Chapter

2.1

Signals and Spectra

Douglas Preis

2.1.1 Introduction

Signals serve several purposes in audio engineering. Primarily, they carry information; for example, electrical analogs of music or speech or numerical data representing such information. Discrete-time signals, formed from sampled values of continuous signals, are now used extensively in digital recording, processing, storage, and reproduction of audio signals. Signals devised and used solely to elicit a response from an audio system are called test signals. Control signals modify the internal operation of signal-processing devices. Certain signals, such as electronic thermal noise, magnetic-tape hiss, or quantization noise in digital systems, may be present but unwanted.

Essential to a deeper understanding of all kinds of signals is the spectrum. The spectrum is defined in slightly different ways for different classes of signals, however. For example, deterministic signals have a mathematical functional relationship to time that can be described by an equation, whereas nondeterministic signals, such as noise generated by a random process, are not predictable but are described only by their statistical properties. Their spectra are defined in different ways. There are also two types of deterministic signals, classified by total energy content or average energy content; and, again, their spectra are defined differently. All spectral representations provide information about the underlying oscillatory content of the signal. This content can be concentrated at specific frequencies or distributed over a continuum of frequencies, or both.

2.1.2 Signal Energy and Power

A deterministic, real-valued signal $f(t)$ is called a *finite-energy* or *transient* signal if

$$0 < \int_{-\infty}^{\infty} f^2(t)\,dt < \infty \tag{2.1.1}$$

where t is time. The integrand $f^2(t)$ can be interpreted as the instantaneous power (energy/time) if $f(t)$ is assumed to be a time-varying voltage across a 1-Ω resistor. The numerical value of the integral in Equation (2.1.1) is the signal's total energy. A *finite power* deterministic signal satisfies

$$0 < \lim_{T \to \infty} \frac{1}{T} \int_{-T/2}^{T/2} f^2(t)\,dt < \infty \qquad (2.1.2)$$

That is, the average energy per time or average power is finite. For example, $f(t)$ could be a constant dc voltage existing for all time across a 1-Ω resistor.

Fundamental to understanding spectral analysis is an elementary periodic signal, that is, one that oscillates with constant frequency and does not decay as time progresses. This simplest oscillating signal is called a *sinusoid*. It predicts, for example, the motion of a swinging pendulum (without friction) or the exchange of energy between inductor and capacitor in a lossless resonant circuit. The sinusoid is the solution to a differential equation that describes a wide variety of physical oscillatory and vibrational phenomena.

2.1.2a Sinusoids and Phasor Representation

The sinusoidal signal illustrated in Figure 2.1.1*a* and described mathematically by

$$f(t) = A\cos(\omega t + \theta) \qquad (2.1.3)$$

is a finite-power signal characterized by its real amplitude A, radian frequency, ω rad/s, and θ, which is a constant phase angle (in radians). The quantity $\omega/(2\pi)$ is the cyclic frequency or number of oscillations per second, called hertz, and equals the reciprocal of the sinusoid's period T, the time taken for one full oscillation. These relationships are illustrated in Figure 2.1.1*a*.

The peak, or maximum, value of the signal is A. Its root-mean-square (rms) value found from taking the square root of the expression in Equation (2.1.2), is $A/(\sqrt{2}) = 0.7071A$, and the average value of $|f(t)|$ over one period is $2A/\pi = 0.6366A$. The average power, from Equation (2.1.2), equals $A^2/2$. or simply the rms value squared. By appropriately changing the phase angle θ, a pure cosine wave or sine wave is realized. For example, $\theta = 0, \pi/2, \pi, 3\pi/2, 2\pi$ corresponds to $A\cos(\omega t)$, $-A\sin(\omega t)$, $-A\cos(\omega t)$, $A\sin(\omega t)$, respectively. If the phase of a sine wave is increased by $\pi/2$ (that is, 90° positive phase shift or phase lead), it becomes a cosine wave. A phase shift of π rad, or 180°, inverts the polarity of a sinusoid.

By using the Euler identity,

$$e^{jx} = \cos(x) + j\sin(x) \qquad (2.1.4)$$

where $j = \sqrt{-1}$. $f(t)$ in Equation (2.1.3), can also be written as the real part of a time-varying complex number (*phasor*), namely

$$f(t) = \text{real}\,[Ae^{j(\omega t + \theta)}] \qquad (2.1.5a)$$

A conceptual picture of Equation (2.1.5*a*), called a phasor diagram, is given in Figure 2.1.1*b*. The tip of the arrow describes the locus of points in the complex plane of the expression in brackets

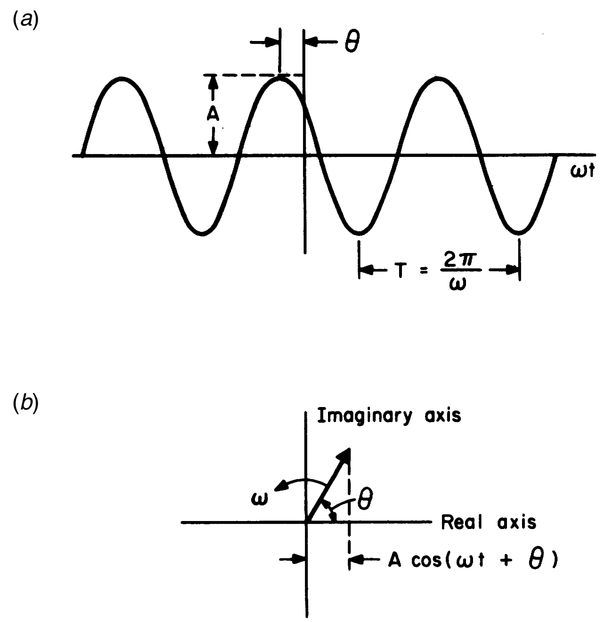


Figure 2.1.1 Sinusoid characteristics: (*a*) sinusoid $A\cos(\omega t + \theta)$, (*b*) phasor diagram representation of sinusoid.

in Equation (2.1.5*a*) as the arrow itself rotates counterclockwise with angular velocity ωrad/s. As time progresses, the projection (or shadow) of the arrow's length A onto the horizontal (real) axis has the values A cos ($\omega t + \theta$). By convention, the phasor diagram is drawn when $t = 0$ or, equivalently, showing the phase θ of the sinusoid relative to a cosine reference phasor, cos (ωt), which would lie on the positive real axis. Note that multiplication of this phasor by $e^{j\psi}$ advances its phase by ψ rad, as can be shown by using Equation (2.1.5*a*). For example, multiplication of the phasor by $j = e^{j\pi/2}$ advances its phase by $\pi/2$, or 90°.

The phasor concept can be extended on the basis of the following two identities derived from Equation (2.1.4)

$$\cos(\omega t) = \frac{1}{2}[e^{j\omega t} + e^{-j\omega t}] \tag{2.1.5b}$$

and

$$\sin(\omega t) = \frac{1}{2}[e^{j\omega t} - e^{j\omega t}] \tag{2.1.5c}$$

where the negative signs in the exponents are associated with ω, implying negative angular velocity $-\omega$. Thus, cos (ωt) and sin (ωt) in Equation (2.1.4) each can be interpreted simply as a sum of two counter-rotating phasors, as shown in Figure 2.1.2. On the left side of this figure, the imaginary parts of the two phasors always cancel because they are equal and opposite, and the real parts add to form cos (ωt). On the right side, imaginary parts cancel at all times, and the real parts add to yield sin (ωt).

There are two ways to represent a sinusoid A cos ($\omega t + \theta$) in terms of counter-rotating phasors. The simpler method is to scale the lengths of the phasors for cos (ωt) in Figure 2.1.2 by the amplitude A and then to change the phase of each phasor by rotating it in its direction of rotation θ rad. This operation corresponds to advancing or leading the phase of the $+\omega$ phasor and advancing or leading the phase of the $-\omega$ phasor by θ rad. The other way makes use of the trigonometric identity

$$A\cos(\omega t + \theta) = A\cos\theta\cos(\omega t) - A\sin\theta\sin(\omega t) \tag{2.1.6}$$

which states that a sinusoid with an arbitrary phase angle can always be represented as the sum of a cosine wave and a sine wave whose coefficients are A cos and $-A\sin\phi$, respectively. Depending on the phase angle θ, these coefficients can be positive, negative, or zero. When either is negative, the polarity of the associated sine wave or cosine wave is just inverted; i.e., its phase is changed by π rad, or 180°, so that the negative sign is absorbed and the coefficients represent nonnegative amplitudes. So, the phasors for cos (ωt) and sin (ωt) in Figure 2.1.2 with their lengths multiplied by the coefficients in Equation (2.1.6) also represent A cos ($\omega t + \theta$).

2.1.2b Line Spectrum

An alternative to counter-rotating phasors is the *line spectrum* (discrete spectrum). Because each phasor shown in Figure 2.1.2 contains only three pieces of information, its length (amplitude), its angular velocity $+\omega$or $-\omega$, and its phase (measured relative to the positive horizontal axis shown in Figure 2.1.2), the same information can be presented differently as shown in Figure 2.1.3. At points $-\omega$ and $+\omega$on the horizontal frequency axis, a vertical line is drawn whose length equals that of its corresponding phasor to give the amplitude. The phase of each phasor is plotted above or below its frequency on a separate graph called the *phase spectrum*. The range of the phase spectrum is from $-\pi$, or -180 °, to $+\pi$, or 180°. Consider, for example, sin (ωt) as represented on the right side of Figure 2.1.2 and Figure 2.1.3. The phase of the counterclockwise phasor is $-\pi/2$, or -90 °, and that of clockwise phasor is π/2, or +90°.

The real advantage of the line-spectrum representation is evident when the signal is composed of several different frequency components $\pm\omega_1$, $\pm\omega_2$, $\pm\omega_3$, . . . , because a separate phasor diagram would be needed for each frequency pair $\pm\omega_n$, whereas all frequencies can be displayed in one line-spectrum plot. A slight and apparent disadvantage is that the concept of negative frequencies is used and signal amplitude is split evenly between the positive and negative frequencies. This representation is a matter of convention, however. If only the total amplitude at each specific frequency is of interest, then phases can be ignored and line spectra drawn at dc (zero frequency) and only at positive values of frequency. The amplitudes are simply doubled at positive frequencies from their values in the "two-sided" line-spectra representation. It is also possible, by a different convention, to define the spectrum by using only positive frequencies and simply giving the amplitude A and phase θ for each sinusoid in the form of Equation (2.1.3). In

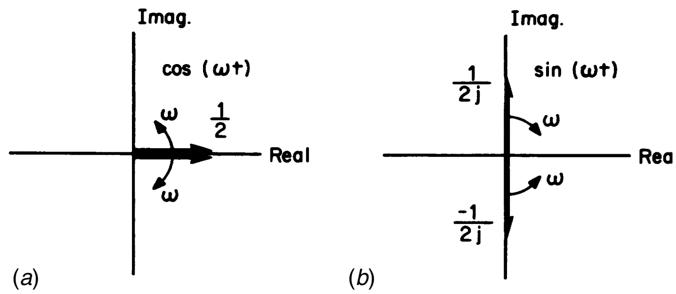


Figure 2.1.2 Waveform representations: (*a*) cosine wave, (*b*) sine wave represented by counter-rotating phasors.

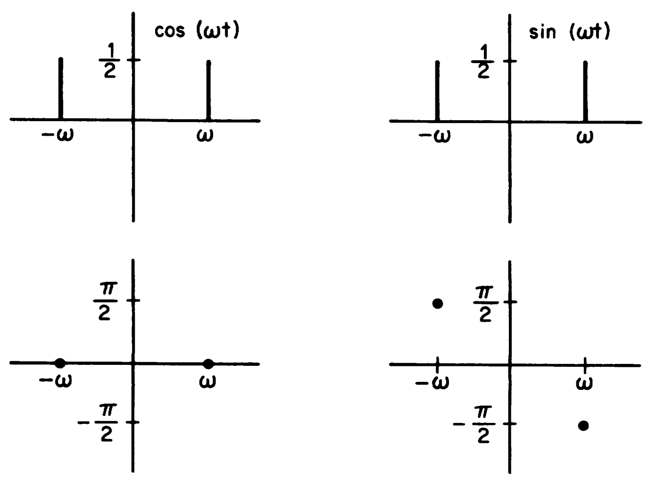


Figure 2.1.3 Line-spectrum representation of $\cos(\omega t)$ at left and $\sin(\omega t)$ at right.

some applications, it is more useful to give the power $A^2/2$, rather than the amplitude A, associated with each spectral line.

The complex exponentials not only have an interesting interpretation as counter-rotating phasors but also have the advantage of representing arbitrary sinusoids without explicitly using the sine and cosine functions themselves, or having to take the real part of a complex number as in Equation (2.1.5*a*). Because complex exponentials form the basis for all mathematical transformations between the time domain and frequency domain, their use has become the method of choice.

While a sinusoid mathematically represents a *pure tone* at a specific frequency $\omega_0/2\pi = 1/T$ Hz, most periodic musical sounds or periodic signal waveforms have *harmonic structure*, meaning that they also contain frequencies that are integer multiples of the lowest *or fundamental* frequency ω_0 which determines the period T. These higher frequencies do, in general, have different amplitudes and phases. Within the period T, each higher frequency or higher harmonic will have an integer number, e.g., 2 or 3 or 4, etc., of full oscillations. The line spectrum is a meaningful way to represent such a signal in the frequency domain. Before the line spectrum can be drawn, the amplitudes and phases of the frequency components in the time waveform are required. These can be found from a specific mathematical analysis of the waveform called Fourier analysis. An understanding of Fourier analysis is essential to the general concept of the spectrum, frequency-domain representation of signals, transformations, and linear signal processing.

2.1.2c Fourier-Series Analysis

To determine whether a periodic $f(t)$ has a sinusoidal component, $\cos(\omega t)$ for example, at frequency ω, the product $f(t)\cos(\omega t)$ is integrated over one full period $T = 2\pi/\omega$. If, for example, $f(t) = A\cos(\omega t)$, then

$$\int_{-T/2}^{T/2} f(t)\cos(\omega t)\,dt = A\int_{T/2}^{T/2} \cos^2(\omega t)\,dt = \frac{AT}{2} \tag{2.1.7}$$

because $\cos^2(\omega t) = 1/2 + 1/2\cos(2\omega t)$ and the second term integrates to zero. Thus, the amplitude A of $\cos(\omega t)$ in $f(t)$ is simply $2/T$ times the integral, Equation (2.1.7). The left integral in Equation (2.1.7) would equal zero if $f(t) = A\cos(n\omega t)$, $n \neq 1$, but $n = 0, 2, 3....$ The same method applies by using $\sin(\omega t)$ if $f(t)$ contains the sinusoidal component $A\sin(\omega t)$. This forms the basis for *Fourier analysis,* that is, a systematic way to determine the harmonic content, in terms of $\sin(n\omega_0 t)$ and $\cos(n\omega_0 t)$, for $n = 0, 1, 2,...,$ of a periodic $f(t)$ whose period $T = 2\pi/\omega_0$.

