neous transmission in both directions. A **half-duplex** (HDX) system allows transmission in either direction but not at the same time.

Fundamental Limitations

An engineer faces two general kinds of constraints when designing a communication system. On the one hand are the **technological problems,** including such diverse considerations as hardware availability, economic factors, federal regulations, and so on. These are problems of feasibility that can be solved in theory, even though perfect solutions may not be practical. On the other hand are the **fundamental physical limitations,** the laws of nature as they pertain to the task in question. These limitations ultimately dictate what can or cannot be accomplished, irrespective of the technological problems. The fundamental limitations of information transmission by electrical means are **bandwidth** and **noise.**

The concept of bandwidth applies to both signals and systems as a measure of **speed.** When a signal changes rapidly with time, its frequency content, or **spectrum,** extends over a wide range and we say that the signal has a large bandwidth. Similarly, the ability of a system to follow signal variations is reflected in its usable frequency response or **transmission bandwidth.** Now all electrical systems contain energy-storage elements, and stored energy cannot be changed instantaneously. Consequently, every communication system has a finite bandwidth B that limits the rate of signal variations.

Communication under real-time conditions requires sufficient transmission bandwidth to accommodate the signal spectrum; otherwise, severe distortion will result. Thus, for example, a bandwidth of several megahertz is needed for a TV video signal, while the much slower variations of a voice signal fit into $B \approx 3$ kHz. For a digital signal with r symbols per second, the bandwidth must be $B \geq r/2$. In the case of information transmission without a real-time constraint, the available bandwidth determines the maximum signal speed. The time required to transmit a given amount of information is therefore inversely proportional to B.

Noise imposes a second limitation on information transmission. Why is noise unavoidable? Rather curiously, the answer comes from kinetic theory. At any temperature above absolute zero, thermal energy causes microscopic particles to exhibit random motion. The random motion of charged particles such as electrons generates random currents or voltages called **thermal noise.** There are also other types of noise, but thermal noise appears in every communication system.

We measure noise relative to an information signal in terms of the **signal-to-noise power ratio** S/N. Thermal noise power is ordinarily quite small, and S/N can be so large that the noise goes unnoticed. At lower values of S/N, however, noise degrades fidelity in analog communication and produces errors in digital communication. These problems become most severe on long-distance links when the transmission loss reduces the received signal power down to the noise level. Amplification at the receiver is then to no avail, because the noise will be amplified along with the signal.

Taking both limitations into account, Shannon (1948)[†] stated that the rate of information transmission cannot exceed the **channel capacity.**

$$C = B \log (1 + S/N)$$

This relationship, known as the **Hartley-Shannon law,** sets an upper limit on the performance of a communication system with a given bandwidth and signal-to-noise ratio.

1.2 MODULATION AND CODING

Modulation and coding are operations performed at the transmitter to achieve efficient and reliable information transmission. So important are these operations that they deserve further consideration here. Subsequently, we'll devote several chapters to modulating and coding techniques.

Modulation Methods

Modulation involves two waveforms: a **modulating signal** that represents the message, and a **carrier wave** that suits the particular application. A modulator systematically alters the carrier wave in correspondence with the variations of the modulating signal. The resulting modulated wave thereby "carries" the message information. We generally require that modulation be a *reversible* operation, so the message can be retrieved by the complementary process of **demodulation.**

Figure 1.2–1 depicts a portion of an analog modulating signal (part *a*) and the corresponding modulated waveform obtained by varying the amplitude of a **sinusoidal** carrier wave (part *b*). This is the familiar amplitude modulation (AM) used for radio broadcasting and other applications. A message may also be impressed on a sinusoidal carrier by frequency modulation (FM) or phase modulation (PM). All methods for sinusoidal carrier modulation are grouped under the heading of **continuous-wave** (CW) modulation.

Incidentally, you act as a CW modulator whenever you speak. The transmission of voice through air is accomplished by generating carrier tones in the vocal cords and modulating these tones with muscular actions of the oral cavity. Thus, what the ear hears as speech is a modulated acoustic wave similar to an AM signal.

Most long-distance transmission systems employ CW modulation with a carrier frequency much higher than the highest frequency component of the modulating signal. The spectrum of the modulated signal then consists of a band of frequency components clustered around the carrier frequency. Under these conditions, we say that CW modulation produces **frequency translation.** In AM broadcasting, for example, the message spectrum typically runs from 100 Hz to 5 kHz; if the carrier frequency is 600 kHz, then the spectrum of the modulated carrier covers 595–605 kHz.

†References are indicated in this fashion throughout the text. Complete citations are listed alphabetically by author in the References at the end of the book.

between AM and FM, and it must be developed further. First the similarity will be stressed.

In phase modulation, the phase deviation is proportional to the amplitude of the modulating signal and therefore independent of its frequency. Also, since the phase-modulated vector sometimes leads and sometimes lags the reference carrier vector, its instantaneous angular velocity must be continually changing between the limits imposed by ϕ_m; thus some form of frequency change must be taking place. In frequency modulation, the frequency deviation is proportional to the amplitude of the modulating voltage. Also, if we take a reference vector, rotating with a constant angular velocity which corresponds to the carrier frequency, then the FM vector will have a phase lead or lag with respect to the reference, since its frequency oscillates between $f_c - \delta$ and $f_c + \delta$. Therefore FM must be a form of PM. With this close similarity of the two forms of angle modulation established, it now remains to explain the difference.

If we consider FM as a form of phase modulation, we must determine what causes the phase change in FM. The larger the frequency deviation, the larger the phase deviation, so that the latter depends at least to a certain extent on the amplitude of the modulation, just as in PM. The difference is shown by comparing the definition of PM, which states in part that the modulation index is proportional to the modulating voltage *only,* with that of the FM, which states that the modulation index *is also inversely proportional to the modulation frequency*. This means that under identical conditions *FM and PM are indistinguishable for a single modulating frequency*. When the modulating frequency is changed the PM modulation index will remain constant, whereas the FM modulation index will increase as modulation frequency is reduced, and vice versa. This is best illustrated with an example.

EXAMPLE 5-4 A 25-MHz carrier is modulated by a 400-Hz audio sine wave. If the carrier voltage is 4 V and the maximum deviation is 10 kHz, write the equation of this modulated wave for (*a*) FM and (*b*) PM. If the modulating frequency is now changed to 2 kHz, all else remaining constant, write a new equation for (*c*) FM and (*d*) PM.

SOLUTION

Calculating the frequencies in radians, we have

$$\omega_c = 2\pi \times 25 \times 10^6 = 1.57 \times 10^8 \text{ rad/s}$$
$$\omega_m = 2\pi \times 400 = 2513 \text{ rad/s}$$

The modulation index will be

$$m = m_f = m_p = \frac{\delta}{f_m} = \frac{10{,}000}{400} = 25$$

This yields the equations

(*a*) $v = 4 \sin (1.57 \times 10^8 t + 25 \sin 2513t)$ (FM)
(*b*) $v = 4 \sin (1.57 \times 10^8 t + 25 \sin 2513t)$ (PM)

Note that the two expressions are identical, as should have been anticipated. Now, when the modulating frequency is multiplied by 5, the equation will show a fivefold increase in the (angular) modulating frequency. While the modulation index in FM is reduced fivefold, for PM the modulation index remains constant. Hence

(*c*) $v = 4 \sin (1.57 \times 10^8 t + 5 \sin 12{,}565t)$ (FM)
(*d*) $v = 4 \sin (1.57 \times 10^8 t + 25 \sin 12{,}565t)$ (PM)

Note that the difference between FM and PM is not apparent at a single modulating frequency. It reveals itself in the *differing behavior of the two systems when the modulating frequency is varied.*

The practical effect of all these considerations is that if an FM transmission were received on a PM receiver, the bass frequencies would have considerably more deviation (*of phase*) than a PM transmitter would have given them. Since the output of a PM receiver would be proportional to phase deviation (or modulation index), the signal would appear unduly bass-boosted. Phase modulation received by an FM system would appear to be *lacking in bass*. This deficiency could be *corrected by bass boosting the modulating signal prior to phase modulation*. This is the practical difference between phase and frequency modulation.

Frequency and amplitude modulation Frequency and amplitude modulation are compared on a different basis from that for FM and PM. These are both practical systems, quite different from each other, and so the performance and characteristics of the two systems will be compared. To begin with, frequency modulation has the following advantages:

1. The amplitude of the frequency-modulated wave is constant. It is thus independent of the modulation depth, whereas in AM modulation depth governs the transmitted power. This means that, in FM transmitters, low-level modulation may be used but all the subsequent amplifiers can be class C and therefore more efficient. Since all these amplifiers will handle constant power, they need not be capable of managing up to four times the average power, as they must in AM. Finally, *all* the transmitted power in FM is useful, whereas in AM most of it is in the transmitted carrier, which contains no useful information.
2. FM receivers can be fitted with amplitude limiters to remove the amplitude variations caused by noise, as shown in Section 5-2.2; this makes FM reception a good deal more immune to noise than AM reception.
3. It is possible to reduce noise still further by increasing the deviation (see Section 5-2.1). This is a feature which AM does not have, since it is not possible to exceed 100 percent modulation without causing severe distortion.
4. Commercial FM broadcasts began in 1940, decades after their AM counterparts. They have a number of advantages due to better planning and other considerations. The following are the most important ones:
   a. Standard frequency allocations (allocated worldwide by the International Radio Consultative Committee (CCIR) of the I.T.U.) provide a guard band between commercial FM stations, so that there is less adjacent-channel interference than in AM;
   b. FM broadcasts operate in the upper VHF and UHF frequency ranges, at which there happens to be less noise than in the MF and HF ranges occupied by AM broadcasts;

c. At the FM broadcast frequencies, the *space wave* is used for propagation, so that the radius of operation is limited to slightly more than line of sight, as shown in Section 8-2.3. It is thus possible to operate several independent transmitters on the same frequency with considerably less interference than would be possible with AM.

The advantages are not all one-sided, or there would be no AM transmissions left. The following are some of the disadvantages of FM:

1. A much wider channel is required by FM, up to 10 times as large as that needed by AM. This is the most significant disadvantage of FM.
2. FM transmitting and receiving equipment tends to be more complex, particularly for modulation and demodulation.
3. Since reception is limited to line of sight, the area of reception for FM is much smaller than for AM. This may be an advantage for cochannel allocations, but it is a disadvantage for FM mobile communications over a wide area. Note that this is due not so much to the intrinsic properties of FM, but rather to the frequencies employed for its transmission.

5-2 NOISE AND FREQUENCY MODULATION

Frequency modulation is much more immune to noise than amplitude modulation and is significantly more immune than phase modulation. In order to establish the reason for this and to determine the extent of the improvement, it is necessary to examine the effect of noise on a carrier.

5-2.1 Effects of Noise on Carrier—Noise Triangle

A single-noise frequency will affect the output of a receiver only if it falls within its bandpass. The carrier and noise voltages will mix, and if the difference is audible, it will naturally interfere with the reception of wanted signals. If such a single-noise voltage is considered vectorially, it is seen that the noise vector is superimposed on the carrier, rotating about it with a relative angular velocity $\omega_n - \omega_c$. This is shown in Figure 5-5. The maximum deviation in amplitude from the average value will be V_n, whereas the maximum phase deviation will be $\phi = \sin^{-1}(V_n/V_c)$.

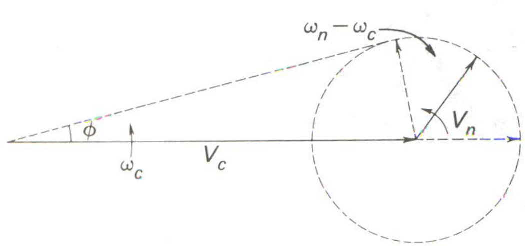

FIGURE 5-5 Vector effect of noise on carrier.

which is readily derived from the message waveform $x(t)$. The system consists of an analog inverter, an integrator, and a Schmitt trigger controlling an electronic switch. The trigger puts the switch in the upper position whenever $x_\Lambda(t)$ increases to $+1$ and puts the switch in the lower position whenever $x_\Lambda(t)$ decreases to -1.

Suppose the system starts operating at $t = 0$ with $x_\Lambda(0) = +1$ and the switch in the upper position. Then, for $0 < t < t_1$,

$$x_\Lambda(t) = 1 - \int_0^t v(\lambda)\,d\lambda = 1 - \frac{2}{\pi}[\theta_c(t) - \theta_c(0)]$$

$$= 1 - \frac{2}{\pi}\theta_c(t) \qquad 0 < t < t_1$$

so $x_\Lambda(t)$ traces out the downward ramp in Fig. 5.3–6a until time t_1 when $x_\Lambda(t_1) = -1$, corresponding to $\theta_c(t_1) = \pi$. Now the trigger throws the switch to the lower position and

$$x_\Lambda(t) = -1 + \int_{t_1}^t v(\lambda)\,d\lambda = -1 + \frac{2}{\pi}[\theta_c(t) - \theta_c(t_1)]$$

$$= -3 + \frac{2}{\pi}\theta_c(t) \qquad t_1 < t < t_2$$

so $x_\Lambda(t)$ traces out the upward ramp in Fig. 5.3–6a. The upward ramp continues until time t_2 when $\theta_c(t_2) = 2\pi$ and $x_\Lambda(t_2) = +1$. The switch then triggers back to the upper position, and the operating cycle goes on periodically for $t > t_2$.

A sinusoidal FM wave is obtained from $x_\Lambda(t)$ using a nonlinear waveshaper with transfer characteristics $T[x_\Lambda(t)] = A_c \sin[(\pi/2)x_\Lambda(t)]$, which performs the inverse of Eq. (5a). Or $x_\Lambda(t)$ can be applied to a hard limiter to produce **square-wave** FM. A laboratory test generator might have all three outputs available.

Frequency Detection

A **frequency detector,** often called a **discriminator,** produces an output voltage that should vary linearly with the instantaneous frequency of the input. There are perhaps as many different circuit designs for frequency detection as there are designers who have considered the problem. However, almost every circuit falls into one of the following four operational categories:

1. FM-to-AM conversion
2. Phase-shift discrimination
3. Zero-crossing detection
4. Frequency feedback

We'll look at illustrative examples from the first three categories, postponing frequency feedback to Sect. 7.3. Analog *phase* detection is not discussed here because

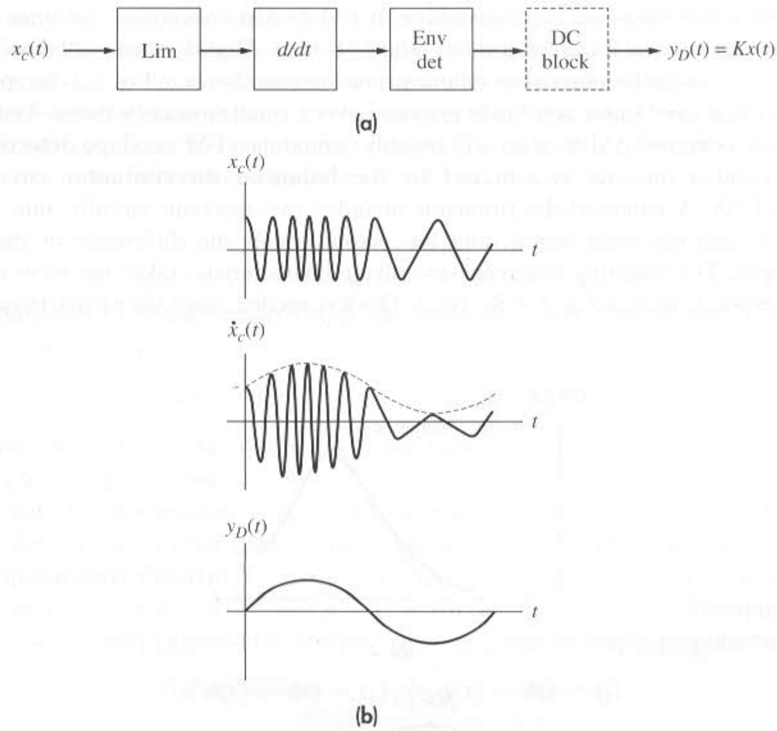

Figure 5.3–7 (*a*) Frequency detector with limiter and FM-to-AM conversion; (*b*) waveforms.

it's seldom needed in practice and, if needed, can be accomplished by integrating the output of a frequency detector.

