AIEE
chnical Paper
40-116

May 1940

FREQUENCY MODULATION

W. L. Everitt
Fellow AIEE
The Ohio State University,
Columbus

A paper recommended by the AIEE committee on communication and scheduled for presentation at the AIEE summer convention, Swampscott, Mass., June 24-28, 1940. Manuscript submitted April 23, 1940; made available for preprinting May 27, 1940.

FREQUENCY MODULATION

A Report on Progress

W.L. Everitt
Fellow A.I.E.E.

INTRODUCTION

Radio communication makes use of a medium common to the whole world for the transmission of many signals simultaneously. In order to accomplish this a high frequency electromagnetic wave has one of its characteristics varied in accordance with the instantaneous variations of the signal to be transmitted. The control of these variations is called modulation. The various simultaneous messages can then be separated by:

a. Differences in the frequency band used;
b. Differences in signal strength;
c. Differences in direction of the source.

The allocation of frequencies and geographic location to stations engaged in different services is now a matter of legislation and international agreement. This regulation is necessary in order to reduce interference to a minimum.

The range of a radio station is limited solely by the point at which undesired interference reduces the quality of the received signal below a certain minimum. The amount of interference which may be tolerated differs with different classes of service, for instance, it would be less on a broadcast program designed to produce pleasure, than on a communication service designed to convey intelligence.

There are five principal sources of interference to radio reception. They are:

a. Interference from other radio stations;

b. Interference from natural electrical disturbances such as thunderstorms. (Static)

c. Interference from electrical equipment not intended for radio purposes.

d. Interference between identical signals traveling from the station originating the desired signal, but over two different paths. Since radio waves are alternating phenomena resolvable into a band of frequencies, the addition of two similar signals traveling over different paths must take account of both magnitude and phase. Distortion in the resultant signal may result due to the varying phase relations between the components with the same frequency in the two signals. One of the paths is usually caused by reflection from some medium such as the ionosphere. If this path varies in length with time, fading will result. When this produces distortion the phenomenon is called selective fading.

e. Interference from random noise produced in the receiver by fluctuations in the motion of the electrons in the early stages of the amplifiers.

The major problems of the radio engineer are:

I. To obtain the maximum range at the minimum cost.

II. To secure the desired quality in the reproduction of signals.

Because the range is determined by the interference, and the quality is greatly affected by it, the reduction of interference becomes of paramount importance.

The reduction of interference must be accomplished by making use of some characteristic which differentiates the desired signal to a greater or less extent from the undesired interference. Four methods of differentiation have been extensively used. Each method in turn has its limitations. These methods and their limitations are:

1. Use of high power in the transmitter so that the strength of the desired signal will dominate the undesired.

   This method is limited by the cost of high powered transmitters and by the interference it introduces to other services. Furthermore it does not affect selective fading since both signals (coming over two paths) are increased by the same amount.

2. Increasing the modulation of the radio wave to the greatest possible value.

   This method is limited in amplitude modulation because it is not possible to vary the magnitude of a radio wave by more than one hundred per cent and interfering signals, including static, will in general be modulated by similar amounts.

3. Use of selective circuits in the receiver so that only energy in the narrow band of frequencies which includes the desired signal will be received.

   This method is limited because a definite band width is necessary for any given quality of reproduction and within this band there may be some portion of the energy in the spectrum of the interference.

4. Use of direction antennas at the receiver so that it is most sensitive to electromagnetic waves coming from the direction of the transmitter creating the desired signal and is insensitive to radiations originating in other directions.

   This method is limited by the expense of directional antennas and by the fact that some interference may be originating in the same direction as the desired signal.

It is apparent that the methods just mentioned, taken individually or in combination, do not offer a complete solution of the problem. In fact no complete solution would appear possible, as the ultimate range of any transmission must be determined by the tolerable interference. However, any method which offers increased possibilities in the differentiation between desired signal and interference may be used to improve transmission. FREQUENCY MODULATION offers such an additional method by which interference may be separated from the desired signal and it is the purpose of this paper to outline the principles and practices by which this may be accomplished.

The use and study of frequency modulation is not new. The Poulsen arc, developed before the war, transmitted continuous wave telegraph signals in which the frequency was shifted from one value to another when the key was depressed. Carson[1] and Roder[2] made theoretical investigations of the effects of frequency modulation on the spectrum of the modulated wave. Carson's investigation was made to analyze the proposal that frequency modulation could be used to reduce the band width required for a given signal. He proved that, on the contrary, frequency modulation never reduced the band width and might greatly increase it. Mathematics provides a correct answer to questions which are asked by its means, but it cannot be expected to provide answers to questions which are not asked. What was overlooked in the early mathematical analyses was the fact, later demonstrated by Armstrong[3], and subsequently by Carson[4] and Fry, that frequency modulation provides an important method of distinguishing between desired and undesired signals which occupy the same portion of the frequency spectrum.

It will be necessary to go into some details of the principles of frequency modulation in order to show the reasons for this effect.

MODULATION

Modulation of a radio wave is the process by which some characteristic of the radio wave is varied in accordance with the time variation of a signal, such as the instantaneous variations associated with speech, music, or the manipulation of a telegraph key. A general alternating wave may be represented by the equation

$$e = A \sin \Theta \tag{1}$$

where Θ is given by the relation

$$\Theta = \omega t + \Phi \tag{1a}$$

and so

$$e = A \sin (\omega t + \Phi) \tag{1b}$$

(In this discussion the word "wave" will be used in one of its accepted meanings to denote a repetitive phenomenon)

Two groups of modulation methods are recognized.

1. Amplitude modulation where A is varied by the signal.
2. Angular modulation where Φ is varied by the signal.

(Frequency modulation is a special form of angular modulation)

AMPLITUDE MODULATION

In an amplitude modulated wave the amplitude is varied about its mean value in proportion to the signal. Let the original signal (such as the sound pressure on the microphone) be represented by the function f(t). Then the amplitude factor A of Eq. (1b) is modified by f(t) to give the amplitude modulated wave

$$e = A\left[1 + b\,f(t)\right]\sin(\omega t + \Phi) \qquad (2)$$

where b is a factor determined by the design and operation of the modulating system and has dimensions such that $[b\,f(t)]$ is a pure numeric. b is usually a constant, but in some cases it is made a function of audio frequency. For example, if it is made to change with frequency in the proper manner compensation may be secured for defects in the frequency characteristic of some other part of the system.

The amplitude variation cannot carry the amplitude below zero. Therefore the factor b should be so chosen by the operator that $[1 + b\,f(t)]$ never becomes negative. Therefore $[b\,f(t)]$ should not exceed an absolute value of unity. This absolute value of the maximum of $[b\,f(t)]$ is called the amplitude modulation factor and is given the notation m_a.

If the signal f(t) is sinusoidal with a frequency $\rho/2\pi$, Eq. (2) becomes

$$e = A\,(1 + m_a \sin \rho t)\sin \omega t \qquad (3)$$

The curve of Eq. (3) is illustrated in Fig. 1 for $m = 0.5$ and $\frac{\omega}{\rho} = 10$. It will be noted that the wave crosses the axis at regular time intervals of $\frac{2\pi}{\omega}$ seconds for both the modulated and unmodulated waves.

In alternating phenomena a single frequency is represented by the projection of a vector of constant length rotating with the constant angular velocity $\omega = 2\pi f$. The wave of Eq. (2) could also be represented by a vector rotating with a constant angular velocity ω, but the length of the vector would be changing at a low frequency rate as given by the equation

$$\text{Length of vector} = A\left[1 + b\,f(t)\right] \qquad (4)$$

The term $A\left[1 + b\,f(t)\right]$ is called the envelope of the wave. In Eq. (3) the envelope would be $A\left[1 + m_a \sin \rho t\right]$ as is illustrated in Fig. 1.

In drawing vectors which represent alternating phenomena it is common practice to consider that the observer is traveling on a platform which is also rotating about the same center with a velocity ω. The original vector would then appear to be stationary and could be represented by a single drawing. However, if either the magnitude or the phase of the vector is changing with time, a series of successive drawings is necessary to illustrate what is happening.

These successive drawings of stationary vectors for the wave of Fig. 1 are shown in Fig. 2 for time intervals of one-eighth the period of the low frequency wave producing the modulation.

At the receiver the detector produces a response which is proportional to the envelope of the modulated wave (except for the constant component).

INTERFERENCE OF TWO AMPLITUDE MODULATED WAVES

If a second amplitude modulated wave of the same carrier frequency and phase is added to the wave of Fig. 1 the resultant wave will have an envelope which is the sum of the envelopes of the two waves, for the vectors will be adding in phase. The interfering effect will be noticeable if the undesired signal is as much as one per cent of the desired signal. Hence it is desirable to make the value of m_a as large as possible, since the operator of a given communication system cannot control the modulation of the interfering wave with the undesired signal.

If the frequency of the interfering wave is slightly different from the desired wave (the difference being too small to eliminate it by selective circuits) then the interfering wave will produce a variation in the envelope which variation has an amplitude equal to the magnitude of the interfering wave (even if it is unmodulated). This additional variation will occur at a frequency which is equal to the difference between the carrier frequencies of the desired and undesired signals, and will produce a squeal which is further superimposed on the resultant envelope. This is illustrated by the vector diagrams in Fig. 3, where the undesired signal has a frequency which exceeds the frequency of the desired signal by $\frac{1.5\rho}{2\pi}$. It is seen that the resultant envelope is modified by an additional component equal to the magnitude of the undesired wave, and so introduces interference proportional to the magnitude of the interfering wave.

Again it is apparent that the amplitude of the envelope of the desired signal should be kept as large as possible in order that the interference may be minimized. If Eq. (2) represents current or voltage, the amplitude of the envelope may be increased by increasing either the power or the amount of modulation (m_a).