A *Fourier series* is a mathematical way to represent a real, periodic, finite power signal in terms of a sum of harmonically related sinusoids. If $f(t)$ is periodic with period T, it repeats itself every T so that

$$f(t+T) = f(t) \tag{2.1.8}$$

and its Fourier series is given by

$$f(t) = \frac{a_0}{2} + \sum_{n=1}^{\infty} a_n \cos(n\omega_0 t) + \sum_{n=1}^{\infty} b_n \sin(n\omega_0 t) \tag{2.1.9}$$

where $\omega_0 = 2\pi/T$ and the *Fourier coefficients* are real numbers given by

$$a_n = \frac{2}{T}\int_{-T/2}^{T/2} f(t)\cos(n\omega_0 t)\,dt \tag{2.1.10}$$

and

$$b_n = \frac{2}{T}\int_{-T/2}^{T/2} f(t)\sin(n\omega_0 t)\, dt \tag{2.1.11}$$

The frequency ω_0 is called the *fundamental frequency* (or first harmonic) and, for $n = 2, 3, 4, \ldots, n_0$ are the second, third, fourth, etc., *harmonics*, respectively. The $a_0/2$ term is the average value or dc (zero-frequency) content of $f(t)$. The Fourier coefficients in Equations (2.1.10) and (2.1.11) can be interpreted as the *average content* of cos ($n\omega_0 t$) and sin ($n\omega_0 t$), respectively, in $f(t)$.

By using the Euler identity, Equations (2.1.4), (2.1.10), and (2.1.11) can be combined to yield

$$c_n = \frac{a_n - jb_n}{2} = \frac{1}{T}\int_{-T/2}^{T/2} f(t)e^{-jn\omega_0 t}\, dt \tag{2.1.12}$$

where each c_n is a *complex* Fourier coefficient whose magnitude and (phase) angle ϕ are, respectively

$$|c_n| = \sqrt{(a_n^2 + b_n^2)/2} \tag{2.1.13a}$$

$$\phi_n = arc\tan\left(-\frac{b_n}{a_n}\right) \tag{2.1.13b}$$

By defining $c_0 = a_0/2$ and, for $n \neq 0$

$$c_{-n} = c_n^* = \frac{a_n + jb_n}{2} \tag{2.1.14}$$

the Fourier series Equation (2.1.9) can be written as

$$f(t) = c_0 + \sum_{n=1}^{\infty} c_n e^{jn\omega_0 t} + \int_{(n=-1)}^{-\infty} c_n e^{jn\omega_0 t} \tag{2.1.15a}$$

$$f(t) = \sum_{n=-\infty}^{\infty} c_n e^{jn\omega_0 t} \tag{2.1.15b}$$

which can be derived from Equation (2.1.9) by using Equations (2.1.12), (2.1.14), and (2.1.4). Equation (2.1.15*b*) is the *complex form* of the Fourier series and can be interpreted by using Equation (2.1.15*a*), which shows the sum of the dc term, positive-frequency terms, and negative-frequency terms (n has only negative values in the second sum), respectively.

When a periodic signal is represented in the form (2.1.15*b*), its corresponding line spectrum can be viewed, first, as pairs of pure sine waves and cosine waves with different real coefficients a_n and b_n, as prescribed by Equations (2.1.10) and (2.1.11) at positive frequencies $n\omega_0$ which sum to form sinusoids like Equation (2.1.6). Using identities (2.1.5), however, these sine waves and cosine waves also can be written as sums of complex exponentials $e^{jn\omega_0 t}$ and $e^{-jn\omega_0 t}$ with amplitudes and phases given by the complex coefficients c_n. Alternatively, the line spectrum can be viewed as sinusoids at positive frequencies $n\omega_0$ in the form (2.1.6) and each represented as the sum of a pair of phasors counter-rotating with angular velocities $+ n\omega_0$ and $-n\omega_0$ whose lengths equal $|c_n|$, and whose initial phase angle is specified by the angle of c_n, In both interpretations negative frequencies are used, but this results in the very compact complex exponential notation in Equation (2.1.15*b*), where sine and cosine functions are no longer needed.

Figure 2.1.4 shows different periodic functions of time and their corresponding line spectra determined from the magnitude and phase (or angle) of the complex Fourier coefficients c_n.

2.1.2d Discrete Fourier Series

If a Fourier series has only a finite number of terms (i.e., its frequency band is limited) and the highest frequency equals $K\omega_0$ rad/s (the *K*th harmonic), then Equation (2.1.15*b*) simplifies to the finite sum

$$f(t) = \sum_{n=-K}^{K} c_n e^{jn\omega_0 t} \tag{2.1.16}$$

In certain applications the values of $f(t)$ are needed only at N instants in time (within the period T) that are spaced T/N apart. The values $f(t)$ at these N points in time, from Equation (2.1.16). are given by

$$f(mT/N) = \sum_{n=-K}^{K} c_n e^{jn\omega_0 mT/N} \tag{2.1.17}$$

where m is an integer. Choosing the number of points $N = 2K$, that is, the time interval between points equal to half of the period of the highest frequency ($T/N = T/2K$) and substituting $2\pi/T$ for ω_0 in Equation (2.1.17) gives

$$f(mT/N) = \sum_{n=-N/2}^{N/2} c_n w_N^{nm} \tag{2.1.18}$$

where

$$W_N = e^{j2\pi/N} \tag{2.1.19}$$

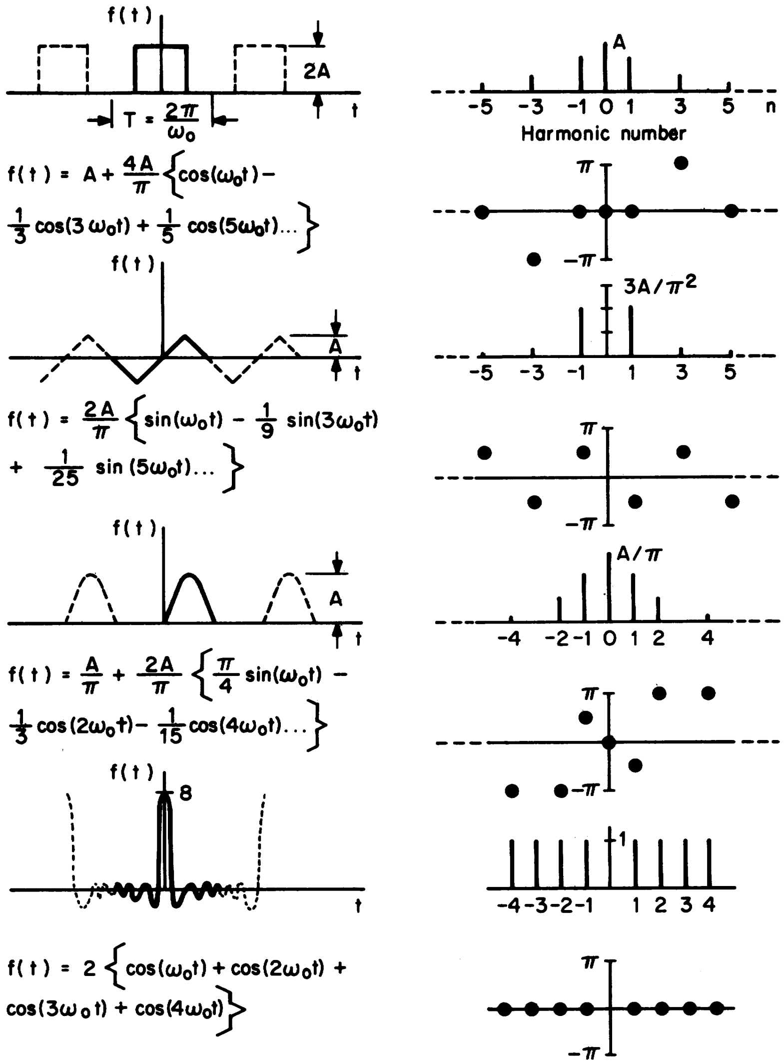


Figure 2.1.4 Periodic functions and their line spectra.

which is called the *N*th root of unity. The quantity w_n^m, for N values of m, mathematically represents specific points on the unit circle in the complex plane spaced $2\pi/N$, or $360°/N$ apart (beginning from $e^{j0} = 1$). Each of these points is physically interpreted as the location of the tip of a phasor of unit length, which rotates through 2π rad, or 360°, during the period T at the instants in time mT/N. Because the phasor for the nth harmonic has an angular velocity n times faster than that of the fundamental ($n = 1$), the term w_N^{nm} appears in Equation (2.1.18). One way to interpret the term $2\pi/N$ in Equation (2.1.19) is that, for any one of the $N/2$ harmonics, only N discrete values of phase are possible after sampling, regardless of frequency. The coefficients c_n in Equation

(2.1.18) are complex numbers of the form $|c_n|e^{j\psi n}$ that scale the unit amplitudes and shift the phases of the sampled harmonics represented by the quantity w_N^{mn} .

The Fourier coefficients are related to the sampled values of $f(t)$ by the similar equation

$$c_n = \frac{1}{N} \sum_{m=-M/2}^{M/2} f(mT/N) w_N^{-nm} \tag{2.1.20}$$

where the integer M is even and numerically equal to N. Equation (2.1.20) can be substituted into Equation (2.1.18), or vice versa, to yield an identity. These two equations define an N-point *discrete Fourier-series pair* which is often called a *discrete Fourier transform* (DFT) pair. They relate the N equally spaced sampled values of $f(t)$ in Equation (2.1.18) to its line spectrum, which contains $N/2$ negative-frequency components, a dc term, and $N/2$ positive-frequency components as well as their phases as prescribed by Equation (2.1.20). For an example, see the last line spectrum in Figure 2.1.4

Equation (2.1.20) is a discrete-time version of Equation (2.1.12) and is valid only if the periodic function $f(t)$ is sampled throughout its period T at a rate that is twice the highest frequency it contains. If $f(t)$ contains harmonics higher than $N/2$, then the c_n coefficients in Equation (2.1.20) are not identical to those in Equation (2.1.12); they will become corrupted or *aliased* because higher frequencies (with a harmonic number greater than $N/2$) in $f(t)$ influence the value of c_n as computed from Equation (2.1.20), but these frequencies cannot be reconstructed by using the harmonics up to harmonic number $N/2$ as in Equation (2.1.18). Furthermore, the resulting line spectrum itself will be incorrect.

The DFT can be computed very efficiently from sampled values of $f(t)$, as contained in Equation (2.1.20), or $f(t)$ can be reconstructed at the sample points, as in Equation (2.1.18), using a numerical algorithm called the *fast Fourier transform* (FFT). The algorithm exploits, in computation, the fact that w_N^{mn} represents only N distinct values of phase.

2.1.2e Spectral Density and Fourier Transformation

Line spectra, where each line $|c_n|$ represents a sinusoid whose average power is $|c_n|^2/2$, cannot be used to represent the spectrum of a finite-energy signal $f(t)$ because the average power of $f(t)$, as given in Equation (2.1.2), is zero by definition. Furthermore, while a finite-energy signal can exist for all time, subject to the constraint given in Equation (2.1.1), it has no finite period T associated with itself. For finite-energy signals, an *amplitude-density spectrum* is defined in the following way

$$F(\omega) = \int_{-\infty}^{\infty} f(t)e^{-j\omega t}dt \tag{2.1.21}$$

The dimensions of $F(\omega)$ are amplitude-multiplied by time or amplitude/(l/time) and are interpreted as amplitude/frequency. The quantity $F(\omega)$ is called *amplitude spectral density* or, simply, *spectral density* of $f(t)$. By comparing Equation (2.1.21) with Equation (2.1.12), $F(\omega)$ is seen to be defined for all rather than discrete frequencies, and the limits of integration include all time rather than one period. Similar to Fourier analysis in concept, the integral (2.1.21) extracts from

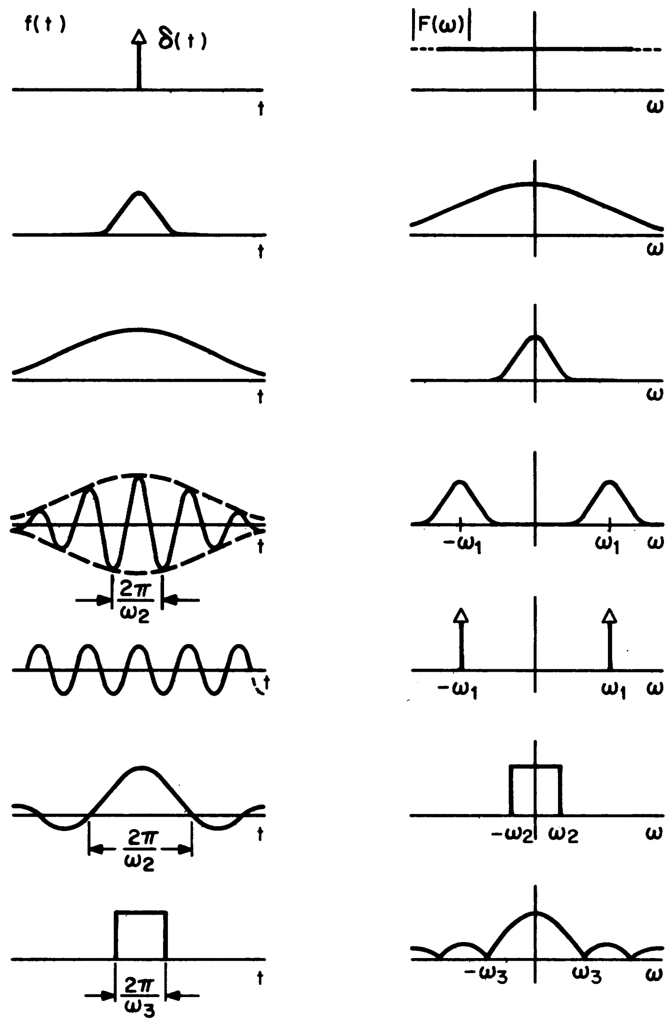


Figure 2.1.5 Amplitude spectral density $|F(\omega)|$ corresponding to time-domain signals $f(t)$.

$f(t)$ its average spectral density $F(\omega)$ and associated phase at each frequency ω throughout a continuum of frequencies. The original signal $f(t)$, again analogously to Fourier analysis, can be reconstructed from its spectral density by using the relation

$$f(t) = \frac{1}{2\pi}\int_{-\infty}^{\infty} F(\omega)e^{j\omega t}d\omega \tag{2.1.22}$$

where all frequencies, instead of discrete frequencies, are used. Equations (2.1.21) and (2.1.22) form a *Fourier-transform pair*. By comparing these two equations it is seen that the value of $f(t)$ at, for example, $t = t_1$ depends on contributions from $F(\omega)$ at all frequencies and, conversely, that the value of the spectral density $F(\omega)$ at $\omega = \omega_1$ depends on contributions from $f(t)$ at all times. In both cases integrals with infinite limits are evaluated. In Figure 2.1.5, several Fourier-transform pairs are illustrated.