Any device or circuit whose output equals the *time derivative* of the input produces **FM-to-AM conversion.** To be more specific, let $x_c(t) = A_c \cos \theta_c(t)$ with $\dot{\theta}_c(t) = 2\pi[f_c + f_\Delta x(t)]$; then

$$\begin{aligned} \dot{x}_c(t) &= -A_c \dot{\theta}_c(t) \sin \theta_c(t) \\ &= 2\pi A_c[f_c + f_\Delta\, x(t)] \sin [\theta_c(t) \pm 180°] \end{aligned} \tag{6}$$

Hence, an **envelope detector** with input $\dot{x}_c(t)$ yields an output proportional to $f(t) = f_c + f_\Delta x(t)$.

Figure 5.3–7*a* diagrams a conceptual frequency detector based on Eq. (6). The diagram includes a limiter at the input to remove any spurious amplitude variations from $x_c(t)$ before they reach the envelope detector. It also includes a dc block to remove the constant carrier-frequency offset from the output signal. Typical waveforms are sketched in Fig. 5.3–7*b* taking the case of tone modulation.

For actual hardware implementation of FM-to-AM conversion, we draw upon the fact that an ideal differentiator has $|H(f)| = 2\pi f$. Slightly above or below resonance, the transfer function of an ordinary tuned circuit shown in Fig. 5.3–8*a* approximates the desired linear amplitude response over a small frequency range. Thus, for instance, a detuned AM receiver will roughly demodulate FM via **slope detection.**

Extended linearity is achieved by the **balanced discriminator** circuit in Fig. 5.3–8*b*. A balanced discriminator includes two resonant circuits, one tuned above f_c and the other below, and the output equals the difference of the two envelopes. The resulting frequency-to-voltage characteristic takes the form of the well-known *S* curve in Fig. 5.3–8*c*. No dc block is needed, since the carrier-frequency

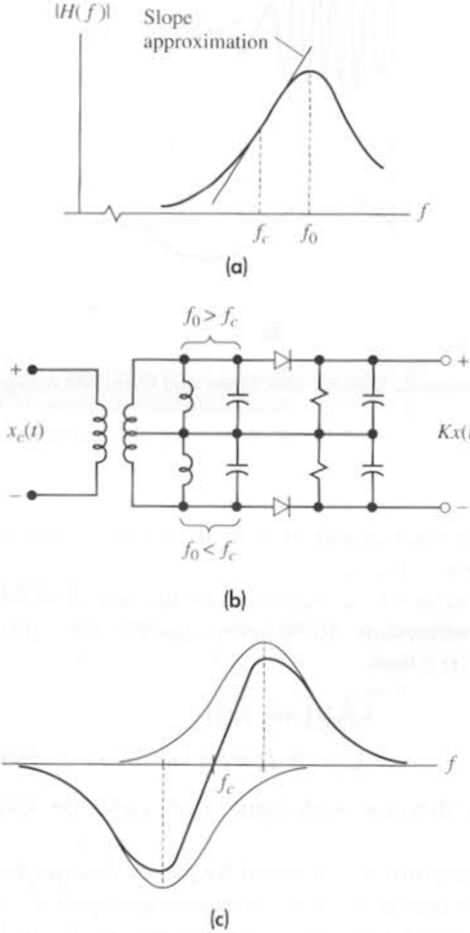

Figure 5.3–8 (*a*) Slope detection with a tuned circuit; (*b*) balanced discriminator circuit; (*c*) frequency-to-voltage characteristic.

4

Noise

4.1 Introduction

Noise, as commonly understood, is a disturbance one "hears," but in telecommunications the word noise is also used as a label for the electrical disturbances that give rise to audible noise in a system. These electrical disturbances also appear as interference in video systems, for example, the white flecks seen on a television picture when the received signal is weak, referred to as a "noisy picture."

Noise can arise in a variety of ways. One obvious example is when a faulty connection exists in a piece of equipment, which, if it is a radio receiver, results in an intermittent or "crackling" type of noise at the output. Such sources of noise can, in principle anyway, be eliminated. Noise also occurs when electrical connections that carry current are made and broken, as, for example, at the brushes of certain types of motors. Again in principle, this type of noise can be suppressed at the source.

Natural phenomena that give rise to noise include electric storms, solar flares, and certain belts of radiation that exist in space. Noise arising from these sources may be more difficult to suppress, and often the only solution is to reposition the receiving antenna to minimize the received noise, while ensuring that reception of the desired signal is not seriously impaired.

Noise is mainly of concern in receiving systems, where it sets a lower limit on the size of signal that can be usefully received. Even when precautions are taken to eliminate noise from faulty connections or that arising from external sources, it is found that certain fundamental sources of noise are present within electronic equipment that limit the receiver sensitivity. One might think that any signal, however small, could simply be amplified up to any desired level. Unfortunately, adding amplifiers to a receiving system also adds noise, and the signal-to-noise ratio, which is the significant quantity, may be degraded by the addition of the amplifiers. Thus the study of the fundamental sources of noise within equipment is essential if the effects of the noise are to be minimized.

4.2 Thermal Noise

It is known that the free electrons within an electrical conductor possess kinetic energy as a result of heat exchange between the conductor and its surroundings. The kinetic energy means that the electrons are in motion, and this motion in turn is randomized through collisions with imperfections in the structure of the conductor. This process occurs in all real conductors and is what gives rise to the conductors' resistance. As a result, the electron density throughout the conductor varies randomly, giving rise to a randomly varying voltage across the ends of the conductor (Fig. 4.2.1). Such a voltage may sometimes be observed in the flickerings of a very sensitive voltmeter. Since the noise arises from thermal causes, it is referred to as *thermal noise* (and also as *Johnson noise,* after its discoverer).

The average or mean noise voltage across the conductor is zero, but the root-mean-square value is finite and can be measured. (It will be recalled that a similar situation occurs for sinusoidal voltage, which has a mean value of zero and a finite rms value.) It is found that the mean-square value of the noise voltage is proportional to the resistance of the conductor, to its absolute temperature, and to the frequency bandwidth of the device measuring (or responding to) the noise. The rms voltage is of course the square root of the mean-square value.

Consider a conductor that has resistance R, across which a true rms measuring voltmeter is connected, and let the voltmeter have an ideal bandpass frequency response of bandwidth B_n as shown in Fig. 4.2.2. The subscript n signifies noise bandwidth, which for the moment may be assumed to be the same as the bandwidth of the ideal filter. The relationship between noise bandwidth and actual frequency response will be developed more fully later. The mean-square voltage measured on the meter is found to be

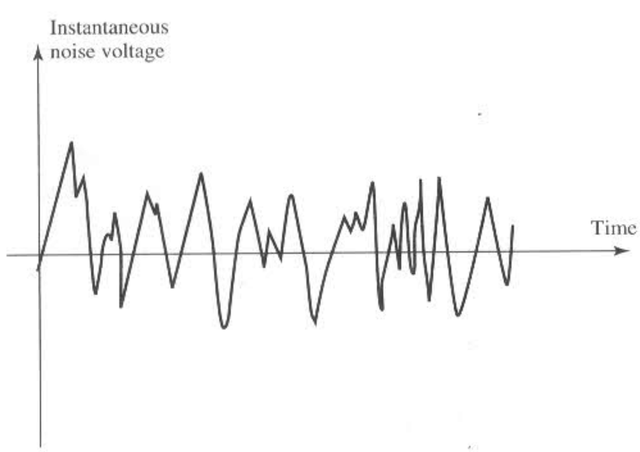

Figure 4.2.1 Thermal noise voltage.

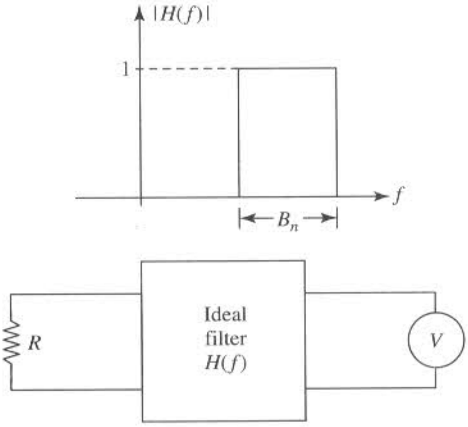

Figure 4.2.2 Measurement of thermal noise.

$$E_n^2 = 4RkTB_n \tag{4.2.1}$$

where E_n = root-mean-square noise voltage, volts
R = resistance of the conductor, ohms
T = conductor temperature, kelvins
B_n = noise bandwidth, hertz
k = Boltzmann's constant
= 1.38×10^{-23} J/K

The equation is given in terms of mean-square voltage rather than root mean square, since this shows the proportionality between the noise power (proportional to E_n^2) and temperature (proportional to kinetic energy).

The rms noise voltage is given by

$$E_n = \sqrt{4RkTB_n} \tag{4.2.2}$$

The presence of the mean-square voltage at the terminals of the resistance R suggests that it may be considered as a generator of electrical noise power. Attractive as the idea may be, thermal noise is not unfortunately a free source of energy. To abstract the noise power, the resistance R would have to be connected to a resistive load, and in thermal equilibrium the load would supply as much energy to R as it receives.

The fact that the noise power cannot be utilized as a free source of energy does not prevent the power being calculated. In analogy with any electrical source, the *available average power* is defined as the maximum average power the source can deliver. For a generator of emf E volts (rms) and internal resistance R, the available power is $E^2/4R$. Applying this to Eq. (4.2.1) gives for the available thermal noise power:

$$P_n = kTB_n \tag{4.2.3}$$

EXAMPLE 4.2.1

Calculate the thermal noise power available from any resistor at room temperature (290 K) for a bandwidth of 1 MHz. Calculate also the corresponding noise voltage, given that $R = 50\ \Omega$.

SOLUTION For a 1-MHz bandwidth, the noise power is

$$P_n = 1.38 \times 10^{-23} \times 290 \times 10^6$$
$$= 4 \times 10^{-15}\ \mathrm{W}$$
$$E_n^2 = 4 \times 50 \times 1.38 \times 10^{-23} \times 290$$
$$= 810^{-13}$$
$$\therefore E_n = \mathbf{0.895\ \mu V}$$

The noise power calculated in Example 4.2.1 may seem to be very small, but it may be of the same order of magnitude as the signal power present. For example, a receiving antenna may typically have an induced signal emf of 1 μV, which is of the same order as the noise voltage.

The thermal noise properties of a resistor R may be represented by the equivalent voltage generator of Fig. 4.2.3(a). This is one of the most useful representations of thermal noise and is widely used in determining the noise performance of equipment. It is best to work initially in terms of E_n^2 rather than E_n, for reasons that will become apparent shortly.

Norton's theorem may be used to find the equivalent current generator and this is shown in Fig. 4.2.3(b). Here, using conductance G (= $1/R$), the rms noise current I_n is given by

$$I_n^2 = 4GkTB_n \tag{4.2.4}$$

It will be recalled that the bandwidth is that of the external circuit, not shown in the source representations, and this must be examined in more detail. Suppose the resistance is left open circuited; then the bandwidth ideally would be infinite, and Eq. (4.2.3) suggests that the open-circuit noise voltage would also be infinite! Two factors prevent this from happening. The first relates to the derivation of the noise energy, which is based on classical thermodynamics and ignores quantum mechanical effects. The quantum mechanical derivation shows that the energy drops off with increasing frequency, and this therefore sets a fundamental limit to the noise power available. However, quantum mechanical effects only become important at

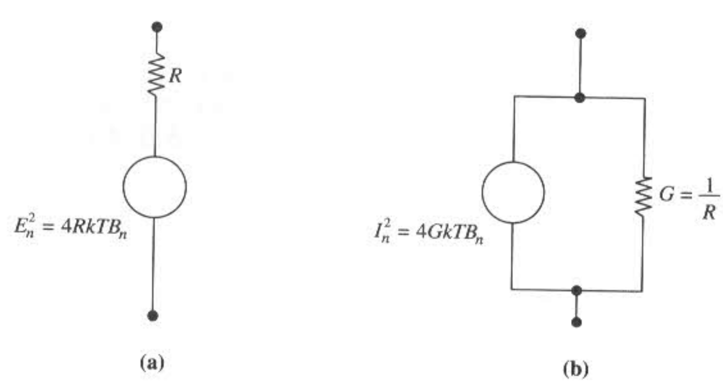

Figure 4.2.3 Equivalent sources for thermal noise: (a) voltage source and (b) current source.

frequencies well into the infrared region. The second and more significant practical factor from the circuit point of view is that *all* real circuits contain reactance (for example, self-inductance and self-capacitance), which sets a finite limit on bandwidth. In the case of the open-circuited resistor, the self-capacitance sets a limit on bandwidth, a situation that is covered in more detail later.

Resistors in Series

Let R_{ser} represent the total resistance of the series chain, where $R_{ser} = R_1 + R_2 + R_3 + \cdots$; then the noise voltage of the equivalent series resistance is

$$\begin{aligned} E_n^2 &= 4R_{ser}kTB_n \\ &= 4(R_1 + R_2 + R_3 + \cdots)kTB_n \\ &= E_{n1}^2 + E_{n2}^2 + E_{n3}^2 + \cdots \end{aligned} \tag{4.2.5}$$

This shows that the total noise voltage *squared* is obtained by summing the mean-square values. Hence the noise voltage of the series chain is given by

$$E_n = \sqrt{E_{n1}^2 + E_{n2}^2 + E_{n3}^2 + \cdots} \tag{4.2.6}$$

Note that simply adding the individual noise voltages would have given the wrong result.

Resistors in Parallel

With resistors in parallel it is best to work in terms of conductance. Thus let G_{par} represent the parallel combination where $G_{par} = G_1 + G_2 + G_3 + \cdots$; then

$$\begin{aligned} I_n^2 &= 4G_{par}kTB_n \\ &= 4(G_1 + G_2 + G_3 + \cdots)kTB_n \\ &= I_{n1}^2 + I_{n2}^2 + I_{n3}^2 + \cdots \end{aligned} \tag{4.2.7}$$

Again, it is to be noted that the mean-square values are added to obtain the total mean-square noise current. Usually, it is more convenient to work in terms of noise voltage rather than current. This is most easily done by first determining the equivalent parallel resistance from $1/R_{par} = 1/R_1 + 1/R_2 + 1/R_3 + \cdots$ and using

$$E_n^2 = 4R_{par}kTB_n \tag{4.2.8}$$

EXAMPLE 4.2.2

Two resistors of 20 and 50 kΩ are at room temperature (290 K). For a bandwidth of 100 kHz, calculate the thermal noise voltage generated by (a) each resistor, (b) the two resistors in series, and (c) the two resistors in parallel.

SOLUTION (a) For the 20-kΩ resistor

$$E_n^2 = 4 \times (20 \times 10^3) \times (4 \times 10^{-21}) \times (100 \times 10^3)$$
$$= 32 \times 10^{-12}\ \text{V}^2$$
$$\therefore E_n = \mathbf{5.66\ \mu V}$$

The voltage for the 50-kΩ resistor may be found by simple proportion:

$$E_n = 5.66 \times \sqrt{\frac{50}{20}}$$
$$= \mathbf{8.95\ \mu V}$$

(b) For the series combination, $R_{\text{ser}} = 20 + 50 = 70$ kΩ. Hence

$$E_n = 5.66 \times \sqrt{\frac{70}{20}}$$
$$= \mathbf{10.59\ \mu V}$$

(c) For the parallel combination, $R_{\text{par}} = \dfrac{20 \times 50}{20 + 50} = 14.29$ kΩ.

$$\therefore E_n = 5.66 \times \sqrt{\frac{14.29}{20}} = \mathbf{4.78\ \mu V}$$

Reactance

Reactances do not generate thermal noise. This follows from the fact that reactance cannot dissipate power. Consider an inductive or capacitive reactance connected in parallel with a resistor R (Fig. 4.2.4). In thermal equilibrium, equal amounts of power must be exchanged; that is, if the resistor supplies thermal noise power P_2 to the reactance, the reactance must supply thermal noise power $P_1 = P_2$ to the resistor. But since the reactance cannot dissipate power, the power P_2 must be zero, and hence P_1 must also be zero.

