# Steady State Delay as Related to Aperiodic Signals

By R. V. L. HARTLEY

The concepts of phase and envelope delay, as applied to any linear system, rather than only to a medium, are discussed. Criteria are set up for the time of occurrence of that part of an aperiodic signal which corresponds to a small segment of the spectrum. The original spectrum of the signal gives the time of entry and this spectrum as modified by the phase characteristic of the system gives the time of exit.

If the amplitude is constant over the segment, it is shown that when the criterion is the time of maximum envelope of the disturbance, the aperiodic delay is identical with the envelope delay. When it is the time of maximum absolute value, the delay depends on the signal spectrum, the phase shift of the system, and the envelope delay, but not on the phase delay.

If the amplitude vaies rapidly with frequency, the component of an aperiodic disturbance which corresponds to a narrow segment of the spectrum persists so long that the resulting over-lapping of neighboring segments makes their interpretation difficult.

IN THE earlier applications of steady state theory to transmission problems the emphasis was placed on the variation of amplitude with frequency. The use of long loaded lines made it necessary to take account of phase distortion[1] as well. With the development of telephotography and television[2], the phase characteristic was found to provide a useful index for predicting the overlapping of adjacent picture elements. For these purposes it has been found convenient to express the phase characteristic in terms of phase or envelope delay. These may be called "steady state delays" since they are defined and measured in terms of sinusoidal disturbances of adjustable frequency. However, the signals for which they are intended to furnish an index are aperiodic in nature. It seemed worthwhile, therefore, to examine more closely the relations existing between "aperiodic delays," defined in terms of such signals, and steady state delays.

Let us first review the development of the concepts of steady state delay. Early in the study of the propagation of sinusoidal waves a distinction was made between phase and group velocity. If we fix on a particular distance of transmission the ratio of this distance to each of these two velocities may be interpreted as a delay associated with the transmission. In the

[1] For discussion and references see "Phase Distortion and Phase Distortion Correction," S. P. Mead, *B. S. T. J.*, Vol. VII, p. 195, 1928.

[2] Symposium on Television, *B. S. T. J.*, Vol. VI, p. 551.

communication art, these delays have been called phase and envelope delay, respectively. If the medium exhibits dispersion they vary with frequency. Let us fix our attention on the conditions throughout the medium at a particular instant during the transmission of a sinusoidal disturbance. We may determine the total change of phase in passing from the input to the output. This may be more than a single cycle. If now we divide this phase shift by the frequency, expressed in the same angular units, we get the time which will be required for the phase at the input to progress to the output, or the phase delay. Also it may readily be shown that the derivative of this phase shift with respect to frequency is equal to the envelope delay as defined above in terms of the group velocity. The simplest treatment of this is based on the consideration of two sinusoidal waves of equal amplitude and slightly different frequencies.

While these delays can be easily interpreted for most media, difficulties arise in the case of those substances which exhibit anomalous dispersion. Here, in the neighborhood of certain frequencies, the phase shift varies rapidly with frequency, and often appears to be discontinuous. Actually the apparent discontinuity is a region of very rapid decrease of phase with frequency, which leads to a negative value of envelope delay. In the same region the transmission varies rapidly with frequency, and selective reflection occurs at the entering boundary. This effect can be explained in terms of resonance in the elements which make up the fine structure of the medium.

The next step was to dissociate the idea of delay from that of velocity in a medium, and associate it with a steady state transfer characteristic between any two points of a linear system. This would permit its application to all sorts of complicated networks in which uniform propagation cannot be readily visualized. Here two types of characteristic are to be distinguished. One, which is associated with what might be called "damped" systems, exhibits a reasonably gradual variation of both phase shift and attenuation with frequency. This is the analog of a medium having normal dispersion. The other, which is associated with "resonant" systems, exhibits the phenomena associated with anomalous dispersion. In the case of filters and hollow wave guides these resonances give rise to regions of high attenuation and reactive impedance, which are the analogs of the regions of high absorption and selective reflection at the boundary of a medium. In applying the idea of delay to networks then, we can expect the results to agree with our intuitions only so long as we keep away from the critical frequencies of resonant systems.