The relations (2.1.21) and (2.1.22) can be derived from the Fourier-analysis Equations (2.1.12) and (2.1.15*b*) by using a limiting process where the period T becomes infinite and the frequency interval between harmonics approaches zero.

The continuous Fourier transform (Equation 2.21) is often approximated or estimated by using the discrete Fourier transform (2.1.18) by assuming that $f(t)$ is periodic, with period T, and that its high-frequency content is zero beyond harmonic number $N/2$, where N equals the number of equally spaced sampled values of $f(t)$ within the assumed period.

The *Laplace transformation* $F(s)$ uses the complex variable $s = \sigma + j\omega$ in place of $j\omega$ in Equation (2.1.21) with the lower limit of integration set equal to zero. The quantity $\sigma > 0$ is a convergence factor that exponentially damps out $f(t)$ as $t \to \infty$. This transform is used in applications where finite-power signals (like sinusoids) that are zero for negative values of time are needed. While the inversion of the Laplace transform is more complicated than Equation (2.1.22), tables of transform pairs are available for commonly used signals.

2.1.2f Impulse Signal

The *unit impulse* is a very important finite-energy signal that is denoted by $\delta(t)$ and defined by the constant spectral density $F(\omega) = 1$ and zero phase for all frequencies. Substituting $F(\omega) = 1$ in Equation (2.1.22) gives

$$\delta(t) = \frac{1}{2\pi}\int_{-\infty}^{\infty} e^{j\omega t} d\omega \tag{2.1.23}$$

Because $e^{j\omega t} = \cos(\omega t) + j\sin(\omega t)$ and the odd function $\sin(\omega t)$ integrates to zero, Equation (2.1.23) can be interpreted as summing (integrating) unit-amplitude cosine waves of every frequency. When $t = 0$, $\cos(\omega t) = 1$ regardless of the frequency ω, so that all the cosine waves reinforce one another, giving $\delta(t)$ infinite value. At times other than $t = 0$, the cosine waves destructively interfere and cancel, so that $\delta(t) = 0$ for $t \neq 0$. Consider, for example, the last example given in Figure 2.1.4 as the number of harmonics and length of the period are increased without limit. By using Equation (2.1.21) with $f(t) = \delta(t)$ and the fact that $F(\omega) = 1$

$$1 = \int_{-\infty}^{\infty} \delta(t) e^{-j\omega t} dt \tag{2.1.24}$$

which implies that the value of the integral is simply the value of the integrand, in this case $e^{-j\omega t}$ evaluated or sampled at $t = 0$. If the numerical value 1 is substituted for $e^{-j\omega t}$ in Equation (2.1.24), then the integral of $\delta(t)$ itself results, which also equals 1. By interpreting this latter integral as the area underneath $\delta(t)$, the entire contribution to the value of the integral must come when $t = 0$ because $\delta(t)$ is zero at other times. The unit impulse $\delta(t)$, also called a *delta* *function*, belongs to a class of *singularity functions*. It has many different interpretations in terms of limits of sequences of functions, one of which is the limit of a rectangularly shaped pulse whose width shrinks to zero and whose height increases to maintain constant area under the pulse so that the integrated value always is unity. The dimensions of $\delta(t)$ are 1/time or frequency.

The delta function provides an interesting link between line spectra and spectral density because finite-power signals can be interpreted as delta functions (in frequency) in the spectral density. For example, the finite-power signal $A\cos(\omega_1 t)$ has spectral density

$$F(\omega) = 2\pi\left[\frac{A}{2}\delta(\omega - \omega_1) + \frac{A}{2}\delta(\omega + \omega_1)\right] \tag{2.1.25}$$

This can be verified by inserting Equation (2.1.25) into Equation (2.1.22) and evaluating the integral by using the sampling property of the delta function. The result is

$$f(t) = \frac{A}{2}(e^{j\omega_1 t} + e^{-j\omega_1 t}) = A\cos(\omega_1 t) \tag{2.1.26}$$

From this example, it is clear that spectral density is a more general concept than line spectra.

2.1.2g Power Spectrum

In Fourier analysis, the product $f(t)e^{-j\omega t}$ is integrated with respect to time to resolve or extract the frequency components, either discrete as in Equation (2.1.12) or continuous as in Equation (2.1.21), from $f(t)$. These frequency components also manifest themselves as a function of τ in the product $f(t)f(t+\tau)$, where τ is a time-shift parameter. For example, if $f(t) = A\cos(\omega_1 t)$, this latter product is

$$A^2\cos(\omega_1 t)\cos(\omega_1 t + \omega_1\tau) = \frac{A^2}{2}\cos(2\omega_1 t + \omega_1\tau) + \frac{A^2}{2}\cos(\omega_1\tau) \tag{2.1.27}$$

For a fixed value of the parameter τ, the first term on the right of Equation (2.1.27) oscillates in time t at $2\omega_1$ and has constant phase angle $\omega_1\tau$ (it is a sinusoid), while the second term does not depend on time t (it is a constant). By defining $r(\tau)$ as the time average of the product

$$r(\tau) = \lim_{T\to\infty}\frac{1}{T}\int_{-T/2}^{T/2} f(t)f(t+\tau)\,dt \tag{2.1.28}$$

and inserting Equation (2.1.27) into Equation (2.1.28), the sinusoid at $2\omega_1$ with fixed phase integrates to zero for every choice of τ, and the second term contributes to yield

$$r(\tau) = \frac{A^2}{2}\cos(\omega_1 t) \tag{2.1.29}$$

The coefficient $A^2/2$ equals the power of $f(t) = A\cos(\omega_1 t)$, and the function $r(\tau)$ oscillates at the same frequency ω_1 but as a function of τ rather than t. Because $f(t)$ generally would contain a multitude of different frequencies, it is worthwhile to examine the spectral properties of $r(\tau)$. The Fourier transform of $r(\tau)$ in Equation (2.1.28), using the variable τ instead of t in Equation (2.1.21), is

$$P(\omega) = \int_{-\infty}^{\infty} r(\tau)e^{-j\omega t}d\tau \tag{2.1.30}$$

and defined as the *power spectrum* or *power spectral density* of the finite-power signal $f(t)$. In Equation (2.1.28), $r(\tau)$ is called the *autocorrelation function* of $f(t)$, and its value for $\tau = 0$ is the average power of $f(t)$ as defined in Equation (2.1.2). The autocorrelation function also equals the inverse Fourier transform of the power spectral density, which, from Equation (2.1.22), is

$$r(\tau) = \frac{1}{2\pi}\int_{-\infty}^{\infty} P(\omega)e^{j\omega t}d\omega \tag{2.1.31}$$

Similarly to Equations (2.1.25) and (2.1.26) for spectral density, the power spectral density associated with Equation (2.1.29) would consist of delta functions of strength $2\pi(A^2/4)$ at frequencies $\pm\omega_1$.

Equations (2.1.28) and (2.1.30) are very useful when spectral properties of a non-deterministic, finite-power signal $f(t) = n(t)$, arising, for example, from a noise process, are studied. By definition, no functional relationship exists between $n(t)$ and t. The signal has finite average power but does not, in general, consist of pure sinusoids at fixed frequencies with constant phase that are representable by line spectra. Therefore, Fourier-analysis methods are not directly applicable. Normally, the autocorrelation function $r(\tau)$ in Equation (2.1.28) is a finite-energy signal which first can be estimated from a sufficient number of samples of $f(t)$, then Fourier-transformed as in Equation (2.1.30). The result is a power spectral density as a function of continuous frequency. The autocorrelation function contains oscillations at all the frequencies contained in $f(t)$ with amplitudes equal to their respective powers, as illustrated by Equation (2.1.29) for a single frequency ω_1, and the power spectral density integral defined in Equation (2.1.30) extracts, via Fourier transformation, the distribution of power in frequency $P(\omega)$ of the underlying noise process. The power spectra density has zero phase because $r(\tau)$ is always an even function of τ.

White noise is a term used to characterize a noise process whose power spectral density is constant or flat as a function of frequency. In contrast to an impulse, whose amplitude spectral density is flat and is interpreted as a sum of cosine waves of all frequencies each with identically zero phase, the phases of the similar cosine waves in white noise would vary randomly in time. The former is a finite-energy signal existing for one instant in time and the latter a finite-power signal existing for all time; both have continuous, flat spectra.

2.1.2h Analytic Signal

The actual waveshapes of certain signals, viewed in the time domain, appear to have identifiable amplitude-modulation (AM) and frequency-modulation (FM) effects that may vary as time progresses. This is not easily recognized in the signal's spectrum unless the modulation and signal to be modulated are simple, such as that generated by AM or FM between two sinusoids. The *analytic signal* is one procedure used to define AM or FM effects and extract quantitative information about them from the spectrum $f(t)$. The analytic signal $f_A(t)$ is a linear transformation of the signal $f(t)$ which also has an interesting spectral interpretation.

Consider the signal

$$f(t) = \cos(\omega t) = \frac{1}{2}(e^{j\omega t} + e^{-j\omega t}) \tag{2.1.32}$$

and from it form a second signal $\hat{f}(t)$ by leading the phase of $e^{j\omega t}$ by 90° and leading the phase of $e^{-j\omega t}$ by 90°, which corresponds to multiplication by j and $-j$, respectively. The second signal is, by using Equation (2.1.5*c*)

$$\hat{f}(t) = \frac{1}{2}(je^{j\omega t} - je^{-j\omega t}) = -\sin(\omega t) \tag{2.1.33}$$

The analytic signal associated with (t) is defined as

$$f_A(t) = f(t) - j\hat{f}(t) = e^{j\omega t} \tag{2.1.34}$$

The analytic signal is a complex function of time, and its (line) spectrum contains no negative-frequency components. Figure 2.1.6 illustrates, with phasors, the successive operations in forming the analytic signal for $\cos(\omega t)$.

The amplitude or *envelope* $E(t)$ and phase $\theta(t)$ of the analytic signal are defined by

$$E(t)e^{j\theta(t)} = f(t) - j\hat{f}(t) \tag{2.1.35}$$

so that, in terms of $f(t)$ and $\hat{f}^2(t)$, they are

$$E(t) = \sqrt{f^2(t) + \hat{f}^2(t)} \tag{2.1.36a}$$

$$\theta(t) = \arctan\,[-\hat{f}(t)/f(t)] \tag{2.1.36b}$$

and, from Equation (2.1.36*b*), the *instantaneous frequency* is defined as

$$\omega_i(t) = \frac{d\theta(t)}{dt} \tag{2.1.37}$$

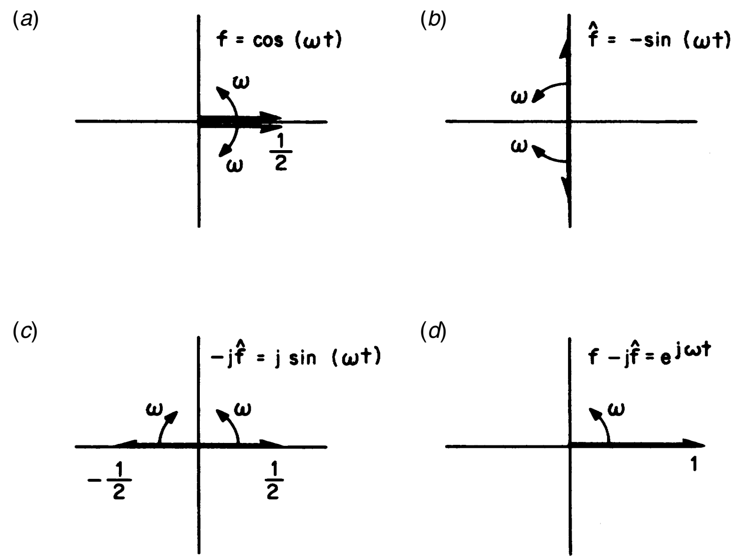


Figure 2.1.6 Formation of analytic signal $e^{j\omega t}$ of $\cos(\omega t)$: (*a*) counterrotating phasors representing $f = \cos(\omega t)$; (*b*) $\hat{f}$, the Hilbert transform of f; (*c*) $-j$ times each phasor in *b;* (*d*) analytic signals given by sum of phasors in *a* and *c*.

In the example $f(t) = \cos(\omega t)$, the amplitude and instantaneous frequency of the associated analytic signal are 1 and ω respectively.

While the analytic signal is a linear transformation of the signal that, in effect, converts negative-frequency components in the spectrum of $f(t)$ to positive-frequency components, the envelope $E(t)$ and instantaneous frequency $\omega_1(t)$ are nonlinear functions of $f(t)$ and $\hat{f}(t)$. For more complicated signals, $\omega_1(t)$ is, in general, not the same as ω in the spectrum.

The spectral operation of advancing the phases of all positive-frequency components by 90° and advancing the phases of all negative-frequency components by 90° is called *Hilbert transformation*; that is, $\hat{f}(t)$ is the Hilbert transform of $f(t)$.

2.1.3 Bibliography

Bracewell, R.: *The Fourier Integral and Its Applications,* McGraw-Hill, New York, N.Y., 1965.

Childers, D. G.: *Modern Spectral Analysis,* IEEE, New York, N.Y., 1978.

Connor, F. R.: *Signals,* Arnold, London, 1972.

Jenkins, G. M., and D. G. Watts: *Spectral Analysis and Its Applications,* Holden-Day, San Francisco, Calif., 1968.

Kharkevich, A. A.: *Spectra and Analysis,* English translation, Consultants Bureau, New York, N.Y., 1960.

Lynn, P. A.: *An Introduction to the Analysis and Processing of Signals,* 2nd ed. Macmillan, London, 1982.

Panter, P. F.: *Modulation, Noise and Spectral Analysis,* McGraw-Hill, New York, N.Y., 1965.

Papoulis, A.: *The Fourier Integral and Its Applications,* McGraw-Hill, New York, N.Y., 1962.

Rabiner, L. R., and C. M. Rader (eds.): *Digital Signal Processing,* IEEE, New York, N.Y., 1972.

Schwartz, M.: *Information Transmission, Modulation and Noise,* McGraw-Hill, New York, N.Y., 1970.

Westman, H. P. (ed.): *ITT Reference Data for Radio Engineers,* Howard W. Sams, New York, N.Y., 1973.

