pick up enough noise to be heard on the average sensitive receiver because static interference and noise levels are so high. Static is not sufficiently localized to make the placing of the antenna of major importance. This statement should not serve as an excuse to ignore the antenna completely and use an arcing refrigerator as the main source of signal energy. However, the problem is usually not at all critical.

In the wavelength bands used for frequency modulation man-made noise is likely to be localized. For good results it is desirable to place the receiving antenna in a location which is relatively free from local interference but has the strongest possible received signal. The field strength of a signal in the frequency bands ordinarily used for frequency modulation increases rapidly with height above the earth. Thus it is desirable to place the receiving antenna as high as possible above the ground.

5.2. Phase Modulation. In amplitude modulation with a unit desired signal the presence of an interfering signal of amplitude, a, introduces an uncertainty of $\pm a$ in our knowledge of the instantaneous amplitude. Thus in Fig. 5.1 we have desired and interfering signals of amplitudes 1 and a respectively, the desired signal making an angle pt with a fixed axis. When the interference is in position No. 1 the apparent instantaneous amplitude is $1 + a$ and the instantaneous phase is pt. In position two we have an amplitude of about 1 and angle $pt + \sin^{-1} a$, at No. 3, $1 - a$ and pt, and at No. 4, 1 and $pt - \sin^{-1} a$. Thus we have an uncertainty of $\pm \sin^{-1} a$ in phase. It is interesting to note that for small values of a, $\sin^{-1} a \cong a$, so that our angular uncertainty is $\pm a$ radians.

Thus, aside from apparatus it is easy to see that with 100 per cent amplitude modulation permissible for our

message, the interference introduces an uncertainty-to-signal ratio of a:1. Similarly, if we choose to signal by means of instantaneous knowledge of the angle and an angular swing of $\pm \theta_{max}$ radians is permitted we have an uncertainty-to-signal ratio of $a:\theta_{max}$. Thus we could expect that with proper apparatus the substitution of phase variations for amplitude variations in a signalling system would improve the signal to

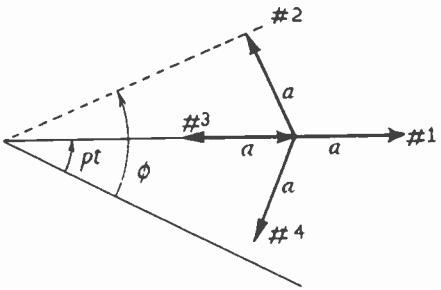

Fig. 5.1. The disturbance on angle and amplitude depends upon instantaneous phase

interference ratio by a factor equal to the maximum permissible angular deviation in radians.

Incidentally, for values of a that are not negligible with respect to 1 and even approach 1, the maximum angular uncertainty is $\pm \sin^{-1} a$, which for values of a near unity means that the angular uncertainty approaches $\frac{1}{2}\pi$, a somewhat worse than proportional value.

We have not previously discussed the possibility of making the instantaneous phase rather than frequency of a radio signal vary with voice pressure, but this system is often used for point-to-point communication systems in mobile work. A transmitter similar to that discussed in Art. 3.4, but with less frequency multiplication, is used. It should be noted that a sinusoidally phase-modulated signal is identical to a frequency-modulated signal and in any case $\omega = d\phi/dt$, $\phi = \int \omega \, dt$, $e = a \cos (\int \omega(t) \, dt)$.

The same type of receiver is used. However, a phase

modulation receiver must give a rectified output proportional to the instantaneous phase deviation. If a frequency-modulation receiver giving a rectified output proportional to instantaneous frequency is used, an integrating circuit must be added after the detector.

Let us assume that in a phase-modulation system we may permit the transmitter to vary sinusoidally by $\pm \theta_M$ radians, at any audio-frequency rate $f_{\max}$. Then

$$\phi = pt + \theta_M \cos 2\pi f_{\max} t \quad . \quad . \quad (5.1)$$

The corresponding frequency is

$$\omega = d\phi/dt = p + 2\pi f_{\max} \theta_M \sin 2\pi f_{\max} t \quad . \quad (5.2)$$

Thus the frequency varies by $\pm \theta_{\max} f_{\max}$.

Suppose, for example, a bandwidth of 30 kc is available for a system using phase modulation and full angular shift is to be permitted up to 3 kc. The instantaneous frequency covers a range of $\pm \theta_M f_{\max}$ so that for a range of $\pm$ 15 kc for instantaneous frequency to be possible θ_M must be limited to 15/3 or 5 radians. If this system is used an improvement of 5:1 over amplitude modulation should be expected.

It should be noted that the work on spectra indicated that the nuisance value of a frequency changing wave is not limited to its range of instantaneous frequencies but extends somewhat beyond this. The angular swing should be limited to somewhat less than the value determined by the instantaneous excursion, the allowance depending in a complicated way upon the statistics of the actual modulating signals.

We have just noted that the frequency range covered by the spectrum is somewhat wider than the instantaneous-frequency excursion. What audio-frequency response restrictions and other difficulties might we expect if we were to include a filter on the transmitter output, artificially limiting the spectrum to the instantaneous excursion? Experimental work has led the writer to a belief that the limitation is not as serious as might at first be expected, but no quantitative answers can be given at present.

5.3. Frequency Modulation. We have seen in the last article that by using a signalling system in which the audio-modulating wave is represented by variations in the instantaneous phase of the transmitted radio wave, the uncertainty can be decreased by a ratio equal to the maximum permissible angular deviation. Our problem of reducing interference is then clear, we must try to provide the maximum possible angular deviations, much greater than those resulting from natural causes. Unfortunately, we have also seen that the use of a wider angle at a high audio frequency results in a large bandwidth. If the optimum improvement is to be made we must provide for using the maximum angular swings consistent with the permissible bandwidth.

On page 5, in discussing the pre-emphasis of the high frequencies in a sound wave before transmitting it, we saw that for most sound waves it is possible to multiply the component frequencies by a factor $\sqrt{1 + (f/f_0)^2}$ where $f_0 \cong 2100$ c.p.s. and yet only increase the peak height by a small percentage. This suggests that we pre-emphasize the audio frequencies and then make the *instantaneous frequency deviation* of the transmitter proportional to the pre-emphasized wave.

Notice that this procedure assures us that if we set the overall transmitter to give about 90 per cent of full permissible deviation on a low frequency wave of given peak voltage and then monitor in such a way as to prevent the composite audio input voltage from exceeding this peak audio value, the instantaneous frequency will keep within the prescribed channel limits. It should be noted that this does not necessarily mean that the spectrum does not spread somewhat outside of the channel. However, this problem is not a serious one if the permissible frequency deviation is much larger than the top audio frequency.