In computing or measuring the phase shift of a system, at a single frequency, the result is indeterminate so far as the addition of multiples of $2\pi$ is concerned. This does not affect the envelope delay, which depends only on the derivative, and so this type of delay can be generalized directly to include the transfer characteristics of arbitrary networks. To give an exact meaning to phase delay some convention would have to be adopted for determining what, if any, multiple of $2\pi$ is to be added to the computed phase for the frequency in question. Apparently no such convention has been agreed upon which is of general application. For damped networks which transmit frequencies down to zero, it is customary to assume the phase shift to be zero at zero frequency, and, for higher frequencies, to add multiples of $2\pi$ so that the phase shift varies continuously with frequency. If, then, $B$ is the computed phase shift, between $-\pi$ and $\pi$, we may represent the continuously varying phase shift by $B + 2m\pi$, where $m$ is the number of discontinuities in $B$ which have been eliminated in passing from zero to the frequency in question. The phase delay may then be defined as

$$D_p = \frac{B + 2m\pi}{\omega}. \tag{1}$$

Any similar convention for resonant systems would be less simple, and since, as will appear below, phase delay has little bearing on aperiodic signals, it seems unwise to attempt to formulate such a convention here.

In contrast with steady state delay, let us now examine the delay of an aperiodic signal. If the signal is transmitted without distortion the concept of delay of the signal as a whole is simple. If, because of distortion, the sent and received signals are different we may still agree upon some recognizable feature of each as determining its time of occurrence. If the distortion is considerable the delay may vary greatly with the distinguishing characteristic chosen. For example, if it depends on the behavior of components of high frequency the delay may be quite different from what it is if it depends on those of low. In the first case the result would be little affected if, before transmission, the signal were sent through a high-pass filter and, in the second, if it went through a low-pass filter. In each case we measure a delay associated with a disturbance which comprises only those Fourier components of the signal which occupy a particular limited range of frequency. We may carry this idea farther and make use of a very narrow band-pass filter. By varying the mid-frequency of this band we obtain a delay which is a function of frequency. Its value, at any frequency, is the delay, as defined by our convention, of a disturbance which corresponds to that part of the spectrum of the signal which is in the immediate neighborhood of the frequency in question. Our problem then is to find recognizable features of a disturbance of this kind such that, when they are used as criteria of delay, the result can be related directly to the phase or envelope delay as defined in terms of periodic disturbances.

Compared with the pair of equal sinusoids used in the derivation of envelope delay, this disturbance differs in that, in any finite range of frequency, there are an infinity of sinusoids, the amplitudes of which need not all be the same. For simplicity, we assume the actual filter to be replaced by an idealized one in which there is no distortion within the band and no transmission outside it. If the signal as a whole be represented by a Fourier integral, we may obtain the desired disturbance, for an angular frequency, $\omega_1$, by integrating from $\omega_1 - \delta$ to $\omega_1 + \delta$. The disturbance may be represented by

$$f(t) = \text{real part of } M \int_{\omega_1-\delta}^{\omega_1+\delta} \exp\left[-\alpha + i(\omega t - \theta)\right] d\omega, \tag{2}$$

where $M$ is a constant dependent on the magnitude of the signal and $\alpha$ and $\theta$ are functions of frequency and position which describe the spectrum of the signal at various points in the system.

The first step is to perform the indicated integration and express the resulting function of time in a convenient form. For this we let

$$\epsilon = \omega - \omega_1 .$$

Since we are interested only in small values of $\epsilon$ we may replace $\alpha$ by

$$\alpha = \alpha_1 + \alpha_1' \epsilon,$$

where $\alpha_1$ and $\alpha_1'$ are the values of $\alpha$ and $\dfrac{\partial \alpha}{\partial \omega}$ at $\omega_1$. Similarly,

$$\theta = \theta_1 + \theta_1' \epsilon.$$

We define an instant, $T_e$, by

$$T_e = \theta_1' , \tag{3}$$

and a time, $\tau$, by

$$\tau = t - T_e . \tag{4}$$

Substituting these in (2) and performing the integration, we get

$$f(t) = \text{real part of } 2M \exp\left[-\alpha_1 + i(\omega_1 \tau - (\theta_1 - \omega_1 \theta_1'))\right] \frac{\sinh\,(-\alpha_1' + i\tau)\delta}{(-\alpha_1' + i\tau)} .$$