Chapter

2.2

Spectral Changes and Linear Distortion

Douglas Preis

2.2.1 Introduction

When a reproduced signal is not a replica of the original signal, it is *distorted.* Distortionless transmission of a time-varying signal through a system requires that the signal's shape be preserved. The two general mechanisms of signal distortion are *nonlinear* and *linear.*

2.2.2 Distortion Mechanisms

In a broad context, *nonlinear distortion* includes all forms of output-signal corruption that are not linearly related to (i.e., statistically linearly dependent on or correlated with) the input signal. *Modulation* (amplitude or frequency) of the signal (or even its time derivatives) by an imperfect system produces a certain amount of up conversion and down conversion of the signal's frequency components. For example, squaring or cubing of the signal resulting from a nonlinear transfer characteristic is a form of (self) amplitude modulation, whereas time-base errors, like speed variations, are equivalent to frequency modulation of the signal. These converted frequencies, like noise, are not linearly related to the input. The coherence function $\gamma^2(\omega)$ is a quantitative measure of the cumulative effect, at each frequency, of these various forms of signal corruption.

Linear distortion implies that even though the output signal is linearly related to the input, the shape of the output signal is different from that of the input signal. The system itself is linear and does process signals linearly (i.e., scale factors are preserved, and superposition is valid), but linear mathematical operations on the input signal such as differentiation or integration are permissible. Linear distortion changes the *relative* relationships among the existing constituents of the signal by altering either intensity or timing, or both, of its different frequency components. As a consequence, the output signal has a different shape. The system function (or complex frequency response) only predicts the spectral changes that the spectrum of the input signal will undergo and not the change of signal shape in time. The actual time-domain signal must be computed from direct convolution or inverse Fourier transformation.

When a single sinusoid is used as an input to a linear system, the corresponding steady-state output is also a sinusoid of the same frequency but, generally, with different amplitude and phase as prescribed by the complex frequency response. This single sinusoid is never linearly distorted because it always is a replica of the input. In contrast, if the frequency content of the input signal is discrete (e.g., a square wave) or continuous (e.g., a single rectangular pulse), then linear distortion is observable as waveshape change. The extent to which an input signal will be linearly distorted depends on both its spectrum and the system function or complex frequency response of the linear system that processes it. Linear distortion encompasses what is sometimes called *transient distortion*, meaning the waveshape change of a finite-energy (transient) input signal (e.g., a short tone burst or pulse) under linear operating conditions. Finite-power input signals such as square waves, random noise, music, or speech also can be linearly distorted, however. Nonlinear effects such as clipping or slew-rate limiting of transient signals are a form of nonlinear distortion, which implies that transient distortion can be amibiguous terminology.

2.2.2a Linear Range

The range of linear operation of a system is usually established with sine-wave signals. The most frequently used procedure is to verify, using a single sine-wave input, whether the magnitude-scale factor is preserved as input amplitude is changed and/or to verify, using two sinusoids of different frequency simultaneously as an input, whether superposition is valid. Both procedures examine the linearity hypothesis under sinusoidal, steady-state operating conditions. Because the input spectrum is as narrow as possible (a line spectrum consisting of one or two discrete frequencies), the spectrum of the output can easily reveal the existence of other, or "new," frequencies which would constitute nonlinear distortion corresponding to the chosen input signal. The ratio of total power contained in these other frequencies to the output power at the input frequency (or frequencies), expressed in percentage or decibels, is often used as both a measure and a specification of nonlinear distortion (e.g., harmonic distortion, intermodulation distortion, or *dynamic intermodulation distortion*, depending on the specific choice of input frequencies). This method of testing linearity is relatively simple and can be very sensitive. With modern, wide-dynamic-range, high-resolution spectrum analyzers and high-purity sine-wave generators, the effects of very small amounts of nonlinearity can be measured for sinusoidal steady-state operation.

Some aspects of this kind of spectral analysis are questionable, however. Incoherent (uncorrelated) power can exist at the test frequency itself (e.g., due to *cubic nonlinearity*, time-base errors, or noise), and—even more important—the system is never excited throughout its full operating bandwidth by the test signal. Because nonlinear effects do not superpose, the percentage of sine-wave nonlinear distortion measured, for example, as a function of test-signal frequency for fixed-output-power level, is not the same as the percentage of uncorrelated output power as a function of frequency for broadband operation with nonsinusoidal input signals. The latter can be expressed by using the coherence function to define a frequency-dependent signal-to-noise ratio as the ratio of coherent power to incoherent power, expressed in percentage or decibels.

2.2.2b Spectra Comparison

Comparison of input and output spectra of a system is worthwhile. Spectral changes can indicate, in the frequency domain, the existence of linear distortion as well as nonlinear distortion. For linear operation, the spectral changes that any input signal undergoes are predicted by the system function or complex frequency response. For nonlinear operation, the portion of the output sig-

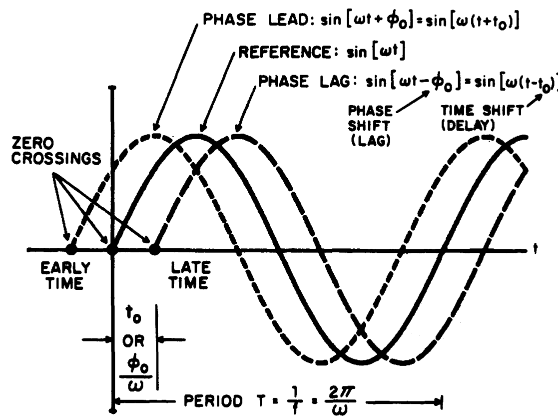


Figure 2.2.1 Standard conventions for sinusoids having steady-state phase shifts relative to a reference sinusoid (solid curve). The leading sinusoid (short dashes) has its first zero crossing at an earlier time than the reference, whereas the lagging sinusoid (long dashes) has its first zero crossing at a later time.

nal's spectrum not linearly related to the input signal depends, in detail, on each specific input signal and how it is nonlinearly processed by the system.

2.2.3 Sinusoidal Steady-State Measurements

The complex frequency response or system function

$$H(\omega) = |H(\omega)|e^{j\phi(\omega)} \quad (2.2.1)$$

of a linear system predicts the *magnitude* (or amplitude) $|H(\omega)|$ and *phase* $\phi(\omega)$ of an output sinusoid relative to an input sinusoid for steady-state operation at the frequency ω. These two frequency-domain quantities are fundamental and form the basis for the discussion of linear distortion of signals by a linear system.

The meaning of the *magnitude response* of a system is well understood and often displayed alone as the frequency response even though it is only the magnitude of $H(\omega)$. The phase-shift-$\phi(\omega)$-versus-frequency characteristic is not commonly shown but is equally important. Some conventions, definitions, and properties of phase shift and the phase-shift characteristic merit discussion.

Figure 2.2.1 illustrates the convention for phase lead or lag between sinusoids of the same frequency relative to a reference. The reference sinusoid (solid line) has an upward-sloping zero crossing at zero time, and that leading (short dashes) has its corresponding zero crossing at an earlier time or is advanced in time, whereas that lagging (long dashes) has its corresponding zero crossing at a later time or is delayed in time.

Aside from this convention, Figure 2.2.1 has two ambiguities and raises a fundamental question about the nature of phase shift itself. First, a phase lead of π is indistinguishable from a phase lag of π; also shifling any of these sinusoids by a multiple of $\pm 2\pi$ would not change the relative phase shifts, and consequently Figure 2.2.1 would not change. Thus, more information than Figure 2.2.1 shows is required to avoid ambiguity. Second, it appears that a phase shift and a time shift are equivalent, but they are not. This is so because $\phi_0 = \omega t_0$ and for a specific phase shift ϕ_0 the corresponding amount of time shift depends upon the sinusoid's frequency ω. To illustrate this point consider Figure 2.2.2*a*, wherein an approximate square wave is constructed from the first three nonzero harmonics of its Fourier series. Figures 2.2.2*b* and *c* show the difference between a constant phase shift for each harmonic and a constant time delay for each harmonic, respectively. The waveform of Figure 2.2.2*b* is severely linearly distorted, whereas that in Figure 2.2.2*c* is not. It is interesting to note that the amount of phase lag necessary to keep the waveform "together" is directly proportional to frequency (that is; the first, third, and fifth harmonics were lagged by $\pi/2$, $3\pi/2$, and $5\pi/2$, respectively). Therefore, in Figure 2.2.2*c* the amount of phase lag varies linearly with frequency, and this corresponds to a uniform time delay.

The fundamental question raised by Figure 2.2.1 is related to causality (i.e., cause and effect). How is it possible for a signal to be "ahead" of the reference signal in time, especially if the reference signal is the input or stimulus to a system and the phase shift of the output signal relative to the input is measured? Equivalently stated, how can the output-signal phase lead or be ahead of the input signal in time? Does this suggest that causality would permit only phase lags to occur in such a situation? It is indeed an interesting question in view of the fact that phase-shift measurements themselves can be somewhat ambiguous. The answer has to do with the fact that these are steady-state measurements, or more precisely, with how the steady state itself is achieved. Both phase lead or phase lag of a system output relative to its input are physically realizable without ambiguity or violation of causality. An example of each case is shown in Figure 2.2.3. Here the actual response of two different circuits to a tone-burst input was captured. In each case the input signal is also shown in time synchronization for reference purposes. It is during the initial transient state that either a phase lag or a phase lead is established. Note that in each case the output zero crossings are initially unequally spaced and that net phase lag or lead is gradually accumulated. Closer inspection of Figure 2.2.3 reveals that, from a mathematical viewpoint, the output in Figure 2.2.3*a* is approximately the integral with respect to time of the input, whereas in Figure 2.2.3*b* the output is nearly the derivative with respect to time of the input. In the steady state, the corresponding phase shifts are seen to be about $-\pi/2$ and $+\pi/2$, respectively. Circuits that have these properties are referred to as phase lag (integrators) or phase lead (differentiators). Another way of interpreting the results of Figure 2.2.3 then (referring also to Figure 2.2.1) is to note that the integral with respect to time of $\sin(\omega t)$ is

$$\frac{-1}{\omega}\cos(\omega t)$$

or

$$\frac{1}{\omega}\sin\left(\omega t - \frac{\pi}{2}\right)$$

whereas the time derivative of $\sin(\omega t)$ is $\omega\cos(\omega t)$ or

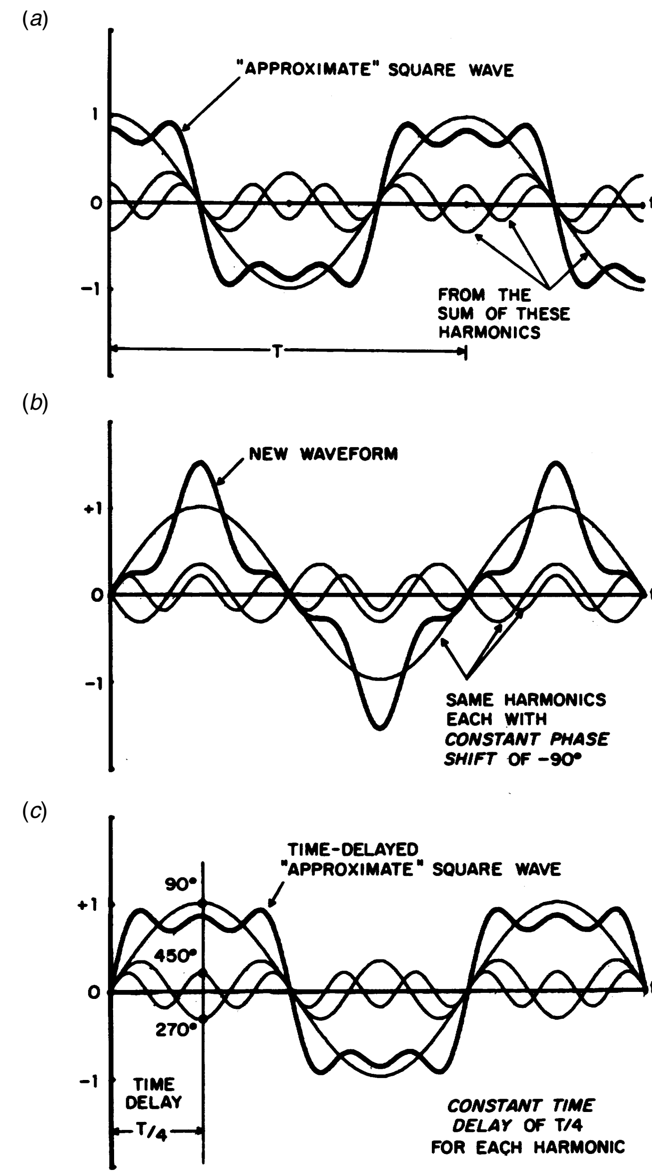


Figure 2.2.2 Characteristics of various waveforms: (*a*) an approximate square wave constructed from the first three (nonzero) Fourier harmonics, (*b*) constant phase shift for each harmonic yields a new waveform that is linearly distorted, (*c*) constant time delay for each harmonic uniformly delays the square wave while preserving its shape.

$$\omega \sin\left(\omega t + \frac{\pi}{2}\right)$$

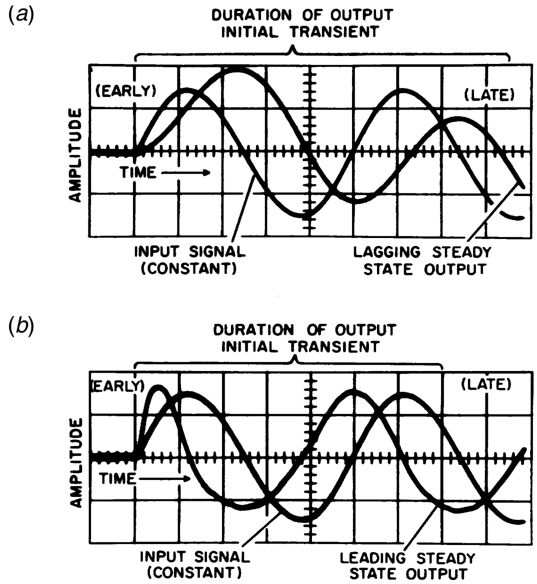


Figure 2.2.3 Steady-state phase lag or lead is established only after an initial transient period: (*a*) a phase lag of about 90º occurs, and the output signal is approximately the *integral* with respect to time of the input; (*b*) a phase lead of 90º occurs, and the output is nearly the time *derivative* of the input. The input sinusoid is the same in each case. The transient buildup of phase lag or lead illustrates that the phase shift of a system is remarkably a property of the steady state.

so that here the steady-state phase lag or lead of $\pi/2$ in the frequency domain corresponds to integration or differentiation, respectively, in the time domain.

An important conclusion to be drawn from Figure 2.2.3 is that the phase shift of a system is remarkably a property of the steady state.

Two significant aspects of a general phase-shift-versus-frequency characteristic $\phi(\omega)$ at a specific frequency ω_0 are its actual numerical value $\phi(\omega_0)$ (positive, zero, or negative) and its behavior in the vicinity of ω_0 (increasing, constant, or decreasing). The reason for this is that most useful signals passed through a system have finite spectral widths which are broad compared with that of any single frequency used to measure the phase characteristic itself.

