NOTES ON THE THEORY OF MODULATION*

By

John R. Carson

(Department of Development and Research, American Telephone and Telegraph Company, New York)

It is a well-known fact that in carrier wave[1] transmission it is necessary to provide for the efficient transmission and reception not only of the carrier frequency itself but also for a band of frequencies of width depending on the frequency and character of the signal itself. This necessity is becoming more and more a serious consideration as the severity of wave length regulation and the necessity of sharp selective tuning are increased. In view of these facts a great deal of inventive thought has been devoted to the problem of narrowing the band of transmission frequencies. Some of the schemes which are directed to this end are very ingenious; all, however, are believed to involve a fundamental fallacy. It is the purpose of this note to discuss briefly the general problem of modulation and to analyze the more ingenious and plausible schemes which have been advanced to solve this problem.

A pure modulated wave may be mathematically defined by the expression

$$f(t)\,A\cos\omega \tag{1}$$

Here $f(t)$ is the low frequency signal, $\omega/2\pi$ is the carrier frequency and A is an amplitude factor which fixes the magnitude of the transmitted wave.

The pure modulated wave is often defined in words as "a carrier wave of constant frequency whose amplitude is proportional to the signal wave." Properly interpreted this definition is correct; however, the inference which is sometimes made, that the *resultant wave* is of constant frequency is erroneous, as may easily be shown.

*Received by the Editor, January 6, 1922.

[1] It is to be understood that the term carrier wave is employed in its generic sense to cover radio as well as carrier wave wire transmission in both of which the carrier wave plays the same role.

[2] See "Carrier Wave Telephony and Telegraphy," by Colpitts and Blackwell, "Journal of the American Institute of Electrical Engineers," April, 1921

Let the signal wave $f(t)$ be represented, as we assume in telephone theory, by a plurality of sinusoidal terms, thus

$$f(t) = \sum_{1}^{m} a_j \cos(p_j t + \theta_j) \tag{2}$$

Substitution in (1) gives for the pure modulated wave:

$$\begin{aligned} &\frac{1}{2} A \sum_{1}^{m} a_j \cos[(\omega - p_j) t - \theta_j] \\ &+ \frac{1}{2} A \sum_{1}^{m} a_j \cos[(\omega + p_j) t + \theta_j] \end{aligned} \tag{3}$$

It follows at once that the frequencies transmitted lie between $(\omega + p_m)/2\pi$ and $(\omega - p_m)/2\pi$; that is the width of the bandis $2\,p_m/2\pi$. For example let us take a carrier wave of 100,000 cycles per second, and modulate this wave with telephone signals which we shall assume contain frequencies up to 2,500 cycles per second. The pure modulated wave then contains frequencies lying between 102,500 and 97,500 cycles per second and a band of 5,000 cycles must be transmitted.

It has, however, been known for several years[2] that it is not necessary to transmit the pure modulated wave which contains a band of frequencies of twice the signal wave range of frequencies and that theoretically perfect transmission can be had by transmitting only one "side band" and suppressing the other. This may be explained as follows: Referring to the expression (3) for the pure modulated wave, suppose that all frequencies below[3] that of the carrier $\omega/2\pi$ are filtered out so that the transmitted wave is

$$\frac{1}{2} A \sum a_j \cos(\omega + p_j)\, t + \theta_j \tag{4}$$

Let the receiving stations be provided with a local generator of frequency $\omega/2\pi$ which is combined in the demodulator with the received wave. If the locally generated wave is represented by $B \cos \omega t$, the demodulated wave is

$$\frac{1}{2} A\, B \cos \omega t \sum a_j \cos(\omega + p_j)\, t + \theta_j$$

which is equivalent to

[3] A precisely similar argument holds if all the frequencies *above* the carrier frequency $\omega 2\pi$ are suppressed.

$$\frac{1}{4}AB\sum a_j \cos(p_j t+\theta_j)$$
$$+\frac{1}{4}AB\sum a_j \cos(2\omega+p_j)\,t+\theta_j$$

The first expression is simply the signal wave $f\ (t)$ multiplied by the factor $\frac{1}{4}AB$ while the second expression is of double radio frequency which is entirely suppressed in the audio frequency circuits.