If we introduce the angles,

$$\beta = \text{arc tan} \frac{-\alpha_1'}{\tau},$$

and

$$\gamma = \text{arc tan} \frac{\tanh\,(-\delta \alpha_1')}{\tan \delta \tau},$$

and take the real part, we get

$$f(t) = 2M \exp(-\alpha_1) \frac{[(\cosh \delta\alpha_1' \sin \delta\tau)^2 + (\sinh \delta\alpha_1' \cos \delta\tau)^2]^{\frac{1}{2}}}{(\alpha_1'^2 + \tau^2)^{\frac{1}{2}}}$$

$$\cos(\omega_1\tau - (\theta_1 - \omega_1\theta_1') + \beta - \gamma). \tag{5}$$

Let us consider first the extreme case where the spectrum of the signal is

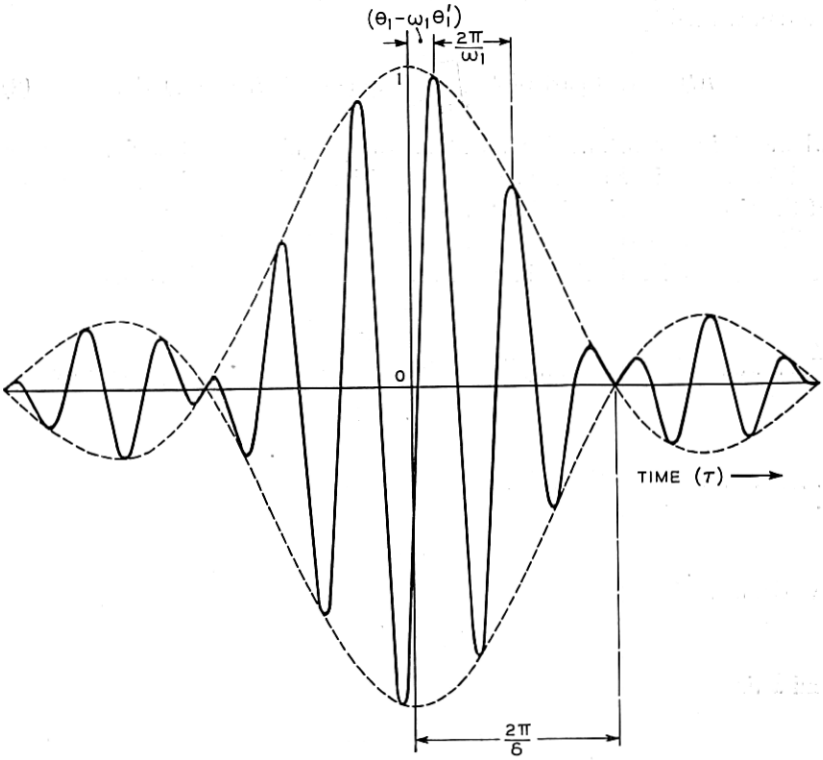


Fig. 1—Elementary disturbance corresponding to a narrow segment of the spectrum

uniform in amplitude in the neighborhood of $\omega_1$, so that $\alpha_1'$ is zero. Then

$$f(t) = 2\delta M \exp(-\alpha_1) \frac{\sin \delta\tau}{\delta\tau} \cos(\omega_1\tau - (\theta_1 - \omega_1\theta_1')). \tag{6}$$

Here the amplitude includes a constant factor which is proportional to the bandwidth, $2\delta$, and to the magnitude, $M \exp(-\alpha_1)$, at the frequency, $\omega_1$, and a function of time, a plot of which is shown in Fig. 1. This function consists of a sinusoidal wave of frequency, $\omega_1$, the amplitude of which varies with time, the envelope being symmetrical about the instant, $T_e = \theta_1'$, at which it is a maximum. $T_e$, the time of maximum envelope, is then a unique instant which is suitable for defining the time at which the disturbance occurs. It is determined solely by the slope of the phase frequency curve for the spectrum.