If we assume that an audio voltage is applied to a pre-emphasized frequency-modulation transmitter we may write

$$\Delta f_{\max} = KE_{\text{audio}}\sqrt{1 + (f/f_0)^2} \quad . \quad . \quad (5.3)$$

Similarly, we can write for the maximum phase variation,

$$\Delta\phi_{max} = \Delta f_{max}/f$$

$$= KE_{audio}\sqrt{\frac{1}{f^2} + \frac{1}{f_0^2}} = \frac{K}{f_0}E_{audio}\sqrt{1 + \left(\frac{f_0}{f}\right)^2} \quad . \quad (5.4)$$

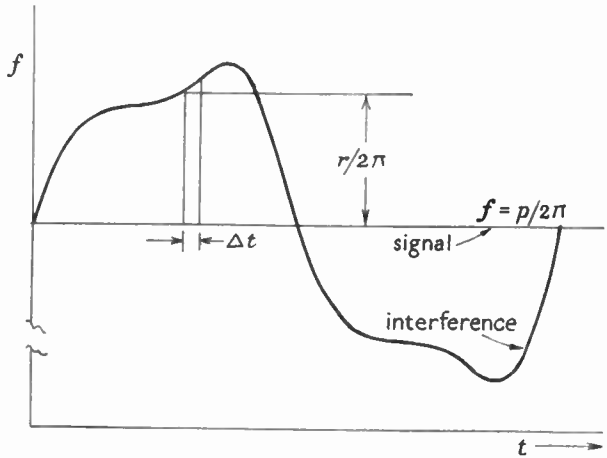

Fig. 5.2. The interfering frequency wave is broken up into small intervals of nearly constant frequency

Eq. 5.3 means that until we approach the pre-emphasis frequency, f_0, the audio characteristics of the system give a frequency deviation proportional to the audio voltage and independent of the audio frequency. On the other hand, Eq. 5.4 means that well above f_0 the audio characteristics are such that the phase variation is proportional to the audio voltage and independent of the frequency until f_0 is approached.

The most troublesome time we can encounter interference is during quiet passages of music where the desired signal is virtually unmodulated. The situation is illustrated by Fig. 5.2. The desired frequency is essentially constant at $p/2\pi$, while

the interference is frequency modulated. Since it is a little difficult to figure the interference resulting from such an arbitrary wave we will break up the time into small intervals and assume that during each of them the difference frequency $r/2\pi$ is essentially constant. We thus set ourselves the problem of computing the interference between two constant-frequency sine waves, of amplitudes 1 and a.

The situation is illustrated by Fig. 5.3, where a desired unit-amplitude signal of frequency p is disturbed by interference of amplitude a and frequency $(p + r)$. We draw the

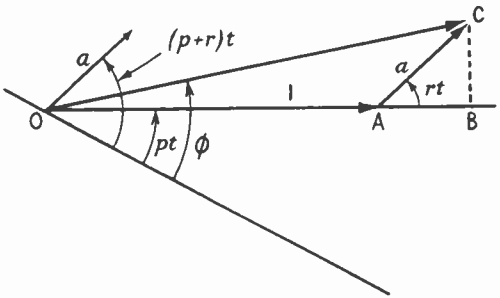

Fig. 5.3. Unit signal and interference a

resultant vector AC and set ourselves the problem of finding its instantaneous frequency, $\omega = d\phi/dt$.

$$\phi = pt + \angle COB = pt + \tan^{-1}[CB/(OA + AB)]$$

$$= pt + \tan^{-1} \frac{a \sin rt}{1 + a \cos rt} \quad . \quad . \quad (5.5)$$

We shall divide this problem into two parts and consider first the relatively simple case where a is much smaller than 1 and we can write the disturbance in frequency, Δf, as

$$\Delta f = (d\phi/dt - p)/2\pi = (ar/2\pi) \cos rt \quad . \quad (5.6)$$

Notice that the disturbance is a beat frequency $r/2\pi$ of magnitude $ar/2\pi$. Since the disturbance cannot be heard if

it has a frequency above the limit of audibility, say 15 kc, we can see that the disturbance is less than a deviation of $a \times 15$ kc.

Notice also that at the output of the audio system this must be multiplied by a de-emphasis factor of $[1 + (f/2100)^2]^{-1/2}$. On the other hand, the maximum audio signal out corresponds to 75 kc occurring at the low audio rate. Hence we can write

$$\frac{\text{Disturbance}}{\text{Maximum signal}} \cong \frac{af_r}{\Delta f_{\max}\sqrt{1 + (f_r/f_0)^2}} \quad . \qquad (5.7)$$

If f_r is considerably larger than f_0 the radical becomes approximately f_r/f_0 and Eq. 5.7 becomes at worst,

$$\frac{\text{Disturbance}}{\text{Maximum signal}} \leqslant \frac{af_0}{f_{\max}} = \frac{a \times 2100}{0{\cdot}9 \times 75{,}000} \cong \frac{a}{32} \quad . \qquad (5.8)$$

Here we have made an allowance of 10 per cent for overmodulation due to pre-emphasis by using a maximum low-audio-frequency swing of only 90 per cent of 75 kc. Doing this we find that interference causes a disturbance in frequency equal to the amplitude ratio times the pre-emphasis frequency. Putting in the most pessimistic values we find that the disturbance to signal ratio on the receiver output is only one thirty-second of the corresponding ratio at the input. *An improvement of* 30 *decibels has been made over the corresponding situation in amplitude-modulation broadcasting an improvement in power ratio of* 1000:1.

There are two important consequences of this large improvement. In the first place static disturbances can be overcome. In the second place, with noise largely absent the full audio frequency range can be tolerated; there is no longer any technical reason for a listener to defend himself from noises by reaching for the "tone control" and cutting out most of the audio spectrum.

It should be noted that the improvement in freedom from noise is proportional to the bandwidth used. If we use 150 kc, as is the practice in broadcasting, we must go to a high frequency. In the United States it is present practice to use

the 88–108 Mc band. This band has markedly different propagation characteristics from the 500–1500 kc band, reception is more or less confined to line of sight or perhaps 50 per cent more. This fact, combined with the noise reduction properties, makes it possible to assign a larger number of channels in a given large area than would be possible with "standard (amplitude-modulation) broadcasting." The larger frequency assignment frees the audio-response problem from spectrum limitations and the fidelity of reproduction is limited mainly by the loud speaker and by the audio-frequency techniques of broadcasting.

If allowance is made for the variation of dielectric constant with height the optical line of sight from a height, h, on flat portions of the earth is given approximately by

$$d \text{ (miles)} = \sqrt{2h \text{ (feet)}} \quad . \quad . \quad (5.9)$$

Thus, for moderate antenna heights at the transmitter and receiver we can expect a range of about 50 miles.

5.4. Frequency Modulation with High Interference Levels. We made the approximation that a be *considerably* less than 1 in Eqs. 5.6, 5.7, and 5.8, in order to demonstrate the possibilities of improvement in interference suppression when using frequency modulation. It has long been observed that commercial receivers fail to give good suppression of interference when the interference gets within 5–10 db of the signal ($a = 0{\cdot}3 - 0{\cdot}5$). This observed fact, together with the mathematical restrictions, led to the assumption that frequency modulation was only capable of reducing interference when the signal is more than about twice the interference. As a matter of fact this is not so. In fact, the improvement is theoretically possible right up to equality between signal and noise (ARGUIMBAU and GRANLUND, 1947; ARGUIMBAU and GRANLUND, 1949; GRANLUND, 1952; MARKO, 1954). Fig. 5.3 is redrawn in Fig. 5.4 with the amplitude ratio a near unity. Now if everything were perfect we should like our receiver output to be determined by the frequency of the desired signal, p. Actually, as we have already seen, the frequency of the resultant is not p alone but contains also a frequency variation

occurring at a beat-frequency rate, r. In a certain sense we can tolerate this, as r is often above audibility, but we should at least like to have the *average* frequency of the resultant equal to p.