The value of the phase shift at an arbitrary frequency that is close to a specific frequency ω_0 can be represented by

$$\phi(\omega) \cong \phi(\omega_0) + \text{correction terms} \tag{2.2.2}$$

and in Equation (2.2.2) the correction terms depend upon the difference $\Delta\omega = \omega - \omega_0$. To a first approximation and with reference to Figure 2.2.4

$$\phi(\omega) \cong \phi(\omega_0) + \Delta\phi \tag{2.2.3a}$$

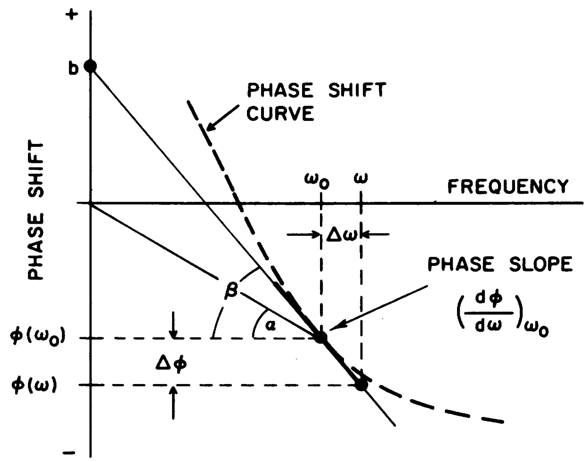


Figure 2.2.4 Arbitrary phase-shift-versus-frequency characteristic (long dashes). For frequencies ω near ω_0 the slope of the phase shift curve is nearly constant and can be approximated by the derivative $d\phi/d\omega$ evaluated at ω_0. The numerical value of the phase slope (or derivative) indicates how the phase shift varies near ω_0.

$$\phi(\omega) \cong \phi(\omega_0) + \frac{\Delta\phi}{\Delta\omega}\Delta\omega \tag{2.2.3b}$$

$$\phi(\omega) \cong \phi(\omega_0) + \left(\frac{d\phi}{d\omega}\right)_{\omega_0}[\omega - \omega_0] \tag{2.2.3c}$$

where

$$\left(\frac{d\phi}{d\omega}\right)_{\omega_0}$$

is the derivative of the phase shift at ω_0 or, equivalently, its slope there. If $\omega = \omega_0$, Equation (2.2.3) is exact, and when ω is near ω_0, it is, in general, approximate. In this approximation it is the slope or first derivative of the phase characteristic at ω_0 that describes the behavior of $\phi(\omega)$ near ω_0. Thus, for signals whose spectra lie in the neighborhood of ω_0 both

$$\phi(\omega_0) \text{ and } \left(\frac{d\phi}{d\omega}\right)_{\omega_0}$$

are important because $\phi(\omega_0)$ gives the steady-state absolute phase shift of the output relative to the input at ω_0, whereas

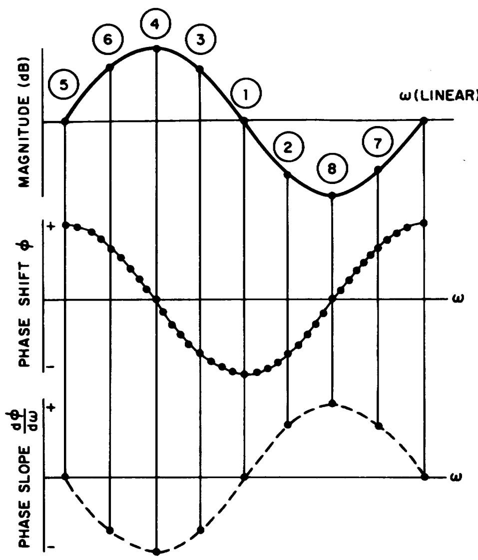


Figure 2.2.5 Possible relationships among magnitude, phase, and phase slope versus linear frequency for a general minimum-phase system. The numbered vertical-line segments each connect a different combination of these three quantities (based on algebraic signs) and indicate important test frequency ranges where transient response may differ.

$$\left(\frac{d\phi}{d\omega}\right)_{\omega_0}[\omega - \omega_0]$$

expresses the phase shift at ω relative to that at ω_0.

Because no restrictions have been placed upon the behavior of $\phi(\omega)$ or its derivative, each may assume positive, zero, or negative values throughout the frequency range of interest. There are, therefore, several different possible combination pairs of phase shift and phase slope. Each of these presumably would affect the response to a transient signal differently. A preliminary discussion of the influence of these aspects of phase response on transient signals is given in the next section together with supporting experimental measurements.

2.2.3a Some Effects of Frequency Response on Transient Signals

An important class of linear systems used in a variety of applications, called *minimum-phase* systems, has mathematically interrelated magnitude and phase responses. These responses are not independent of one another. Specifying one determines the other.

Figure 2.2.5 summarizes the possible relationships between magnitude response, phase shift, and phase slope versus linear frequency for a general minimum-phase system. For example, a narrow bandpass system could have a magnitude response characteristic like that from 5-6-4-3-1, and the associated phase-shift and phase-slope characteristics would be the corresponding portions of those curves below. In a very wideband system the distance between points 5 and 1 on the frequency axis would be considerably greater and the curve 6-4-3 much flatter; also the corresponding sections of the phase and phase-slope curves would require appropriate modification. In a similar way, 1-2-8-7 could represent the magnitude response for a narrowband-reject filter which would have the appropriately corresponding phase and phase-slope curves as shown.

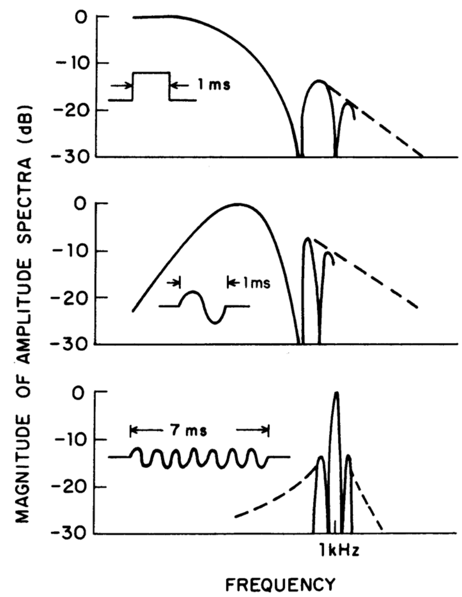


Figure 2.2.6 Magnitude of the spectral density in decibels versus frequency for test signals used to evaluate the response of a general minimum-phase system to transient signals. The dashed lines indicate the upper bound for the continuing spectral peaks.

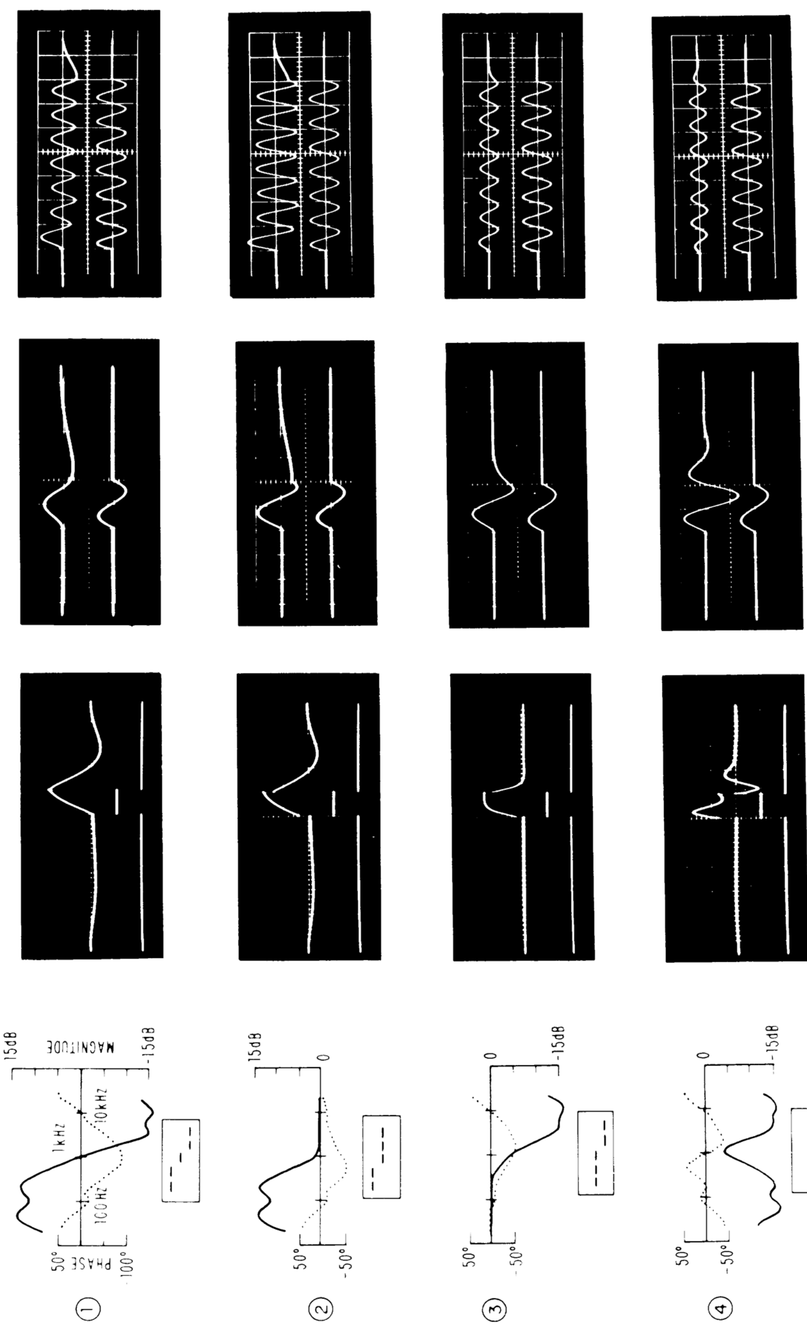
The steady-state characteristics of the minimum-phase system shown in Figure 2.2.5 are also interconnected by eight numbered vertical lines. These numbers indicate important test points (or frequency ranges) where phase shift and phase slope are, in various combinations, positive, zero, and negative. Experimentally measured effects of each of these eight different combinations of $\phi(\omega)$ and $d\phi/d\omega$ on three different transient signals, illustrated along with their spectra in Figure 2.2.6, are shown in Figure 2.2.7. The three test signals were a 1-ms-wide rectangular pulse, a single cycle of a l-kHz sine wave, and a 7-ms tone burst at 1 kHz.

The eight different combinations of phase shift and phase slope in Figure 2.2.4 were simulated, in each case, near 1 kHz by using a five-band graphic equalizer as the minimum-phase system. In each of these eight cases, the linear distortion of transient signals is quite different. Here, the influence of complex frequency response on the system response to transient signals can be explained, qualitatively, in terms of the algebraic signs of phase shift and phase slope. Table 2.2.1 categorizes the eight test cases on this basis. Tables 2.2.2 and 2.2.3 summarize some general effects which the algebraic sign of the phase shift and the phase slope (at 1 kHz) has on response to transient signals.

For the tone bursts, it appears that the actual numerical value of ϕ influences the “inner” structure of the waveform, whereas $d\phi/d\omega$ mostly affects the envelope, or “outer” structure. This is consistent with the approximation in Equation. (2.2.3*c*)

$$\phi(\omega) = \phi(\omega_0) + \left(\frac{d\phi}{d\omega}\right)_{\omega_0}(\omega - \omega_0) \text{ with } \omega_0/2\pi = 1 \text{ kHz}$$

Positive values for both ϕ and $d\phi/d\omega$ imply that ϕ is positive and increasing near ω_0, and this results in sharp and abrupt transient behavior. If ϕ and $d\phi/d\omega$ are both negative, then ϕ is


15dB
MAGNITUDE
-15dB
1kHz
10kHz
100Hz
50°
PHASE
-100°
0
-50°
①
②
③
④

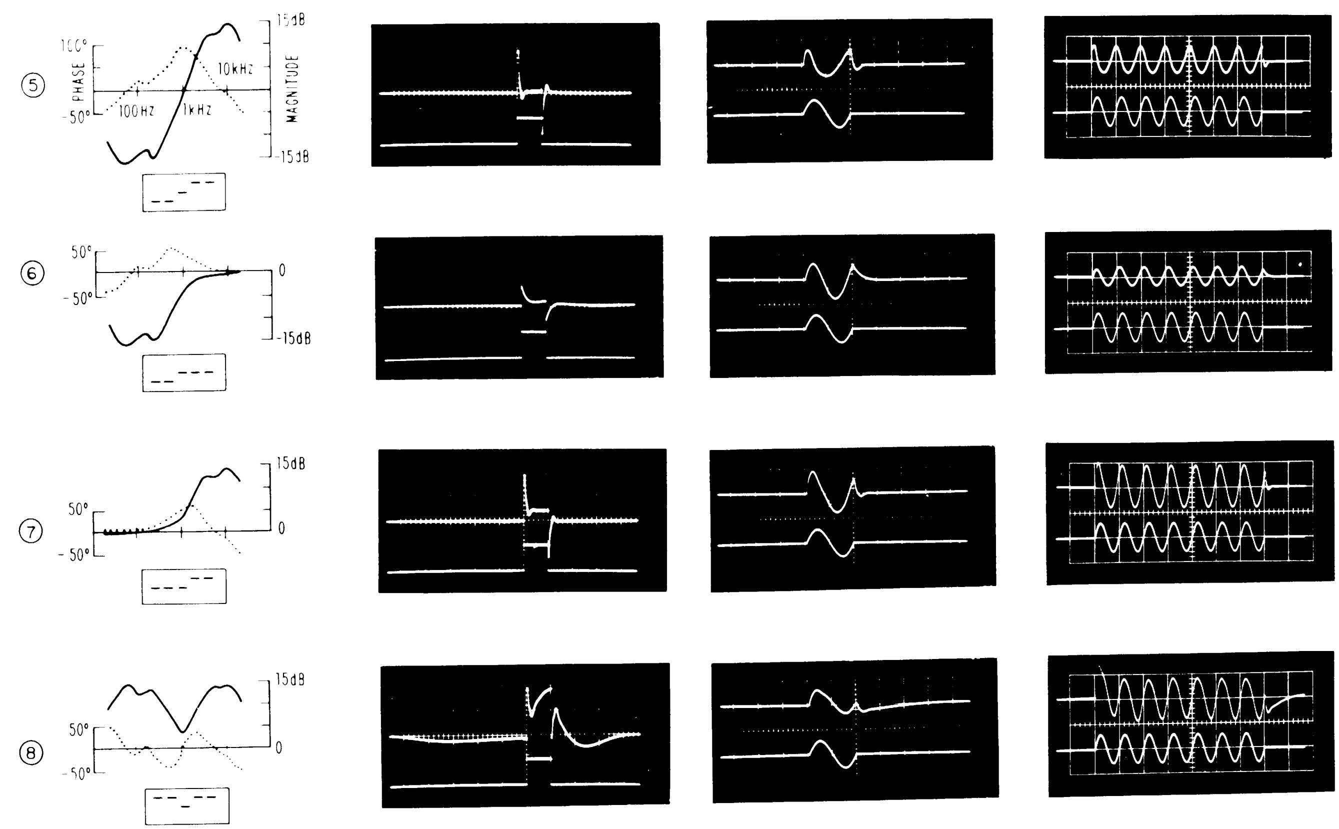


Figure 2.2.7 Comparison of steady-state magnitude and phase response measurements (of a five-band graphic equalizer) with responses to the transient signals whose spectra are given in Figure 2.2.6. These eight cases correspond to the eight test points for the minimum-phase system in Figure 2.2.5 normalized to a frequency range near approximately 1 kHz. (System output and input signals are shown in the upper and lower traces, respectively.)