In the system of modulation discussed above it will be observed that the *amplitude* of the carrier wave is varied by and in accordance with the low frequencies signal wave and that this process inherently requires the transmission of a band of frequencies at least equal to the range of essential frequencies in the signal itself.[4] In order to eliminate this necessity which is inherent in all actual systems of modulation it has been proposed a number of times to employ an apparently radically different system of modulation which may be termed *frequency* modulation as distinguished from *amplitude* modulation, in the belief that the former system makes possible the transmission of signals by a narrower range of transmitted frequencies. This belief is erroneous; the suggestion is, however, quite ingenious, and the reasoning on which the supposed advantage is based is very plausible, and indeed requires some mathematical analysis before its incorrectness can be satisfactorily established. The system of *frequency* modulation will now be explained and analyzed in terms of the specific physical system in which the idea was first called to the attention of the writer.

Suppose that we have an ideal non-dissipative oscillation circuit of inductance L and capacity C. Such a circuit is of course ideal and unrealizable, but the analysis of the actual vacuum tube oscillator, as regards frequency modulation, may be safely based on a consideration of the ideal circuit. This circuit when once energized, will continue to oscillate at frequency $\omega/2\pi$ when $\omega=1/\sqrt{LC}$. Now suppose that the capacity (or inductance) is varied in accordance with an audio frequency signal. For the present we shall simplify the discussion by assuming that the signal is a pure tone of frequency $p/2\,\pi$, and that the instantaneous value of the capacity is $C_o(1-2\,h\ sin\ pt)$ where h is an amplitude factor proportional to the signal intensity. We

[4] It should be noted that after either "side band" is suppressed, the resulting wave does not fall within the definition of a pure modulated wave of equation (1).

shall later consider the general case where the signal is represented by $f(t)$. Assuming that h is small compared with unity we can write

$$\omega = \frac{1}{\sqrt{LC}} = \frac{1}{\sqrt{LC_o}}(1+h\,sin\,pt)$$
$$=\omega_o(1+h\,sin\,pt)$$

From the foregoing reasoning it has frequently been concluded that the oscillation circuit generates a continuously varying frequency of instantaneous value

$$\frac{\omega_o}{2\pi}(1+h\,sin\,pt)$$

so that the generated frequency varies between the limits $\omega_o(1-h)/2\pi$ and $\omega_o(1+h)/2\pi$. According to this theory, if $2h\omega_o$ is made less than p the range of frequencies transmitted $2h\omega_o/2\pi$ will be smaller than $p/2\pi$, which is the minimum range required in *amplitude* modulation.

The foregoing gives, very briefly, the essential reasoning underlying the idea of *frequency* modulations. We shall now analyze the scheme more closely: The differential equation of the circuit may be written as

$$L\frac{d^2I}{dt^2}+\frac{1}{C}I=0$$

$$\frac{d^2I}{dt^2}+\frac{1}{LC_o}(1+2h\,sin\,pt)I=0 \qquad (5)$$

$$\frac{d^2I}{dt^2}+\omega^2I=0$$

Now if ω is treated as a constant, a particular solution is

$$I=A\,cos\,\omega t$$

If we now substitute for ω the expression

$$\frac{1}{\sqrt{LC_o}}(1+h\,sin\,pt)=\omega_o(1+h\,sin\,pt)$$

we get

$$I=A\,cos\left[\omega_o(1+h\,sin\,pt)\,t\right] \qquad (6)$$

which is interpreted as representing a wave of instantaneous frequency

$$\frac{\omega_o}{2\pi}(1+h\,sin\,pt)$$

Both the solution of the equation and the interpretation of this solution are incorrect. Equation (5) by a simple transformation of variables is reducible to the form

$$\frac{d^2 I}{d x^2}+(a+16\,q\,cos\,x)I=0$$

which is the canonical form of Mathieu's Equation (see Whittaker and Watson, "Modern Analysis," page 402), and it is easily shown from the theory of this equation that the solution of (5) is *the real part of*

$$e^{i\omega_n t}\sum_{-\infty}^{+\infty} a_n\, e^{inpt} \tag{7}$$

Consequently the solution is a series of the form

$$\sum_{-\infty}^{+\infty} b_n\, cos\left[(\omega_o+n\,p)\,t+\theta_n\right] \tag{8}$$