When we look at Fig. 5.4 we recognize that for $a = 1$ the resultant bisects the angle DOB, COA is an isosceles triangle, and $\angle COA = rt/2$. Thus it looks as though our resultant frequency is very nearly $p + \frac{1}{2}r$ instead of p. On the other hand, if we draw a circle of radius a about A we notice that

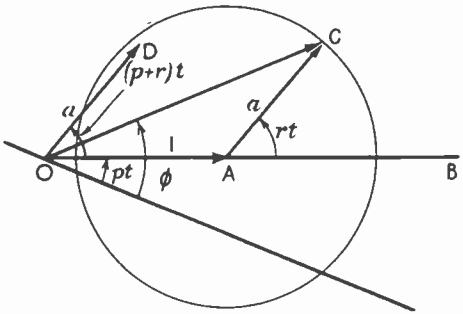

Fig. 5.4. Interference for $a \cong 1$

the tip of the resultant moves around this circle. As rt, ($\angle CAB$), increases from zero the $\angle COA$ at first increases at a rate $r/2$, but the $\angle COA$ never quite reaches 90° as rt swings toward 180°. Instead, after nearly reaching 90°, $\angle COA$ snaps back fairly quickly to zero and then goes negative, nearly reaching $-90°$. It then resumes its positive motion at a speed of about $r/2$. Notice that in a full revolution of the interfering signal about the desired signal's tip the resultant gets *precisely nowhere*. Since the resultant never differs in angle from pt by more than 90° we can write,

$$\lim_{t \to \infty} \frac{\phi - pt}{t} = 0 \qquad . \quad . \quad (5.10)$$

and we recognize that *the resultant's frequency averaged over a long enough time is precisely the frequency of the desired*

signal. The long forward motion at a rate of about $r/2$ is exactly balanced by short spurts backward.

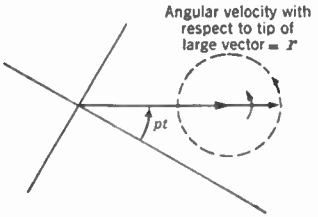

Fig. 5.5. Fig. 5.4 is redrawn for instant at which $rt = 2\pi n$

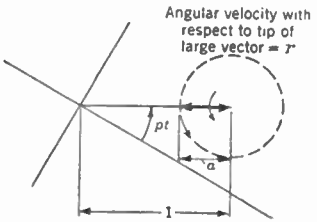

Fig. 5.6. Fig. 5.5 is redrawn after rt has increased by π

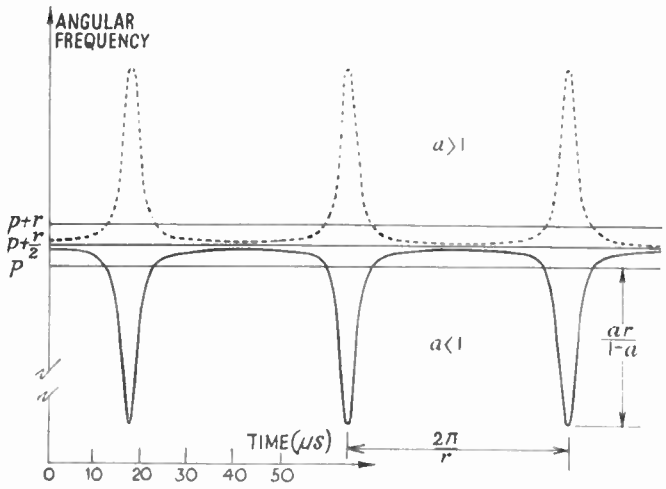

Fig. 5.7. Solid line shows variation if instantaneous frequency for $a = 0{\cdot}8$ and 150 kc beat frequency

It is useful to study two particular positions of the interference for $rt = 0$, and $rt = \pi$. Thus, in Fig. 5.5 we compute the speed of the resultant vector by figuring the top speed

relative to pt and dividing by the radius. The top speed is ar, the radius $1 + a$. Hence,

$$\omega_{\max} = p + \frac{ar}{1 + a} . \quad . \quad . \quad (5.11)$$

Similarly for Fig. 5.6,

$$\omega_{\min} = p - \frac{ar}{1 - a} . \quad . \quad . \quad (5.12)$$

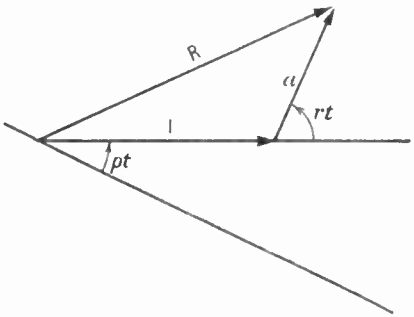

Fig. 5.8. The resultant $R = \varepsilon^{ipt}(1 + a\varepsilon^{irt})$

Fig. 5.7 is a plot of the frequency variation for $a = 0{\cdot}8$. Notice that the forward motion at a relative speed $r/2$ is counteracted once per beat cycle by a sharp negative frequency *spike* of large height, short duration, and having an area approaching π.

Thus we see that the interference does not change the average frequency but does generate a series of spikes. As two nearly equal signals interfere, the instantaneous resultant frequency lingers most of the time near the average of the two frequencies but spikes extend from this *in the direction of the stronger signal* and make the average frequency equal to the frequency of the larger signal.

It should be noticed that when p and $p + r$ are constant the resultant frequency varies periodically at a difference frequency rate. It is interesting to expand this periodically

varying resultant frequency as a Fourier series. This is readily done by means of complex notation. Thus, in Fig. 5.8 the resultant can be written

$$R = 1\varepsilon^{jpt} + a\varepsilon^{j(p+r)t}$$

$$= \varepsilon^{jpt}(1 + a\varepsilon^{jrt}) \quad . \quad . \quad . \quad (5.13)$$

If we recall that the natural logarithm of a complex quantity has a real component equal to the logarithm of its magnitude and an imaginary component equal to its angle, we are led to look for the imaginary part of the logarithm of the resultant

$$\log R = jpt + \log (1 + a\varepsilon^{jrt}) \quad . \quad (5.14)$$

To find the imaginary part we note that for $x < 1$

$$\frac{1}{1+x} = 1 - x + x^2 - x^3 + \ldots$$

$$\therefore \log (1 + x) = x - \tfrac{1}{2}x^2 + \tfrac{1}{3}x^3 - \ldots$$

and since $a < 1$ we have

$$\log R = jpt + a\varepsilon^{jrt} - \tfrac{1}{2}a^2\varepsilon^{2jrt} + \ldots$$

Hence the phase is given by

$$\phi = Im\,\{\log R\} = pt + a \sin rt - \tfrac{1}{2}a^2 \sin 2rt + \ldots$$

and the frequency

$$\omega = \frac{d\phi}{dt} = p + ar \cos rt - a^2 r \cos 2rt + \ldots$$

or

$$\omega = p + r[a \cos rt - a^2 \cos 2rt + a^3 \cos 3rt - \ldots] \quad (5.15)$$