Table 2.2.1 Categories of Test Cases

		$\frac{d\phi}{d\omega}$ at 1 kHz		
		+	0	−
	+	7	5	6
ϕ at 1 kHz	0	8	*	4
	−	2	1	3

Table 2.2.2 Phase Shift Effects

ϕ (phase shift)	Effect
+	Totally or partially differentiates waveform Phase lead within first few cycles Cusplike transients
0	No phase shift at 1 kHz Tone-burst envelope shape subject to change
−	Totally or partially integrates waveform Phase lag within first few cycles Smooth transients

Table 2.2.3 Phase Shift Effects

$\frac{d\phi}{d\omega}$ (phase slope)	Effect
+	Very rapid tone-burst buildup Overshoot Some long decays
0	Little or no effect Mostly differentiation or integration due to $\phi \lesseqgtr 0$
−	Slower tone-burst buildup Slight elongation of tone burst Smooth decays

becoming more negative, and this yields smoother, slower transients. When the phase shift and phase slope have opposite signs, these opposing effects are seen to combine.

Perhaps the most interesting case occurs when $\phi(\omega_0) = 0$ but $[d\phi/d\omega]_{\omega_0} \neq 0$ (see Figure 2.2.7, cases 4 and 8). For the very-narrow-spectrum tone burst there is no steady-state phase shift as expected; however, the response to the wide-spectrum rectangular pulse in these two cases is remarkable. These results may be interpreted in the following way. Because the rectangular pulse has a considerably wider spectrum than the tone burst, it has many spectral components both above and below $\omega_0/2\pi = 1\,\text{kHz}$. In case 8, $[d\phi/d\omega]_{\omega_0} > 0$ and from the approximation (2.2.3c)

$$\phi(\omega) = [d\phi/d\omega]_{\omega_0}(\omega - \omega_0)$$

so that for $\omega > \omega_0$, $\phi(\omega)$ is positive, whereas for $\omega < \omega_0$, $\phi(\omega)$ is negative. Therefore, and in simple terms, higher frequencies tend to be differentiated in time, whereas lower ones are integrated. This is seen to occur. Just the opposite occurs in case 4 because $[d\phi/d\omega]_{\omega_0} < 0$.

From the foregoing examples it is clear that the response of a system to transient signals depends on its frequency response. For minimum-phase systems, the phase and derivative of phase with respect to frequency can be used to interpret, qualitatively, important aspects of linear distortion of signals in the time domain.

2.2.4 Phase Delay and Group Delay

Phase delay and group delay are useful quantities related to the phase shift $\phi(\omega)$ and defined as

$$\tau_p(\omega) = -\phi(\omega)/\omega \tag{2.2.4}$$

and

$$\tau_g(\omega) = -\frac{d\phi(\omega)}{d\omega} \tag{2.2.5}$$

respectively. The negative signs are required because, according to the conventions for sinusoids in Figure 2.2.1, negative values of phase shift correspond to positive time delays. At a specific frequency ω_0, these two quantities are constants in the two-term Taylor-series expansion of $\phi(\omega)$ Equation (2.2.3c), valid near and at ω_0, which can be rewritten as

$$\phi(\omega) \cong -\omega_0\tau_p(\omega_0) - \tau_g(\omega_0)[\omega - \omega_0] \tag{2.2.6}$$

Equation (2.2.6) restates the fact that the phase shift at ω is equal to the phase shift at ω_0 plus the phase shift at ω relative to ω_0. The steady-state phase shift for the components of a narrowband signal near ω_0 is given by Equation (2.2.6), and the effect of the two terms in this equation can be interpreted in the following way. First, each component in the band undergoes a fixed phase shift

$$-\omega_0\tau_p(\omega_0) = \phi(\omega_0)$$

then those components at frequencies different from ω_0 are subjected to additional phase shift $-\tau_g(\omega_0)[\omega - \omega_0]$. This additional phase shift is one which varies linearly with frequency, so it does not alter the waveshape (see Figure 2.2.2).

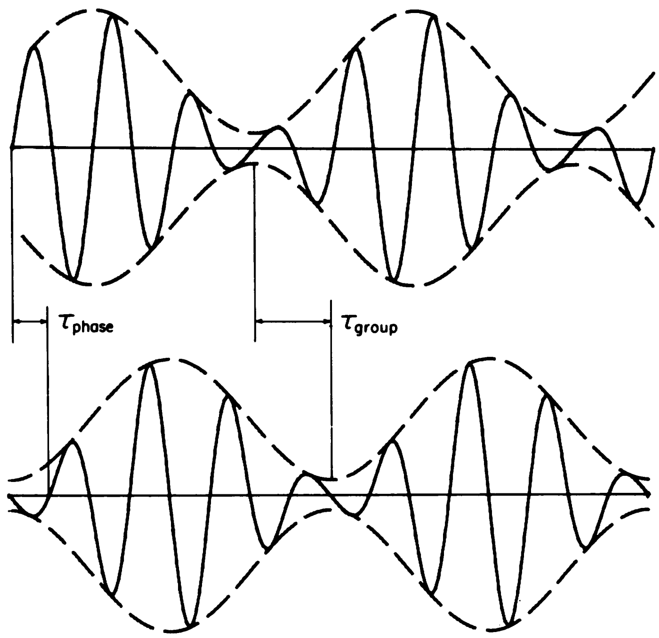


Figure 2.2.8 The difference between phase delay τ_{phase} and group delay τ_{group} is illustrated by comparing these two amplitude-modulated waveforms. The lower waveform has positive phase delay and positive group delay relative to the upper waveform. Because the envelope of the high-frequency oscillation is delayed by an amount of time τ_{group} group delay is sometimes referred to as *envelope delay.*

An amplitude-modulated (AM) sinusoid is a narrowband signal, and the effects of phase delay and group delay on such a signal are illustrated in Figure 2.2.8. Phase delay phase-lags (delays) the high-frequency carrier (inner structure), while the envelope (outer structure) is delayed by an amount equal to the group delay.

Geometrically, Figure 2.2.4 shows that

$$\tau_g(\omega_0) = \tan\beta \text{ and } \tau_p(\omega_0) = \tan\alpha$$

Note that $\tau_g = \tau_p$ only when $\alpha = \beta$ and the intercept $b = 0$. In this special case $\phi(0) = 0$ and $\phi(\omega)$ varies linearly as a function of ω (that is, the entire phase-shift characteristic is a straight line which passes through the origin having slope $-\tau_g$).

Generally both τ_p and τ_g can assume positive, zero, or negative values depending upon the detailed behavior of the phase-shift characteristic. Referring to the special minimum-phase system in Figure 2.2.5 and in view of definition (2.2.5), the group-delay characteristic is the negative of the phase slope, and therefore its shape is the same as the magnitude characteristic for this special case. It is also interesting to note from the same figure that for the bandpass system 5-6-4-3-1

$$\tau_p(\omega) = -\phi(\omega)/\omega$$

can be positive, negative, or zero, but

$$\tau_g(\omega) = -d\phi/d\omega \geq 0$$

For the system 1-2-8-7, $\tau_g(\omega) \leq 0$.

The approximate nature of Equation (2.2.6) deserves particular emphasis because either is valid only over a narrow range of frequencies $\Delta\omega$, and outside this range the correction terms mentioned in Equation (2.2.2) contain higher-order derivatives in the Taylor series that generally cannot be neglected.

2.2.4a Distortionless Processing of Signals

In the time domain, the requirement for *distortionless* linear signal processing (i.e., no wave-shape change) is that the system impulse response $h(t)$ have the form

$$h(t) = K\hat{o}(t-T) \tag{2.2.7a}$$

where $\delta(t)$ is the unit impulse, and the constants $K > 0$ and $T \geq 0$. Equation (2.2.7a) and the convolution theorem together imply that the output signal $g(t)$ is related to the input $f(t)$ by

$$g(t) = Kf(t-T) \tag{2.2.7b}$$

The distortionless system scales any input signal by a constant factor K and delays the signal as a whole by T seconds. The output is a delayed replica of the input. Through substitution, Equation (2.2.7b) gives the corresponding restrictions on the frequency response, namely

$$H(\omega) = Ke^{-jwT} \tag{2.2.8}$$

Comparison of Equation (2.2.8) with Equation (2.2.1) indicates that the frequency-domain requirements are twofold: constant magnitude response $|H(\omega)| = K$ and phase response proportional to frequency $\phi(\omega) = \omega T$. Waveform distortion or *linear distortion* is caused by deviations of $|H(\omega)|$ from a constant value K as well as departures of $\phi(\omega)$ from the linearly decreasing characteristic $-\omega T$. The former is called *amplitude distortion* and the latter *phase distortion.* From Equation (2.2.8) absence of phase distortion requires that the phase and group delays in Equations (2.2.4) and (2.2.5) each equal the overall time delay $T \geq 0$

$$\tau_p(\omega) = \tau_g(\omega) = T \tag{2.2.9}$$

Some experimentally measured effects of the deviations of $|H(\omega)|$ and $\tau_g(\omega)$ from a constant value are illustrated in Figure 2.15. In the experiment, four bandpass filters were connected in cascade to give the attenuation magnitude (reciprocal of gain magnitude) and group-delay characteristics shown in Figure 2.2.9*a*. The group delay is reasonably flat at midband, having a

(*a*)

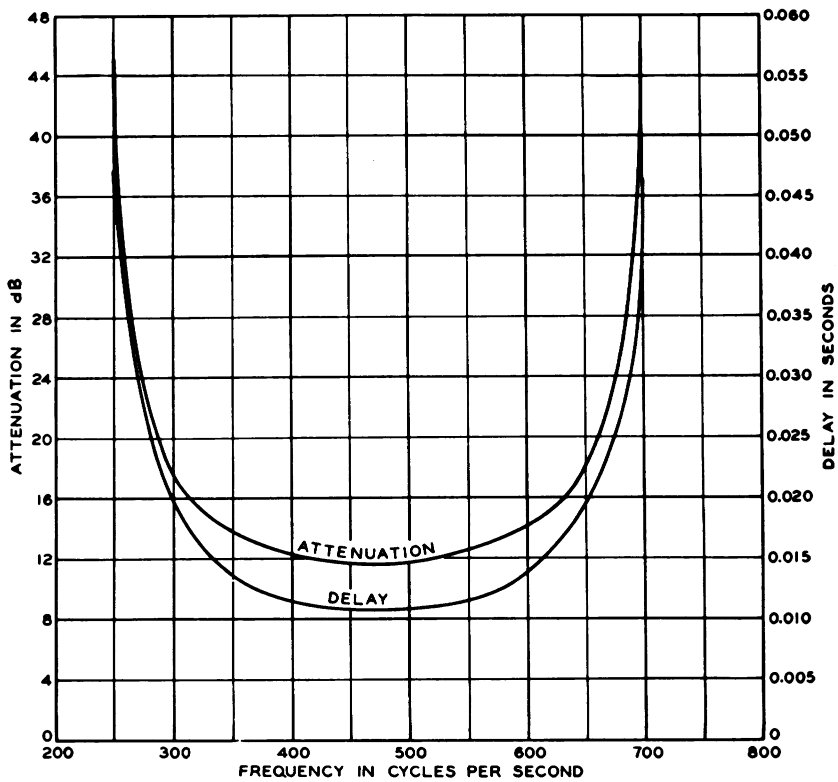


Figure 2.2.9 Attenuation and group delay characteristics: (*a*) attention (reciprocal of gain) magnitude in decibels and group delay in seconds versus frequency characteristics of four cascaded bandpass filters; (*b, next page*) experimentally measured responses to transient input signal (tone burst) of filtering network whose steady-state characteristics are shown in *a*. For this bandpass system, each output signal is delayed by the minimum value of the group delay. When the tone-burst spectrum lies near either passband edge, significant amounts of linear distortion occur in the form of waveform elongation. This is due, for the most part, to the departure of the group-delay characteristic from its flat value in the midband and is called *group-delay distortion*.

minimum value there of $\tau_g = 10.9$ ms. Near and at the passband edges τ_g deviates considerably from its minimum value. The effects in the time domain are shown in Figure 2.2.9*b*. Here, tone bursts at frequencies of 260, 300, 480, and 680 Hz were applied to the filter, and both input and output oscillographs were obtained. In each of these cases, the oscillations start to build up after a time equal to the minimum value of τ_g. There is significant linear distortion for the tone bursts whose spectra lie at the passband edges. Some of this distortion can be ascribed to nonconstant attenuation, but the waveform elongation is primarily due to the group delay $\tau_g(\omega)$ deviating from its minimum value.