ANGULAR MODULATION

In angular modulation (of which frequency modulation is a subdivision) the angle Φ of Eq. (1b) is given by a function of time which is related, but not in all cases, directly proportional, to the signal function f(t). The two principal subdivisions of angular modulation which have been extensively studied are Phase Modulation and Frequency Modulation.

(a) Phase Modulation.

In this type of modulation the phase angle Φ is made to v~~e~~ary in accordance with the signal. That is

$$\Phi = b_1 f(t) \qquad (5)$$

where b_1 is a constant determined by the design and operation of the modulating system. When Eq. (5) is inserted in Eq. (1b) the wave becomes

$$e = A \sin \left[\omega t + b_1 f(t)\right] \qquad (6)$$

The maximum value of $b_1 f(t)$ is called the phase modulation factor m_p. It is the maximum number of radians by which the phase of the carrier is altered during modulation. If the signal is sinusoidal with a frequency $\rho/2\pi$, Eq. (6) becomes

$$e = A \sin \left[\omega t + m_p \sin \rho t \right] \tag{7}$$

(b) Frequency Modulation.

In this type of modulation the instantaneous frequency is varied about the average value $\omega/2\pi$ in proportion to the instantaneous value of the signal. By definition, the use of the word "frequency" is extended to the general equations (1) and(1b) by the relation

$$2\pi\, f_{inst} = \frac{d\Theta}{dt} = \omega + \frac{d\Phi}{dt} \tag{8}$$

Since ω is a constant (2π times the carrier frequency) the signal must modify $\frac{d\Phi}{dt}$ so that the instantaneous frequency is given by the relation

$$f_{inst} = \frac{\omega}{2\pi} + b_2\, f(t) \tag{9}$$

where b_2 is a design and operating constant. The maximum value of $b_2\, f(t)$ is the maximum deviation in instantaneous frequency of the modulated wave from the unmodulated one and is called the frequency modulation factor, or frequency deviation, m_f. If $f(t)$ is a sine wave with a frequency $\rho/2\pi$ then

$$b_2\, f(t) = m_f \sin \rho t \tag{10}$$

If Eq. (10) is combined with Eqs. (8) and (9)

$$2\pi\, f_{inst} = \omega + 2\pi\, m_f \sin \rho t = \omega + \frac{d\Phi}{dt}$$

which gives

$$\Phi = \int 2\pi\, m_f \sin \rho t\, dt = - \frac{m_f}{f_\rho} \cos \rho t$$

where f_ρ is the frequency of the modulating signal. If this phase angle is inserted in Eq. (1b) the result will be

$$e = A \sin \left[\omega t - \frac{m_f}{f_\rho} \cos \rho t \right] \tag{11}$$

Eqs. (11) and (7), which apply to a signal with a single frequency, do not differ appreciably (except for a 90° shift in the modulation phase). In Eq. (11) the maximum shift in phase (corresponding to the phase modulation factor m_p) will be

$$m_p = \frac{m_f}{f_\rho} \tag{12}$$

where m_f m_p is the frequency deviation and f_ρ the modulating audio frequency. The value of m_f when f_ρ is the maximum audio or signal frequency to be transmitted is called the deviation ratio.

m_p in phase modulation and m_f in frequency modulation are arbitrary design factors. Unlike amplitude modulation they are not restricted to a maximum value of unity, for m_p may be hundreds of radians or m_f thousands of cycles per second if desired. The limitations on m_p and m_f will be determined by the allowable frequency spectrum and will be discussed later.

The distinction between phase and frequency modulation is as follows: if the frequency, but not the intensity of the modulating signal changes

m_p is constant in phase modulation

m_f is constant in frequency modulation.

It follows from Eq. (12) that in frequency modulation the phase deviation m_p is inversely proportional to the modulating frequency. On the other hand in phase modulation the frequency deviation is directly proportional to the modulating frequency.

Fig. 4 is an illustration of the angular modulation as represented by Eq. 7 for the case where $m_p = 0.5$ and $\omega/\rho = 12$. On a casual examination this would appear to be a single frequency wave. However, the intervals at which it crosses the axis vary throughout the audio cycle. In order to show this the first, fourth, seventh, tenth and thirteenth cycles are expanded and shown in Fig. 5. It is seen that the varying shift in phase also produces a change in frequency which varies throughout the low frequency cycle.

The successive vector diagrams for the angular modulated wave of Fig. 4 (corresponding to the diagrams of Fig. 2 for an amplitude modulated wave) are shown in Fig. 6. The signal wave is included for identification of the various instants.

The difference between phase and frequency modulation may be illustrated by the way the motion of the resultant vector would appear to an observer riding with the carrier vector. In phase modulation, two audio signals of equal amplitude, but of different frequencies, would produce equal angular amplitudes in the apparent swing of the resultant vector. In frequency modulation two audio signals of equal amplitude would produce equal maximum angular velocities in the apparent swing of the resultant vector. In this latter case (frequency modulation) the maximum angle of swing would be inversely proportional to the audio frequency (as is indicated by Eq. 12). This is illustrated by Fig. 7 where the vectors for both frequency and phase modulation are drawn for two signals with an audio frequency ratio of two to one. Note that in phase modulation the maximum angle Φ_m is the same for both signals while for frequency modulation the maximum angle Φ_m for signal A (the lower frequency) is twice that for signal B. Since the angular velocity is proportional to the instantaneous value of the signal in frequency modulation the vector reaches its maximum angle of deviation when the signal is zero while in phase modulation it reaches its maximum angle of deviation when the signal is a maximum.

OTHER TYPES OF ANGULAR MODULATION

Phase and frequency modulation are not the only possible types of angular modulation, but are only two members of an infinite group. Other possible types are:

(c) Angular Acceleration Modulation.

In this type of modulation the second time derivative of Φ is directly proportional to the signal function

$$\frac{d^2\Phi}{dt^2} = b_3\, f(t)$$

In this type m_p would be inversely proportional to the square of the audio frequency.

(d) N^{th} Order Modulation.

In this general type of angular modulation the n^{th} derivative of Φ is directly proportional to the signal function.

$$\frac{d^n\Phi}{dt^n} = b_{n+1}\, f(t) \tag{13}$$

In this type m_p would be inversely proportional to the n^{th} power of the frequency for modulating signals of equal intensity.

In radio transmission by angular modulation means are provided at the receiver so that the detected signal is proportional to the angular modulation (of the particular sub-division selected) and at the same time this detected signal is made unresponsive to amplitude variations. These means will be discussed in more detail later.

INTERFERENCE OF TWO ANGULAR MODULATED WAVES

When two angular modulated waves of the same carrier frequency are added together, the total angular modulation is not the sum of the two individual modulations. This is in distinct contrast to amplitude modulation where the resultant envelope is the sum of the individual envelopes.

This can be illustrated by Fig. 8 where an angular modulated wave B is represented by a vector whose angle is changing with time. This is added to a larger vector A which for the moment will be assumed to be unmodulated. The resultant vector R will be the sum of the two vectors.

It is apparent that if B is less than A, then no matter what the total angular variation of B may be (even if it is hundreds of radians) the total angular variation between R and A cannot exceed $\tan^{-1}\frac{B}{A}$. For instance if $\sin^{-1}\frac{B}{A}$
$\frac{B}{A} = 0.5$ the maximum value of m_p for the vector R when A is unmodulated is
$m_p = 0.46$. If $\frac{B}{A} = 0.5$ and A in turn has its angle modulated, then the difference 0.52
between the angle of A and that of R cannot exceed 0.46 radians at any instant. 0.52
If the modulation factor (m_p) of A is made large in comparison with 0.46, the 0.52
interference of B becomes negligible, in spite of the fact that the magnitude of B is by no means negligible in comparison with A.

This analysis justifies the experimental results which show that when two frequency modulated signals are picked up by a receiver, there is no appreciable interference between the two signals if the stronger exceeds the weaker by a ratio of two to one or more.

It will be seen that the greater the value of m_p used for the desired signal the greater is the discrimination against the undesired signal, but this discrimination is not affected by the value of m_p used in the undesired signal.

The discrimination against interference obtained by angular modulation applies to all five types of interference enumerated in the early part of the paper. In particular static may be represented as a vector of varying phase and magnitude. The selective circuits of the receiver admit only those components within the band to which it is receptive. If the amplitude of the admitted noise does not exceed half the amplitude of the desired wave, a very small amount of noise will be introduced into the output. The greater the average phase 0.52 deviation in comparison with the angle 0.46 (approximately 0.5) the greater will be the discrimination against the noise. It should also be observed that components of the noise vector which differ in frequency from the carrier by superaudible frequencies, will produce superimposed angular velocities above audibility and so do not contribute to the noise, as long as the noise is small compared with the signal.

In radio operation it will be found that if a portable receiver is driven in an automobile away from a frequency modulated transmitting station, no appreciable noise will be experienced until the desired field strength drops to twice the noise field strength (taking into account only those components of noise accepted by the selective circuits of the receiver). The noise then rises rapidly, so that a sharp threshhold is experienced.

Within the distance limited by the threshhold, the signal to noise ratio can be improved by either increasing the power or increasing the modulation factor (either phase or frequency). Since power is proportional to the square of voltage or current in a given system, doubling the frequency deviation in FREQUENCY MODULATION has the same effect on the signal to noise ratio as increasing the transmitted power four times. In general an increase in the maximum frequency deviation by a ratio n would be equivalent in its effect on the signal to noise ratio to an increase in power by the ratio n^2.