This equation shows among other things that the fundamental beat frequency component of the frequency variation has a magnitude ar and hence is linearly proportional to the interference. The r.m.s. value of the frequency deviation is given by

$$\sqrt{[\tfrac{1}{2}r^2(a^2 + a^4 + a^6 + \ldots)]}$$

In order to find the r.m.s. detected output from the receiver we must weight each term in this expression by a factor representing the audio response of the receiver. The characteristic of the de-emphasis circuit is given by $\left\{1 + \left(\frac{nr}{\omega_0}\right)^2\right\}^{-1/2}$ where ω_0 is the de-emphasis frequency. To prevent the influence of frequencies near and beyond the limit of audibility from being unduly emphasized, some kind of low pass filter should be included. We have arbitrarily used a constant-K filter characteristic $\left\{1 + \left(\frac{nr}{\omega_1}\right)^6\right\}^{-1/2}$ and for the purposes of numerical calculation assumed a cut-off frequency corresponding to 13 kc. The r.m.s. detected output is therefore given by

$$\sqrt{\sum_{n=1}^{\infty} \frac{1}{2} \frac{a^{2n} r^2}{\left\{1 + \left(\frac{nr}{\omega_0}\right)^2\right\}\left\{1 + \left(\frac{nr}{\omega_1}\right)^6\right\}}}$$

For a sinusoidal variation of difference frequency $r = r_m \sin\phi$ we have an output

$$\sqrt{\frac{1}{2\pi}\int_0^{2\pi} \sum_{n=1}^{\infty} \frac{1}{2} \frac{a^{2n}(r_m \sin\phi)^2 \, d\phi}{\left\{1 + \left(\frac{nr_m \sin\phi}{\omega_0}\right)^2\right\}\left\{1 + \left(\frac{nr_m \sin\phi}{\omega_1}\right)^6\right\}}} \quad . \quad . \quad . \quad (5.16)$$

The results for a de-emphasis frequency of 2100 c.p.s. and a cut-off frequency of 13 kc are plotted in Fig. 5.9. Notice that the maximum audio interference in equivalent peak kilocycles occurs when the maximum deviation is near the upper audio limit. Notice also that the ratio of peak permissible signal swing (75 kc) to peak interference is approximately

$$\frac{\text{Maximum signal}}{\text{Maximum interference}} = \frac{\Delta f_{\max}}{a f_0} = \frac{75}{2{\cdot}12a} = \frac{35}{a} \quad . \quad (5.17)$$

5.5. Impulsive Interference. Some electrical devices, when not properly shielded, radiate disturbances of short duration, disturbances approximating impulses. Such impulses when impressed on a selective receiver cause the receiver tuned circuits to "ring" generating in effect pulses of intermediate frequency. These act like short pulses having the exact

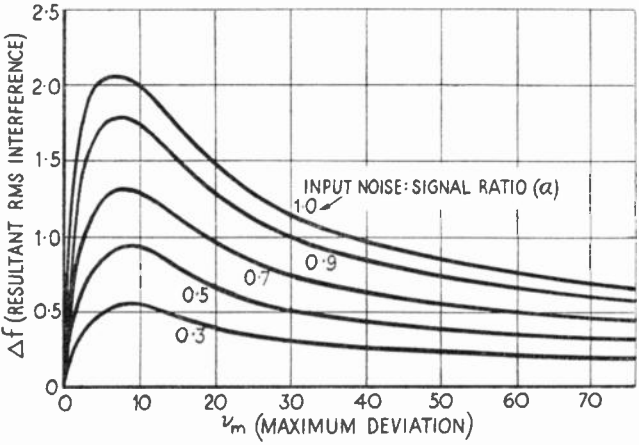

Fig. 5.9. R.M.S. interference as a function of interfering frequency swing for various input noise : signal ratios (a)

frequency to which the receiver is tuned. The theory of this type of interference has been discussed by SMITH and BRADLEY (1946).

We shall calculate the form of these pulses assuming the receiver to have four symmetrical cascaded tuned circuits with a composite bandwidth of 150 kc. The bandwidth of each separate circuit figured from the ideas of uniform stretch functions would have to be $\sqrt{n}B = \sqrt{4} \times 150$ kc $= 300$ kc. A single parallel tuned circuit of L, C, R is equivalent in envelope response to a single parallel circuit of $2C$ and R. Note that the bandwidth of each tuned circuit is given by

$\sqrt{4B}$ or $2B$ and is $1/(2\pi CR)$. Thus $2CR = 1/(2\pi B)$, and the ringing tone across the first circuit has an envelope amplitude, $a_0\varepsilon^{-2\pi Bt}$. By applying the superposition integral three times we find an envelope, $a = a_0 \dfrac{(2\pi Bt)^3}{3!}\varepsilon^{-2\pi Bt}$. This is the amplitude envelope of an interfering pulse having the frequency to which the receiver is tuned. If the desired signal is at that moment

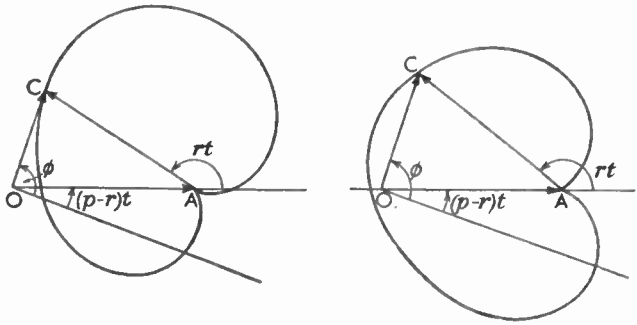

Fig. 5.10.

(*a*) The interference produces no net change in angle and a "click" results

(*b*) The interference produces a net change in angle of 2π and a "pop" results

modulated away from the centre value the two differ in frequency by at most 75 kc. Fig. 5.10 is a vector representation of the resultant of the signal and pulse for an arbitrary initial relative phase. Here it is assumed that the incoming signal has an instantaneous value of 75 kc below the centre frequency. Notice that the resultant starts at an angle $(p - r)t$ and travels around the closed loop. In the case illustrated the angle first advances then decreases, and finally comes back to the starting point. As it does so the resultant amplitude varies but more important the angle is modulated. Notice that in Fig. 5.10*a* there is no net change in angle and hence

the curve for the frequency shift, $d\phi/dt$, has no area. Consideration of such a curve from a Fourier point of view shows that it is lacking in low frequency components and causes a negligible audio disturbance, a "click," as SMITH and BRADLEY have expressed it. On the other hand, for a larger impulse the curve of Fig. 5.10*b* results, and the *origin is encircled.* As a result the resultant gains in angle on the desired signal by 2π. The curve of frequency deviation has an area of 2π and the audio interference has appreciable low frequency components, a "pop" is said to result. The curves of resultant