The effect of reactance on the noise bandwidth must, however, be taken into account, as shown in the next section.

Spectral Densities

Thermal noise falls into the category of power signals as described in Section 2.17, and hence it has a spectral density. As pointed out previously, the bandwidth B_n is a property of the external measuring or receiving system and is

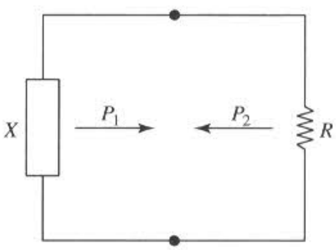

Figure 4.2.4 Power exchange between a reactance and a resistance is $P_1 = P_2 = 0$.

assumed flat so that, from Eq. (4.2.3), the available power spectral density, in watts per hertz, or joules, is

$$\begin{aligned} G_a(f) &= \frac{P_n}{B_n} \\ &= kT \end{aligned} \tag{4.2.9}$$

The spectral density for the mean-square voltage is also a useful function. This has units of volts^2 per hertz and is given by

$$\begin{aligned} G_v(f) &= \frac{E_n^2}{B_n} \\ &= 4RkT \end{aligned} \tag{4.2.10}$$

The spectral densities are flat, that is, independent of frequency, as shown in Fig. 4.2.5, and as a result thermal noise is sometimes referred to as *white noise,* in analogy to white light, which has a flat spectrum. When white noise is passed through a network, the spectral density will be altered by the shape of the network frequency response. The total noise power at the output is found by summing the noise contributions over the complete frequency range, taking into account the shape of the frequency response.

Consider a power spectral response as shown in Fig. 4.2.6. At frequency f_1, the available noise power for an infinitesimally small bandwidth δf about f_1 is $\delta P_{n1} = S_p(f_1)\delta f$. This is so because the bandwidth δf may be assumed flat about f_1, and the available power is given as the product of spectral density (watts/hertz) × bandwidth (hertz). The available noise power is therefore seen to be equal to the area of the shaded strip about f_1. Similar arguments can be applied at frequencies $f_2, f_3, \ldots$, and the total power, given by the sum of all these contributions, is equal to the sum of all these small areas, which is the total area under the curve. More formally, this is equal to the integral of the spectral density function over the frequency range $f = 0$ to $f = \infty$.

A similar argument can be applied to mean-square voltage. The spectral density curve in this case has units of V^2/Hz, and multiplying this by bandwidth δf Hz results in units of V^2, so the area under the curve gives the total mean-square voltage.

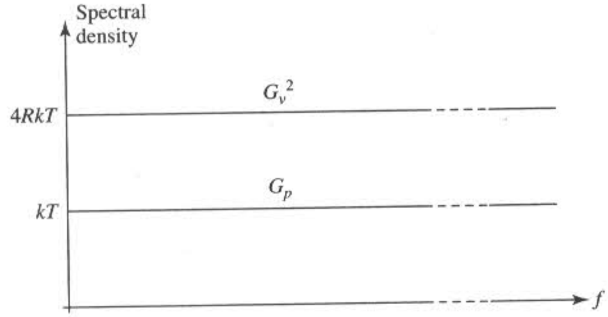

Figure 4.2.5 Thermal noise spectral densities.

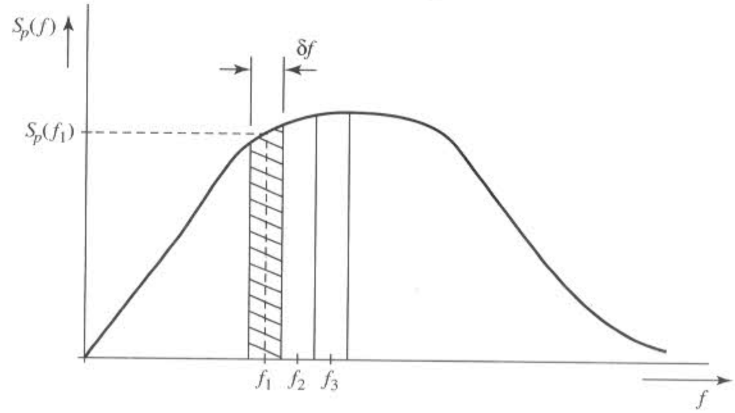

Figure 4.2.6 Nonuniform noise spectral density.

Equivalent Noise Bandwidth

Suppose that a resistor R is connected to the input of an LC filter, as shown in Fig. 4.2.7(a). This represents an input generator of mean-square voltage spectral density $4RkT$ feeding a network consisting of R and the LC filter. Let the transfer function of the network including R be $H(f)$, as shown in Fig. 4.2.7(b). The spectral density for the mean-square output voltage is therefore $4RkT|H(f)|^2$. This follows since $H(f)$ is the ratio of output to input voltage, and here mean-square values are being considered.

From what was shown previously, the total mean-square output voltage is given by the area under the output spectral density curve

$$\begin{aligned} V_n^2 &= \int_0^\infty 4RkT|H(f)|^2 df \\ &= 4RkT \times (\text{area under } |H(f)|^2 \text{ curve}) \end{aligned} \tag{4.2.11}$$

Now the total mean-square voltage at the output can be stated as $V_n^2 = 4RkTB_n$, and equating this with Eq. (4.2.11) gives, for the equivalent noise bandwidth of the network,

$$\begin{aligned} B_n &= \int_0^\infty |H(f)|^2 df \\ &= (\text{area under } |H(f)|^2 \text{ curve}) \end{aligned} \tag{4.2.12}$$

As a simple example consider the circuit of Fig. 4.2.8, which consists of a resistor in parallel with a capacitor. The capacitor may in fact be the self-capacitance of the resistor, or an external capacitor, for example, the input capacitance of the voltmeter used to measure the noise voltage across R.

The transfer function of the RC network is

$$|H(f)| = \frac{1}{\sqrt{1 + (\omega CR)^2}} \tag{4.2.13}$$

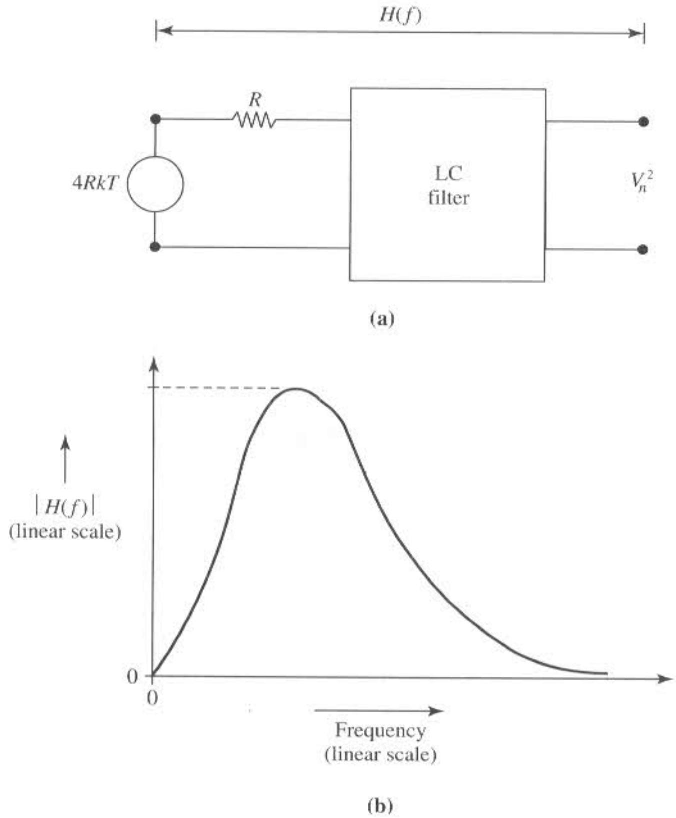

Figure 4.2.7 (a) Filtered noise and (b) the transfer function of the filter including R.

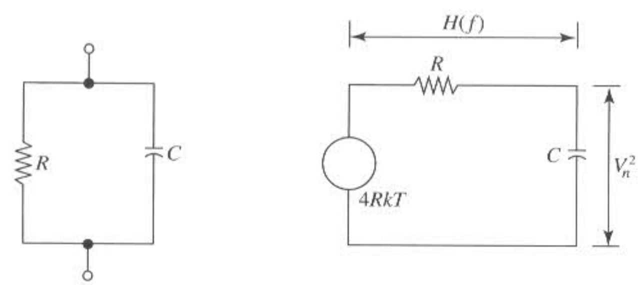

Figure 4.2.8 *RC* network and its transfer function used in determining noise bandwidth.

The equivalent noise bandwidth of the RC network is found using Eq. (4.2.12) as

$$B_n = \int_0^\infty |H(f)|^2 df$$

$$= \frac{1}{4RC} \tag{4.2.14}$$

(Details of the integration are left as an exercise for the reader.) The mean-square output voltage is given by

$$V_n^2 = 4RkT \times \frac{1}{4RC}$$
$$= \frac{kT}{C} \qquad (4.2.15)$$

This is a surprising result. It shows that the mean-square output voltage is independent of R, even though it originates from R, and it is inversely proportional to C, even though C does not generate noise.

A second example is that of the tuned circuit shown in Fig. 4.2.9. Here the capacitor is assumed lossless, and the inductor has a series resistance r that generates thermal noise.

The transfer function in this case is

$$|H(f)| = \left|\frac{X_c}{Z_s}\right| \qquad (4.2.16)$$

where $Z_s = r(1 + jyQ)$ is the impedance of the series tuned circuit as given by Eq. (1.3.10) and $X_c = 1/j\omega C$ is the reactance of C. As before, the equivalent noise bandwidth is found by solving Eq. (4.2.12).

Consider first the situation where the circuit is resonant at f_0, and the noise is restricted to a small bandwidth $\Delta f \ll f_0$ about the resonant frequency. The transfer function is then approximated by $|H(f)| \cong 1/\omega_0 Cr = Q$, and the area under the $|H(f)|^2$ curve over a small constant bandwidth δf is $Q^2\delta f$. Hence the mean-square noise voltage is

$$V_n^2 = 4rkTB_n \qquad (4.2.17)$$
$$= 4rQ^2kT\delta f$$
$$= 4R_D kT\Delta f$$

Here, use is made of the relationship $Q^2 r = R_D$ developed in Section 1.4. This is an important result, because the bandwidth is often limited in practice to some small percentage about f_0. An example will illustrate this.

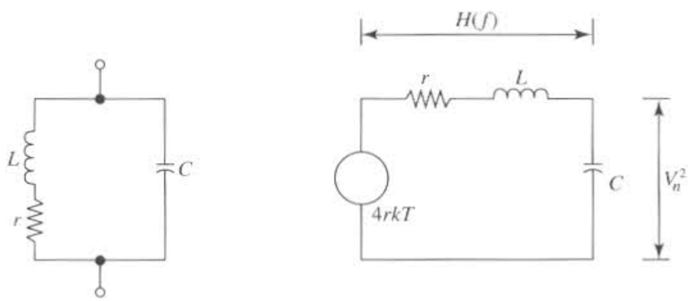

Figure 4.2.9 Tuned circuit and its transfer function used in determining noise bandwidth.

EXAMPLE 4.2.3

The parallel tuned circuit at the input of a radio receiver is tuned to resonate at 120 MHz by a capacitance of 25 pF. The Q-factor of the circuit is 30. The channel bandwidth of the receiver is limited to 10 kHz by the audio sections. Calculate the effective noise voltage appearing at the input at room temperature.

SOLUTION

$$R_D = \frac{Q}{\omega_o C}$$

$$= \frac{30}{2 \times \pi \times 120 \times 10^6 \times 25 \times 10^{-12}}$$

$$= 1.59 \text{ k}\Omega$$

$$\therefore V_n = \sqrt{4 \times 1.59 \times 10^3 \times 4 \times 10^{-21} \times 10^4}$$

$$= \mathbf{0.5\ \mu V}$$

Where the complete frequency range 0 to ∞ has to be taken into account, the integral becomes much more difficult to solve, and only the result will be given here. This is

$$B_n = \frac{1}{4R_D C} \tag{4.2.18}$$

where R_D is the dynamic resistance of the tuned circuit.

The noise bandwidth can be expressed as a function of the -3-dB bandwidth of the circuit. From Eq. (1.3.17), $B_{3\text{ dB}} = f_o/Q$, and from Eq. (1.4.4) $R_D = Q/\omega_o C$. Combining these expressions along with that for the noise bandwidth gives

$$B_n = \frac{\pi}{2} B_{3\text{ dB}} \tag{4.2.19}$$

By postulating that the noise originates from a resistor R_D and is limited by the bandwidth B_n, the mean-square voltage at the output can be expressed as

$$V_n^2 = 4R_D kT \times \frac{1}{4R_D C}$$

$$= \frac{kT}{C} \tag{4.2.20}$$

In the foregoing, to simplify the analysis it was assumed that the Q-factor remained constant, independent of frequency. This certainly would not be true for the range zero to infinity, but the end result still gives a good indication of the noise expected in practice.

For most radio receivers the noise is generated at the front end (antenna input) of the receiver, while the output noise bandwidth is determined by the audio sections of the receiver. The equivalent noise bandwidth is equal to the area under the normalized power-gain/frequency curve for the low-frequency sections. By normalized is meant that the curve is scaled such that the maximum value is equal to unity. Usually this information is available in the form of a frequency response curve showing output in decibels relative to maximum and with frequency plotted on a logarithmic scale, as sketched in Fig. 4.2.10(a).

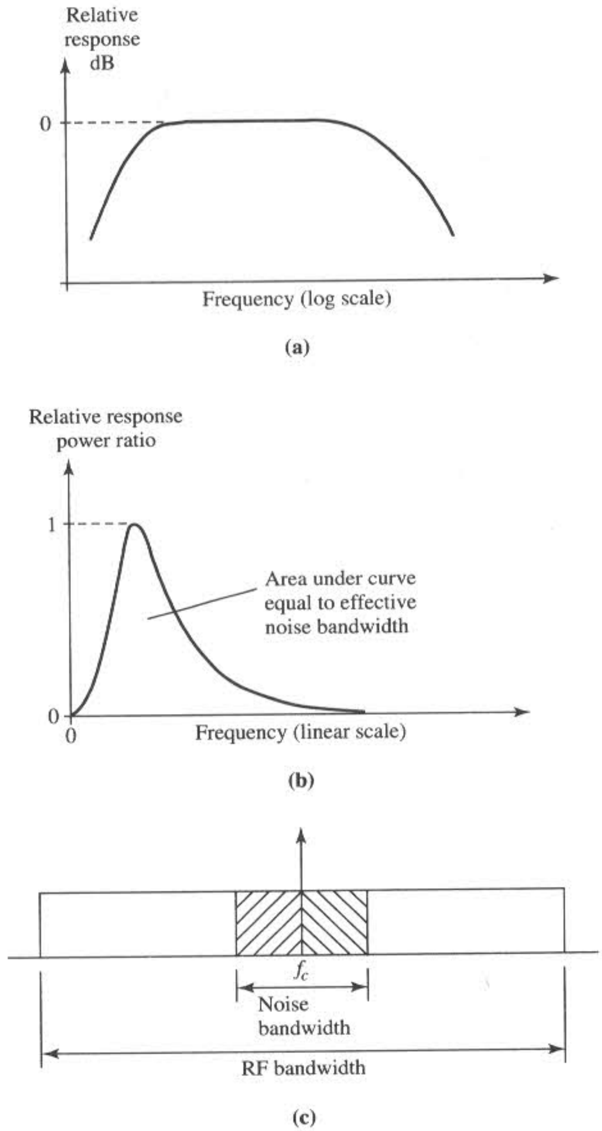

Figure 4.2.10 (a) Amplifier frequency response curve. (b) Curve of (a) using linear scales. (c) Noise bandwidth of a double-sideband receiver.