The actual voltage produced by noise in the amplifier of a radio receiver increases with the width of the band accepted, the rate of increase depending on the type of noise. This introduces some disadvantage to the use of a wide band, because a stronger desired wave is necessary to insure that the desired voltage shall exceed the noise voltage.

R.F. Guy of the National Broadcasting Company reported before the Federal Communications Commission that in experiments with a 1 kw transmitter and an antenna 1000 feet high, the threshhold for a value of m_f equal to 75 kilocycles was at 86 miles, while with an m_f of 15 kilocycles the threshhold was at 100 miles due to the smaller noise voltage accepted by a more selective receiver. Within the threshhold distances, however, there is a greater discrimination against noise with the greater frequency deviation.

The second major objection to the use of a large frequency deviation is that it would limit the number of stations which can serve a given area if a fixed total band width is allowed for the service.

A compromise must be reached and standards set before the use of frequency modulation is generally adopted so that the receivers may work with the transmitters. In order to study the possible or required allotment, a spectrum analysis must be made of the different classes of modulation.

SPECTRUM ANALYSIS OF AMPLITUDE MODULATION

The wave of Eq. ~~(2)~~ (3) may be expanded by the use of simple trigonometric identities. This equation becomes

$$e = A \sin \omega t + \frac{mA}{2} \cos (\omega - \rho)t - \frac{mA}{2} \cos (\omega + \rho) t \qquad (14)$$

Eq. (14) shows that the wave which is amplitude modulated by a single frequency may be analyzed into three component frequencies with the following designations:

$A \sin \omega t$	The carrier
$\frac{mA}{2} \cos (\omega - \rho)t$	The lower side frequency
$\frac{mA}{2} \cos (\omega + \rho)t$	The upper side frequency.

The three components may be represented by three vectors rotating at different angular velocities. Again if the observer were rotating with the carrier vector, this vector would appear to be stationary. The upper side frequency vector would appear to be rotating counter-clockwise at a velocity ρ and the lower side frequency would appear to be rotating clockwise at the same velocity ρ.

The three vectors corresponding to the wave of Fig. 1 are shown in Fig. 9. It will be observed that the upper and lower side band vectors add together to form a vector M called the modulation vector, which is always in phase with the carrier vector, but which varies in magnitude.

The three component frequencies of Eq. (6) are represented graphically in Fig. 10. This is shown primarily so that it may be compared later with the frequency spectrum used in frequency modulation.

Carrier amplitude
is [illegible]
& ma

If the original signal were a complicated sound wave instead of a single frequency, a spectrum analysis would show it to be represented by a band of frequencies. The lower and upper side frequencies would expand into two bands of frequencies each as wide as the band of the original audio signal. For instance if the signal were restricted to a band of 0-5000 cycles the two side bands would extend from 5000 cycles below to 5000 cycles above the carrier frequency. Since the quality of a signal depends upon the width of the band which may be transmitted, an improvement in the quality of transmission would require an extension of the frequency spectrum occupied by the radio wave. However, the narrower the frequency band which is used the greater will be the number of stations which can be accommodated. In practice a compromise must be made. Standard broadcasting stations in North America are assigned carrier frequencies in the range of 550 to 1600 kilocycles, these assignments being separated at intervals of ten kilocycles. In order to prevent interference, selective circuits are required in the receiver which are so sharp in most commercial models that side-band components more than 3000 cycles away from the carrier are greatly attenuated. Hence the quality which is permissible in practical operation is limited by the major problem of interference.

SPECTRUM ANALYSIS OF ANGULAR MODULATION

The angular modulated wave of Eq. (7) may be expanded by the use of the identities

$$\sin (p \sin x) = 2 \left[J_1(p) \sin x + J_3(p) \sin 3x + J_5(p) \sin 5x + \dots \right] \qquad (15a)$$

$$\cos (p \sin x) = J_0(p) + 2 \left[J_2(p) \cos 2x + J_4(p) \cos 4x + J_6(p) \cos 6x + \dots \right] \qquad (15b)$$

where $J_n(p)$ is the n^{th} order Bessel Function of the first kind. Eq. (7) may be written

$$e = A\left[\sin \omega t \cos (m_p \sin \rho t) + \cos \omega t \sin (m_p \sin \rho t)\right] \qquad (16)$$

If Eqs. (15a) and (15b) are inserted in Eq. (16) the following result will be obtained

$$\begin{aligned} e = A\Big\{ J_0(m_p) \sin \omega t &+ J_1(m_p)\left[\sin (\omega + \rho)t - \sin (\omega - \rho) t\right] \\ &+ J_2(m_p)\left[\sin (\omega + 2\rho)t + \sin (\omega - 2\rho)t\right] \\ &+ J_3(m_p)\left[\sin (\omega + 3\rho)t - \sin (\omega - 3\rho)t\right] \\ &+ J_4(m_p)\left[\sin (\omega + 4\rho)t + \sin (\omega - 4\rho)t\right] \\ &+ \cdots\cdots\cdots\cdots\cdots\cdots \\ &+ J_n(m_p)\left[\sin (\omega + n\rho)t + (-1)^n \sin (\omega - n\rho)t\right] \\ &+ \cdots\cdots\cdots\cdots\cdots\cdots \Big\} \end{aligned} \qquad (17)$$

This indicates that there are an infinite number of side frequencies for a single frequency signal. However this is not as bad as might at first appear because for any given value of m_p, there will be a value of n above which the coefficients $J_n(m_p)$ fall off rapidly and become negligible. This is shown in Fig. 11. For example if m_p is one-half radian or less, only the first pair of side frequencies are important. On the other hand if m_p is equal to 20 radians, side frequencies out to the 24th pair would be appreciable. For large values of n this rapid falling off of $J_n(m_p)$ occurs just beyond $n = m_p$. Observe also that the value of the carrier component is always reduced when modulation occurs since $J_0(m_p)$ is less than one for all values of m_p different from zero. This is in contrast with amplitude modulation where the value of the carrier is not affected by modulation.

Fig. 4 was drawn for a phase modulation factor of 0.5 and so the first pair of side-bands are the only ones of importance. If all other side-bands are neglected the vector diagrams including the side-bands for different instants of Fig. 4 can be shown as in Fig. 12. The signal wave is shown for identification. The carrier and resultant vectors are the same as those shown in Fig. 6. The modulation vector, which is the sum of the two side-band vectors, is always 90° out of phase with the carrier and varies in magnitude in the same way that the modulation vector varies in amplitude modulation. The neglect of higher order side frequencies is the same as an assumption that there is a negligible difference between the arc and a tangent line of the same length when the angle is small.

When the modulation vector is added to the carrier vector it causes the resultant vector to alternately advance beyond and retard behind the carrier vector. The maximum advance and retardation is approximately one-half radian. The length of the resultant vector is substantially constant. If the additional side-bands were included the length of R would be exactly constant.

If the phase modulation exceeds one-half a radian additional side-bands must be included because the arc and chord are no longer substantially the same. The addition of the vectors corresponding to these side-bands is illustrated in Fig. 13 for $m_p = 1$ and for one-quarter of an audio cycle, the other three-quarters being similar. It will be noticed that each pair of side-bands has associated with it a modulation vector which maintains a constant phase with respect to carrier (assuming that phase reversals are taken care of by negative signs).

If the modulated wave represents a quantity whose square is proportional to power in a given system the average power in an angular modulated wave is not changed by the modulation, as the r.m.s. value of the wave is not modified if the amplitude remains constant. Therefore the square root of the sum of the squares of the carrier and all the side-band components remains constant for all values of m_p. The side-band power is obtained by a reduction in carrier power. This is also proven by the well known relation

$$J_o^2\,(m_p) + 2 \sum_{n=1}^{n=\infty} J_n^2\,(m_p) = 1$$

for all values of m_p. The number of terms which are of importance in the infinite series can be evaluated by setting

$$J_o^2\,(\overset{m_p}{\cancel{x}}) + 2 \sum_{n=1}^{n=s} J_n^2\,(m_p) \geq \lambda$$

Then if λ is taken as some value less than unity, the sum can terminate with a finite value of n equal to s. If λ is equal to 0.999 then 99.9 per cent of the energy in the wave would be due to side-band components corresponding to values of n equal to or less than s.

For example, if $m_p = 4$ and six components are taken in each side-band

$$J_o^2(m_p) + 2 \sum_{n=1}^{n=6} J_n^2(m_p) = 0.157688 + 2(0.004356 + 0.132569 + 0.185072 + 0.079017 + 0.017450 + 0.002411)$$

$$= 0.999438$$

and all the components corresponding to $n > 6$ would contain only 0.0562 per cent of the energy.

The constancy of power output is in marked contrast to amplitude modulation where the carrier power remains constant and the side-band power is added. For that reason certain problems in design are simplified in a phase or frequency modulated transmitter. This will be discussed in the section on transmitters.

COMPARISON OF THE SPECTRA OF PHASE AND FREQUENCY MODULATION

In PHASE MODULATION the value of m_p is made directly proportional to the maximum value of the signal. If two different audio frequencies have equal amplitudes, and modulate the signal in succession, the same number of side-band components would be necessary for each case and these components would have the same relative magnitude. It has been shown that the advantage of angular modulation in the reduction of interference requires the use of large values of m_p for the desired signal. If the value of m_p for a special case is taken equal to 20, then by Fig. 11 it is apparent that approximately 24 side-band components would be desirable for the upper side-band and a similar number for the lower side-band. Therefore if an audio signal of high quality containing components up to 15,000 cycles were to be transmitted, a band width of approximately 2 x 24 x 15,000 or 720,000 cycles would be required. This is obviously impracticable. For this reason phase modulation (as distinguished from frequency modulation) has not been used for radio transmission.