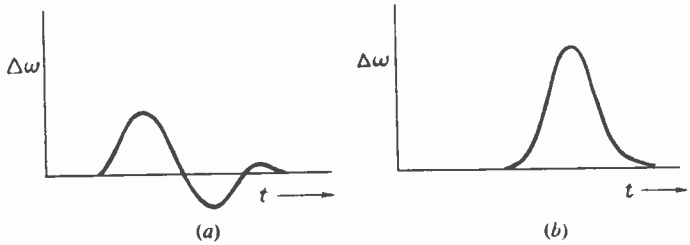

Fig. 5.11. (*a*) Area of "click" is zero. (*b*) Area of "pop" is not zero

instantaneous frequency deviation from the desired signal are drawn in Figs. 5.11*a* and 5.11*b*. It should be noticed that a "pop" occurs whenever the origin is linked. This happens when the ringing tone has an amplitude higher than that of the incoming signal at an instant of time when the signal is 180° out of phase with the ringing tone. In order to locate "pops" we can plot the envelope of the ringing tone against time as indicated in Fig. 5.12. We superimpose the signal amplitude on this and draw vertical lines at points of phase opposition. If one of these lines falls below the envelope of a ringing train a "pop" occurs. Notice that the probability of a "pop" occurring increases as the beat-frequency increases. There is no possibility of a "pop" if an ideal receiver is properly tuned and the signal is not deviated. Since "pops" are most objectionable during low levels of sound the

desirability of careful tuning becomes obvious. In fact, a good receiver is best tuned by adjusting it to minimize noise.

Random noise after passing through a narrow-band filter results in a signal whose envelope value and frequency both fluctuate. If such a disturbance is superposed upon a constant-amplitude frequency-modulated signal two cases must be distinguished. When the noise envelope rarely if ever exceeds the signal the noise causes small disturbance in the resultant

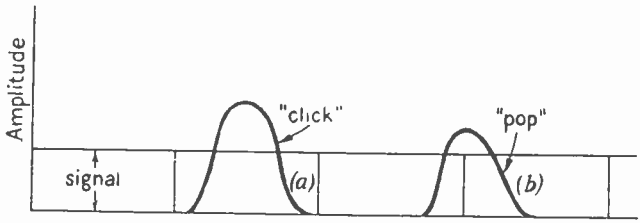

Fig. 5.12. The "pops" are located by the existence of a ringing tone larger than the signal when the phase difference is 180°

instantaneous frequency. The effective signal to noise output of the receiver is improved over the input ratio in about the same way as in the case where the disturbance is of constant amplitude. However, as the mean disturbance amplitude approaches the signal, the peak heights of the disturbance occasionally exceed the signal and when this happens in an unfavourable relative phase a "pop" results. As the signal is decreased relative to noise, the probability of "pops" increases and they occur more and more frequently. These "pops" have a *much* greater nuisance value than the interference when signals are large enough to suppress this random capture. Finally, when the signal drops below noise for a large percentage of the time the receiver output is fairly pure noise and experimentally we find that it has a peak value corresponding to a frequency swing of about plus and minus 15 kc.

5.6. Receiver Design. In case we wish to design a receiver capable of suppressing high percentages of interference, certain precautions are necessary.

1. *The linear portions must have very flat frequency response over the modulation band.* The necessity of this requirement is readily seen by inspecting Fig. 5.13. Here the receiver tuning response curve is arbitrarily distorted. A noise signal $N = aS$

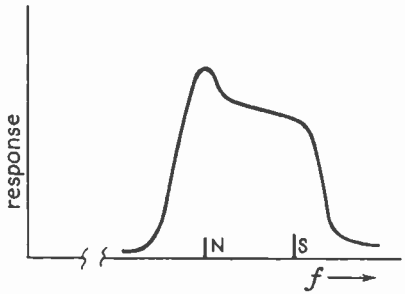

Fig. 5.13. The linear stages must have a flat enough response to prevent the signal S and the noise N interchanging their relative amplitudes

is added to the desired signal. If the response curve amplifies the noise frequency more efficiently than it amplifies the signal frequency the response to aS may be greater than that to S and the relative amplitudes of the signal and interference interchange. As the signal and interference swing back and forth over the band they interchange in amplitude one or more times per cycle and the resultant bedlam is intolerable.

2. *The response must be independent of amplitude over the interfering beat-frequency cycle.* Over a beat-frequency cycle the amplitude varies from the sum to the difference of the signal and noise voltages at a beat-frequency rate. Thus, using the same notation as before, the resultant varies from $1 + a$ to $1 - a$. For $a = 0{\cdot}9$ this is a ratio of 19:1. When the signal and interference are on the same channel they may be separated by a full channel width—by 150 kc for broadcast

purposes. Thus a limiter should be capable of suppressing large swings in amplitude very rapidly. In addition, of course, the receiver should be able to do this for a wide range of signal levels. PAANANEN (1953) has found that for really good broadcast practice these requirements are met by a combination of automatic volume control with two stages of gated-beam-tube limiters and a ratio detector.

3. *The variation of rectified output with frequency should be linear over a wide frequency range.* We saw in Art. 5.4 that during heavy interference the instantaneous frequency swings over a wide range in sharp frequency spikes. In fact, if in

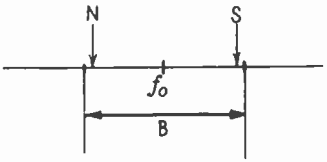

Fig. 5.14. Signal and noise at opposite ends of the receiver band

Fig. 5.14 we have a signal of unit amplitude at one end of the band and a noise interference of amplitude a at the other end, the resultant frequency lingers near the centre of the band during most of the beat cycle but swings violently upward at the phase opposition points. At those points the frequency swings to a maximum of $f_0 + B/2 + aB/(1 - a)$. Similarly, when S and N are interchanged the instantaneous frequency of a spike reaches a minimum of $f_0 - B/2 - aB/(1 - a)$. Thus, the frequency swings over a band of

$$W = f_0 + \frac{B}{2} + \frac{aB}{1-a} - \left(f_0 - \frac{B}{2} - \frac{aB}{1-a}\right)$$

$$= B + \frac{2aB}{1-a} = \frac{1+a}{1-a}B \quad . \quad . \quad (5.18)$$

For $B = 150$ kc and $a = 0{\cdot}9$ a bandwidth of 2·85 *megacycles* is covered. If the rectified output of a receiver is to be a measure of the desired signal-frequency deviations, and its

average is not to be changed by the presence of interference, the frequency detector should be linear over the entire range covered by the instantaneous frequency variations impressed on the detector. Thus, for $a = 0{\cdot}9$ a detector linear over a 2·85 Mc band is called for.

4. *The detector should be quick acting.* It should be recognized that a spike train may occur at a repetition rate of 150 kc. More serious than this even, the instantaneous variation of the amplitude along the steep sides of a spike such as that shown in Fig. 5.7 is very high. To avoid detector diagonal clipping the detector must be very quick acting. The Granlund-type ratio detector was developed primarily with this in mind. Thus the output condenser in Fig. 4.17 is alternately charged and discharged by the two diodes; it is not allowed to coast on the discharge portion of the cycle as in a single-diode amplitude-modulation detector.