The instant, $T_e$, may be interpreted, in accordance with the principle of stationary phase, as the one at which the sinusoidal components of (2) are most nearly in the same phase, and so have the least destructive inter-

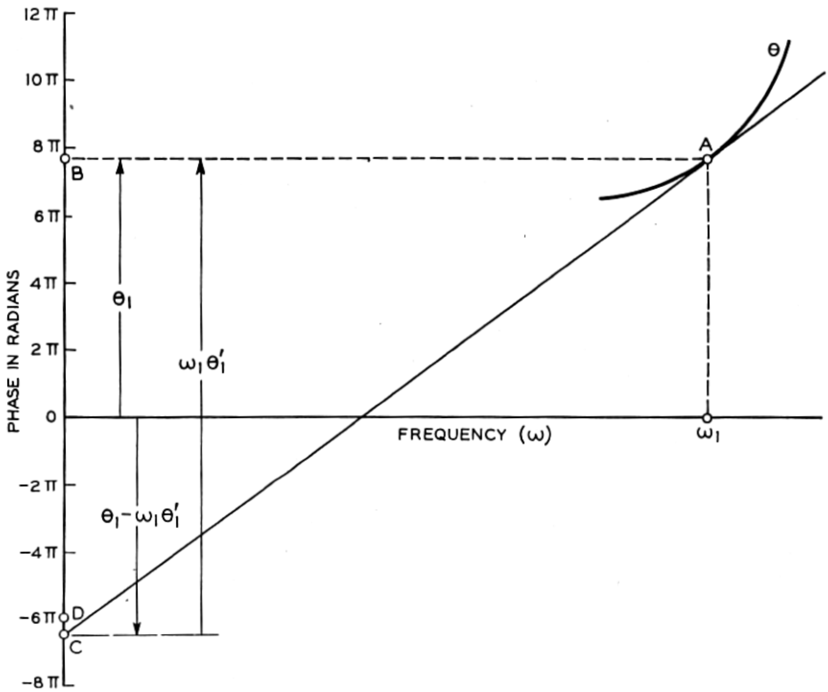


Fig. 2—Graphical representation of the phase of an elementary disturbance

ference. This condition will hold when the instantaneous phase angle is changing least rapidly with frequency, that is, when

$$\frac{\partial}{\partial \omega}(\omega t - \theta) = 0,$$

from which

$$t = \theta_1'.$$

The angle, $\theta_1 - \omega_1\theta_1'$, in (6), gives the phase of the wave at the instant, $T_e$, when its envelope is a maximum. The interpretation of this angle will be aided by the geometrical construction of Fig. 2 which is similar to that employed for phase and group velocity[3]. The abscissae are values of $\omega$ and the ordinates are values of phase in radians. A portion of the function, $\theta$, in the neighborhood of $\omega_1$ is shown. The distance, $OB$, is $\theta_1$. The slope of the tangent, $CA$, to the curve at $A$ is $\theta_1'$. The distance, $CB$, is $\omega_1\theta_1'$. Consequently, $OC$, or the intercept of this tangent on the phase axis, is $\theta_1 - \omega_1\theta_1'$. If, as shown in the figure, the absolute value of this intercept is greater than $\pi$, we may transform (6) to a form in which the angle is less than $\pi$, by the substitution

$$\varphi = \theta_1 - \omega_1\theta_1' + 2n\pi, \tag{7}$$

where $n$ is an integer and

$$|\varphi| < \pi.$$

In Fig. 2, $n$ is 3, and $\varphi$ is the distance $DC$. (6) then becomes

$$f(t) = 2\delta M \exp(-\alpha_1) \frac{\sin \delta\tau}{\delta\tau} \cos(\omega_1\tau - \varphi),$$

and $\varphi$ is the ordinary phase lag of the sinusoid, relative to an origin of time given by the instant of maximum envelope.

We may choose as the instant at which the disturbance occurs, not $T_e$, at which the envelope is a maximum, but $T_a$, at which the instantaneous value of the function has its maximum absolute value. Since $\delta$ is small compared with $\omega_1$, this will occur very nearly at the smallest absolute value of $\tau$ for which $\cos(\omega_1\tau - \varphi)$ is $\pm 1$. This will occur for

$$\tau = \frac{\varphi}{\omega_1}, \quad \text{when} \quad -\frac{\pi}{2} < \varphi < \frac{\pi}{2},$$

and for

$$\tau = \frac{\varphi \pm \pi}{\omega_1} \quad \text{when} \quad -\pi < \varphi < -\frac{\pi}{2} \quad \text{or} \quad \frac{\pi}{2} < \varphi < \pi.$$