In FREQUENCY MODULATION the value of m_f is made directly proportional to the maximum value of the signal. If two different audio frequencies have equal amplitudes and modulate the signal in succession with equal values of m_f, by Eq. (12) the values of m_p for the two cases will be inversely proportional to the audio frequency. Thus if m_p is equal to 4 for 15,000 cycles it would be equal to 40 for 1500 cycles and equal to 400 for 150 cycles. A study of Fig. 11 shows that the number of components of appreciable magnitude in each side band is slightly in excess of m_p. Therefore as the modulating frequency is reduced, the number of components necessary increases, and the modulated wave occupies almost a constant band width in the spectrum. As an example consider a case where the maximum frequency deviation is assumed to be 60,000 cycles. Then if a high quality signal is to be transmitted, frequency components in this signal up to 15,000 cycles might be desired. If the wave were frequency modulated with a 60,000 cycle deviation (m_f = 60,000) at 15,000 cycles m_p would equal $\frac{60,000}{15,000} = 4$ radians. For this case Fig. 11 shows that approximately six components in each side-band separated at intervals of 15,000 cycles are desirable and the corresponding band width would be 2 x 6 x 15,000 or 180,000 cycles. On the other hand if the wave were to be frequency modulated with a 60,000 cycle deviation by a 3000 cycle wave $m_p = \frac{60,000}{3,000} = 20$ and approximately 24 components in each side-band separated at intervals of 3000 cycles would be desirable. The band width for this signal would be 2 x 24 x 3000 = 144,000 cycles which is somewhat less than that needed for $m_p = 4$ at 15,000 cycles. The following table may be constructed for a maximum deviation of 60,000 cycles.

TABLE I

Audio Frequency	m_p for 60,000 cycle deviation	Approx. number of side-band components required	Approx. band width in kilocycles
30	2000	4030	120.06
60	1000	2020	120.20
600	100	208	124.8
2500	24	46	140
3000	20	24	144
5000	12	30	150
10000	6	16	160
15000	4	12	180

The spectrum analysis for a deviation of 60 kilocycles and modulating frequencies of 2500, 5000, 10,000, and 15,000 cycles is shown in Fig. 14 and it is apparent that the signal is contained within a band width of approximately 200 kilocycles in all cases. In practice, no audio signal would contain only the higher audio frequencies and so it is found practicable to use a deviation of m_f of 75,000 cycles for a band width of 200 kilocycles.

The spectrum analyses for a modulating frequency of 15,000 cycles and deviation frequencies of 30, 15, 7.5 and 3.0 kilocycles are shown in Fig. 15.

It is apparent from Figs. 14 and 15 that when the deviation frequency is large compared with the audio frequency, the band width required is approximately twice the deviation frequency while when the audio frequency is large compared with the deviation frequency the band width is twice the audio frequency. The latter case coincides with the situation in amplitude modulation. In other words the band width required is approximately twice the larger of the two frequencies (audio or deviation). If the audio and deviation frequencies are approximately equal the band width required is approximately four times that of the larger frequency. (See Fig. 15 for m_p = 0.5, 1 and 2)

The spectra of Figs. 14 and 15 may be used for any other combinations of audio and deviation frequencies which have the same deviation ratio m_p by modifying the scale of abscissa so that the interval between adjacent components is equal to the audio frequency.

The reader must be cautioned that if a signal contains two or more audio frequencies, the resultant spectra cannot be obtained by adding the spectra resulting from each audio frequency alone (as can be done in amplitude modulation). However, the total spectra will remain approximately within the limits set by the maximum frequency deviation when the latter is large.

Although the discrimination against noise is proportional to m_p, it is impracticable to use large values of m_p at all audio frequencies because of the band width involved. However, noise and interference is the composite result of a larger number of noise components. If frequency modulation is employed, the maximum value of m_p is obtained for each audio component in the signal which will at the same time keep the side-band components within the limits in the spectrum assigned to the transmission. Therefore frequency modulation is the type of angular modulation which reduces the composite noise effect to the greatest practicable extent.

If the maximum modulation factor m_p (deviation ratio) is low, m_p less than 0.5, frequency modulation does not appear to have any advantages over amplitude modulation. The early proposals for the use of frequency modulation envisioned this method of operation and so were discarded after the analyses of Carson[1] and Roder[2].

PRACTICAL CONSIDERATIONS IN FREQUENCY MODULATION

It has been shown that the reduction of interference makes frequency modulated transmitters most desirable for the transmission of high quality signals. However these transmitters require relatively large frequency bands. Therefore frequency modulation, or F-M, does not appear to be feasible in the present broadcast band. For high quality broadcasting they should be allocated to high frequencies where band widths of 200 kilocycles are available. The Federal Communications Commission has assigned the frequencies from 43 to 50 megacycles for this service or a total of 40 channels. This is in the range of the so-called ultra-high frequencies.

Radio waves with frequencies of the order of 40 megacycles and above are not reflected from the ionosphere, and so their range is limited by the curvature of the earth. For the same reason static is greatly reduced at these frequencies because the energy which lies in the ultra-high frequency spectrum and which is originated by electrical disturbances at distant points on the earth's surface cannot travel to the receiver over long distances. Other factors also reduce static at high frequencies. As an average, the static voltage producing interference in a receiver with a given band width and tuned to 40 megacycles is about one-fortieth of the static voltage which would be picked up by a receiver of the same band width tuned to 1000 kilocycles. It is also possible to transmit a wide band audio signal which requires corresponding wide side-bands at ultra-high frequencies because for practical reasons it is not desirable to assign carrier frequencies as close together as in the standard broadcast band and so sufficient band width is available for high quality transmission. Therefore ultra-high frequencies inherently offer an improvement in quality whether amplitude or frequency modulation is used.

However, there are three important difficulties with ultra-high frequency transmission for broadcasting purposes. The first difficulty is the effect of the curvature of the earth on ultra-high frequencies. Because there is no reflection from the ionosphere, it is frequently stated that the limit of transmission for these frequencies is the distance from the transmitting antenna to the horizon. The equation for this distance is

$$d = 1.22\sqrt{H} \qquad (18)$$

where d is the distance to the horizon in miles
H is the height of the transmitting antenna in feet.

For a height of 400 feet the distance to the horizon would be only 24.4 miles.

Within this distance the field strength of the signal is given approximately by the equation

$$E = \frac{0.0105\sqrt{W}\,G\,H\,A\,F}{D^2} \qquad (19)$$

where E = field strength in microvolts per meter
W = power in watts
G = gain of the antenna over a half wave dipole
A = receiver antenna height in feet
F = frequency in megacycles
D = distance in miles.

The formula and the statement that the transmission is limited to the horizon are not strictly true because it is found that diffraction and refraction of the waves produces a signal[7] beyond the horizon, but this signal falls off more rapidly than the inverse distance squared term of Eq. (19).

The second difficulty with ultra-high frequency transmission is the noise produced locally by electrical apparatus. The principal sources of this noise are automobile ignition and fever therapy machines. These sources might be eliminated in time by legislation requiring adequate shielding.

A third difficulty with ultra-high frequency transmission is the sharp shadows thrown by buildings, hills, etc. Sharp shadows are produced in the field of any wave motion when the interfering bodies have dimensions large in comparison with the wave length, and so are particularly apparent at ultra-high frequencies (short wave lengths).

Because of these difficulties the transmission of amplitude modulated waves at ultra-high frequencies has not made very much progress for broadcast transmission. However, these difficulties are only apsects of the fundamental problem pointed out at the beginning of the paper, that radio transmission is limited by the ratio of interference to signal. It has been shown that frequency modulation offers an important improvement in the solution of this problem. By its means the signal to noise ratio at any fixed distance may be increased. It seems probable an increase in transmission range may be secured by frequency modulation due to the greater signal to noise ratio and this may make economically feasible the use of ultra-high frequencies for broadcasting purposes.

In field tests, adequate signals for the operation of an F-M receiver at two to three times the horizon distance are regularly reported when the transmitter is a high powered one.

The wide band width which is required for frequency modulation is compensated for by the reduction in interference between two stations on the same channel. In amplitude modulation, interference is caused if an undesired signal is one per cent of the desired signal. It has been shown that in F-M a ratio of 2 to 1 between desired and undesired signals is sufficient. If two stations of equal power and located in neighboring cities should operate on the same channel, there would be only a small territory about half way between the two stations where there would be any interference. Even in this territory, relatively simple directional antennas would be sufficient to pick out one station or the other. Therefore numerous stations could be located throughout the nation on the same carrier frequencies.

I.R. Weir, of the General Electric Company reported before the Federal Communications Commission that in their experiments they used a 150 watt transmitter at Albany and a 50 watt transmitter at Schenectady, a distance of 14.5 miles. They operated both transmitters on the same frequency with both frequency and amplitude modulation. In driving a car equipped with a receiver along a line between the two stations the following results were obtained.

Type of Modulation	Interference free range of Albany station	Transitional distance with interference	Interference free range of Schenectady station
Frequency	10.5	1.0	3.0
Amplitude	3.3	11.7	0.5

A further advantage of frequency modulation in the operation of the transmitter and receiver is that non-linear distortion in frequency modulation is not affected by the non-linearity of the tubes, as their non-linearity is a function of amplitude. In F-M non-linear distortion depends only on circuit design, i.e., on non-linear relations which are a function of frequency. For that reason it is claimed that high quality reproduction is more practicable.

FREQUENCY MODULATION RECEIVERS

In the discussion of F-M receivers and transmitters, a certain familiarity with communication theory must be assumed in order to conserve space. The discussion will deal with present commercial practice.

Frequency modulated receivers differ from amplitude modulation receivers in three important respects. They are:

a. The inclusion of a limiter to remove any amplitude modulation resulting from an interfering signal.

b. The use of a special detector circuit called the discriminator to change the frequency modulation into a variable amplitude signal.

c. The use of a wide band intermediate amplifier.