Before determining the area under the curve, the decibel axis must be converted to a linear power-ratio scale and the frequency axis to a linear frequency scale, as shown in Fig. 4.2.10(b). The equivalent noise bandwidth is then equal to the area under this curve for a single-sideband receiver. Where the receiver is of the double-sideband type, then the noise bandwidth appears on both sides of the carrier and is effectively doubled. This is shown in Fig. 4.2.10(c).

4.3 Shot Noise

Shot noise is a random fluctuation that accompanies any direct current crossing a potential barrier. The effect occurs because the carriers (holes and electrons in semiconductors) do not cross the barrier simultaneously, but rather with a random distribution in the timing for each carrier, which gives rise to a random component of current superimposed on the steady current. In the case of bipolar junction transistors, the bias current crossing the forward biased emitter–base junction carries shot noise. With vacuum tubes the electrons emitted from the cathode have to overcome a potential barrier that exists between cathode and vacuum. The name *shot noise* was first coined in connection with tubes, where the analogy was made between the electrons striking the plate and lead shot from a gun striking a target.

Although it is always present, shot noise is not normally observed during measurement of direct current because it is small compared to the dc value; however, it does contribute significantly to the noise in amplifier circuits. The idea of shot noise is illustrated in Fig. 4.3.1.

Shot noise is similar to thermal noise in that its spectrum is flat (except in the high microwave frequency range). The mean-square noise component is proportional to the dc flowing, and for most devices the mean-square, shot-noise current is given by

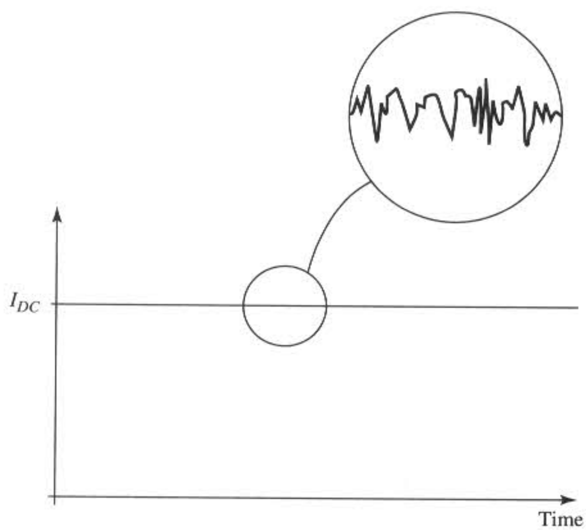

Figure 4.3.1 Shot noise.

$$I_n^2 = 2 I_{dc} q_e B_n \quad \text{amperes}^2 \tag{4.3.1}$$

where I_{dc} is the direct current in amperes, q_e the magnitude of electron charge ($= 1.6 \times 10^{-19}$ C), and B_n is the equivalent noise bandwidth in hertz.

EXAMPLE 4.3.1

Calculate the shot noise component of current present on a direct current of 1 mA flowing across a semiconductor junction, given that the effective noise bandwidth is 1 MHz.

SOLUTION

$$\begin{aligned} I_n^2 &= 2 \times 10^{-3} \times 1.6 \times 10^{-19} \times 10^6 \\ &= 3.2 \times 10^{-16}\ \text{A}^2 \\ \therefore I_n &= \mathbf{18\ nA} \end{aligned}$$

4.4 Partition Noise

Partition noise occurs wherever current has to divide between two or more electrodes and results from the random fluctuations in the division. It would be expected therefore that a diode would be less noisy than a transistor (other factors being equal) if the third electrode draws current (such as base or gate current). It is for this reason that the input stage of microwave receivers is often a diode circuit, although, more recently, gallium arsenide field-effect transistors, which draw zero gate current, have been developed for low-noise microwave amplification. The spectrum for partition noise is flat.

4.5 Low Frequency or Flicker Noise

Below frequencies of a few kilohertz, a component of noise appears, the spectral density of which increases as the frequency decreases. This is known as *flicker noise* (and sometimes as $1/f$ noise). In vacuum tubes it arises from slow changes in the oxide structure of oxide-coated cathodes and from the migration of impurity ions through the oxide. In semiconductors, flicker noise arises from fluctuations in the carrier densities (holes and electrons), which in turn give rise to fluctuations in the conductivity of the material. It follows therefore that a noise voltage will be developed whenever direct current flows through the semiconductor, and the mean-square voltage will be proportional to the square of the direct current. Interestingly enough, although flicker noise is a low-frequency effect, it plays an important part in limiting the sensitivity of microwave diode mixers used for Doppler radar systems. This is because, although the input frequencies to the mixer are in the microwave range, the Doppler frequency output is in the low (audio-frequency) range, where flicker noise is significant.

4.6 Burst Noise

Another type of low-frequency noise observed in bipolar transistors is known as *burst noise,* the name arising because the noise appears as a series of bursts at two or more levels (rather like noisy pulses). When present in an audio system, the noise produces popping sounds, and for this reason is also known as "popcorn" noise. The source of burst noise is not clearly understood at present, but the spectral density is known to increase as the frequency decreases.

4.7 Avalanche Noise

The reverse-bias characteristics of a diode exhibit a region where the reverse current, normally very small, increases extremely rapidly with a slight increase in the magnitude of the reverse-bias voltage. This is known as the *avalanche region* and comes about because the holes and electrons in the diode's depletion region gain sufficient energy from the reverse-bias field to ionize atoms by collision. The ionizing process means that additional holes and electrons are produced, which in turn contribute to the ionization process, and thus the descriptive term *avalanche.*

The collisions that result in the avalanching occur at random, with the result that large noise spikes are present in the avalanche current. In diodes such as zener diodes, which are used as voltage reference sources, the avalanche noise is a nuisance to be avoided. However, avalanche noise is put to good use in noise measurements, as described in Section 4.19. The spectral density of avalanche noise is flat.

4.8 Bipolar Transistor Noise

Bipolar transistors exhibit all the sources of noise discussed previously, that is, thermal, shot, partition, flicker, and burst noise. The thermal noise is generated by the bulk or extrinsic resistances of the electrodes, but the only significant component is that generated by the extrinsic base resistance. It should be emphasised at this point that the small-signal equivalent resistances for the base – emitter and the base – collector junctions do not generate thermal noise, but they do enter into the noise calculations made using the small-signal equivalent circuit for the transistor.

The bias currents in the transistor show shot noise and partition noise, and, in addition, the flicker and burst noise components are usually associated with the base current.

4.9 Field-effect Transistor Noise

In field-effect transistors (both JFETs and MOSFETs), the main source of noise is the thermal noise generated by the physical resistance of the drain–source channel. Flicker noise also originates in this channel. Additionally, there will be shot noise associated with the gate leakage current. This will develop a noise component of voltage across the signal–source impedance and is only significant where this impedance is very high (in the megohm range).

4.10 Equivalent Input Noise Generators and Comparison of BJTs and FETs

An amplifier may be represented by the block schematic of Fig. 4.10.1(a), in which a noisy amplifier is shown and where the source and load resistances generate thermal noise. The circuit may be redrawn as shown in Fig. 4.10.1(b) in which the amplifier itself is considered to be noiseless, the amplifier noise being represented by *fictitious noise generators* $V_{na} = \sqrt{4R_n kT_o B_n}$ and $I_{na} = \sqrt{2q_e I_{EQ} B_n}$ at the input. Here, B_n is the equivalent noise bandwidth of the amplifier in hertz, T_o is room temperature in kelvins, k is Boltzmann's constant = 1.38×10^{-23} J/K, and $q_e = 1.6 \times 10^{-19}$ C is the magnitude of the electron charge. These terms have all been defined previously. What is new here is the *fictitious* resistance R_n ohms, known as the *equivalent input noise resistance* of the amplifier, and I_{EQ} amperes, the *equivalent input shot noise current.* Both these parameters have to be calculated or specified for a transistor under given operating conditions.

The noise generated by the load resistance R_L is generally very small compared to the other sources and is assumed to be negligible, so this is dropped from the equivalent circuit. The thermal noise generated by the signal–source resistance R_s is generally significant and must be taken into account as shown in Fig. 4.10.1(b).

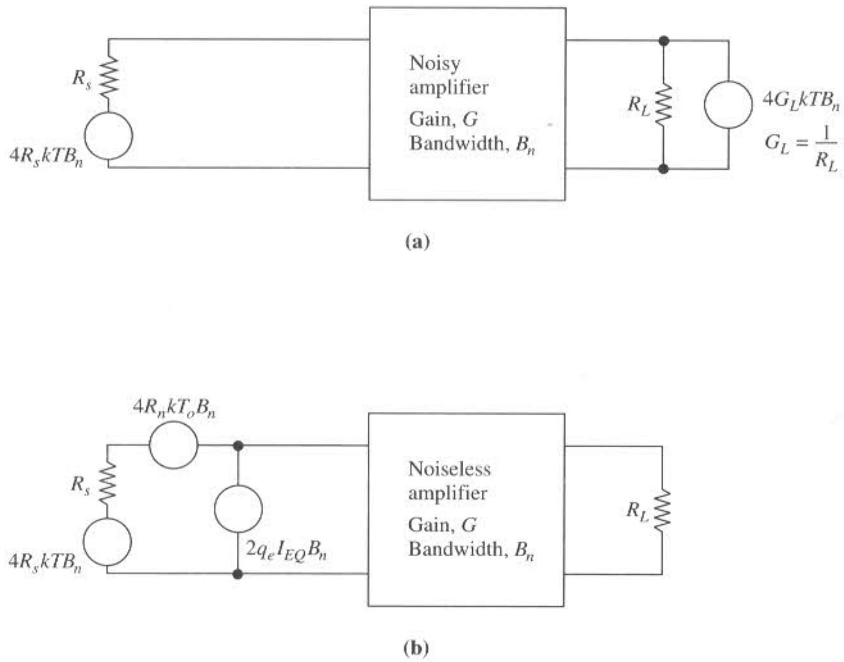

Figure 4.10.1 (a) Noisy amplifier and (b) the equivalent input noise generators.

4.20 Narrowband Band-pass Noise

Band-pass filtering of signals arises in many situations, the basic arrangement being shown in Fig. 4.20.1. The filter has an equivalent noise bandwidth B_N (see Section 4.2) and a center frequency f_c. A narrowband system is one in which the center frequency is much greater than the bandwidth, which is the situation to be considered here.

The signal source is shown as a voltage generator of internal resistance R_S. System noise is referred to the input as a thermal noise source at a noise temperature T_s. The available power spectral density is, from Eq. (4.2.9),

$$G_a(f) = kT_s \tag{4.20.1}$$

For the ideal band-pass system shown, the spectral density is not altered by transmission through the filter, but the filter bandwidth determines the available noise power as kT_sB_N. So far, this is a result that has already been encountered in general. An alternative description of the output noise, however, turns out to be very useful, especially in connection with the modulation systems described in later chapters. The waveforms of input and output noise voltages are shown in Fig. 4.20.2.

The output waveform has the form of a modulated wave and can be expressed mathematically as

$$n(t) = A_n(t)\cos(\omega_c t + \phi_n(t)) \tag{4.20.2}$$

This represents the noise in terms of a randomly varying voltage envelope $A_n(t)$ and a random phase angle $\phi_n(t)$. These components are readily identified as part of the waveform, as shown in Fig. 4.20.2, but an equivalent although not so apparent expression can be obtained by trigonometric expansion of the output waveform as

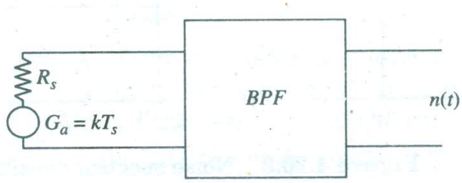

Figure 4.20.1 Noise in a band-pass system.

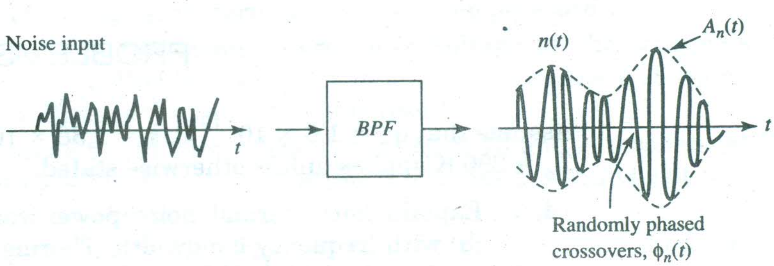

Figure 4.20.2 Input and output noise waveforms for a band-pass system.

$$n(t) = n_I(t) \cos \omega_c t - n_Q(t) \sin \omega_c t \tag{4.20.3}$$

Here, $n_I(t)$ is a random noise voltage termed the *in-phase component* because it multiplies a cosine term used as a reference phasor, and $n_Q(t)$ is a similar random voltage termed the *quadrature component* because it multiplies a sine term, which is therefore 90° out of phase, or in quadrature with, the reference phasor. The reason for using this form of equation is that, when dealing with modulated signals, the output noise voltage is determined by these two components (this is described in detail in later chapters on modulation). The two noise voltages $n_I(t)$ and $n_Q(t)$ appear to modulate a carrier at frequency f_c and are known as the *low-pass equivalent noise voltages*. The carrier f_c may be chosen anywhere within the passband, but the analysis is simplified by placing it at the center as shown. This is illustrated in Fig. 4.20.3

A number of important relationships exist between $n_I(t)$ and $n_Q(t)$ and $n(t)$, some of which will be stated here without proof. All three have similar noise characteristics and $n_I(t)$ and $n_Q(t)$ are uncorrelated. Of particular importance in later work on modulation is that where the power spectral density of $n(t)$ is $G_a(f) = kT_s$ the power spectral densities for $n_I(t)$ and $n_Q(t)$ are

$$G_I(f) = G_Q(f) = 2kT_s \tag{4.20.4}$$

This important result, which is illustrated in Fig. 4.20.3, will be encountered again in relation to modulated signals.

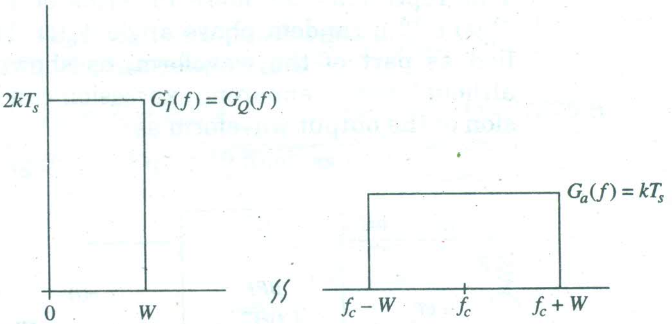

Figure 4.20.3 Noise spectral densities.

PROBLEMS

Assume that $q_e = 1.6 \times 10^{-19}$ C, $k = 1.38 \times 10^{-23}$ J/K and room temperature $T_o = 290$ K applies unless otherwise stated.

4.1. Explain how thermal noise power varies (a) with temperature and (b) with frequency bandwidth. Thermal noise from a resistor is measured as 4×10^{-17} W for a given bandwidth and at a temperature of 20°C. What will the noise power be when the temperature is changed to (c) 50°C; (d) 70 K?

8.14 Noise in AM Systems

For AM systems that carry analog-type message signals, the signal-to-noise ratio is the most commonly used measure of performance. As discussed in Chapter 4, all the noise generated within the receiver can be referred to the receiver input, which makes it easy to compare receiver noise, antenna noise, and received signal. The antenna and receiver noise powers can be added, so the receiving system can be modeled as shown in Fig. 8.14.1.