The frequencies present in the wave form an infinite series spaced at the interval $p/2\pi$ of the signal frequency. They may be tabulated as

$$\begin{array}{ccc} & \omega_o & \\ \omega_o+p & & \omega_o-p \\ \omega_o+2\,p & & \omega_o-2\,p \\ \omega_o+3\,p & & \omega_o-3\,p \\ \cdots\cdots & & \cdots\cdots \\ \omega_o+n\,p & & \omega_o-n\,p \end{array}$$

It follows at once that *the transmission of the signal by frequency modulation requires the transmission of a band of frequencies* at least $2p/2\,\pi$ *in width*; that is *a band of width equal to twice that of the signal itself.*

If the solution (7) is substituted in the differential equation (5), we get the following system of difference equations for the determination of the constants.

$$-i\,(2\,np\,\omega_o+n^2\,p^2)\,a_n+h\,\omega_o^2(a_{n-1}-a_{n+1})=0 \tag{9}$$

In the practically important case where p is so small compared with ω_o that np may be neglected in comparison with ω_o, this is satisfied by

$$A_n=(i)^{-n}\,J_n\,(h\,\omega_o/p) \tag{10}$$

where $J_n\,(h\,\omega_o/p)$ is the Bessel function of order n and argument $h\,\omega_o/p$. In this case the series sums up to

$$I=A\,cos\left(\omega_o\,t-\frac{h\,\omega_o}{p}\,cos\,pt\right) \tag{11}$$

The ratio of the term of frequency $({}_o\omega+p)$ to the fundamental of frequency ω_o is $J_1(h\,\omega_o/p)/J_o(h\,\omega_o/p)$, which in case $h\,\omega_o/p$ is less than unity is approximately equal to $h\,\omega/2\,p$. This system of modulation, therefore, discriminates against high frequencies and therefore inherently introduces distortion.

In analyzing this system of frequency modulation consideration has been limited to the case of a signal consisting of a pure tone of frequency $p/2\,\pi$. In the more general case where the signal must be represented by an arbitrary $f\,(t)$, as is the case in telephonic transmission, a general solution can only be gotten when $f\,(t)$ is periodic and analyzable into a Fourier series. In this case the differential equation of the problem is reducible to Hill's equation (Whittaker and Watson, "Modern Analysis," page 406), and the theory of this equation shows that the frequencies present in the wave are exactly the same as those given above if $p/2\,\pi$ is the fundamental frequency of $f\,(t)$.

However if we introduce the approximations indicated by physical considerations, a much simpler and more instructive approximate solution is obtainable without analyzing $f\,(t)$. Let the instantaneous capacity be represented by $C_o(1-2h\,f\,(t))$; then assuming $2h\,f\,(t)$ small compared with unity the differential equation of the problem is

$$\frac{d^2 I}{dt^2}+\omega_o^2\,(1+2\,h f\,(t))\,I=0$$

Assuming that the solution is the real part of $e^{i\omega_o t}\phi\,(t)$ and substituting in the differential equation we get

$$i\,2\,\omega_o\phi'\,(t)+\phi\,(t)+2\,h\,\omega_o^2 f\,(t)\,\phi\,(t)=0.$$

Now if $f\,(t)$ is a relatively slowly varying function compared with the carrier wave, the term $\phi''\,(t)$ may be neglected and we get

$$\phi\,(t)=A\,e^{i\,\omega_o h\int f(t)\,dt}$$

whence

$$I=A\,cos\left[\omega_o\,(t+h\textstyle\int f\,(t)\,dt)\right] \qquad (12)$$

If $\omega_o\,h\int f\,(t)\,dt$ is small compared with unity, as it would be in practice, this gives approximately

$$I=A\,cos\,\omega_o\,t-\omega_o\,h A\textstyle\int f\,(t)\,dt\,sin\,\omega_o\,t$$

The second term is a modulated wave, but the amplitude instead of being proportional to the signal wave is proportional to its integral. Consequently this type of modulation inherently distorts without any compensating advantages whatsoever.