In addition it is much more important in a frequency modulation receiver to have a high gain intermediate frequency amplifier, because the operation of the limiter is dependent upon a certain minimum signal being applied to its input.

Also, if advantage is to be taken of the high quality transmission which is possible with F-M, a better than average audio system and loud speaker should be included.

Except for these differences, the F-M receiver will follow the practice of amplitude modulated receivers. The use of super-heterodyne receivers is universal.

If a tuned radio frequency amplifier has impressed upon its grid a high frequency voltage exceeding a certain minimum amplitude, the radio frequency current in the tuned plate circuit will be practically independent of the magnitude of the input. This is because after the radio frequency component of voltage in the plate circuit has reached an amplitude equal to the d-c component of the plate voltage no further increase in the radio frequency component can be obtained for the instantaneous value of plate voltage cannot be driven negative. By the use of a high gain tube and low values of plate voltage, saturation at relatively low values of grid excitation may be obtained. Fig. 16 shows the circuit and the limiting action for a typical commercial F-M receiver. The operation of the limiter is the same as that of a Class C amplifier[5] and the use of a bias obtained by a resistance in the grid circuit assists in securing a flat curve.

The use of a limiter is necessary because any amplitude modulation which reaches the detector will also produce amplitude variations in the reproduced signal. The use of the limiter or its equivalent at the receiver is the most important component in an F-M system, for no reduction in interference will occur without its operation. The limiter is possible in frequency modulation because the saturation does not affect the instantaneous frequency of the output.

The elimination of hum due to the use of a-c supplies in receivers has also been a problem for many years. The principal source of this hum is amplitude modulation produced in the receiver by the hum component of the rectified d-c plate supply and by the heating of the cathodes by alternating current. If this occurs in the early stages of the receiver it is amplified along with the signal. While this hum level is reasonably well controlled in modern receivers, the limiter in a frequency modulated receiver provides an additional improvement. This is particularly important because the elimination of other sources of interference makes the reduction of hum to an extremely low level much more desirable.

Because the audio amplifier and loudspeaker must reproduce the amplitude variations of the original signal, the discriminator or detector circuit must change the frequency variations into amplitude variations. The most common discriminator circuit in use is that of Fig. 17. The tuned transformer $L_1 - L_2$ has a split secondary feeding two diode plates with the resistance loads R_1 and R_2. The voltage E_3 will be 90° out of phase with the voltage across L_2 at the resonant frequency of $L_2 - C_2$. The radio frequency across Diode 1 in series with resistance R_1 is represented by the vector $E_a = E_1 + E_3$ while that across Diode 2 in series with resistance R_2 is $E_b = E_2 + E_3$. When the output of the discriminator is at the intermediate frequency the phase relations will be those shown in diagram A. However, when the frequency shifts, the voltage across L_2, $(E_1 - E_2)$, shifts in phase. It can be shown that this voltage will follow the locus indicated by the circles. Hence E_a and E_b change in magnitude, one decreasing and the other increasing. By detector theory[6] the instantaneous value of the voltage across R_1 is proportional to E_a while that across R_2 is proportional to E_b. Therefore the instantaneous voltage E_4 supplied to the audio amplifier changes. Increases in frequency make E_4 positive and decreases make it negative.

The relation between the frequency deviation and the instantaneous value of E_4 is also shown in Fig. 17 for a typical receiver. The circuit constants must be selected so that the straight line portion of the characteristic will accommodate the frequency shift which is used at the transmitter. The linearity of this characteristic affects the non-linear distortion in the reproduced signal. It should be observed that its linearity is not a function of the tube characteristic. This makes it apparent that it is necessary to adopt standards at both receiver and transmitter which will be coordinated.

After a hearing before the Federal Communications Commission where all the problems were thoroughly discussed, the band width of 200 kilocycles per channel was adopted as standard. The allocation of channels adjacent to each other in the range of 43 to 50 megacycles simplifies the problem of receiver design, since no band switching equipment will be required.

Because the intermediate amplifier must amplify a wider band of frequencies than is necessary in amplitude modulation a higher intermediate frequency is used. Present practice of one manufacturer is to use 2.1 megacycles for the intermediate carrier frequency.

A system which uses negative frequency feedback in the receiver[8] has been proposed. It accomplishes similar results to those obtained by the use of the limiter, but as it is not at present available commercially, it will not be discussed here, although additional advantages are claimed for it.

Receivers are being manufactured which receive both amplitude and frequency modulation. In these combined receivers separate intermediate frequency transformers and detectors must be used. This makes the combination receiver somewhat more complicated, but complicated circuits have never been a bar to commercial production. This is exemplified by super-heterodyne receivers retailing for ten dollars which have more extensive circuits than two hundred dollar sets of fifteen years ago.

FREQUENCY MODULATION TRANSMITTERS

Two important and quite different methods of obtaining frequency modulation are in use at the present time. A third method has been announced, but its details have not been disclosed.

The method proposed by Armstrong[3] makes use of the fact that for low values of m_p (m_p less than 0.5) the fundamental difference between amplitude and phase modulation lies in the phase relation between the carrier and side-band vectors. This is illustrated by a comparison of Figs. 9 and 12. Balanced modulators have been extensively used[5] in carrier current systems to obtain the two side-bands of amplitude modulation and eliminate the carrier. In the Armstrong system the carrier output of the oscillator is shifted 90° and then used in a balanced modulator to obtain side-band components corresponding to amplitude modulation with this carrier phase. The output is then added back to a carrier component with the phase of the oscillator to form a phase modulation system. The maximum allowable modulation factor m_p produced with these components is 0.5 since only one pair of side frequencies is available.

A block diagram showing the Armstrong system is shown in Fig. 18. The several components required will be discussed in turn.

This method is fundamentally a phase modulation system, but it may be converted into a frequency modulation system if an audio equalizer is used before the modulation. The required characteristic of this equalizer is that the output voltage must be inversely proportional to the audio frequency of the input voltage. With this equalizer in the audio system the phase modulation will be inversely proportional to the audio frequency of the original signal. It has been shown that this is the requirement for frequency modulation and so the combination produces true frequency modulation.

It has been explained that with this system m_p is limited to a maximum value of 0.5, in the original modulation. On account of the characteristic of F-M which has been discussed, this maximum value of m_p can only be secured at the lowest audio frequency to be transmitted. The value of m_p at higher frequencies will be inversely proportional to the audio frequency. Then if a range from 40 to 15,000 cycles is to be transmitted the maximum value of m_p at 15,000 cycles will be

$$\frac{40}{15{,}000} \times 0.5 = 0.0013.$$

Such a method of obtaining frequency modulation, if used alone, would be entirely impractical since operation would be restricted to such low values of m_p and the advantages of frequency modulation lie in the use of large values.

If a radio frequency amplifier has a large grid bias and relatively large r.f. voltage applied to its grid, the plate current will flow in pulses[5] containing harmonics of the grid exciting voltage. If the plate circuit is tuned to a harmonic of the grid voltage, the voltage across the tuned circuit will have a high value at this harmonic. For efficient operation this harmonic should be a low one, say the second or third. When tuned to a second harmonic the combination is called a frequency doubler.

It has been found that a frequency doubler also doubles the phase or frequency shift if an angular modulated wave is applied to the grid, and therefore doubles the value of m_p and also m_f for such a wave. A series of n doublers would multiply m_p by 2^n.

If the method of obtaining frequency modulation now being discussed is operated with a relatively low carrier frequency and the output is passed through a sequence of doublers or triplers, the value of m_p can be correspondingly increased. It has been shown that if a deviation of 75,000 cycles is desired, the value of m_p at an audio frequency of 15,000 cycles would be $\frac{75{,}000}{15{,}000} = 5$. It has also been shown that for the wide audio range of 40-15,000 cycles the m_p which can be initially produced by the Armstrong method at an audio frequency of 15,000 cycles is 0.0013. Therefore the amount of multiplication of m_p required is approximately $\frac{5}{0.0013} = 3750$. The number of doublers required is then obtained by the solution of the equation

$$2^n = 3750$$

or

$$n = 12 \quad \text{nearly.}$$

If a final carrier frequency of 41 megacycles is to be used and straight multiplication were to be employed, the carrier frequency at which the initial modulation should take place would be determined by the frequency multiplication required. In this case

$$\text{Initial carrier frequency} = \frac{\text{Final carrier frequency}}{\text{Multiplication of doublers}} = \frac{41{,}000{,}000}{2^{12}}$$

$$= 10{,}000 \text{ c.p.s.}$$

It is not possible to modulate a 10,000 cycle carrier by an audio frequency of 15,000 cycles, since the audio frequency must be less than the carrier frequency. Therefore some modification of the system must be made. The modification adopted is to perform the modulation at an initial carrier frequency of the order of 200 kilocycles and pass it through a first group of say six doublers. The carrier frequency has then been multiplied 2^6 times and has reached a value of 12.8 megacycles. If now a final carrier frequency of 41 megacycles is desired, the 12.8 megacycle signal is combined with the output of a second crystal oscillator whose frequency is selected so that a beat note of $\frac{41}{2^6}$ or 0.6406 megacycle is obtained. This does not affect m_p. The 640.6 kilocycle wave is then passed through a second group of six doublers to obtain the final 41 megacycle output. Since the initial frequency modulated wave has now passed through twelve doublers, the multiplication of 2^{12} or 4096 has been obtained.

The output of the second group of doublers is then amplified up to the final power required.

While this circuit seems complicated, it should be remembered that these operations can be performed at low power and with receiving type tubes. The cost is not prohibitive for a transmitter station, since the investment in other equipment would be much greater.