From the point of view of determining the signal-to-noise ratio, the additional information needed is the bandwidth of the system. From Fig. 8.14.1, it is seen that between the antenna and the detector stage there is the bandwidth of the RF stages, followed by the bandwidth of the IF stages. Normally, the bandwidth of the IF stages is very much smaller than the RF bandwidth, and this will be the bandwidth that determines the noise reaching the detector. Following the detector, the bandwidth is that of the *baseband,* which is that required by the modulating signal. To distinguish these clearly, the baseband bandwidth will be denoted by W and the IF bandwidth by B_{IF}. Also, for AM systems it may be assumed that $B_{IF} \cong 2\,W$. In Section 4.2, the concept of equivalent noise bandwidth was explained, and it will be further assumed that B_{IF} and W refer to the equivalent noise bandwidths.

Equation (4.20.3), which gives the noise output from a bandpass system, is rewritten here as

$$n_{IF}(t) = n_I(t)\cos 2\pi f_{IF}t - n_Q(t)\sin \omega_{IF}t \tag{8.14.1}$$

When the noise waveform is passed through the detector, the resulting noise output is very much dependent on whether or not a carrier is present and on the size of the carrier. When evaluating the signal-to-noise ratio, a carrier must be present. Let the modulated carrier be represented by

$$\begin{aligned} e(t) &= E_{c\,\max}(1 + m\cos 2\pi f_m t)\cos 2\pi f_{IF}t \\ &= A_c(t)\cos 2\pi f_{IF}t \end{aligned} \tag{8.14.2}$$

The input to the detector is therefore

$$\begin{aligned} e_{\det}(t) &= e(t) + n_{IF}(t) \\ &= (A_c(t) + n_I(t))\cos 2\pi f_{IF}t - n_Q(t)\sin 2\pi f_{IF}t \end{aligned} \tag{8.14.3}$$

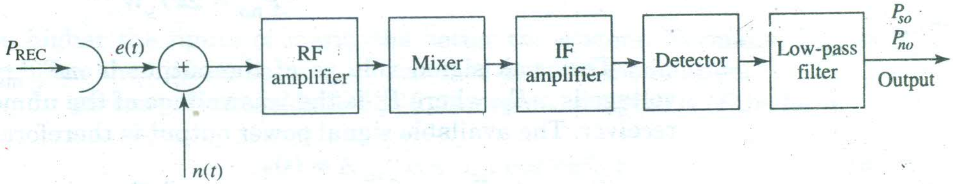

Figure 8.14.1 Model of an AM receiver showing noise added.

The AM envelope detector recovers the envelope of this waveform, and therefore the waveform needs to be expressed as a cosine carrier wave of the form $R(t)\cos(\omega_{IF}t + \psi(t))$. Derivation of the amplitude and phase angle terms is left as an exercise for the student (see Problem 8.51). For the present application, the phase angle $\psi(t)$ can be ignored, and the amplitude term is given by

$$R(t) = \sqrt{[(A_c(t) + n_I(t)]^2 + [n_Q(t)]^2} \tag{8.14.4}$$

Although complete, this expression for $R(t)$ needs to be simplified to get a clearer picture of how the noise adds to the output. For many situations, it can be assumed that the AM carrier is much greater than the noise voltage for most of the time (remembering that the noise is random and that occasional large spikes will be encountered that are greater than the carrier). With this assumption, $R(t)$ simplifies to

$$\begin{aligned} R(t) &= \sqrt{[A_c(t)]^2 + 2A_c(t)n_I(t) + [n_I(t)]^2 + [n_Q(t)]^2} \\ &\cong \sqrt{[A_c(t)]^2 + 2A_c(t)n_I(t)} \end{aligned} \tag{8.14.5}$$

A further expansion and simplification of the square-root term can be made using the binomial theorem.

$$\begin{aligned} R(t) &= \sqrt{[A_c(t)]^2 + 2A_c(t)n_I(t)} \\ &= A_c(t)\left(1 + 2\frac{n_I(t)}{A_c(t)}\right)^{1/2} \\ &\cong A_c(t) + n_I(t) \end{aligned} \tag{8.14.6}$$

For sinusoidal modulation, this becomes

$$R(t) = E_{c\,\max} + mE_{c\,\max}\cos 2\pi f_m t + n_I(t) \tag{8.14.7}$$

The envelope is seen to consist of a dc term, the modulating signal voltage, and the noise voltage $n_I(t)$. As shown in Section 8.11, the dc output from the detector is blocked so that only the ac components contribute to the final output. For the noise, the available power spectral density is given by Eq. (4.20.4) as $2kT_s$, and hence the available noise power output is

$$P_{no} = 2kT_sW \tag{8.14.8}$$

The peak signal voltage at the output is $mE_{c\,\max}$, and hence the rms voltage is mE_c, where E_c is the rms voltage of the unmodulated carrier at the receiver. The available signal power output is therefore

$$P_{so} = \frac{m^2E_c^2}{4R_{\text{out}}} \tag{8.14.9}$$

The output signal-to-noise ratio is

$$\left(\frac{\mathrm{S}}{\mathrm{N}}\right)_o = \frac{P_{so}}{P_{no}}$$

$$= \frac{m^2E_c^2}{8R_{\text{out}}kT_sW} \tag{8.14.10}$$

Standard practice is to compare the output signal-to-noise ratio to a reference ratio, which is the signal-to-noise ratio at the detector input *but with the noise calculated for the baseband bandwidth* (W in this case). The noise power spectral density at the detector input is kT_s, and so the reference noise power

$$P_{n\,\text{REF}} = kT_sW \tag{8.14.11}$$

The available signal power from a source with internal resistance R_s is [see Eq. (8.6.3)]

$$P_R = \frac{E_c^2}{4R_s}\left(1 + \frac{m^2}{2}\right) \tag{8.14.12}$$

Hence the reference signal-to-noise ratio is

$$\left(\frac{\mathrm{S}}{\mathrm{N}}\right)_{\text{REF}} = \frac{P_R}{P_{n\,\text{REF}}}$$

$$= \frac{E_c^2(1 + m^2/2)}{4R_SkT_sW} \tag{8.14.13}$$

A figure of merit that is used is the ratio of these two signal-to-noise ratios. Denoting this by R_{AM}, then

$$R_{\text{AM}} = \frac{(S/N)_o}{(S/N)_{\text{REF}}}$$

$$= \frac{m^2}{(2 + m^2)}\frac{R_s}{R_{\text{out}}} \tag{8.14.14}$$

The higher the figure of merit, the better the system. Normally, $R_{out} \cong R_s$ and hence the highest value is $\frac{1}{3}$, achieved at 100% modulation.

In the case of sinusoidal DSBSC, the received signal is of the form

$$e(t) = E_{\max} \cos \omega_m t \cos 2\pi f_{IF}t \tag{8.14.15}$$

where $E_{\max}$ is the peak value of the received signal. The input to the detector is therefore

$$
\begin{aligned}
e_{\det}(t) &= e(t) + n(t) \\
&= (E_{\max} \cos \omega_m t + n_I(t)) \cos 2\pi f_{IF} t - n_Q(t) \sin \omega_{IF} t \\
&= A(t) \cos \omega_{IF} t - n_Q(t) \sin \omega_{IF} t
\end{aligned} \tag{8.14.16}
$$

where $A(t) = (E_{\max} \cos \omega_m t + n_I(t))$. With DSBSC a different type of detection is used, known as *coherent detection*. Demodulation of the DSBSC signal utilizes a balanced mixer (or similar circuit) as described in Section 5.10. A locally generated carrier is required that is exactly locked onto the incoming carrier $\cos \omega_{IF} t$, and the two signals are fed into the balanced mixer. One complication with DSBSC detection (which also applies to SSB detection as described in Chapter 9) is generating the local carrier. However, methods are available for achieving this, and as a result the output of the balanced demodulator is

$$
e_{\text{out}}(t) = k e_{\det}(t) \cos \omega_{IF} t \tag{8.14.17}
$$

where k is a constant of the multiplier circuit. Multiplying this out in full, which is left as an exercise for the student, results in

$$
e_{\text{out}}(t) = \frac{k}{2}(E_{\max} \cos \omega_m t + n_I(t) + \text{high frequency terms} \tag{8.14.18}
$$

Low-pass filtering following the balanced demodulator removes the high-frequency terms, leaving, as the baseband output,

$$
e_{BB}(t) = \frac{k}{2}(E_{\max} \cos \omega_m t + n_I(t)) \tag{8.14.19}
$$

Thus, apart from the multiplying constant $k/2$ and the absence of a dc term, this output is the same as that given in Eq. (8.14.7) for the standard AM case. The $k/2$ factor is common to signal and noise and can be ignored. It should be noted, however, that, whereas the standard AM result requires that the carrier be much greater than the noise, this approximation is not required for the DSBSC case. Equation (8.14.10) applies for the output signal-to-noise ratio, but with $E_{\max}$ replacing $mE_{c\max}$, as seen by comparing eqs. (8.14.19) and (8.14.7). The result is

$$
\begin{aligned}
\left(\frac{\mathrm{S}}{\mathrm{N}}\right)_o &= \frac{(E_{\max}/\sqrt{2})^2}{8R_{\text{out}}\, kT_s W} \\
&= \frac{E^2_{\max}}{16R_{\text{out}}\, kT_s W}
\end{aligned} \tag{8.14.20}
$$

For the reference signal-to-noise ratio, the reference noise is $P_{n\,\text{REF}} = kT_s W$ as given by Eq. (8.14.11). The rms voltage of the received (DSBSC) signal $E_{\max} \cos \omega_m t \cos \omega_{IF} t$ is $E_{\max}/2$ (note *not* $E_{\max}/\sqrt{2}$,) as is readily checked from ac circuit theory. The available signal power at the input is therefore

$$P_R = \frac{(E_{\max}/2)^2}{4R_s}$$

$$= \frac{E^2_{\max}}{16\,R_s} \tag{8.14.21}$$

The reference signal-to-noise is therefore

$$\left(\frac{S}{N}\right)_{REF} = \frac{E^2_{\max}}{16R_S kT_s W} \tag{8.14.22}$$

The figure of merit is therefore

$$R_{DSBSC} = \frac{(S/N)_o}{(S/N)_{REF}}$$

$$= \frac{R_s}{R_{out}} \tag{8.14.23}$$

For $R_{out} \cong R_s$, the figure of merit is unity, which is three times better than the best that can be achieved for R_{AM}. It will be shown in Chapter 9 that single-sideband (SSB) transmission also has a unity figure of merit, and since this requires half the bandwidth of a DSBSC signal, it is the preferred method of AM carrier transmission (apart from AM broadcast applications, where simplicity of receiver design has established standard AM as the preferred method).

PROBLEMS

8.1 A sinusoidal carrier is amplitude modulated by a square wave that has zero dc component and a peak-to-peak value of 2 V. The periodic time of the square wave is 0.5 ms. The carrier amplitude is 2.5 V, and its frequency is 10 kHz. Write out the equations for the modulating signal, the carrier, and the modulated wave, and plot these functions over a time base equal to twice the periodic time of the square wave.

8.2. A sinusoidal carrier is amplitude modulated by a triangular wave. The triangular wave has zero mean value and is an even function, the first quarter-cycle being described by $-1 + 8t$ volts, with t in milliseconds. The periodic time of the triangular wave is 0.5 ms. The carrier amplitude is 2.5 V, and its frequency is 100 kHz. Write out the equations for the modulating signal, the carrier, and the modulated wave, and plot these functions over a time base equal to twice the periodic time of the triangular wave.

8.3. Calculate the modulation index for each of the modulated waves in Problems 8.1 and 8.2.

It was stated that most types of FM detectors require amplitude limiting in order to function properly. The ratio detector (see Section 10.14) is an exception to this, because it has a degree of inherent limiting built into it. In critical applications it is still necessary to provide additional limiting, but the ratio detector performs well enough otherwise.

10.17 Noise in FM Systems

Noise in an FM receiver can be referred to the input as shown in Fig. 10.17.1(a). With FM receivers the limiter stages described in Section 10.16 help to reduce impulse-type noise, so FM has this advantage over AM.

Another advantage arises from the nature of the noise modulation process. The noise at the receiver input cannot directly frequency modulate the incoming carrier since its frequency is fixed at a distant transmitter, which may in fact be crystal controlled. The noise phase modulates the carrier at the receiver and, as will be shown, this leads to a reduction in output compared to the AM situation.

An important advantage with FM reception is that an improvement in signal-to-noise ratio can be achieved by increasing the frequency deviation. This requires an increased bandwidth, but at least the option is there for an exchange of bandwidth for signal-to-noise ratio. This aspect of FM reception will be explained in detail in this section.

Figure 10.17.1(b) shows the signal and noise voltages at the FM detector. The noise, having passed through a band-pass filter, can be represented by narrow-band noise as described in Section 4.20. This will be analyzed shortly. At this point it need only be noted that the power spectral density as given by Eq. (4.20.1) is kT_s, and hence the available noise power at the detector input for a bandwidth W is $P_{n\text{REF}} = kT_sW$. The rms signal voltage is $E_c = E_{c\,\max}/\sqrt{2}$, and therefore the signal power at the detector input is $P_R = E_{c\,\max}^2/(8R_s)$. The reference signal-to-noise ratio, (introduced in Eq. [8.14.13]), is for the FM case

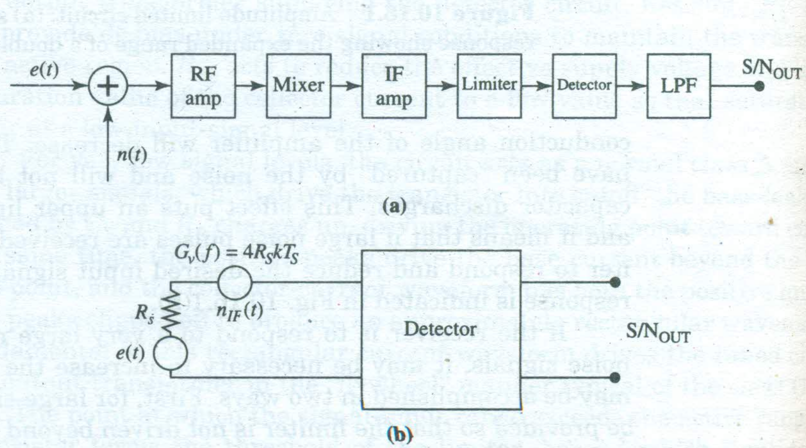

Figure 10.17.1 (a) Block schematic for an FM receiver. (b) Signal and noise at the detector.

$$\left(\frac{\mathrm{S}}{\mathrm{N}}\right)_{\mathrm{REF}} = \frac{P_R}{P_{n\mathrm{REF}}}$$

$$= \frac{E_{c\,\max}^2}{8R_S kT_s W} \tag{10.17.1}$$

Here again it is emphasized that, although the reference signal-to-noise ratio is determined at the input side of the detector, the bandwidth W is that determined by the LPF at the output side.