These experimental results indicate that, for distortionless processing of signals, the band of frequencies throughout which both magnitude response and group delay of the system are con-

(*b*)

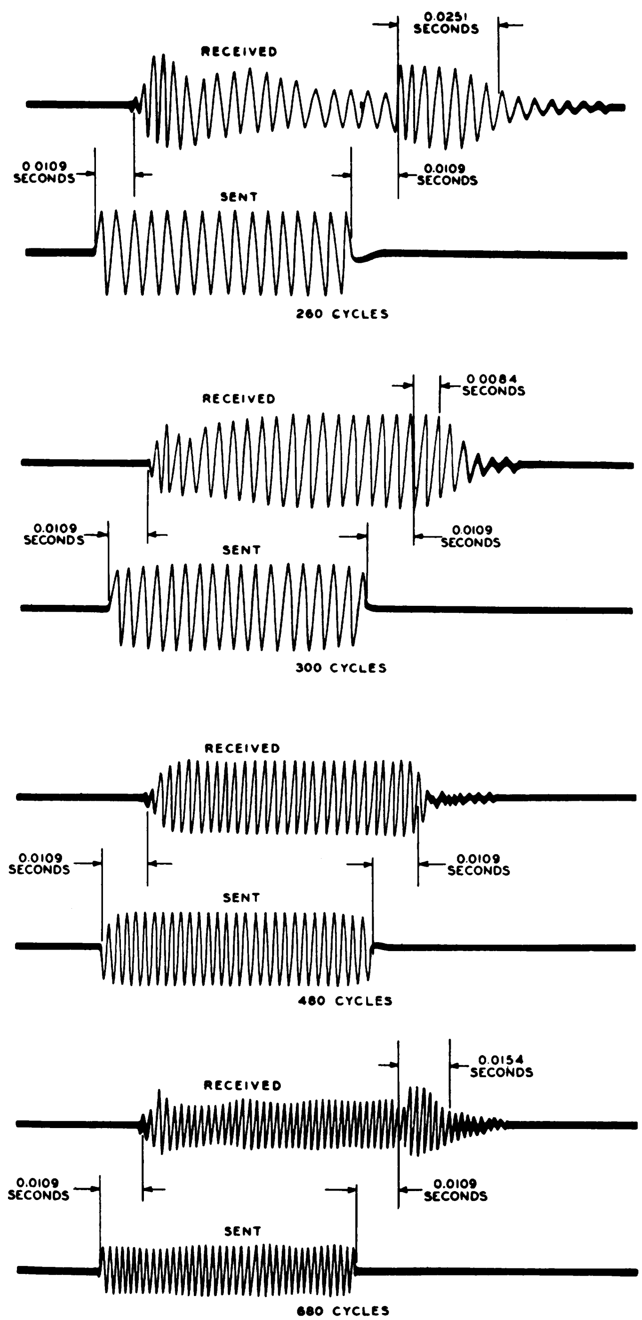


Figure 2.2.9 *Continued*

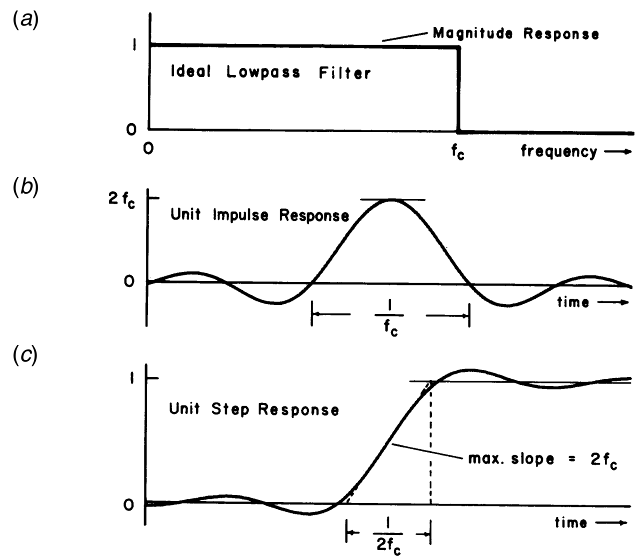


Figure 2.2.10 Response characteristics: (*a*) magnitude response; (*b*) impulse response; (*c*) step response of an ideal linear-phase, band-limited (low-pass) system with cutoff frequency f_c Hz.

stant or flat is important *relative* to the spectral bandwidth of signals to be processed by the system.

2.2.4b Linear Phase and Minimum Phase

The impulse response of a (distortionless) unity-gain, *linear-phase*, band-limited (low-pass) system is symmetrical in time about its central peak value and described mathematically by

$$h(t) = \frac{\sin(\omega_c t)}{\pi t} \qquad (2.2.10)$$

where $\omega_c/2\pi = f_c$ is the cutoff frequency in hertz. The magnitude response is $|H(\omega)| = 1$, and the phase response $\phi(\omega) = 0$ (a special case of linear phase). This result can be interpreted as a cophase superposition of cosine waves up to frequency ω_c. In general, the group delay for such a linear-phase system is constant but otherwise arbitrary; that is, $\tau_g(\omega) = T$ s because the phase shift $\varphi(\omega) = -\omega T$ is linear but can have arbitrary slope. Figure 2.10*b* illustrates $h(t)$ in Equation (2.2.10) delayed, shifted to the right in time, so that its peak value occurs at, say, a positive time $t = T$ rather than $t = 0$. Regardless of the value of T, the impulse response is not causal; that is, it will have finite values for negative time. So if an impulse excitation $\delta(t)$ were applied to the system when $t = 0$, the response to that impulse would exist for negative time. Such *anticipatory transients* violate cause (stimulus) and effect (response). In practice, a causal approximation to the ideal $h(t)$ in Equation (2.2.10) is realized by introducing sufficient delay T and truncating or "windowing" the response so that it is zero for negative time. This latter process would produce ripples in the magnitude response, however. Also shown in Figure 2.2.10 are the corresponding magnitude response (*a*) and step response (*c*), which is the integral with respect to time of $h(t)$. In principle, this ideal system would not linearly distort signals whose spectra are zero for $\omega > \omega_c$. In practice, only approximations to this ideal response are realizable.

A minimum-phase system is causal and has the least amount of phase shift possible corresponding to its specific magnitude response $|H(\omega)|$. The phase is given by the (Hilbert-transform) relationship

$$\phi_m(\omega) = \frac{1}{\pi}\int_{-\infty}^{\infty} \frac{\ln|H(\omega')|}{\omega' - \omega} d\omega' \tag{2.2.11}$$

and the minimum-phase group delay associated with Equation (2.2.11) is

$$\tau_{gm}(\omega) = -\frac{d\phi_m(\omega)}{d\omega} \tag{2.2.12}$$

While minimum-phase systems have impulse responses that are zero for negative time (causal), they are not distortionless. Because magnitude and phase responses are interrelated, the linear distortion they introduce can often be interpreted by using group delay—Equation (2.2.12).

In contrast to the preceding example, consider approximations to band limiting using realizable minimum-phase, maximally flat, low-pass systems of successively higher order. The frequency-domain and time-domain responses for three-, six-, and nine-pole systems are plotted in normalized form in Figure 2.2.11. Here the impulse responses are causal but not symmetrical. The loss of symmetry is due to group-delay distortion. The group delay changes as frequency increases, and this implies that phase shift is not proportional to frequency, especially near the cutoff frequency f_c. In this example, deviations of τ_g from its low-frequency value are a measure of group-delay distortion. Note that the maximum deviation of τ_g, indicated by the length of the solid vertical bars in Figure 2.2.11*b*, is quantitatively related to the broadening of the impulse response, while the low-frequency value of τ_g predicts the arrival time of the main portion of the impulse response, as indicated by the position and length of the corresponding solid horizontal bars in Figure 2.2.11*c*. Actually, the low-frequency value of τ_g equals the time delay of the center of gravity of the impulse response. By increasing the rate of attenuation above f_c, the overall delay of the impulse response increases, initial buildup is slower, ringing is more pronounced, and the response becomes more dispersed and less symmetrical in time.

The minimum-phase group delay $\tau_{gm}(\omega)$ can be evaluated, in theory, for an arbitrary magnitude response $|H(\omega)|$ by using Equations (2.2.11) and (2.2.12). Consider, for example, an interesting limiting case of the previous maximally flat low-pass system where an ideal "brick wall" magnitude response is assumed, as shown in Figure 2.2.12. Here the system has unity gain below the cutoff ω_c, and A dB of attenuation above ω_c. The normalized frequency $\Omega = \omega/\omega_c$. Although this magnitude response cannot be realized exactly, it could be approximated closely with an elliptic filter. The group delay is

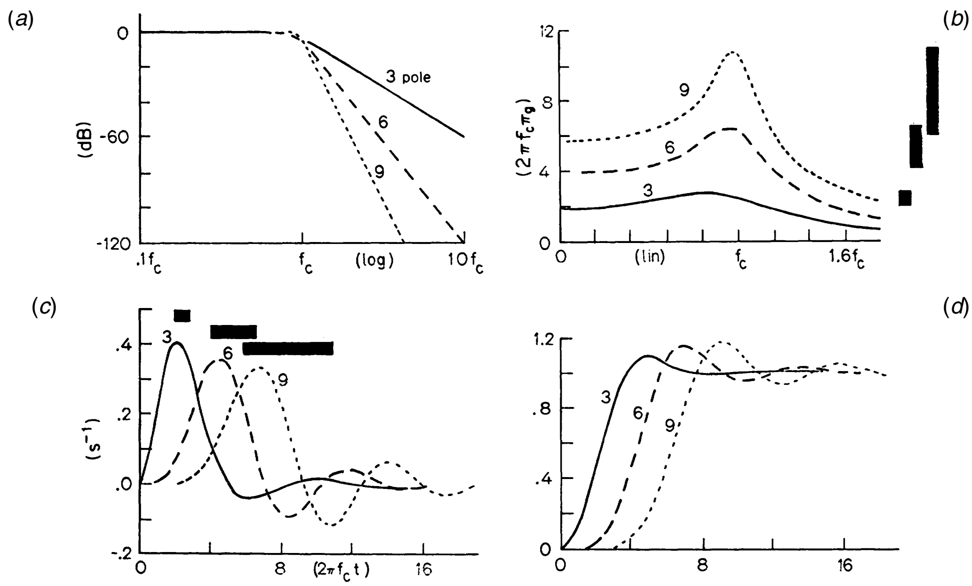


Figure 2.2.11 Response of maximally flat, minimum-phase, low-pass systems: (*a*) magnitude, (*b*) group delay, (*c*) impulse, (*d*) step.

$$\tau_g(\Omega) = \frac{T_0}{1-\Omega^2} \tag{2.2.13}$$

where the constant

$$T_0 = \frac{A \ln 10}{10\pi\omega_c} \tag{2.2.14}$$

As predicted by Equation (2.2.13) and seen in Figure 2.2.12, the group delay becomes infinitely large at the band edge where $\omega = \omega_c$. This is a consequence of the assumed rectangular magnitude response. As a specific numerical example, let A = 80 dB and $\omega_c/2\pi = 14$ kHz. Then T_0 = 62 μs and $\tau_g(\omega) = 0.5$ ms when $\omega/(2\pi) \cong 14$ kHz. It is interesting to note that demanding greater (stop-band) attenuation for $\omega > \omega_c$, requires a larger value for the attenuation parameter *A,* and $\tau_g(0)$ increases, as does the deviation of the group delay within the passband.

A minimum-phase system is also a minimum-delay system since it has the least amount of phase change for a given magnitude response. (The group delay equals the negative rate of change of phase with respect to frequency.) Thus, signal energy is released as fast as is physically possible without violating causality. However, in doing so, certain frequency components are released sooner than others, and this constitutes a form of phase distortion, sometimes called *dispersion*. Its presence is indicated by deviations of the group delay from a constant value. A lin-

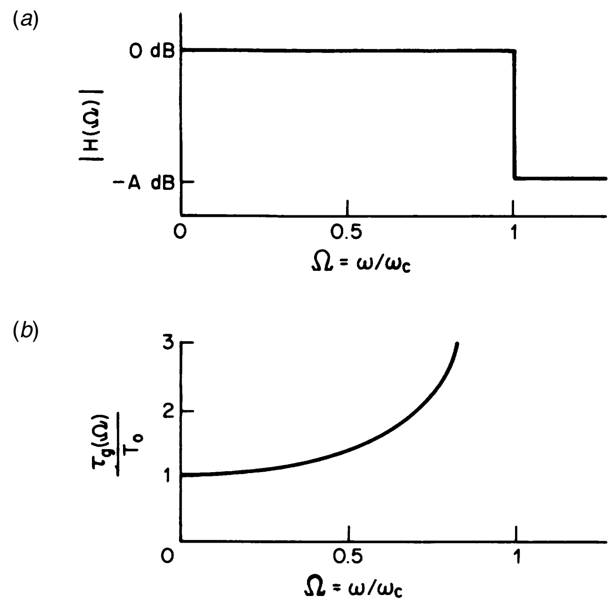


Figure 2.2.12 Magnitude and group delay characteristics: (*a*) magnitude response of a minimum-phase low-pass filter with stop-band attenuation of –*A* db versus normalized frequency $\Omega = \omega/\omega_c$, (*b*) normalized group delay of minimum-phase filter in *a*.

ear-phase system necessarily introduces greater delay than a minimum-phase system with the same magnitude response. However, it has the advantage that there is no dispersion. This is accomplished by delaying all frequency components by the same amount of time.

2.2.5 Bandwidth and Rise Time

An important parameter associated with a band-limited low-pass system is the *rise time*. As illustrated in Figures 2.2.10 and 2.2.11, eliminating high frequencies broadens signals in time and reduces transition times. The *step response* (response of the system to an input that changes from 0 to 1 when $t = 0$) shows, in the case of perfect band limiting illustrated in Figure 2.2.10, that the transition from 0 to 1 requires $\pi/\omega_c = 1/(2f_c)\text{s}$. Defining this as the rise time gives the useful result that the product of bandwidth in hertz and rise time in seconds is $f_c \times 1(2f_c) = 0.5$. For example, with $f_c = 20\,\text{kHz}$ the rise time is 25 μs. The rise time equals half of the period of the cutoff frequency (for this perfectly band-limited system).

Practical low-pass systems, such as the minimum-phase systems shown in Figure 2.2.11, do not have a sharp cutoff frequency, nor do they have perfectly flat group delay like the ideal low-pass model in Figure 2.2.10. The product of the –3-dB bandwidth and the rise time for real systems usually lies within the range of 0.3 to 0.45. The reason that the rise time is somewhat

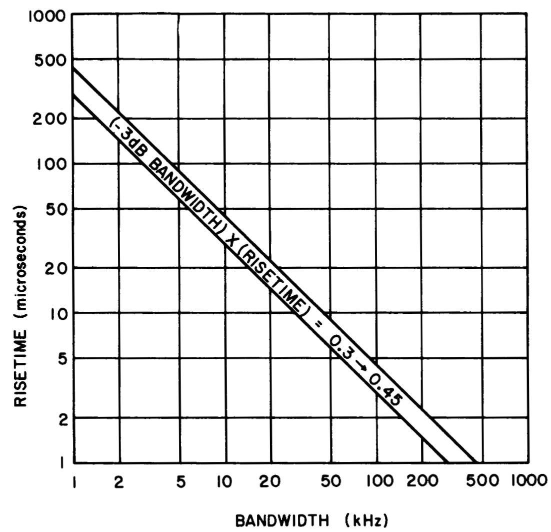


Figure 2.2.13 Relationship between rise time and bandwidth of practical linear systems. For a given bandwidth, the rise time will lie within the tolerance strip shown; conversely, the bandwidth requirements for a specific rise time also can be found.

shorter (faster) is twofold. Because the cutoff is more gradual, some frequencies beyond the –3-dB point contribute to the total output response. Also, the rise time to a unit step is conventionally defined as the time for the output to change from 10 to 90 percent of its final value.

Figure 2.2.13 displays this important relationship between rise time and bandwidth. Given the –3-dB bandwidth of a system, the corresponding range of typically expected rise times can be read. Conversely, knowing the rise time directly indicates nominal bandwidth requirements. Within the tolerance strip shown in Figure 2.2.13, fixing the bandwidth always determines the rise time, and conversely. This figure is a useful guide that relates a frequency-domain measurement (bandwidth) to a time-domain measurement (rise time). This fundamental relationship suggests that testing a practical band-limited linear system with signals having rise times significantly shorter than the rise time of the linear system itself cannot yield new information about its transient response. In fact, the system may not be able to process such signals linearly.