5. *The linear stages of the receiver should have adequate selectivity.* The fact that the instantaneous frequency during frequency spikes may vary by megacycles is at first very disconcerting. Does this mean that the tuned circuits in the receiver must be megacycles wide? If so, what happens to the selectivity? As a matter of fact, the spectrum of two superposed signals is merely the sum of their individual spectra. For example, the spectrum of the sum of a unit-amplitude 100 Mc signal and a 0·95 amplitude, 99·85 Mc interfering signal consists of two lines at 100 and at 99·85 Mc respectively. This is a simple fact. What, then, becomes of the argument we have just given which would indicate that the *instantaneous* resultant frequency swings from about 99·925 to 103 Mc? Although we are entitled to think of the *instantaneous* frequency as covering a 3 Mc band, it is perfectly passed by a filter 150 kc wide. The apparent contradiction is removed when we recognize the fact that the resultant signal is simultaneously *amplitude* and frequency modulated and this reduces the spectrum to two lines. However, when the amplitude variations are removed by limiting, the spectrum not only covers the full range indicated by the excursion in instantaneous frequency but actually has

sidebands considerably beyond this due to the speed with which the frequency is changing.

A typical spectrum for the case $a = 0{\cdot}8$ is shown in Fig. 5.15. Excepting the wanted signal the largest spectral component is at the frequency of the interfering signal. However, calculation shows that for all values of a the magnitude of this component relative to the magnitude of the wanted signal component is less than a, the strength of the interference

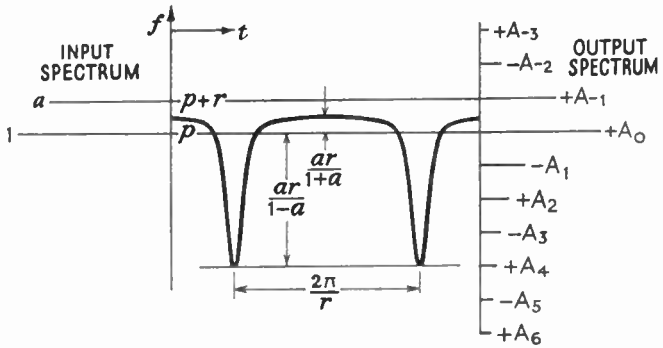

Fig. 5.15. The bandwidth of the output spectrum is much greater than that of the input spectrum

relative to the wanted signal at the input. Thus, it would appear that we could obtain a reduction of the interference by using a filter after the limiter which would pass only these two spectral components. However, the separation of these two components is variable, and when the signal and interference are on nearby frequencies a larger number of spectral components would be passed by the filter than when they were more widely separated. The main requirement is that the average frequency be maintained at precisely the frequency of the larger signal or in other words that the resultant signal can never be made to encircle the origin by the interference. This is known to be the case if all the spectral components are included, but if an arbitrary number of them is removed by the filter then there is no guarantee that this will remain

true. In practice, the use of a filter of bandwidth sufficient to pass all the spectral components has proved most satisfactory up to the present, but future work may show that improvements can be made by using somewhat narrower bandwidths (BAGHDADY, 1955).

The schemes just outlined are not the only means of suppressing high-level interference. CORRINGTON (1951) has obtained good results using an interlocked oscillator and the Bradley detector can be made to capture in the presence of heavy interference. In these cases, however, the principles are similar, as the rectified output must be independent of amplitude and must vary linearly with instantaneous frequency over a range considerably wider than the modulation band.

In addition to meeting the above requirements a frequency modulation receiver must be free from overloading effects caused by strong signals on channels other than that to which the receiver is tuned. This requires good selectivity ahead of the mixer and also suggests class A operation of the mixer in which the mutual conductance can be made to vary sinusoidally at the local oscillator frequency (CROSS, 1953). Frequency modulation is usually allocated high frequency regions of the radio spectrum, regions in which the atmospheric noise is small and may be less than that caused by thermal agitation and the statistics of electron emission. For this reason it is usually profitable to use great care in achieving a low noise figure in the low-level portions of the receiver. Since this problem is not unique to frequency modulation its detailed discussion seems out of place here.

Receivers built in accordance with the above principles have a broad tuning characteristic in the absence of interference unless the wanted signal is very weak. This is a result of the very broad detector; if the receiver is mistuned or if the local oscillator drifts the output remains unaffected so long as the signal remains within the linear range of the detector and the I.F. amplifier has sufficient gain to provide a large enough signal to saturate the limiter (PAANANEN, 1953). The chief practical difficulty with this type of receiver

is to achieve a sufficiently stable characteristic in the I.F. amplifier. A lower intermediate frequency would assist with stability but would make it more difficult to obtain a satisfactory image ratio. A double superheterodyne design can overcome the difficulty.

REFERENCES

ARGUIMBAU and GRANLUND: *Proc. National Electronics Conference*, **3**, 644 (1947)

ARGUIMBAU and GRANLUND: *Electronics*, **12**, 101 (1949)

ARGUIMBAU, GRANLUND, MANNA, and STUTT: *Transatlantic Frequency-modulation Experiments*, Research Lab. of Electronics (M.I.T.), Technical Report, 278 (1954)

BAGHDADY: *Proc. I.R.E.*, **43**, 51 (1955).

CORRINGTON: *Electronics*, **24**, 120 (1951)

CROSS: *TV and Radio Engineering*, **23**, No. 1 (Feb.–Mar.) (1953)

GRANLUND: Research Lab. of Electronics (M.I.T.), Technical Report, 42 (1949)

MARKO: *Archiv. der Electrischen Ubertrangung*, **8**, 111 (1954)

PAANANEN: *TV and Radio Engineering*, **23**, Nos. 2, 3, and 4 (Apl.–May, June–July, Aug.–Sept.) (1953)

SMITH and BRADLEY: *Proc. I.R.E.*, **34**, 743 (1946)

CHAPTER VI

APPLICATION TO TELEVISION

6.1. Introduction. The basic problem in the case of television transmission as in any other communication system is to achieve faithful reproduction of the signal with the highest possible signal to noise ratio and the best discrimination against interference. The fundamental advantages and disadvantages of frequency modulation remain the same, regardless of the type of information to be transmitted. However, sometimes the particular characteristics of the information can be used to offset drawbacks or to give further improvement. For example, the particular characteristics of speech and music and of the noise encountered, made it possible to increase the signal to noise ratio by the use of pre-emphasis. In this chapter we shall consider one or two peculiar properties of television signals in relation to their transmission by frequency modulation.

6.2. Bandwidth Requirements. One of the major differences between a television signal and speech or music signals is the bandwidth. For music, 15 kc is adequate for high quality reproduction, whereas for television, bandwidths of the order of four or five megacycles are customarily used. The analysis of a carrier frequency-modulated by a sine wave shows that there is an infinite number of sidebands which occur at frequencies removed from the carrier by multiples of the modulating frequency. Thus, in order to get faithful reproduction, it would appear that a bandwidth of at least several times the highest modulation frequency would be necessary. In frequency-modulation sound-broadcast practice a bandwidth of the order of $\pm$ 75 kc has been used, which is five times the highest modulation frequency. A corresponding system for television would mean the allocation of channels some 40 Mc wide.