The second method of frequency modulation operates in a distinctly different manner. If the capacity of a condenser could be varied at an audio rate, and if this condenser were included in the tuned circuit of an oscillator, it is apparent that the output would be frequency modulated.

The fundamental characteristic of a reactance is that the current flowing into the two terminals is 90° out of phase with the voltage applied across these terminals. This same effect can be secured with a tube circuit. The method is illustrated in Fig. 19. The resistance R_1 is made small in comparison with the reactance of C_1. Then the a-c voltage between the control grid and cathode is substantially 90° out of phase with the voltage impressed across the terminals a-b. But the plate current which flows in the tube is determined largely by the grid voltage. The choke L_3 provides a large impedance to alternating current so that the a-c component of the plate current will flow in the terminals a-b. The current which flows into the terminals a-b will be 90° out of phase with the voltage across it because of the grid control and the circuit appears like a reactance at these terminals.

The magnitude of the reactance is a function of the amplification constant of the tube. In a variable-mu tube the amplification constant may be controlled by the voltage on the grid. Therefore if an audio voltage is also impressed on the control grid of the tube the effective reactance may be varied at an audio rate.

The terminals a-b are connected in parallel with the tuned circuit L_2 - C_2 of a conventional oscillator whose output frequency is determined by the resonance of that circuit. With this combination a variation in the audio voltage on the grid of the reactance tube will produce a frequency modulated wave in the output circuit of the oscillator L_4 - C_4.

The circuit of Fig. 19 does not have the inherent stability of the system of Fig. 18 because the carrier frequency or frequency for zero modulation is not crystal controlled. Stability equivalent to crystal control is necessary in modern radio operation. In order to secure crystal control a more elaborate system is necessary. This is illustrated by the block diagram of Fig. 20.

The oscillator and reactance circuit are similar to those of Fig. 19. The oscillator usually operates at some sub-multiple of the desired frequency, say one-fourth or one-sixth. A frequency multiplier is used before the operation of the final power amplifier, as is standard practice at ultra-high frequencies. A sample of the output is brought back into a frequency converter where it is mixed with the output of a crystal oscillator for comparison purposes. The resultant beat note is then passed through a frequency discriminator of the same general type as used in receivers and illustrated in Fig. 17. If the final output changes frequency, the beat-frequency passing into the discriminator will change. This change is used to produce a d-c voltage in the discriminator which in turn is applied to the reactance tube to provide a correction on the frequency drift which has occurred.

This method of increasing stability has a marked similarity to the use of inverse feedback in audio amplifiers to increase their stability.

A delay or filter must be introduced in the frequency feedback circuit so that it is unresponsive to the variations in frequency produced by the audio modulation, but will make corrections for the long period drifts associated with oscillators which are not crystal controlled.

The two methods of frequency modulation transmitters each have their proponents. The Armstrong method is claimed to be more stable because the carrier frequency is directly controlled by the two crystals. The reactance tube method is simpler and would seem particularly applicable to low powers and portable equipment. It is too soon in terms of practical operation to be sure which will find the most general application.

The operation of the final amplifiers in both systems is essentially simpler than is the case in amplitude modulation, since attention need not be paid to the linearity of input-output amplitude curves.

In amplitude modulation the final high power stage presents certain difficulties in operation. A transmitter normally requires[5] either a high power audio amplifier with an output about 75 per cent of the rated carrier power to perform the modulation, or else the final stage must be operated at half its maximum efficiency. This is because in amplitude modulation the power output must be increased during modulation by an amount equal to the side-band power (50 per cent of the carrier power for $m = 1.0$). The increase in power must be supplied by either an audio amplifier or by an increase in the efficiency of the output stage during the audio cycle. An increase in efficiency can only be obtained if the efficiency in the absence of modulation is limited to half its maximum possible value. Both of these expedients increase the cost of high power stations materially. Since the amplitude and power output of an F-M transmitter is constant during modulation, the final stage can operate at its maximum efficiency at all times even though the original modulation is performed at a low power level where only receiving type tubes are required.

The question is frequently raised whether the wide band which must be transmitted for F-M does not introduce difficulties in the tuned circuits. However, it should be remembered that the selectivity of any tuned circuits is given in terms of frequency ratio rather than absolute band width. A 200 kilocycle band width at 40 megacycles is only one-half of one per cent of the carrier frequency, while the 10 kilocycle band width used in Standard Broadcasting at 1000 kilocycles is one per cent of the carrier frequency. Therefore no difficulties are introduced in the transmitter by the required absolute width of the band when the carrier frequency is high.

APPLICATION OF FREQUENCY MODULATION TO SERVICES OTHER THAN BROADCASTING

Frequency Modulation seems to have inherent advantages for other services than broadcasting. Important among these are its application to police and airway communication. The sharp limitation of the range obtained with F-M is an advantage and it would appear that many more police transmitters covering specific areas could be used without interference. In airway service some of the most important contacts are needed during severe electrical storms and F-M could make an important contribution in this field.

F-M would also appear to have advantages for military purposes where limited ranges are desired and interference is a particularly severe problem.

In police, airway and military communication, high quality reproduction is not necessary and so a more limited band would serve the purpose.

F-M has also been proposed for longer distance communication by the use of relay or repeater stations apaced at intervals determined by the range of the transmitters. By the use of directional antennas this range can be increased. Previous proposals to use relay stations for broadcast station interconnection have not met with favor because of the limited range and additional interference which they would cause, but it appears that a large part of this objection is eliminated when F-M is used.

It is also possible to multiplex other services on F-M transmission, such as facsimile, transmitting this service along with a sound signal, but with a reduction in the signal to noise ratio. Another suggestion has been to transmit two sound signals to obtain binaural reproduction. The possibilities of these uses have been only partially explored.

CONCLUSION

It would appear that frequency modulation is capable of producing a marked change in broadcasting within the next few years. It is very doubtful if it will eliminate the use of the standard broadcast frequencies, as they are still capable of covering larger distances when (and only when) a transmitter is given exclusive use of a channel. It seems probable that many local stations designed to cover a limited area will be transferred to F-M operation, and the number of cleared channels increased so as to take better advantage of the limited band of standard broadcast frequencies available. The development in these lines must eventually respond to the laws of economics and engineering.

BIBLIOGRAPHY

1. Notes on Theory of Modulation. J.R. Carson, Proc. I.R.E., Vol. 10, February 1922, p. 57.

2. Amplitude Phase and Frequency Modulation. H. Roder, Proc. I.R.E., Col. 19, December 1931, p. 2145.

3. A Method of Reducing Disturbances in Radio Signalling by a System of Frequency Modulation. E.H. Armstrong, Proc. I.R.E., Vol. 24, May 1936, p. 689.

4. Variable Frequency Electrical Circuit Theory, with Application to the Theory of Frequency Modulation. J.R. Carson and T.C. Fry, Bell System Technical Journal, Vol. 16, October 1937, p. 516.

5. Communication Engineering. W.L. Everitt, McGraw-Hill Book Company, 1937.

6. Ultra-High-Frequency Propagation Formulas. H.O. Peterson, R.C.A. Review, Vol. 4, October 1939, p. 163.

7. Ground Wave Field Intensity Over a Finitely Conducting Spherical Earth. A Theory of Tropospheric Wave Propagation. Summary of Statements on Ultra-High-Frequency Wave Propagation. K.A. Norton, Reports before the Federal Communications Commission, April 1940.

8. The Application of Negative Feedback to Frequency Modulation Systems. J.G. Chaffee, Bell System Technical Journal, Vol. 18, July 1939, p. 404.

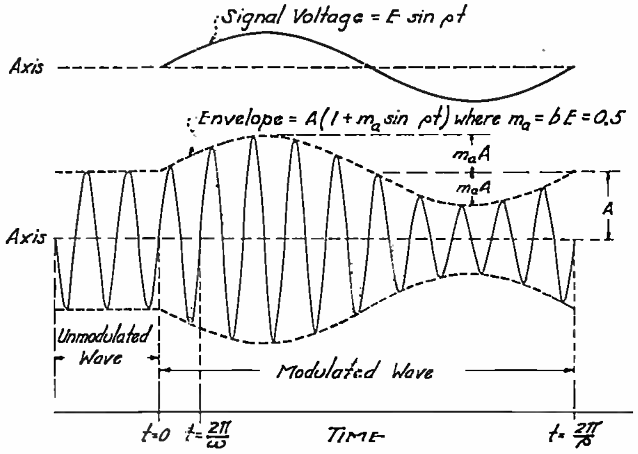
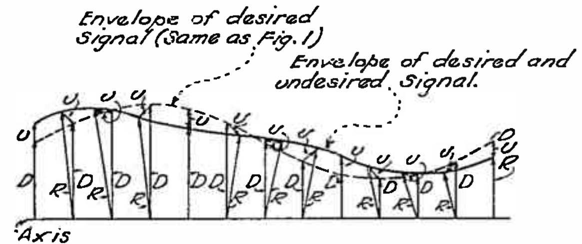
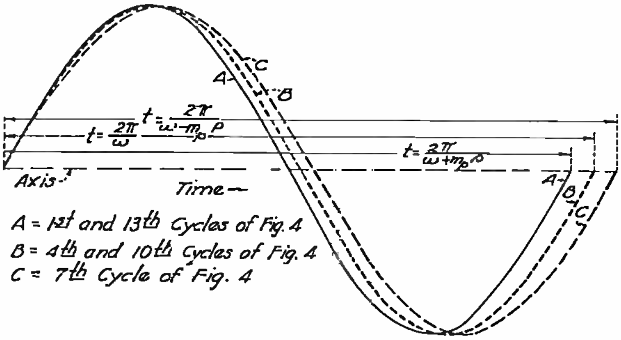


Fig. 1 An Amplitude-modulated Wave.