The foregoing solutions, tho unquestionably mathematically correct, are somewhat difficult to reconcile with our physical intuitions, and our physical concepts of such "variable frequency" mechanisms as, for example, the siren. Upon closer analysis it is seen, however that the difficulty arises in connection with what we mean by frequency, and can be cleared up satisfactorily, it is believed, by the following generalized concept and definition of frequency.

Suppose we have a function $sin\,(\Omega\,(t))$ where $\Omega\,(t)$ is any specified function of time: Its derivative with respect to time is $\Omega'\,(t)\,cos\,\Omega(t)$ where $\Omega'\,(t)=d/dt\Omega(t)$. We define the *generalized trequency* of such a function as equal to $\frac{1}{2\pi}\,\Omega'\,(t)$. This definition, while formally arbitrary, has considerable physical significance, and is believed to be a useful concept. In the case where $\Omega(t)=\omega t$ it agrees with the usual definition of frequency $\omega/2\pi$. Furthermore, if we apply this definition to formulas (11) and (12) the *generalized frequencies* are respectively $\omega_o\,(1+h\,sin\,pt)$ and $\omega_o\,(1+hf\,(t))$, which agree with our physical intuitions. It agrees also with the fact that in the case of the siren the mathematical analysis of which differs in no essential way from that of the "variable frequency" oscillator just discussed, the *generalized frequency,* as defined above, corresponds with the "instantaneous frequency" which the ear apperceives. This may be shown as follows:

In the neighborhood of time $t=\tau$, $\Omega\,(t)$ may be expanded as

$$\Omega\,(t)=\Omega\,(\tau)+\frac{t-\tau}{1\,!}\,\Omega'\,(\tau)+\frac{(t-\tau)^2}{2\,!}\,\Omega''\,(\tau)+\;\cdot\;\cdot\;\cdot$$

Now the function $sin\,[\Omega\,(t)]$ alternates when the function $\Omega\,(t)$ changes by the amount π; or otherwise stated, the interval between zeros corresponds to the time intervals during which $\Omega\,(t)$ changes by the amount π. From the foregoing expansion this interval is approximately $\pi/\Omega'\,(t)$ in the neighborhood of the time $t=\tau$. That is to say, *the rate of alternation of the function* $sin\,[\Omega\,(t)]$ is approximately *the same at any time t, as that of the function* $sin\,\omega t$, *where* $\omega=\Omega'\,(t)$. In this sense and this sense only do "variable frequency mechanisms" generate a continously varying frequency over the range of frequencies corresponding to the extreme values of $\Omega'\,(t)$.

Exactly the same conclusions are reached by the analysis of another theoretically possible scheme of frequency modulation which suggests itself. This is to vary the speed of a radio frequency alternator in accordance with the signal so that its instantaneous angular velocity is representable by

$$\omega\,(1+hf\,(t))$$

A superficial consideration of this scheme would lead to the erroneous conclusion that the frequency generated is $\frac{\omega}{2\pi}(1+hf\,(t))$ and for a sinusoidal signal varies between $\frac{\omega}{2\pi}(1-h)$ and $\frac{\omega}{2\pi}(1+h)$. A mathematical analysis shows, however, that it differs in no essential way from the arrangement analyzed above and that the frequency band which must be transmitted is at least equal to that required in *amplitude* modulation.

The foregoing discussion is immediately applicable to the analysis of the systêm of continuous wave radio telegraphy which employs the so-called "spacing wave." In its essentials this system merely employs "frequency modulation" instead of "amplitude modulation" in the sense employed above and formula (12) is directly applicable. It follows therefore at best, as regards the necessary range of frequencies, the "spacing wave" system is inferior to that in which the dot and dash correspond to modulation of amplitude of a constant frequency carrier. Superiority, however, has been claimed for the former on the alleged ground that, since "the amplitude is constant," transient disturbances are minimized. This claim is seen to be quite invalid when the real significance of "frequency modulation" is analyzed, and no such superiority exists.

SUMMARY: The transmission system of "frequency modulation" (transmission by variation of the frequency of the radiated wave) is mathematically analyzed, and the width of the band of frequencies occupied by this method of transmission at a given speed is compared with the width of the corresponding band for transmission by amplitude variation. It is proved that the frequency modulation system using a spacing or compensating wave is inferior to the amplitude variation system both as to the width of the frequency band occupied and as to distortion of signal wave form.