There is a further point to be considered in the case of

television, because the signal may have a steady d.c. component. Suppose we use a deviation of ± 2·5 Mc. When the picture is black the signal is removed from centre frequency by 2·5 Mc. Now suppose there is some fine detail in the black part of the picture corresponding to a frequency of, say, 5 Mc. This will produce sidebands on each side of the steady black signal spaced at 5 Mc intervals. One of the first order sidebands is now removed from centre frequency by 2·5 + 5 or 7·5 Mc. Similarly, fine detail in a white region would produce a sideband 7·5 Mc on the other side of the centre frequency, so that a total bandwidth of 15 Mc would be required if only the first order sidebands were transmitted. The bandwidth required on this basis is given by the deviation plus twice the highest modulation frequency. Conversely, it would indicate that the maximum deviation allowable is given by the available bandwidth minus twice the highest modulation frequency. Since large deviations are desirable (at least as high as the maximum modulation frequency) it would seem that a channel several times wider than those in use at present would be required.

Although the use of such a wide channel is sufficient to ensure faithful reproduction it is by no means obvious that its use is necessary to obtain acceptable results. The major part of the information, the outline of the principal objects forming the picture will be contained in the waveform as a few rapid transitions of high contrast (or large amplitude). The detail will consist of a large number of transitions generally of considerably lower contrast. Thus, a transient analysis seems more appropriate to the television signal case than the analysis for sinusoidal modulation. The mathematical treatment of the response of even a simple tuned circuit to a frequency step is by no means easy, though several authors have written on it (SALINGER, 1942; EAGLESFIELD, 1946; HATTON, 1951). HATTON considers the frequency-time response of a simple tuned circuit when the driving frequency is changed suddenly. He calculates that the rise time of the system for a sudden incremental jump in frequency is the same as the rise time of the system for double side band

amplitude modulation. When the frequency jump is of the same order of magnitude as the bandwidth of the tuned circuit which effectively corresponds to using the whole of the available bandwidth for deviation, the rise time is somewhat shorter. Thus the situation is by no means as bad as we might have supposed on the basis of spectral analysis of a sinusoidally modulated wave. Some experimental results made by the authors using a 3-element filter and conditions approximating to the standards used for present day amplitude modulation television broadcasting showed that the rise time in the worst case may be increased by a factor of 2–2·5 (STUART, ARGUIMBAU, MANNA, and RODGERS, 1953).

Television does suffer from multi-path transmission interference which results in a ghost image. From what has been said in Chapter V it would seem that this interference could be reduced by using frequency modulation. However, we saw that it is only the average frequency which is preserved and the interference does produce a spike pattern superimposed upon the wanted signal. The success of the system depends upon how well this spike interference can be filtered out. The question will now be considered for the television case.

6.3. The Effects of Multi-path Transmission. We shall first discuss the effect of the spike interference on the picture. Consider a simple black-to-white transition as shown in Fig. 6.1*a*, and suppose that we have two-path transmission. Suppose, also, that the direct signal is the larger. Before the transition takes place, the frequency of the signals arriving by both paths is that corresponding to black, f_b; therefore, the frequency of the resultant is f_b (Fig. 6.1*c*). For a period equal to the time delay of the weaker signal with respect to the stronger, the frequency of the stronger signal is that corresponding to white f_w, whereas that of the weaker signal is still f_b. Throughout most of this period, then, the resultant signal has a frequency $\frac{1}{2}(f_b + f_w)$ which corresponds to some shade of grey. However, since the average frequency of the resultant must be f_w, there will be frequency spikes in the direction corresponding to white. These spikes will exceed

the white level, but since a signal corresponding to "whiter than white" will be registered as white by the system, the

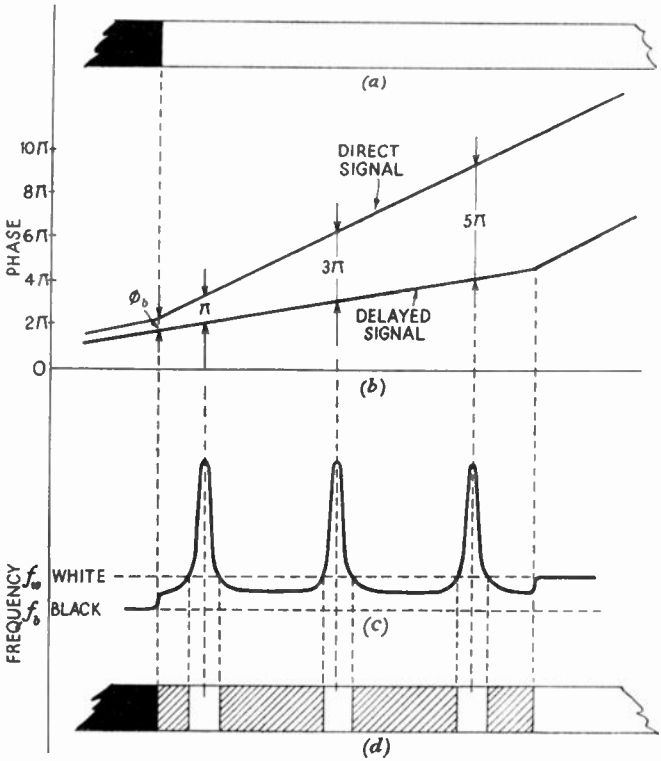

Fig. 6.1. The formation of frequency modulation television ghosts

resulting pattern will consist of white spots on a grey ground. (See Fig. 6.1*d*.) (Note that in Fig. 6.1 the abscissa can be thought of as either time or distance across the picture; the two are related by the speed of scanning, which is constant.)

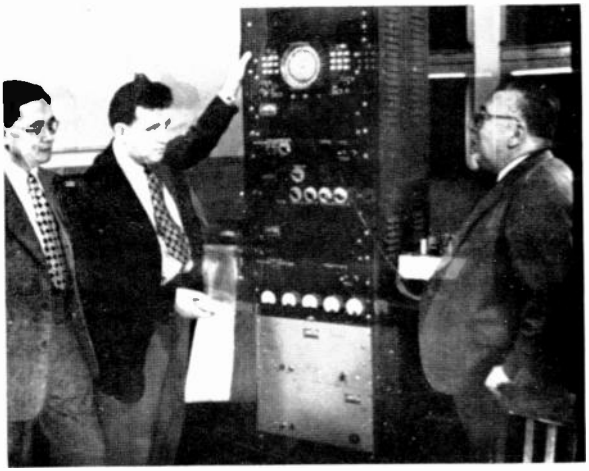

Fig. 6.2. Picture showing ghost due to multipath transmission—amplitude modulation

Fig. 6.3. Picture showing ghost due to multipath transmission—narrow-band frequency modulation

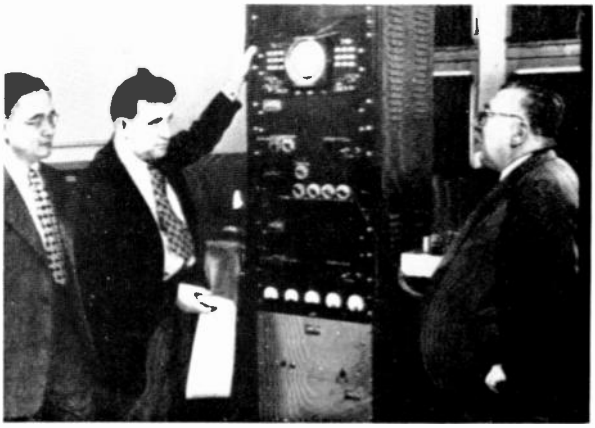

Fig. 6.4. Picture showing ghost due to multipath transmission—wide-band frequency modulation