Consider now the signal output from the band-pass filter in the absence of noise, when a sinusoidally modulated carrier is received. The instantaneous frequency is

$$f_i = f_{IF} + \Delta f \cos \omega_m t \tag{10.17.2}$$

The FM detector converts this to a signal output given by

$$v_m(t) = C\Delta f \cos \omega_m t \tag{10.17.3}$$

where C, a constant, is the frequency-to-voltage coefficient of the detector. The rms voltage output is $E_m = C\Delta f/\sqrt{2}$, and hence the available power output is

$$P_{so} = \frac{(C\Delta f)^2}{8R_{\mathrm{out}}} \tag{10.17.4}$$

The next part of the analysis requires finding the noise at the output of the detector. The noise voltage output from a band-pass system is given by Eq. (4.20.3), which is rewritten here as

$$n_{IF}(t) = n_1(t) \cos \omega_{IF} t - n_Q(t) \sin \omega_{IF} t \tag{10.17.5}$$

The incoming carrier in this case is frequency modulated and can be written as

$$e(t) = E_{c\,\max} \cos (\omega_{IF} t + \phi_m(t)) \tag{10.17.6}$$

where $\phi_m(t)$ is the equivalent phase modulation produced by the signal (see Section 10.9). An analysis of the detector output when noise and modulating signal are present *simultaneously* is very complicated, but fortunately the additional noise terms in the output, resulting from the interaction of $\phi_m(t)$ and the noise modulation, lie outside the baseband. As a result, the noise analysis can be made assuming that an unmodulated carrier is received. In this case, the input to the detector is

$$\begin{aligned} e_{\mathrm{det}}(t) &= e_c(t) + n_{IF}(t) \\ &= (E_{c\,\max} + n_I(t)) \cos \omega_{IF} t - n_Q(t) \sin \omega_{IF} t \end{aligned} \tag{10.17.7}$$

This should be compared with Eq. (8.14.3) for the AM case. As with the AM case, the waveform can be expressed in an equivalent form: $e_{det}(t) = R(t)\cos(\omega_{IF}t + \psi(t))$; but in this situation it is the phase angle $\psi(t)$, rather than the envelope $R(t)$, that is of interest. A trigonometric analysis of the equation gives

$$\psi(t) = \tan^{-1}\frac{n_Q(t)}{E_{c\max} + n_I(t)} \tag{10.17.8}$$

Also, as in the AM case, the analysis will be limited to the situation where the carrier amplitude is much greater than the noise voltage for most of the time. Under these circumstances the noise phase angle becomes

$$\psi(t) \cong \tan^{-1}\frac{n_Q(t)}{E_{c\max}}$$
$$\cong \frac{n_Q(t)}{E_{c\max}} \tag{10.17.9}$$

The total carrier angle is $\theta(t) = \omega_{IF}t + \psi(t)$, and from Eq. (10.9.2) the equivalent frequency modulation is

$$f_{ieq}(t) = f_{IF} + \frac{1}{2\pi}\frac{d\psi(t)}{dt}$$
$$= f_{IF} + \frac{\dot{n}_Q(t)}{2\pi E_{c\max}} \tag{10.17.10}$$

where the dot notation is used for the differential coefficient. The noise voltage output is therefore

$$v_n(t) = \frac{C\dot{n}_Q(t)}{2\pi E_{c\max}} \tag{10.17.11}$$

To find the noise power output, it is best to work in the frequency domain. A result of Fourier analysis shows that, if a voltage waveform $v(t)$ has a power spectral density $G(f)$, then the power spectral density for $\dot{v}(t)$ is $\omega^2 G(f)$. Equation (4.20.4) gives the spectral density for $n_Q(t)$ as $2kT_s$, and hence the spectral density for $\dot{n}_Q(t)$ is $\omega^2 2kT_s$. Combining this with Eq. (10.17.11) and simplifying gives, for the power spectral density of $v_n(t)$,

$$G_v(f) = \frac{C^2 f^2 2kT_s}{E^2_{c\max}} \tag{10.17.12}$$

From the definition of power spectral density (see Section 2.17), the noise power output is

$$P_{no} = \int_0^W G_v(f)df$$

$$= \frac{W^3C^22kT_s}{3E_{c\,\max}^2} \tag{10.17.13}$$

Combining this with Eq. (10.17.4) and simplifying gives, for the output signal-to-noise ratio,

$$\left(\frac{\mathrm{S}}{\mathrm{N}}\right)_o = \frac{P_{so}}{P_{no}}$$

$$= \frac{3\Delta f^2 E_{c\,\max}^2}{16R_{\text{out}}kT_sW^3} \tag{10.17.14}$$

The signal-to-noise figure of merit introduced in Section 8.14 is, for FM,

$$R_{\text{FM}} = \frac{(\mathrm{S/N})_o}{(\mathrm{S/N})_{\text{REF}}}$$

$$= 1.5 \cdot \left(\frac{\Delta f}{W}\right)^2 \frac{R_s}{R_{\text{out}}}$$

$$= 1.5\beta_W^2 \frac{R_s}{R_{\text{out}}} \tag{10.17.15}$$

Here β_W is the modulation index calculated for the highest baseband frequency W. Note that this is not the same in general as the modulation index for the signal, which is $\beta = \Delta f/f_m$.

The information on FM signal-to-noise ratio is often presented in another way, using the carrier-to-noise ratio as the input parameter at the detector. The carrier-to-noise ratio is similar to the $(\mathrm{S/N})_{\text{REF}}$ ratio except that the total IF bandwidth is used rather than W. Denoting the carrier-to-noise ratio by (C/N), Eq. (10.17.1) is modified to

$$\left(\frac{\mathrm{C}}{\mathrm{N}}\right) = \frac{E_{c\,\max}^2}{8R_SkT_sB_{IF}} \tag{10.17.16}$$

Also, applying Carson's rule, $B_{IF} = 2(\beta_W + 1)W$ gives

$$\left(\frac{\mathrm{C}}{\mathrm{N}}\right) = \frac{E_{c\,\max}^2}{16R_SkT_s(\beta_W + 1)W} \tag{10.17.17}$$

The ratio of output signal-to-noise ratio to carrier-to-noise ratio is known as the receiver (or detector) processing gain and is

CHAPTER 10

Transmission Lines

Transmission lines are used to transmit electric energy and signals from one point to another, specifically from a source to a load. Examples include the connection between a transmitter and an antenna, connections between computers in a network, or connections between a hydroelectric generating plant and a substation several hundred miles away. Other familiar examples include the interconnects between components of a stereo system and the connection between a cable service provider and your television set. Examples that are less familiar include the connections between devices on a circuit board that are designed to operate at high frequencies.

What all of these examples have in common is that the devices to be connected are separated by distances on the order of a wavelength or much larger, whereas in basic circuit analysis methods, connections between elements are assumed to have negligible length. The latter condition enabled us, for example, to take for granted that the voltage across a resistor on one side of a circuit was exactly in phase with the voltage source on the other side, or, more generally, that the time measured at the source location is precisely the same time as measured at all other points in the circuit. When distances are sufficiently large between source and receiver, time delay effects become appreciable, leading to delay-induced phase differences. In short, we deal with *wave phenomena* on transmission lines in the same manner that we deal with point-to-point energy propagation in free space or in dielectrics.

The basic elements in a circuit, such as resistors, capacitors, inductors, and the connections between them, are considered *lumped* elements if the time delay in traversing the elements is negligible. On the other hand, if the elements or interconnections are large enough, it may be necessary to consider them as *distributed* elements. This means that their resistive, capacitive, and inductive characteristics must be evaluated on a per-unit-distance basis. Transmission lines have this property in general, and thus they become circuit elements in themselves, possessing impedances that contribute to the circuit problem. The basic rule is that one must consider elements as distributed if the propagation delay across the element dimension is on the order of the shortest time interval of interest. In the time-harmonic case, this condition would lead to a measurable phase difference between each end of the device in question.

In this chapter, we investigate wave phenomena in transmission lines. Our objectives include (1) to understand how to treat transmission lines as circuit elements possessing complex impedances that are functions of line length and frequency, (2) to understand wave propagation on lines, including cases in which losses may occur, (3) to learn methods of combining different transmission lines to accomplish a desired objective, and (4) to understand transient phenomena on lines. ■

10.1 PHYSICAL DESCRIPTION OF TRANSMISSION LINE PROPAGATION

To obtain a feel for the manner in which waves propagate on transmission lines, the following demonstration may be helpful. Consider a *lossless* line, as shown in Figure 10.1. By lossless, we mean that all power that is launched into the line at the input end eventually arrives at the output end. A battery having voltage V_0 is connected to the input by closing switch S_1 at time $t = 0$. When the switch is closed, the effect is to launch voltage, $V^+ = V_0$. This voltage does not instantaneously appear everywhere on the line, but rather begins to travel from the battery toward the load resistor, R, at a certain velocity. The *wavefront,* represented by the vertical dashed line in Figure 10.1, represents the instantaneous boundary between the section of the line that has been charged to V_0 and the remaining section that is yet to be charged. It also represents the boundary between the section of the line that carries the charging current, I^+, and the remaining section that carries no current. Both current and voltage are discontinuous across the wavefront.

As the line charges, the wavefront moves from left to right at velocity ν, which is to be determined. On reaching the far end, all or a fraction of the wave voltage and current will reflect, depending on what the line is attached to. For example, if the resistor at the far end is left disconnected (switch S_2 is open), then all of the wavefront voltage will be reflected. If the resistor is connected, then some fraction of the incident voltage will reflect. The details of this will be treated in Section 10.9. Of interest at the moment are the factors that determine the wave velocity. The key

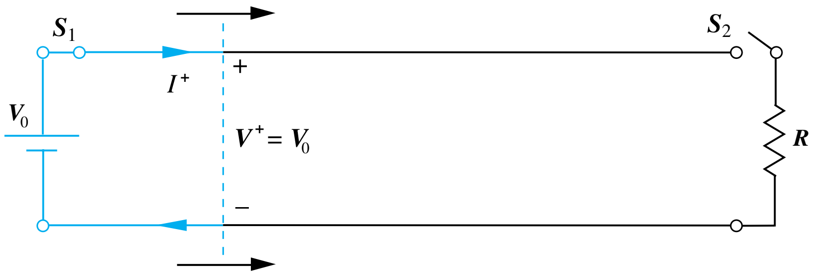

Figure 10.1 Basic transmission line circuit, showing voltage and current waves initiated by closing switch S_1.

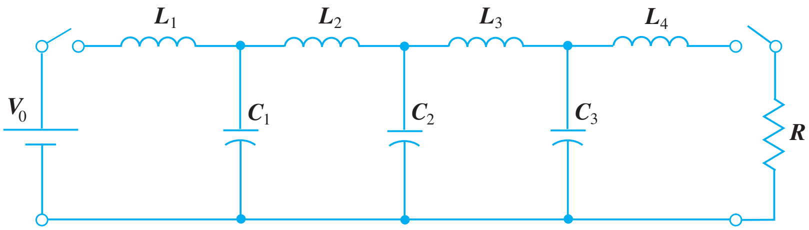

Figure 10.2 Lumped-element model of a transmission line. All inductance values are equal, as are all capacitance values.

to understanding and quantifying this is to note that the conducting transmission line will possess capacitance and inductance that are expressed on a per-unit-length basis. We have already derived expressions for these and evaluated them in Chapters 6 and 8 for certain transmission line geometries. Knowing these line characteristics, we can construct a model for the transmission line using lumped capacitors and inductors, as shown in Figure 10.2. The ladder network thus formed is referred to as a *pulse-forming network,* for reasons that will soon become clear.[1]

Consider now what happens when connecting the same switched voltage source to the network. Referring to Figure 10.2, on closing the switch at the battery location, current begins to increase in L_1, allowing C_1 to charge. As C_1 approaches full charge, current in L_2 begins to increase, allowing C_2 to charge next. This progressive charging process continues down the network, until all three capacitors are fully charged. In the network, a "wavefront" location can be identified as the point between two adjacent capacitors that exhibit the most difference between their charge levels. As the charging process continues, the wavefront moves from left to right. Its speed depends on how fast each inductor can reach its full-current state and, simultaneously, by how fast each capacitor is able to charge to full voltage. The wave is faster if the values of L_i and C_i are lower. We therefore expect the wave velocity to be inversely proportional to a function involving the product of inductance and capacitance. In the lossless transmission line, it turns out (as will be shown) that the wave velocity is given by $\nu = 1/\sqrt{LC}$, where L and C are specified per unit length.

Similar behavior is seen in the line and network when either is *initially charged*. In this case, the battery remains connected, and a resistor can be connected (by a switch) across the output end, as shown in Figure 10.2. In the case of the ladder network, the capacitor nearest the shunted end (C_3) will discharge through the resistor first, followed by the next-nearest capacitor, and so on. When the network is completely discharged, a voltage pulse has been formed across the resistor, and so we see why this ladder configuration is called a pulse-forming network. Essentially identical behavior is seen in a charged transmission line when connecting a resistor between conductors at the output end. The switched voltage exercises, as used in these discussions, are examples of transient problems on transmission lines. Transients will be treated in detail in Section 10.14. In the beginning, line responses to sinusoidal signals are emphasized.

[1] Designs and applications of pulse-forming networks are discussed in Reference 1.

Finally, we surmise that the existence of voltage and current across and within the transmission line conductors implies the existence of electric and magnetic fields in the space around the conductors. Consequently, we have two possible approaches to the analysis of transmission lines: (1) We can solve Maxwell's equations subject to the line configuration to obtain the fields, and with these find general expressions for the wave power, velocity, and other parameters of interest. (2) Or we can (for now) avoid the fields and solve for the voltage and current using an appropriate circuit model. It is the latter approach that we use in this chapter; the contribution of field theory is solely in the prior (and assumed) evaluation of the inductance and capacitance parameters. We will find, however, that circuit models become inconvenient or useless when losses in transmission lines are to be fully characterized, or when analyzing more complicated wave behavior (i.e., moding) which may occur as frequencies get high. The loss issues will be taken up in Section 10.5. Moding phenomena will be considered in Chapter 13.

10.2 THE TRANSMISSION LINE EQUATIONS

Our first goal is to obtain the differential equations, known as the *wave equations,* which the voltage or current must satisfy on a uniform transmission line. To do this, we construct a circuit model for an incremental length of line, write two circuit equations, and use these to obtain the wave equations.

Our circuit model contains the *primary constants* of the transmission line. These include the inductance, L, and capacitance, C, as well as the shunt conductance, G, and series resistance, R—all of which have values that are specified *per unit length*. The shunt conductance is used to model leakage current through the dielectric that may occur throughout the line length; the assumption is that the dielectric may possess conductivity, σ_d, in addition to a dielectric constant, ϵ_r, where the latter affects the capacitance. The series resistance is associated with any finite conductivity, σ_c, in the conductors. Either one of the latter parameters, R and G, will be responsible for power loss in transmission. In general, both are functions of frequency. Knowing the frequency and the dimensions, we can determine the values of R, G, L, and C by using formulas developed in earlier chapters.

We assume propagation in the $\mathbf{a}_z$ direction. Our model consists of a line section of length Δz containing resistance $R\Delta z$, inductance $L\Delta z$, conductance $G\Delta z$, and capacitance $C\Delta z$, as indicated in Figure 10.3. Because the section of the line looks the same from either end, we divide the series elements in half to produce a symmetrical network. We could equally well have placed half the conductance and half the capacitance at each end.

Our objective is to determine the manner and extent to which the output voltage and current are changed from their input values in the limit as the length approaches a very small value. We will consequently obtain a pair of differential equations that describe the rates of change of voltage and current with respect to z. In Figure 10.3, the input and output voltages and currents differ respectively by quantities ΔV and ΔI, which are to be determined. The two equations are obtained by successive applications of Kirchoff's voltage law (KVL) and Kirchoff's current law (KCL).

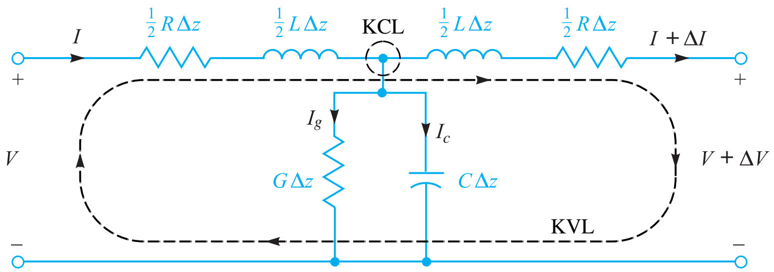

Figure 10.3 Lumped-element model of a short transmission line section with losses. The length of the section is Δz. Analysis involves applying Kirchoff's voltage and current laws (KVL and KCL) to the indicated loop and node, respectively.