From (4), (3) and (7),

$$T_a = \frac{\theta_1 + k\pi}{\omega_1},$$

where $k$ is an integer such that

$$-\frac{\pi}{2} < \Psi = \theta_1 - \omega_1\theta_1' + k\pi < \frac{\pi}{2}.$$

The significance of this can be seen from Fig. 3. Here, in addition to the $\theta$ curve of Fig. 2, there are plotted a series of curves whose ordinates differ

[3] Lamb, "Hydrodynamics," Cambridge U. Press 1916, p. 371.

from it by multiples of $\pi$. In so far as any one purely sinusoidal component of the disturbance is concerned, values of phase determined by those curves which differ by an even multiple of $\pi$ would be indistinguishable. Those differing by an odd multiple would represent a reversal of sign. Let us

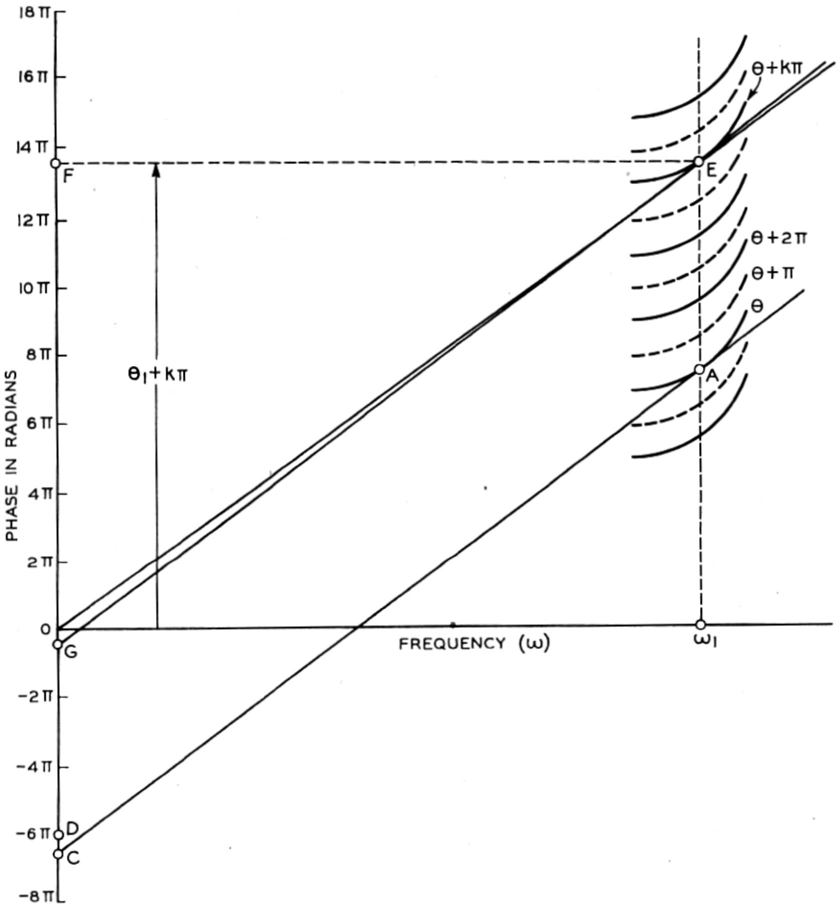


Fig. 3—Graphical representation of the time of maximum absolute value

now select that curve for which the tangent at $\omega_1$ intersects the phase axis nearest the origin, and call it $\theta + k\pi$. Since, for the case drawn,

$$|DC| < \frac{\pi}{2},$$

$$k = 2n.$$

If it were greater we should have

$$k = 2n \pm 1.$$

It is then obvious that the time of maximum absolute value, $T_a$, is given by the slope of the line OE. It differs from $T_e$ by the difference in slope of the lines $OE$ and $GE$.

We have then deduced from the spectrum of the disturbance its time of occurrence in terms of two definitions of the latter. The next step is to compare these times for the input and output and determine the corresponding delays. Let us consider first the case where the attenuation is independent of frequency, so that $\alpha_1'$ is zero in the output signal also. We may then confine our attention to the phase, $\theta$. Let us represent its value at the input by $b$, and the phase shift of the system by $B$. Then at the output $\theta$ will be equal to $b + B$. If we take the time of occurrence as determined by the maximum envelope, these times at the input and output are

$$T_{e_0} = b_1',$$
$$T_{e_1} = b_1' + B_1'.$$

The delay is then

$$D_e = T_{e_1} - T_{e_0} = B_1',$$

which is by definition the envelope delay of the system.