Now suppose that the time delay of the interfering signal is τ. Then during a period of "black" transmission, the phase difference between the two signals is $\phi_b = 2\pi\tau f_b + 2\pi k$ (where k is an integer). When the direct path changes frequency the phase difference between the two signals increases linearly at a rate given by $2\pi(f_w - f_b)$ as shown in Fig. 6.1*b*. Thus, if the instant at which the direct path changes corresponds to $t = 0$, the phase difference at time t later is given by

$$2\pi\tau f_b + 2\pi k + 2\pi(f_w - f_b)t$$

A spike is formed when this expression has a value $(2n - 1)\pi$, where n is an integer

$$(2n - 1)\pi = 2\pi\tau f_b + 2\pi k + 2\pi(f_w - f_b)t$$

or
$$t = \frac{1}{2\pi(f_w - f_b)} [\{2(n - k) - 1\}\pi - 2\pi\tau f_b]$$

$$= \frac{1}{2(f_w - f_b)} [2(n - k) - 1 - 2\tau f_b] \quad . \quad (6.1)$$

The first spike is formed when

$$t = \frac{1 - 2\tau f_b}{2(f_w - f_b)} \quad . \quad . \quad . \quad (6.2)$$

which is dependent only upon τ, f_w, and f_b. Therefore the distance of the spot from the transition is the same for all transitions between two given frequencies. And on any line scan, the white spots will fall adjacent to the ones formed on the previous line and the resulting pattern will consist of white stripes on a grey background, as shown in Fig. 6.1*d*. Furthermore, the pattern formed on any particular frame will superimpose exactly on that formed in the previous frame.

From Chapter V we know that the spike recurrence rate is equal to the frequency difference between the two signals. When we use a 5 Mc deviation, the spacing of the stripes following a black-to-white transition, when viewed on a 12 in. screen, will be approximately 0·04 in., which will be

quite noticeable. A transition of lower contrast will give wider spacing and will lead to a more objectionable effect, although the magnitude of the spike (which is also proportional to the frequency difference) is smaller, until the spike becomes small enough to make the variation in intensity negligible.

The problem of removing this effect by filtering is difficult. The major effect of a low-pass filter would be to reduce the height of the sharp spike. This reduction will not make any difference to the picture, since it will still be registered as white. The filter will also cause only a relatively small shift in the grey level of the background. A more promising method of removing the effect would be to increase the spike recurrence rate above the resolving power of the system. In fact, if this resolving power is just sufficient to resolve the highest frequency we desire to transmit, this solution automatically demands a bandwidth assignment greater than that determined from the frequency components in the modulation waveform. (This is the situation existing in frequency-modulation sound broadcasting where the highest audio-frequency transmitted is about 15 kc/sec, and the bandwidth assignment is $\pm$ 75 kc.)

If we retain the original 5 Mc bandwidth, the spike frequency will always be in the same region as the picture information. It would appear, therefore, that the spikes cannot be removed without removing some of the picture information, with consequent deterioration of the quality of the final picture. However, the spikes are formed at the receiver, they are not present in the original modulation waveform at the transmitter. Thus it is possible to design a filter for reducing the effect of the spikes without affecting the picture quality, if the inverse filter is used at the transmitter to pre-correct for its effect. We are therefore led to a consideration of pre-emphasis, commonly used in frequency-modulation sound transmission.

Consider, again, the case of a rapid transition from black to white. The resulting waveform will approximate a ramp function. If this is applied to a pre-emphasis circuit, which

is essentially a differentiating circuit, an overshoot is produced. If the normal black-to-white transition shifts the transmitter frequency from one end of the band to the other, the pre-emphasized transition would, because of the overshoot, take the frequency outside the bandwidth. To prevent this from happening, the frequency difference corresponding to a black-to-white transition would have to be reduced. This, however, would reduce the repetition frequency of the spikes, making them more difficult to remove. An investigation into the problems of video pre-emphasis has been made previously (HATTON, 1951). The solution adopted here was to pass the pre-emphasized wave through a clipping circuit which removes all overshoots that would take the transmitter outside the prescribed range.

Since a standard de-emphasis circuit is used, some deterioration of picture quality results. The characteristic of the pre-emphasis circuit is given approximately by $[1 + (f/f_p)^2]^{1/2}$. By using a ramp function with a rise time of 0·08 μsec this circuit gives an overshoot of about 100 per cent, with f_p adjusted to 800 kc. From earlier numerical calculations, [using the uniform stretch function (KALLMAN, SPENCER, and SINGER, 1945) to represent the transition], we can estimate that removing the whole of this overshoot and then using standard de-emphasis will increase the rise time by a factor of about 8. Hence, there may be some blurring of high-contrast transitions. But this is the worst case that can occur. The rise time of a transition less than full black-to-white will not suffer so much deterioration, since only part of the overshoot is removed. (If the contrast is less than a certain level, none of the overshoot is removed.) However, the blurring of high-contrast transitions will always be present, even under single-path transmission conditions. The characteristic of the de-emphasis circuit is given by $[1 + (f/f_p)^2]^{-1/2}$. The lower the frequency f_p, the more effective the de-emphasis circuit becomes in reducing the spikes; but the overshoot produced by the corresponding pre-emphasis circuit will increase. This overshoot must be removed to prevent the given bandwidth being exceeded; and the larger the

overshoot is, the greater the deterioration of picture quality caused by its removal. Thus the retention of picture quality and the reduction of the spikes are conflicting requirements.

Some typical results are shown in Figs. 6.2, 6.3, and 6.4. In each case the interfering signal is about 50 per cent of the wanted signal. Figs. 6.2 and 6.3 compare the ghosts obtained with amplitude modulation and frequency modulation using a bandwidth of 5 Mc. In the frequency modulation case the whole of the available bandwidth was used for deviation, a pre-emphasis frequency of 800 kc was used and any overshoots which would take the transmitter frequency outside the assigned bandwidth were removed. Fig. 6.4 shows the improvement in interference suppression which can be obtained by using a 10 Mc bandwidth for the frequency modulation system. Again the whole bandwidth was used for deviation.

Under conditions of multipath transmission it is possible for the stronger signal to capture in a frequency-modulation system even in the presence of a strong interfering signal. This does not necessarily result in an undistorted signal. Spike interference is present and its repetition rate depends upon the frequency swing. If the assigned bandwidth is equal to the highest frequency component of the waveform to be transmitted, then the spike frequency must always fall in a region of the spectrum where there is also wanted information and effective filtering is difficult. If wider bandwidths are employed, then a more effective reduction of the interference is possible.

REFERENCES

EAGLESFIELD: *Wireless Engineer*, **23**, 96 (1946)

HATTON: Research Lab. of Electronics (M.I.T.), Technical Report, 196 (1951)

HATTON: Research Lab. of Electronics (M.I.T.), Technical Report, 202 (1951)

KALLMAN, SPENCER, and SINGER: *Proc. I.R.E.*, **33**, 169 (1945)

SALINGER: *Proc. I.R.E.*, **30**, 378 (1942)

STUART, ARGUIMBAU, MANNA, and RODGERS: Research Lab. of Electronics (M.I.T), Technical Report, 259 (1953)

INDEX