First, KVL is applied to the loop that encompasses the entire section length, as shown in Figure 10.3:

$$V = \frac{1}{2}RI\Delta z + \frac{1}{2}L\frac{\partial I}{\partial t}\Delta z + \frac{1}{2}L\left(\frac{\partial I}{\partial t} + \frac{\partial \Delta I}{\partial t}\right)\Delta z$$
$$+ \frac{1}{2}R(I + \Delta I)\Delta z + (V + \Delta V) \qquad (1)$$

We can solve Eq. (1) for the ratio, $\Delta V/\Delta z$, obtaining:

$$\frac{\Delta V}{\Delta z} = -\left(RI + L\frac{\partial I}{\partial t} + \frac{1}{2}L\frac{\partial \Delta I}{\partial t} + \frac{1}{2}R\Delta I\right) \qquad (2)$$

Next, we write:

$$\Delta I = \frac{\partial I}{\partial z}\Delta z \qquad \text{and} \qquad \Delta V = \frac{\partial V}{\partial z}\Delta z \qquad (3)$$

which are then substituted into (2) to result in

$$\frac{\partial V}{\partial z} = -\left(1 + \frac{\Delta z}{2}\frac{\partial}{\partial z}\right)\left(RI + L\frac{\partial I}{\partial t}\right) \qquad (4)$$

Now, in the limit as Δz approaches zero (or a value small enough to be negligible), (4) simplifies to the final form:

$$\boxed{\frac{\partial V}{\partial z} = -\left(RI + L\frac{\partial I}{\partial t}\right)} \qquad (5)$$

Equation (5) is the first of the two equations that we are looking for. To find the second equation, we apply KCL to the upper central node in the circuit of Figure 10.3, noting from the symmetry that the voltage at the node will be $V + \Delta V/2$:

$$I = I_g + I_c + (I + \Delta I) = G\Delta z\left(V + \frac{\Delta V}{2}\right)$$
$$+ C\Delta z\frac{\partial}{\partial t}\left(V + \frac{\Delta V}{2}\right) + (I + \Delta I) \qquad (6)$$

Then, using (3) and simplifying, we obtain

$$\frac{\partial I}{\partial z} = -\left(1 + \frac{\Delta z}{2}\frac{\partial}{\partial z}\right)\left(GV + C\frac{\partial V}{\partial t}\right) \tag{7}$$

Again, we obtain the final form by allowing Δz to be reduced to a negligible magnitude. The result is

$$\boxed{\frac{\partial I}{\partial z} = -\left(GV + C\frac{\partial V}{\partial t}\right)} \tag{8}$$

The coupled differential equations, (5) and (8), describe the evolution of current and voltage in any transmission line. Historically, they have been referred to as the *telegraphist's equations*. Their solution leads to the wave equation for the transmission line, which we now undertake. We begin by differentiating Eq. (5) with respect to z and Eq. (8) with respect to t, obtaining:

$$\frac{\partial^2 V}{\partial z^2} = -R\frac{\partial I}{\partial z} - L\frac{\partial^2 I}{\partial t \partial z} \tag{9}$$

and

$$\frac{\partial I}{\partial z \partial t} = -G\frac{\partial V}{\partial t} - C\frac{\partial^2 V}{\partial t^2} \tag{10}$$

Next, Eqs. (8) and (10) are substituted into (9). After rearranging terms, the result is:

$$\boxed{\frac{\partial^2 V}{\partial z^2} = LC\frac{\partial^2 V}{\partial t^2} + (LG + RC)\frac{\partial V}{\partial t} + RGV} \tag{11}$$

An analogous procedure involves differentiating Eq. (5) with respect to t and Eq. (8) with respect to z. Then, Eq. (5) and its derivative are substituted into the derivative of (8) to obtain an equation for the current that is in identical form to that of (11):

$$\boxed{\frac{\partial^2 I}{\partial z^2} = LC\frac{\partial^2 I}{\partial t^2} + (LG + RC)\frac{\partial I}{\partial t} + RGI} \tag{12}$$

Equations (11) and (12) are the *general wave equations* for the transmission line. Their solutions under various conditions form a major part of our study.

10.3 LOSSLESS PROPAGATION

Lossless propagation means that power is not dissipated or otherwise deviated as the wave travels down the transmission line; all power at the input end eventually reaches the output end. More realistically, any mechanisms that would cause losses to occur have negligible effect. In our model, lossless propagation occurs when $R = G = 0$.

Under this condition, only the first term on the right-hand side of either Eq. (11) or Eq. (12) survives. Eq. (11), for example, becomes

$$\boxed{\frac{\partial^2 V}{\partial z^2} = LC\frac{\partial^2 V}{\partial t^2}} \tag{13}$$

In considering the voltage function that will satisfy (13), it is most expedient to simply state the solution, and then show that it is correct. The solution of (13) is of the form:

$$\boxed{V(z,t) = f_1\left(t - \frac{z}{\nu}\right) + f_2\left(t + \frac{z}{\nu}\right) = V^+ + V^-} \tag{14}$$

where ν, the *wave velocity,* is a constant. The expressions $(t \pm z/\nu)$ are the arguments of functions f_1 and f_2. The identities of the functions themselves are not critical to the solution of (13). Therefore, f_1 and f_2 can be *any* function.

Animations

The arguments of f_1 and f_2 indicate, respectively, travel of the functions in the forward and backward z directions. We assign the symbols V^+ and V^- to identify the forward and backward voltage wave components. To understand the behavior, consider for example the value of f_1 (whatever this might be) at the zero value of its argument, occurring when $z = t = 0$. Now, as time increases to positive values (as it must), and if we are to keep track of $f_1(0)$, then the value of z must also increase to keep the argument $(t - z/\nu)$ equal to zero. The function f_1 therefore moves (or propagates) in the positive z direction. Using similar reasoning, the function f_2 will propagate in the *negative* z direction, as z in the argument $(t + z/\nu)$ must *decrease* to offset the increase in t. Therefore we associate the argument $(t - z/\nu)$ with *forward* z propagation, and the argument $(t + z/\nu)$ with *backward* z travel. This behavior occurs irrespective of what f_1 and f_2 are. As is evident in the argument forms, the propagation velocity is ν in both cases.

We next verify that functions having the argument forms expressed in (14) are solutions to (13). First, we take partial derivatives of f_1, for example with respect to z and t. Using the chain rule, the z partial derivative is

$$\frac{\partial f_1}{\partial z} = \frac{\partial f_1}{\partial(t - z/\nu)}\frac{\partial(t - z/\nu)}{\partial z} = -\frac{1}{\nu}f_1' \tag{15}$$

where it is apparent that the primed function, f_1', denotes the derivative of f_1 with respect to its argument. The partial derivative with respect to time is

$$\frac{\partial f_1}{\partial t} = \frac{\partial f_1}{\partial(t - z/\nu)}\frac{\partial(t - z/\nu)}{\partial t} = f_1' \tag{16}$$

Next, the second partial derivatives with respect to z and t can be taken using similar reasoning:

$$\frac{\partial^2 f_1}{\partial z^2} = \frac{1}{\nu^2}f_1'' \qquad \text{and} \qquad \frac{\partial^2 f_1}{\partial t^2} = f_1'' \tag{17}$$

where f_1'' is the second derivative of f_1 with respect to its argument. The results in (17) can now be substituted into (13), obtaining

$$\frac{1}{\nu^2} f_1'' = LC f_1'' \tag{18}$$

We now identify the wave velocity for lossless propagation, which is the condition for equality in (18):

$$\boxed{\nu = \frac{1}{\sqrt{LC}}} \tag{19}$$

Performing the same procedure using f_2 (and its argument) leads to the same expression for ν.

The form of ν as expressed in Eq. (19) confirms our original expectation that the wave velocity would be in some inverse proportion to L and C. The same result will be true for current, as Eq. (12) under lossless conditions would lead to a solution of the form identical to that of (14), with velocity given by (19). What is not known yet, however, is the relation *between* voltage and current.

We have already found that voltage and current are related through the telegraphist's equations, (5) and (8). These, under lossless conditions ($R = G = 0$), become

$$\boxed{\frac{\partial V}{\partial z} = -L\frac{\partial I}{\partial t}} \tag{20}$$

$$\boxed{\frac{\partial I}{\partial z} = -C\frac{\partial V}{\partial t}} \tag{21}$$

Using the voltage function, we can substitute (14) into (20) and use the methods demonstrated in (15) to write

$$\frac{\partial I}{\partial t} = -\frac{1}{L}\frac{\partial V}{\partial z} = \frac{1}{L\nu}(f_1' - f_2') \tag{22}$$

We next integrate (22) over time, obtaining the current in terms of its forward and backward propagating components:

$$\boxed{I(z,t) = \frac{1}{L\nu}\left[f_1\left(t - \frac{z}{\nu}\right) - f_2\left(t + \frac{z}{\nu}\right)\right] = I^+ + I^-} \tag{23}$$

In performing this integration, all integration constants are set to zero. The reason for this, as demonstrated by (20) and (21), is that a time-varying voltage must lead to a time-varying current, with the reverse also true. The factor $1/L\nu$ appearing in (23) multiplies voltage to obtain current, and so we identify the product $L\nu$ as the *characteristic impedance,* Z_0, of the lossless line. Z_0 is defined as the ratio of

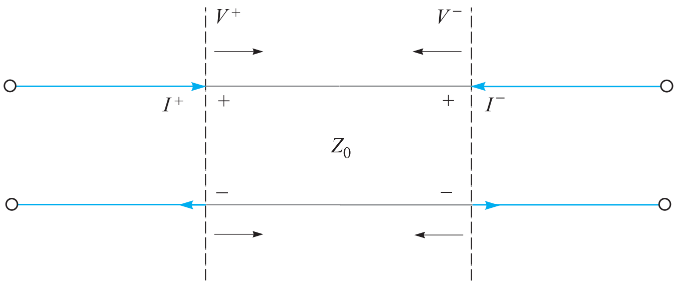

Figure 10.4 Current directions in waves having positive voltage polarity.

the voltage to the current in a single propagating wave. Using (19), we write the characteristic impedance as

$$\boxed{Z_0 = Lv = \sqrt{\frac{L}{C}}} \tag{24}$$

By inspecting (14) and (23), we now note that

$$\boxed{V^+ = Z_0 I^+} \tag{25a}$$

and

$$\boxed{V^- = -Z_0 I^-} \tag{25b}$$

The significance of the preceding relations can be seen in Figure 10.4. The figure shows forward- and backward-propagating voltage waves, V^+ and V^-, both of which have positive polarity. The currents that are associated with these voltages will flow in opposite directions. We define *positive current* as having a *clockwise* flow in the line, and *negative current* as having a *counterclockwise* flow. The minus sign in (25*b*) thus assures that negative current will be associated with a backward-propagating wave that has positive polarity. This is a general convention, applying to lines with losses also. Propagation with losses is studied by solving (11) under the assumption that either R or G (or both) are not zero. We will do this in Section 10.7 under the special case of sinusoidal voltages and currents. Sinusoids in lossless transmission lines are considered in Section 10.4.

10.4 LOSSLESS PROPAGATION OF SINUSOIDAL VOLTAGES

An understanding of sinusoidal waves on transmission lines is important because any signal that is transmitted in practice can be decomposed into a discrete or continuous summation of sinusoids. This is the basis of *frequency domain* analysis of signals on lines. In such studies, the effect of the transmission line on any signal can be determined by noting the effects on the frequency components. This means that one can effectively propagate the spectrum of a given signal, using frequency-dependent line parameters, and then reassemble the frequency components into the resultant signal in time domain. Our objective in this section is to obtain an understanding of sinusoidal propagation and the implications on signal behavior for the lossless line case.

We begin by assigning sinusoidal functions to the voltage functions in Eq. (14). Specifically, we consider a specific frequency, $f = \omega/2\pi$, and write $f_1 = f_2 = V_0 \cos(\omega t + \phi)$. By convention, the cosine function is chosen; the sine is obtainable, as we know, by setting $\phi = -\pi/2$. We next replace t with $(t \pm z/\nu_p)$, obtaining

$$\mathcal{V}(z,t) = |V_0| \cos[\omega(t \pm z/\nu_p) + \phi] = |V_0| \cos[\omega t \pm \beta z + \phi] \tag{26}$$

where we have assigned a new notation to the velocity, which is now called the *phase velocity*, ν_p. This is applicable to a pure sinusoid (having a single frequency) and will be found to depend on frequency in some cases. Choosing, for the moment, $\phi = 0$, we obtain the two possibilities of forward or backward z travel by choosing the minus or plus sign in (26). The two cases are:

$$\boxed{\mathcal{V}_f(z,t) = |V_0| \cos(\omega t - \beta z)} \quad \text{(forward } z \text{ propagation)} \tag{27a}$$

and

$$\boxed{\mathcal{V}_b(z,t) = |V_0| \cos(\omega t + \beta z)} \quad \text{(backward } z \text{ propagation)} \tag{27b}$$

where the magnitude factor, $|V_0|$, is the value of $\mathcal{V}$ at $z = 0, t = 0$. We define the *phase constant* β, obtained from (26), as

$$\boxed{\beta \equiv \frac{\omega}{\nu_p}} \tag{28}$$

We refer to the solutions expressed in (27a) and (27b) as the *real instantaneous* forms of the transmission-line voltage. They are the mathematical representations of what one would experimentally measure. The terms ωt and βz, appearing in these equations, have units of angle and are usually expressed in radians. We know that ω is the radian time frequency, measuring phase shift *per unit time,* and it has units of rad/s. In a similar way, we see that β will be interpreted as a *spatial* frequency, which in the present case measures the phase shift *per unit distance* along the z direction. Its units are rad/m. If we were to fix the time at $t = 0$, Eqs. (27a) and (27b) would become

$$\mathcal{V}_f(z,0) = \mathcal{V}_b(z,0) = |V_0| \cos(\beta z) \tag{29}$$

which we identify as a simple periodic function that repeats every incremental distance λ, known as the *wavelength*. The requirement is that $\beta\lambda = 2\pi$, and so

$$\boxed{\lambda = \frac{2\pi}{\beta} = \frac{\nu_p}{f}} \tag{30}$$

We next consider a point (such as a wave crest) on the cosine function of Eq. (27a), the occurrence of which requires the argument of the cosine to be an integer multiple of 2π. Considering the mth crest of the wave, the condition at $t = 0$ becomes

$$\beta z = 2m\pi$$

To keep track of this point on the wave, we require that the entire cosine argument be the same multiple of 2π for all time. From (27a) the condition becomes

$$\omega t - \beta z = \omega(t - z/v_p) = 2m\pi \tag{31}$$

Again, with increasing time, the position z must also increase in order to satisfy (31). Consequently the wave crest (and the entire wave) travels in the positive z direction at velocity v_p. Eq. (27b), having cosine argument $(\omega t + \beta z)$, describes a wave that travels in the *negative* z direction, since as time increases, z must now *decrease* to keep the argument constant. Similar behavior is found for the wave current, but complications arise from line-dependent phase differences that occur between current and voltage. These issues are best addressed once we are familiar with complex analysis of sinusoidal signals.

10.5 COMPLEX ANALYSIS OF SINUSOIDAL WAVES

Expressing sinusoidal waves as complex functions is useful (and essentially indispensable) because it greatly eases the evaluation and visualization of phase that will be found to accumulate by way of many mechanisms. In addition, we will find many cases in which two or more sinusoidal waves must be combined to form a resultant wave—a task made much easier if complex analysis is used.

Expressing sinusoidal functions in complex form is based on the Euler identity:

$$e^{\pm jx} = \cos(x) \pm j\sin(x) \tag{32}$$

from which we may write the cosine and sine, respectively, as the real and imaginary parts of the complex exponent:

$$\cos(x) = \mathrm{Re}[e^{\pm jx}] = \frac{1}{2}(e^{jx} + e^{-jx}) = \frac{1}{2}e^{jx} + c.c. \tag{33a}$$

$$\sin(x) = \pm\mathrm{Im}[e^{\pm jx}] = \frac{1}{2j}(e^{jx} - e^{-jx}) = \frac{1}{2j}e^{jx} + c.c. \tag{33b}$$

where $j \equiv \sqrt{-1}$, and where $c.c.$ denotes the complex conjugate of the preceding term. The conjugate is formed by changing the sign of j wherever it appears in the complex expression.

We may next apply (33a) to our voltage wave function, Eq. (26):

$$\mathcal{V}(z,t) = |V_0|\cos[\omega t \pm \beta z + \phi] = \frac{1}{2}\underbrace{(|V_0|e^{j\phi})}_{V_0} e^{\pm j\beta z} e^{j\omega t} + c.c. \tag{34}$$

Note that we have arranged the phases in (34) such that we identify the *complex amplitude* of the wave as $V_0 = (|V_0|e^{j\phi})$. In future usage, a single symbol (V_0 in the present example) will usually be used for the voltage or current amplitudes, with the understanding that these will generally be complex (having magnitude and phase).