If we take the time of occurrence based on the maximum absolute value, we have, at the input,

$$T_{a_0} = \frac{b_1 + k_0\pi}{\omega_1},$$

where

$$-\frac{\pi}{2} < \Psi_0 = b_1 - \omega_1 b_1' + k_0\pi < \frac{\pi}{2}.$$

At the output,

$$T_{a_1} = \frac{b_1 + B_1 + (k_0 + k_3)\pi}{\omega_1},$$

where

$$-\frac{\pi}{2} < \Psi_3 = b_1 + B_1 - \omega_1(b_1' + B_1') + (k_0 + k_3)\pi < \frac{\pi}{2}.$$

The delay,

$$D_a = T_{a_1} - T_{a_0} = \frac{B_1 + k_3\pi}{\omega_1}.$$

While there is a superficial similarity between this and the phase delay (1), it is of little real significance; $m$, in (1), is determined by the aggregate increase in phase shift with frequency, while $k$, is determined mainly by the rate of increase at $\omega_1$. An example of a situation in which the two delays are very different, is furnished by a wave guide when the frequency only just exceeds the cutoff. The phase delay is then almost zero while the rate of change of phase shift with frequency is very large.

Thus the delay based on maximum absolute value depends on both the envelope delay and the phase shift of the system, but not on the phase delay. There remains to examine this dependence in more detail. The value of $k_3$ depends on the spectrum of the signal as well as the characteristic of the system. It is of interest to see if it can be replaced by a quantity derived from the system characteristic alone. The most obvious thing to try is a delay which is derived from the phase shift of the system in the same way that the time of absolute maximum is derived from that of the signal spectrum. This would be

$$D_S = \frac{B_1 + k_2\pi}{\omega_1},$$

where

$$-\frac{\pi}{2} < \Psi_2 = B_1 - \omega_1 B_1' + k_2\pi < \frac{\pi}{2}.$$

The difference between this and the aperiodic delay based on absolute value is

$$D_S - D_A = \frac{\pi}{\omega_1}(k_2 - k_3),$$

$$= \frac{1}{\omega_1}(\Psi_0 + \Psi_2 - \Psi_3).$$

Since $k_2 - k_3$ is either zero or an integer and $|\Psi_3|$ is less than $\frac{\pi}{2}$, if

$$-\frac{\pi}{2} < \Psi_0 + \Psi_2 < \frac{\pi}{2},$$

$$D_S - D_A = 0.$$

If

$$-\pi < \Psi_0 + \Psi_2 < -\frac{\pi}{2},$$

$$D_S - D_A = -\frac{\pi}{\omega_1}.$$

If

$$\frac{\pi}{2} < \Psi_0 + \Psi_2 < \pi,$$

$$D_S - D_A = \frac{\pi}{\omega_1}.$$

Thus the delay as derived from the system characteristic alone may be identical with the aperiodic delay based on maximum absolute value or it may differ from it by $\pm\frac{\pi}{\omega_1}$, that is by half a period. Which condition holds depends on the interrelation of the phase functions which characterize the signal spectrum at the input and the transmission of the system, and not on either of these functions alone.

If the attenuation is not uniform, $\alpha_1'$ cannot be neglected and the expression for the output signal becomes more complicated. Both the amplitude and phase in (5) vary with time in a manner which depends on the value chosen for $\delta$. The expression becomes fairly simple, however, for the case where $\alpha_1'$ is very large, as in anomalous dispersion and in highly resonant systems. Then, even when $\delta$ is small, we may assume that

$$\cosh(\delta\alpha_1') = \exp(\pm\,\delta\alpha_1'),$$
$$\sinh(\delta\alpha_1') = \pm\exp(\pm\,\delta\alpha_1'),$$

according as $\alpha_1' \gtrless 0$.