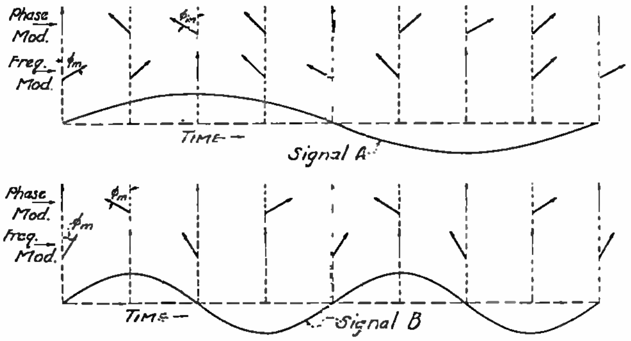
D= Vector of Desired Signal
U= Vector of Undesired or Interfering Signal
R= Resultant Vector (D+U) of sum of desired and undesired Signals.

Fig. 3 Interference with an amplitude modulated wave of a oarrier of slightly different frequency.



A = 1st and 13th Cycles of Fig. 4
B = 4th and 10th Cycles of Fig. 4
C = 7th Cycle of Fig. 4

Fig. 5 Expansion of individual cyoles in Fig. 4.



Note: ϕ_m is same for both signals for Phase Modulation. ϕ_m is inversely proportional to signal frequency for Frequency Modulation Maximum velocities of vectors are same for both signals for Frequency Modulation.

Fig. 7 Comparison of phase and frequency modulation by means of vector diagrams.

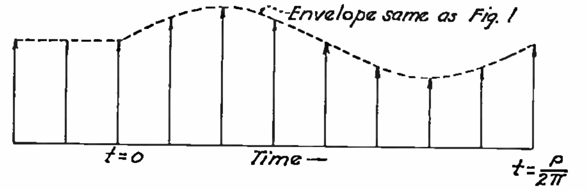
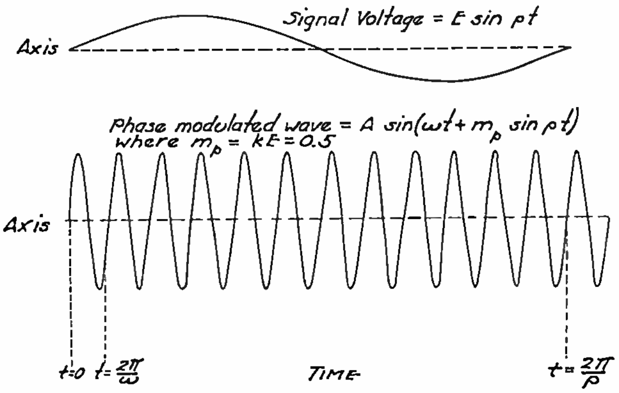
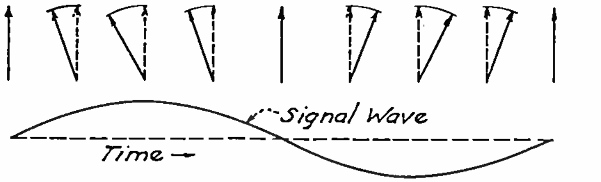


Fig. 2 Vector diagrams of the amplitude modulated wave of Fig. 1 for suocessive instants.



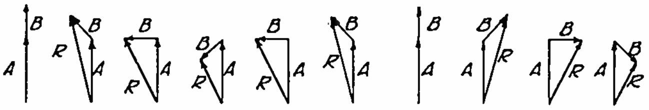
Note:- 1st, 4th, 7th, 10th and 13th cycles shown in detail in Fig. 5

Fig. 4 An angular-modulated wave.



Solid line vectors are vectors of modulated wave
Dotted line vectors are vectors of unmodulated wave

Fig. 6 Vector diagrams of the angular modulated wave of Fig. 4 for successive instants.



A is vector of desired signal-Unmodulated
B is vector of interfering signal (Same carrier)
R is vector of total wave (A+B)

Fig. 8 Vector diagrams showing interference in angular modulation.

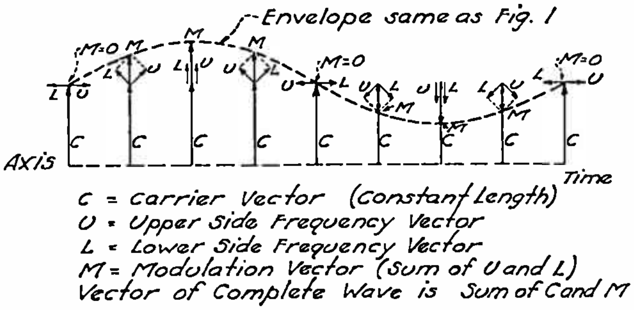


Fig. 9 Vector diagram of carrier and side frequencies in an amplitude modulated wave.

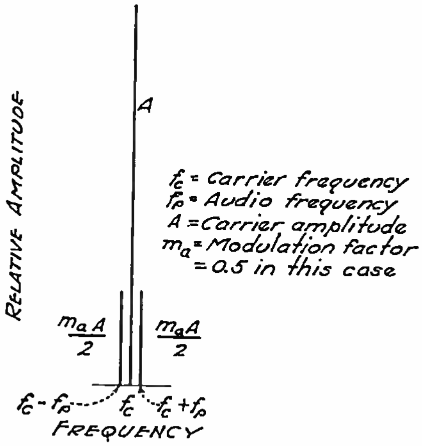


Fig. 10 Spectrum analysis of an amplitude modulated wave.

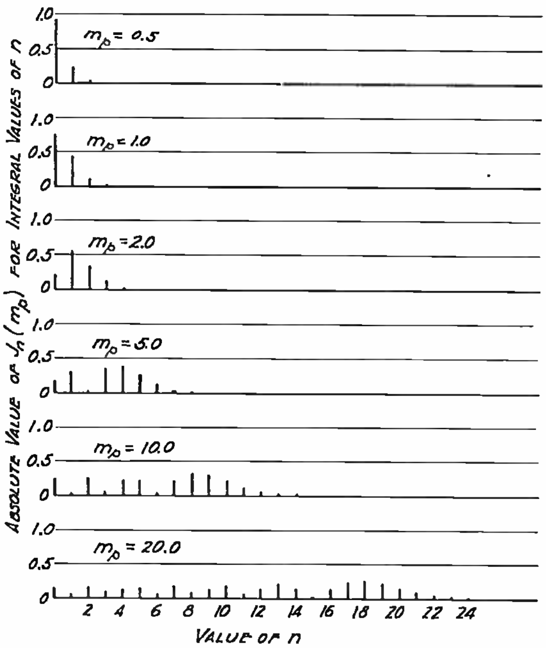


Fig. 11 Values of the Bessel Function of the first kind for integral orders.

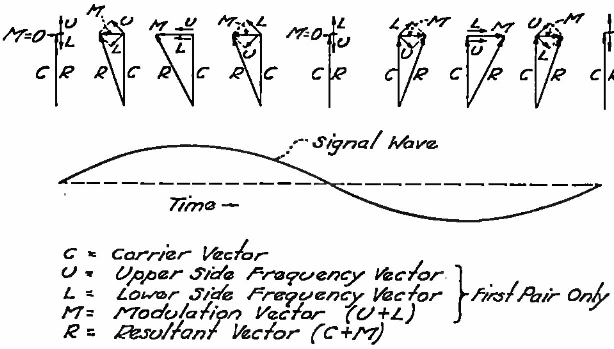


Fig. 12 Vector diagram of carrier and side frequencies in an angular modulated wave for low values of modulation.

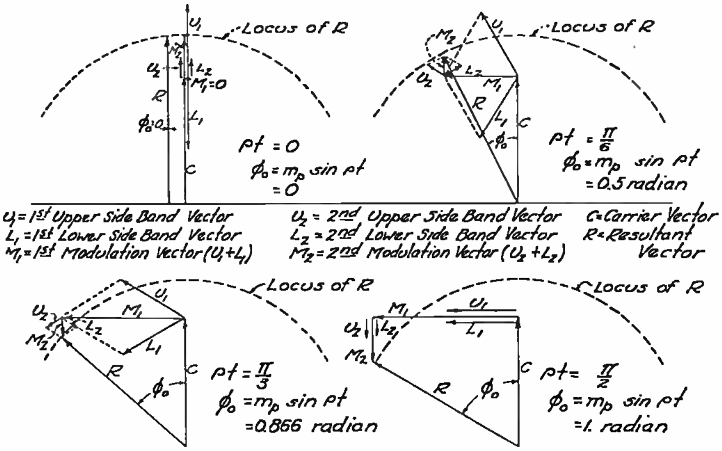


Fig. 13 Vector diagram of carrier and two pairs of side frequencies in an angular modulated wave where $m_p = 1$.

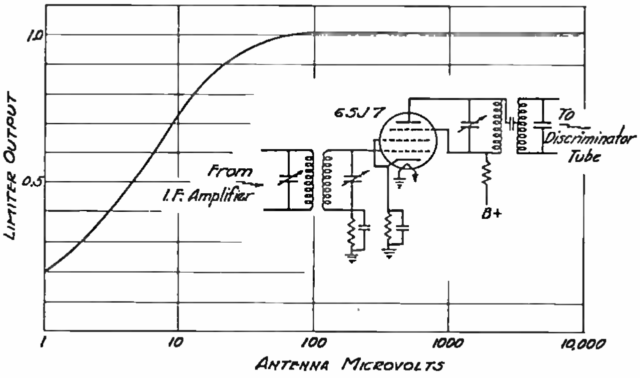


Fig. 16 Characteristic and circuit of a limiter.