The *slew rate* of a system is the maximum time rate of change (output units/time) for large-signal nonlinear operation when the output is required to change between extreme minimum and maximum values. It is not the same as rise time, which is a parameter defined for linear operation.

2.2.5a Echo Distortion

Ripples in the magnitude and/or phase of the system function $H(\omega)$ produce an interesting form of linear distortion of pulse signals called *echo distortion*. Assuming that these ripples are small and sinusoidal, a model for the system function

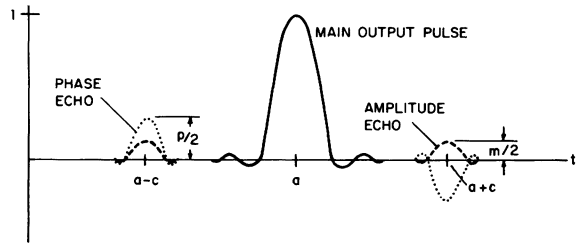


Figure 2.2.14 Small preechoes and postechoes are produced at the output of a linear system having linear distortion in response to a pulse-like input. The main output pulse is delayed by *a* seconds (minimum value of τ_g) and undistorted. Nonflat magnitude response produces symmetrical (dashed curves) "amplitude" echoes, whereas group-delay distortion produces unsymmetrical (dotted curves) "phase" echoes. These echoes *are* the linear distortion. In minimum-phase systems, the noncasual echoes at $t = a - c$ are equal and opposite and cancel one another. In practical systems, the echoes may overlap and change the shape of (linearly distort) the main output pulse.

$$H(\omega) = |H(\omega)|e^{j\phi(\omega)}$$

is

$$|H(\omega)| = 1 - m\cos(\omega c) \tag{2.2.15a}$$

$$\phi(\omega) = -p\sin(\omega c) - \omega a \tag{2.2.15b}$$

where c is the number of ripples per unit bandwidth, in hertz, and m and p are the maximum values of the magnitude and phase ripples, respectively. If $m = p = 0$, then there is no linear distortion of signals, just unity gain, and uniform time delay of $T = \tau_g(\omega) = a$ due to the linear-phase term $-\omega a$. By using Fourier-transform methods, it can be shown that the output $g(t)$ corresponding to an arbitrary input $f(t)$ has the form

$$g(t) = f(t-a) + \frac{m-p}{2}f(t-a-c) + \frac{m+p}{2}f(t-a+c) \tag{2.2.16}$$

Equation (2.2.16) states that the main portion of the output signal is delayed by a seconds and is undistorted, but there are, in addition, small preechoes and postechoes (replicas) which flank it, being advanced and delayed in time (relative to $t = a$) by c seconds. This is shown in Figure 2.2.14. Amplitude echoes are symmetrical (+ + or – –), but phase echoes are asymmetrical (+ – or –+). These echoes *are* the linearly distorted portion of the output and are called echo distortion. The detection of linear distortion by observing paired echoes is possible when the echoes do not overlap and combine with the undistorted part of the signal to form a new (and linearly distorted) waveshape that may be asymmetrical and have a shifted peak time.

In connection with minimum-phase systems, if the magnitude response varies in frequency as a cosine function, then the phase response varies as a negative sine function (see Figure 2.2.5

beginning at point 4) as the Hilbert-transform relationship—Equation (2.2.11)—would predict. (Also, the group delay varies, like the magnitude response, as a cosine function.) This result implies that m and p in Equation (2.2.15) would be equal and opposite; that is, $m = -p$. In this case the preecho vanishes because the last term in Equation (2.2.16) is zero, but the postechoes reinforce. The impulse responses of many minimum-phase systems can be interpreted on this basis.

2.2.5b Classifications of Phase Distortion

When a system is causal, the minimum amount of phase shift $\phi_m(\omega)$ that it can have is prescribed by the Hilbert-transform relation, Equation (2.2.11). There can be additional or *excess* phase shift $\phi_x(\omega)$ as well, so that in general the total phase shift is the sum

$$\phi(\omega) = \phi_m(\omega) + \phi_x(\omega) \tag{2.2.17}$$

A practical definition for the excess phase is

$$\phi_x(\omega) = \theta_a(\omega) - (\omega T + \theta_0) \tag{2.2.18}$$

where θ_0 is a constant and $\theta_a(0) = 0$. In Equation (2.2.18) $-\omega T$ represents pure time delay, $\theta_a(\omega)$ is the frequency-dependent phase shift of an all-pass filter, and θ_0 represents a frequency-independent phase shift caused by, for example, polarity reversal between input and output or a Hilbert transformer which introduces a constant phase shift for all frequencies. The group delay, defined in Equation (2.2.5), is found by substituting Equation (2.2.18) into Equation (2.2.17) and differentiating. The result is

$$\tau_g(\omega) = T - \frac{d\phi_m(\omega)}{d\omega} - \frac{d\theta_a(\omega)}{d\omega} \tag{2.2.19a}$$

$$= T + \tau_{gm}(\omega) + \tau_{ga}(\omega) \tag{2.2.19b}$$

Because deviations of group delay from the constant value T indicate the presence of phase distortion, *group-delay distortion* is defined as

$$\Delta\tau_g(\omega) = \tau_{gm}(\omega) + \tau_{ga}(\omega) \tag{2.2.20}$$

This definition implies that $\Delta T_g(\omega) = 0$ is a *necessary* condition for no phase distortion and, furthermore, that the peak-to-peak excursions of $\Delta T_g(\omega)$ are both a useful indication and a quantitative measure of phase distortion. Although the all-pass group delay $\tau_{ga}(\omega) \geq 0$, the minimum-phase group delay $\tau_{mp}(\omega)$ can be negative, zero, or positive (as can be inferred from the phase responses in Figure 2.2.7 by examining their negative derivatives).

Note that when $\tau_g(\omega)$ is calculated from $\phi(\omega)$ by using Equation (2.2.5), only phase-slope information is preserved. The phase intercept

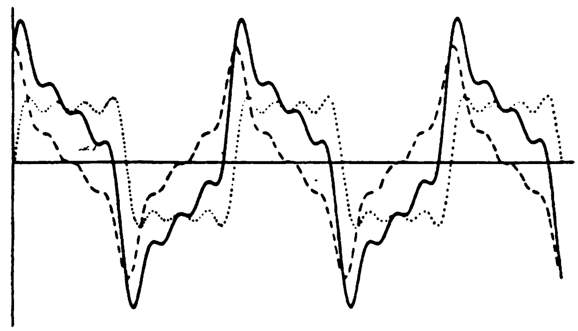


Figure 2.2.15 Band-limited square wave (dotted curve), its Hilbert transform (dashed curve), and the sum of dotted and dashed curves (solid curve).

$$\phi(0) = \phi_m(0) + \theta_0$$

is lost through differentiation. This result implies that when $\Delta\tau_g(\omega) = 0$ in Equation (2.2.20), some phase distortion is possible if, for example

$$\phi(\omega) = \phi_m(\omega) = \frac{-\pi}{2}$$

[$H(\omega)$ is an ideal integrator] or

$$\phi(0) = \theta_0 = \frac{\pi}{2}$$

[$H(\omega)$ contains a Hilbert transformer]. Thus $\Delta\tau_g(\omega) = 0$ and $\phi(0) \neq 0$ (or a multiple of π) implies no group-delay distortion but a form of phase distortion known as *phase-intercept distortion*. With reference to Figure 2.2.4, the phase intercept b is zero when the phase delay and group delay are equal, as stated in Equation (2.2.9), which is the *sufficient* condition for no phase distortion. Generally, the total phase distortion produced by a linear system consists of both group-delay and phase-intercept distortion.

Figure 2.2.15 illustrates phase distortion caused by a frequency-independent phase shift or phase-intercept distortion. The dotted curve represents a band-limited square wave (sum of the first four nonzero harmonics), and the dashed curve is the Hilbert transform of the square wave obtained by shifting the phase of each harmonic $\pi/2$ rad, or 90°. This constant phase shift of each harmonic yields a linearly distorted waveshape having a significantly greater peak factor, as shown. The solid curve is the sum of the square wave and its Hilbert transform. Because corresponding harmonics in this sum are of equal amplitude and in phase quadrature, the solid curve could have been obtained by scaling the magnitude of the amplitude spectrum of the original square wave by $\sqrt{2}$ and rotating its phase spectrum by 45°. For this example

$$\phi_x(\omega) = \theta_0 = \pi/4 \text{ rad}$$

in Equation (2.2.18).

In summary, there are two classifications of phase distortion: 1) group-delay distortion, which is due to the minimum-phase response and/or the frequency-dependent all-pass portion of the excess phase response; and 2) phase-intercept distortion, which is caused by a fixed or constant (frequency-independent) phase shift for all frequencies.

2.2.6 Bibliography

Bendat, J. S., and A. G. Riersol: *Engineering Applications of Correlation and Spectral Analysis*, Wiley, New York, N.Y., 1980.

Bendat, J. S., and A. G. Piersol: *Random Data: Analysis and Measurement Procedures,* Wiley-Interscience, New York, N.Y., 1971.

Blinchikoff, H. J., and A. I. Zverev: *Filtering in the Time and Frequency Domains*, Wiley, New York, N.Y., 1976.

Bloom, P. J., and D. Preis: "Perceptual Identification and Discrimination of Phase Distortions," *IEEE ICASSP Proc.*, pp. 1396–1399, April 1983.

Bode, H. W.: *Network Analysis and Feedback Amplifier Design,* Van Nostrand, New York, N.Y., 1945.

Cheng, D. K.: *Analysis of Linear Systems*, Addison-Wesley, Reading, Mass., 1961.

Deer, J. A., P. J. Bloom, and D. Preis: "Perception of Phase Distortion in All-Pass Filters," *J. Audio Eng. Soc.*, vol. 33, no. 10, pp. 782–786, October 1985.

Di Toro. M. J.: "Phase and Amplitude Distortion in Linear Networks," *Proc. IRE*, vol. 36, pp. 24–36, January 1948.

Guillemin, E. A.: *Communication Networks*, vol. 11, Wiley, New York, N.Y., 1935.

Henderson. K. W., and W. H. Kautz: "Transient Response of Conventional Filters," *IRE Trans. Circuit Theory*, CT-5, pp. 333–347, December 1958.

Hewlett-Packard: "Application Note 63—Section II, Appendix A, "Table of Important Transforms," Hewlett-Packard, Palo Alto, Calif, pp. 37, 38, 1954.

Kupfmuller, K.: *Die Systemtheorie der elektrischen Nachrichtenuhertragung,* S. Hirzel Verlag, Stuttgart, 1968.

Lane, C. E.: "Phase Distortion in Telephone Apparatus," *Bell* Syst. *Tech. J.*, vol. 9, pp. 493–521, July 1930.

Lathi, B. P.: *Signals, Systems and Communications,* Wiley, New York, N.Y., 1965.

Mallinson, J. C.: "Tutorial Review of Magnetic Recording." *Proc. IEEE*, vol. 62, pp. 196–208, February 1976.

Members of the Technical Staff of Bell Telephone Laboratories: *Transmission Systems for Communications,* 4th ed., Western Electric Company, Technical Publications, Winston-Salem, N.C., 197 I.

Oppenheim, A. V., and R. W. Schafer: *Digital Signal Processing*, Prentice-Hall, Englewood Cliffs, N.J., 1975.

Papoulis, A.: *Signal Analysis*, McGraw-Hill, New York, N.Y., 1977.

Peus, S.: “Microphones and Transients,” *db Mag.*, translated from *Radio Mentor* by S. Temmer, vol. 11, pp. 35–38, May 1977.

Preis, D: “A Catalog of Frequency and Transient Responses,” *J. Audio Eng. Soc.*, vol. 25, no. 12, pp. 990–1007, December 1977.

Pries, D.: “Audio Signal Processing with Transversal Filters,” *IEEE Conf. Proc.*, 1979 ICASSP, pp. 310–313, April 1979.

Preis, D.: “Hilbert-Transformer Side-Chain Phase Equalizer for Analogue Magnetic Recording,” *Electron. Lett.*, vol. 13, pp. 616–617, September 1977.

Preis, D.: “Impulse Testing and Peak Clipping,” *J. Audio Eng. Soc.*, vol. 25, no. 1, pp. 2–l4, January 1977.

Preis, D.: “Least-Squares Time-Domain Deconvolution for Transversal-Filter Equalizers,” *Electron. Lett.*, vol. 13, no. 12, pp. 356–357, June 1977.

Preis, D.: “Linear Distortion,” *J. Audio Eng. Soc.*, vol. 24, no. 5, pp. 346–367, June 1976.

Pries, D.: “Measures and Perception of Phase Distortion in Electroacoustical Systems,” *IEEE Conf. Proc.*, 1980 ICASSP, pp. 490–493, 1980.

Pries, D.: “Phase Equalization for Analogue Magnetic Recorders by Transversal Filtering,” *Electron. Lett.*, vol. 13, pp. 127–128, March 1977.

Pries, D.: “Phase Equalization for Magnetic Recording,” *IEEE Conf. Proc.*, 198l ICASSP, pp. 790–795, March 1981.

Preis, D.: “Phase Distortion and Phase Equalization in Audio Signal Processing—A Tutorial Review,” *J. Audio Eng Soc.*, vol. 30, no. 11, pp. 774–794, November 1982.

Pries, D., and C. Bunks: “Three Algorithms for the Design of Transversal-Filter Equalizers,” *Proc. 1981 IEEE Int. Symp. Circuits Sys.*, pp. 536–539, 1981.

Pries, D., and P. J. Bloom: “Perception of Phase Distortion in Anti-Alias Filters,” *J. Audio Eng. Soc.*, vol. 32, no. 11, pp. 842–848, November 1984.

Preis, D., F. Hlawatsch, P. J. Bloom, and J. A. Deer: “Wigner Distribution Analysis of Filters with Perceptible Phase Distortion,” *J. Audio Eng. Soc.*, December 1987.

Small, R. H.: “Closed-Box Loudspeaker Systems, Part 1: Analysis,” *J. Audio Eng. Soc.*, vol. 20, pp. 798–808, December 1972.

Totzek, U., and D. Press: “How to Measure and Interpret Coherence Loss in Magnetic Recording,” *J. Audio Eng. Soc.*, December 1987.

Totzek, U., D. Preis, and J. F. Boebme: “A Spectral Model for Time-Base Distortions and Magnetic Recording,” *Archiv. fur Elektronik und Ubertragungstechnik*, vol. 41, no. 4, pp. 223–231, July-August 1987.

Wheeler, H. A.: "The Interpretation of Amplitude and Phase Distortion in Terms of Paired Echoes," *Proc. IRE*, vol. 27, pp. 359–385, June 1939.

Williams, A. B.: *Active Filter Design,* Artech House. Dedham, Mass., 1975.

Zverev, A. 1.: *Handbook of Filter Synthesis*, Wiley, New York, N.Y., 1967.