The amplitude factor in (5) then becomes

$$\frac{M\exp(-\alpha_1 \pm \delta\alpha_1')}{(\alpha_1'^2 + \tau^2)^{\frac{1}{2}}}.$$

Here the exponent is equal to the value of $\alpha$ at that edge of the segment of the spectrum where the amplitude is greatest. The amplitude is symmetrical about $\tau = 0$, that is, about $t = \theta_1'$, at which point it has its maximum value. Hence the instant of maximum envelope is still given by the slope of the phase, frequency curve, as when $\alpha_1'$ is small. However, the maximum is now extremely flat and its sharpness no longer depends directly on $\delta$. Over the range of values of $\tau$ for which $\tau^2 << \alpha_1'^2$, the amplitude is sensibly constant. When $\tau = \pm\alpha_1'$, it is reduced to $\frac{1}{\sqrt{2}}$ times its maximum. For $\tau^2 >> \alpha_1'^2$, it varies inversely as $|\tau|$.

To investigate the oscillating factor of (5) we note that now

$$\gamma = \pm\delta\tau \pm \frac{\pi}{2},$$

where the sign of $\delta\tau$ depends on that of $\alpha_1'$ and that of $\frac{\pi}{2}$ does not. The oscillating factor then is

$$\cos\left[(\omega_1 \mp \delta)\tau - (\theta_1 - \omega_1\theta_1') - \eta\right],$$

where

$$\eta = \arctan\frac{\alpha_1'}{\tau} \pm \frac{\pi}{2}. \qquad (8)$$

The frequency, $(\omega_1 \mp \delta)$, is that of the edge of the segment of the spectrum where the amplitude is relatively very large. The phase differs from that for small values of $\alpha_1'$ by a quantity $\eta$ which is an ambiguous function of the time $\tau$. This ambiguity may be removed if we assume that the phase varies continuously and that, for very small values of $\tau$, the amplitude has the same sign as the spectrum component corresponding to an infinitesimal value of $\delta$. As $\tau$ increases through zero, $\arctan\frac{\alpha_1'}{\tau}$ changes discontinuously from $\mp\frac{\pi}{2}$ to $\pm\frac{\pi}{2}$ according as $\alpha_1' \lessgtr 0$. To avoid a similar discontinuity, in $\eta$ we say that the sign of $\frac{\pi}{2}$ in (8) is to be taken opposite for positive and negative values of $\tau$. If we make it $\pm$ for $\tau < 0$, and $\mp$ for $\tau > 0$, according as $\alpha_1' \gtrless 0$, then $\eta$ is zero in the neighborhood of $\tau = 0$. Since the amplitude factor is always positive, this corresponds to a spectral component of positive amplitude. If we make the sign of $\frac{\pi}{2}$ $\mp$ for $\tau < 0$, and $\pm$ for $\tau > 0$, $\eta$ becomes $\pm\,\pi$, which is the equivalent of a negative amplitude. Hence a knowledge of the spectral component of frequency $\omega_1$ enables us to determine the sign in (8). For large values of $(\tau)$, $\eta$ reduces to $\pm\frac{\pi}{2}$.

Here we have assumed the amplitude of the input signal to be independent of frequency. If this is not the case the same conditions hold at the input as have just been discussed for the output of a resonant system.

The main conclusion to be drawn from the foregoing is that when the amplitude is changing rapidly with frequency, the component of an aperiodic disturbance which corresponds to a narrow segment of the spectrum persists for a considerable period so that there is much overlapping of the contributions of neighboring segments. It is therefore difficult to deduce the nature of the disturbance at any particular time from any narrow region of its spectrum. For the same reason it is difficult to associate the delay experienced by an aperiodic signal with the steady state characteristic of a network when the attenuation of the latter is changing rapidly with frequency.

The net result of our study then is that steady state phase delay has no direct relation to the particular types of delay of an aperiodic signal which we have chosen to investigate. When the amplitude does not change rapidly with frequency, envelope delay is identical with the delay produced in the maximum value of the envelope of a disturbance corresponding to that part of the signal spectrum which is in the immediate neighborhood of the frequency in question. The envelope delay, together with the phase shift, determines the delay in the maximum absolute value of this disturbance, subject to an uncertainty of half a period. This uncertainty depends on the particular combination of signal spectrum and system characteristic. When the amplitude does change rapidly with frequency, the envelope delay still gives the delay in the maximum value of the envelope. However, this maximum is so flat that the interpretation of the results is very difficult.