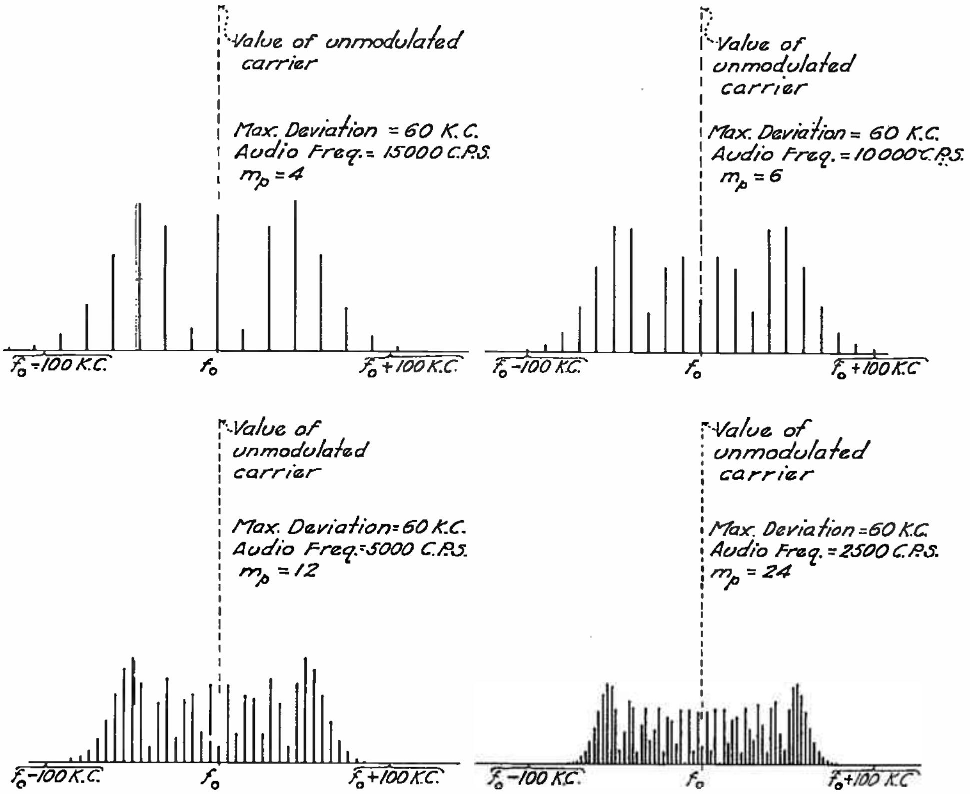


Fig. 14 Spectrum analysis of frequency modulation for a constant deviation and variable modulating frequencies.

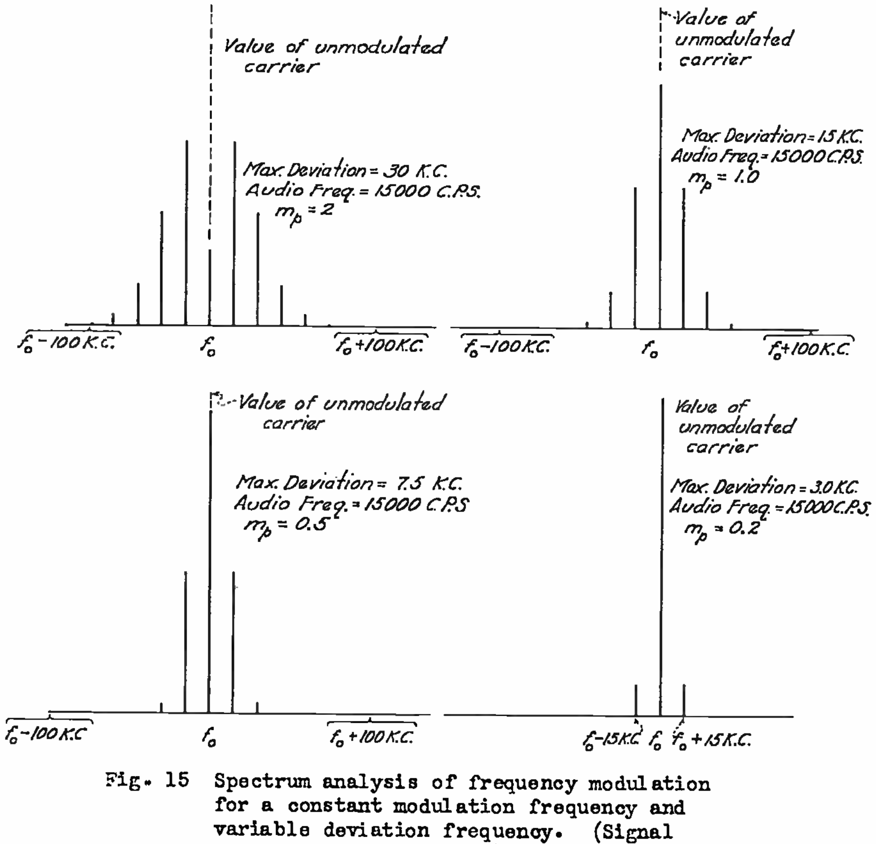


Fig. 15 Spectrum analysis of frequency modulation for a constant modulation frequency and variable deviation frequency. (Signal is for phase modulation)

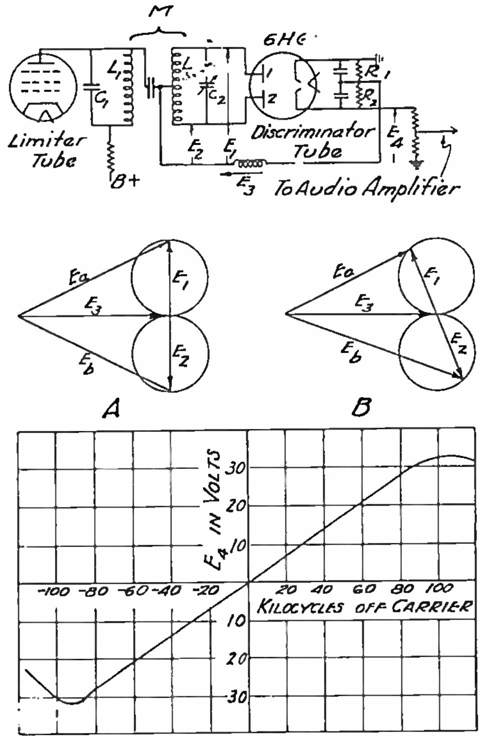


Fig. 17 Characteristics and circuit of a frequency discriminator.

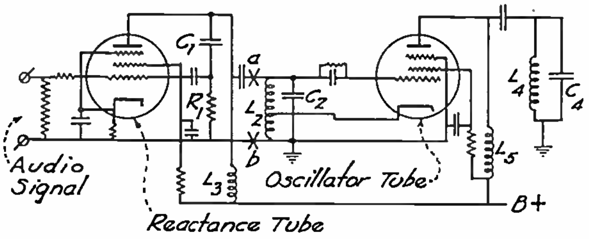


Fig. 19 Circuit of an oscillator which is frequency-modulated by a reactance tube.

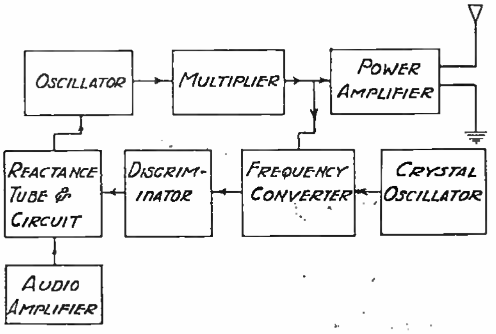


Fig. 20 Block diagram of a reactance tube controlled frequency-modulated transmitter with crystal stabilization.

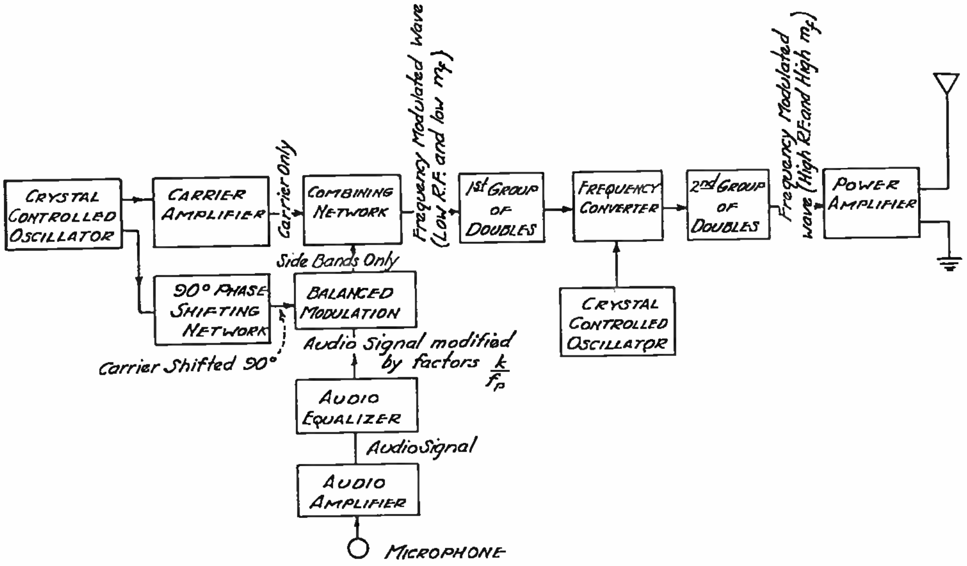


Fig. 18 Block diagram of Armstrong frequency modulated transmitter.