Two additional definitions follow from Eq. (34). First, we define the *complex instantaneous* voltage as:

$$\boxed{V_c(z,t) = V_0 e^{\pm j\beta z} e^{j\omega t}} \tag{35}$$

The *phasor* voltage is then formed by dropping the $e^{j\omega t}$ factor from the complex instantaneous form:

$$\boxed{V_s(z) = V_0 e^{\pm j\beta z}} \tag{36}$$

The phasor voltage can be defined provided we have *sinusoidal steady-state* conditions—meaning that V_0 is independent of time. This has in fact been our assumption all along, because a time-varying amplitude would imply the existence of other frequency components in our signal. Again, we are treating only a single-frequency wave. The significance of the phasor voltage is that we are effectively letting time stand still and observing the stationary wave in space at $t = 0$. The processes of evaluating relative phases between various line positions and of combining multiple waves is made much simpler in phasor form. Again, this works only if all waves under consideration have the same frequency. With the definitions in (35) and (36), the real instantaneous voltage can be constructed using (34):

$$\mathcal{V}(z,t) = |V_0| \cos[\omega t \pm \beta z + \phi] = \text{Re}[V_c(z,t)] = \frac{1}{2} V_c + c.c. \tag{37a}$$

Or, in terms of the phasor voltage:

$$\boxed{\mathcal{V}(z,t) = |V_0| \cos[\omega t \pm \beta z + \phi] = \text{Re}[V_s(z) e^{j\omega t}] = \frac{1}{2} V_s(z) e^{j\omega t} + c.c.} \tag{37b}$$

In words, we may obtain our real sinusoidal voltage wave by multiplying the phasor voltage by $e^{j\omega t}$ (reincorporating the time dependence) and then taking the real part of the resulting expression. It is imperative that one becomes familiar with these relations and their meaning before proceeding further.

EXAMPLE 10.1

Two voltage waves having equal frequencies and amplitudes propagate in opposite directions on a lossless transmission line. Determine the total voltage as a function of time and position.

Solution. Because the waves have the same frequency, we can write their combination using their phasor forms. Assuming phase constant, β, and real amplitude, V_0, the two wave voltages combine in this way:

$$V_{sT}(z) = V_0 e^{-j\beta z} + V_0 e^{+j\beta z} = 2V_0 \cos(\beta z)$$

In real instantaneous form, this becomes

$$\mathcal{V}(z,t) = \mathrm{Re}[2V_0 \cos(\beta z) e^{j\omega t}] = 2V_0 \cos(\beta z)\cos(\omega t)$$

We recognize this as a *standing wave,* in which the amplitude varies, as $\cos(\beta z)$, and oscillates in time, as $\cos(\omega t)$. Zeros in the amplitude (nulls) occur at fixed locations, $z_n = (m\pi)/(2\beta)$ where m is an odd integer. We extend the concept in Section 10.10, where we explore the *voltage standing wave ratio* as a measurement technique.

10.6 TRANSMISSION LINE EQUATIONS AND THEIR SOLUTIONS IN PHASOR FORM

We now apply our results of the previous section to the transmission line equations, beginning with the general wave equation, (11). This is rewritten as follows, for the real instantaneous voltage, $\mathcal{V}(z,t)$:

$$\frac{\partial^2 \mathcal{V}}{\partial z^2} = LC\frac{\partial^2 \mathcal{V}}{\partial t^2} + (LG + RC)\frac{\partial \mathcal{V}}{\partial t} + RG\mathcal{V} \tag{38}$$

We next substitute $\mathcal{V}(z,t)$ as given by the far right-hand side of (37b), noting that the complex conjugate term ($c.c.$) will form a separate redundant equation. We also use the fact that the operator $\partial/\partial t$, when applied to the complex form, is equivalent to multiplying by a factor of $j\omega$. After substitution, and after all time derivatives are taken, the factor $e^{j\omega t}$ divides out. We are left with the wave equation in terms of the phasor voltage:

$$\frac{d^2 V_s}{dz^2} = -\omega^2 LCV_s + j\omega(LG + RC)V_s + RGV_s \tag{39}$$

Rearranging terms leads to the simplified form:

$$\boxed{\frac{d^2 V_s}{dz^2} = \underbrace{(R + j\omega L)}_{Z}\underbrace{(G + j\omega C)}_{Y}V_s = \gamma^2 V_s} \tag{40}$$

where Z and Y, as indicated, are respectively the *net series impedance* and the *net shunt admittance* in the transmission line—both as per-unit-distance measures. The *propagation constant* in the line is defined as

$$\boxed{\gamma = \sqrt{(R + j\omega L)(G + j\omega C)} = \sqrt{ZY} = \alpha + j\beta} \tag{41}$$

The significance of the term will be explained in Section 10.7. For our immediate purposes, the solution of (40) will be

$$\boxed{V_s(z) = V_0^+ e^{-\gamma z} + V_0^- e^{+\gamma z}} \tag{42a}$$

The wave equation for current will be identical in form to (40). We therefore expect the phasor current to be in the form:

$$\boxed{I_s(z) = I_0^+ e^{-\gamma z} + I_0^- e^{\gamma z}} \tag{42b}$$

The relation between the current and voltage waves is now found, as before, through the telegraphist's equations, (5) and (8). In a manner consistent with Eq. (37*b*), we write the sinusoidal current as

$$\mathcal{I}(z,t) = |I_0|\cos(\omega t \pm \beta z + \xi) = \frac{1}{2}\underbrace{(|I_0|e^{j\xi})}_{I_0} e^{\pm j\beta z} e^{j\omega t} + c.c. = \frac{1}{2} I_s(z) e^{j\omega t} + c.c. \tag{43}$$

Substituting the far right-hand sides of (37*b*) and (43) into (5) and (8) transforms the latter equations as follows:

$$\frac{\partial \mathcal{V}}{\partial z} = -\left(R\mathcal{I} + L\frac{\partial \mathcal{I}}{\partial t}\right) \quad \Rightarrow \quad \boxed{\frac{dV_s}{dz} = -(R + j\omega L)I_s = -ZI_s} \tag{44a}$$

and

$$\frac{\partial \mathcal{I}}{\partial z} = -\left(G\mathcal{V} + C\frac{\partial \mathcal{V}}{\partial t}\right) \quad \Rightarrow \quad \boxed{\frac{dI_s}{dz} = -(G + j\omega C)V_s = -YV_s} \tag{44b}$$

We can now substitute (42*a*) and (42*b*) into either (44*a*) or (44*b*) [we will use (44a)] to find:

$$-\gamma V_0^+ e^{-\gamma z} + \gamma V_0^- e^{\gamma z} = -Z(I_0^+ e^{-\gamma z} + I_0^- e^{\gamma z}) \tag{45}$$

Next, equating coefficients of $e^{-\gamma z}$ and $e^{\gamma z}$, we find the general expression for the line characteristic impedance:

$$Z_0 = \frac{V_0^+}{I_0^+} = -\frac{V_0^-}{I_0^-} = \frac{Z}{\gamma} = \frac{Z}{\sqrt{ZY}} = \sqrt{\frac{Z}{Y}} \tag{46}$$

Incorporating the expressions for Z and Y, we find the characteristic impedance in terms of our known line parameters:

$$\boxed{Z_0 = \sqrt{\frac{R + j\omega L}{G + j\omega C}} = |Z_0|e^{j\theta}} \tag{47}$$

Note that with the voltage and current as given in (37*b*) and (43), we would identify the phase of the characteristic impedance, $\theta = \phi - \xi$.

EXAMPLE 10.2

A lossless transmission line is 80 cm long and operates at a frequency of 600 MHz. The line parameters are $L = 0.25\ \mu$H/m and $C = 100$ pF/m. Find the characteristic impedance, the phase constant, and the phase velocity.

Solution. Because the line is lossless, both R and G are zero. The characteristic impedance is

$$Z_0 = \sqrt{\frac{L}{C}} = \sqrt{\frac{0.25 \times 10^{-6}}{100 \times 10^{-12}}} = 50\ \Omega$$

Because $\gamma = \alpha + j\beta = \sqrt{(R + j\omega L)(G + j\omega C)} = j\omega\sqrt{LC}$, we see that

$$\beta = \omega\sqrt{LC} = 2\pi(600 \times 10^6)\sqrt{(0.25 \times 10^{-6})(100 \times 10^{-12})} = 18.85 \text{ rad/m}$$

Also,

$$v_p = \frac{\omega}{\beta} = \frac{2\pi(600 \times 10^6)}{18.85} = 2 \times 10^8 \text{ m/s}$$

10.7 LOW-LOSS PROPAGATION

Having obtained the phasor forms of voltage and current in a general transmission line [Eqs. (42*a*) and (42*b*)], we can now look more closely at the significance of these results. First we incorporate (41) into (42*a*) to obtain

$$\boxed{V_s(z) = V_0^+ e^{-\alpha z} e^{-j\beta z} + V_0^- e^{\alpha z} e^{j\beta z}} \tag{48}$$

Next, multiplying (48) by $e^{j\omega t}$ and taking the real part gives the real instantaneous voltage:

$$\boxed{\mathcal{V}(z, t) = V_0^+ e^{-\alpha z} \cos(\omega t - \beta z) + V_0^- e^{\alpha z} \cos(\omega t + \beta z)} \tag{49}$$

In this exercise, we have assigned V_0^+ and V_0^- to be real. Eq. (49) is recognized as describing forward- and backward-propagating waves that diminish in amplitude with distance according to $e^{-\alpha z}$ for the forward wave, and $e^{\alpha z}$ for the backward wave. Both waves are said to *attenuate* with propagation distance at a rate determined by the *attenuation coefficient,* α, expressed in units of nepers/m [Np/m].[2]

The phase constant, β, found by taking the imaginary part of (41), is likely to be a somewhat complicated function, and will in general depend on R and G. Nevertheless, β is still defined as the ratio ω/v_p, and the wavelength is still defined as the distance that provides a phase shift of 2π rad, so that $\lambda = 2\pi/\beta$. By inspecting (41), we observe that losses in propagation are avoided (or $\alpha = 0$) only when $R = G = 0$. In that case, (41) gives $\gamma = j\beta = j\omega\sqrt{LC}$, and so $v_p = 1/\sqrt{LC}$, as we found before.

Expressions for α and β when losses are small can be readily obtained from (41). In the *low-loss approximation,* we require $R \ll \omega L$ and $G \ll \omega C$, a condition that

[2] The term *neper* was selected (by some poor speller) to honor John Napier, a Scottish mathematician who first proposed the use of logarithms.

is often true in practice. Before we apply these conditions, Eq. (41) can be written in the form:

$$\gamma = \alpha + j\beta = [(R + j\omega L)(G + j\omega C)]^{1/2}$$
$$= j\omega\sqrt{LC}\left[\left(1 + \frac{R}{j\omega L}\right)^{1/2}\left(1 + \frac{G}{j\omega C}\right)^{1/2}\right] \tag{50}$$

The low-loss approximation then allows us to use the first three terms in the binomial series:

$$\sqrt{1+x} \doteq 1 + \frac{x}{2} - \frac{x^2}{8} \quad (x \ll 1) \tag{51}$$

We use (51) to expand the terms in large parentheses in (50), obtaining:

$$\gamma \doteq j\omega\sqrt{LC}\left[\left(1 + \frac{R}{j2\omega L} + \frac{R^2}{8\omega^2 L^2}\right)\left(1 + \frac{G}{j2\omega C} + \frac{G^2}{8\omega^2 C^2}\right)\right] \tag{52}$$

All products in (52) are then carried out, neglecting the terms involving RG^2, R^2G, and R^2G^2, as these will be negligible compared to all others. The result is

$$\gamma = \alpha + j\beta \doteq j\omega\sqrt{LC}\left[1 + \frac{1}{j2\omega}\left(\frac{R}{L} + \frac{G}{C}\right) + \frac{1}{8\omega^2}\left(\frac{R^2}{L^2} - \frac{2RG}{LC} + \frac{G^2}{C^2}\right)\right] \tag{53}$$

Now, separating real and imaginary parts of (53) yields α and β:

$$\boxed{\alpha \doteq \frac{1}{2}\left(R\sqrt{\frac{C}{L}} + G\sqrt{\frac{L}{C}}\right)} \tag{54a}$$

and

$$\boxed{\beta \doteq \omega\sqrt{LC}\left[1 + \frac{1}{8}\left(\frac{G}{\omega C} - \frac{R}{\omega L}\right)^2\right]} \tag{54b}$$

We note that α scales in direct proportion to R and G, as would be expected. We also note that the terms in (54b) that involve R and G lead to a phase velocity, $\nu_p = \omega/\beta$, that is frequency-dependent. Moreover, the *group velocity*, $\nu_g = d\omega/d\beta$, will also depend on frequency, and will lead to signal distortion, as we will explore in Chapter 12. Note that with nonzero R and G, phase and group velocities that are constant with frequency can be obtained when $R/L = G/C$, known as *Heaviside's condition*. In this case, (54b) becomes $\beta \doteq \omega\sqrt{LC}$, and the line is said to be *distortionless*. Further complications occur when accounting for possible frequency dependencies within R, G, L, and C. Consequently, conditions of low-loss or distortion-free propagation will usually occur over limited frequency ranges. As a rule, loss increases with increasing frequency, mostly because of the increase in R with frequency. The nature of this latter effect, known as *skin effect* loss, requires field theory to understand

and quantify. We will study this in Chapter 11, and we will apply it to transmission line structures in Chapter 13.

Finally, we can apply the low-loss approximation to the characteristic impedance, Eq. (47). Using (51), we find

$$Z_0 = \sqrt{\frac{R + j\omega L}{G + j\omega C}} = \sqrt{\frac{j\omega L\left(1 + \frac{R}{j\omega L}\right)}{j\omega C\left(1 + \frac{G}{j\omega C}\right)}} \doteq \sqrt{\frac{L}{C}}\left[\frac{\left(1 + \frac{R}{j2\omega L} + \frac{R^2}{8\omega^2 L^2}\right)}{\left(1 + \frac{G}{j2\omega C} + \frac{G^2}{8\omega^2 C^2}\right)}\right] \tag{55}$$

Next, we multiply (55) by a factor of 1, in the form of the complex conjugate of the denominator of (55) divided by itself. The resulting expression is simplified by neglecting all terms on the order of R^2G, G^2R, and higher. Additionally, the approximation, $1/(1 + x) \doteq 1 - x$, where $x \ll 1$ is used. The result is

$$Z_0 \doteq \sqrt{\frac{L}{C}}\left\{1 + \frac{1}{2\omega^2}\left[\frac{1}{4}\left(\frac{R}{L} + \frac{G}{C}\right)^2 - \frac{G^2}{C^2}\right] + \frac{j}{2\omega}\left(\frac{G}{C} - \frac{R}{L}\right)\right\} \tag{56}$$

Note that when Heaviside's condition (again, $R/L = G/C$) holds, Z_0 simplifies to just $\sqrt{L/C}$, as is true when both R and G are zero.

EXAMPLE 10.3

Suppose in a certain transmission line $G = 0$, but R is finite valued and satisfies the low-loss requirement, $R \ll \omega L$. Use Eq. (56) to write the approximate magnitude and phase of Z_0.

Solution. With $G = 0$, the imaginary part of (56) is much greater than the second term in the real part [proportional to $(R/\omega L)^2$]. Therefore, the characteristic impedance becomes

$$Z_0(G = 0) \doteq \sqrt{\frac{L}{C}}\left(1 - j\frac{R}{2\omega L}\right) = |Z_0|e^{j\theta}$$

where $|Z_0| \doteq \sqrt{L/C}$, and $\theta = \tan^{-1}(-R/2\omega L)$.

D10.1. At an operating radian frequency of 500 Mrad/s, typical circuit values for a certain transmission line are: $R = 0.2\ \Omega$/m, $L = 0.25\ \mu$H/m, $G = 10\ \mu$S/m, and $C = 100$ pF/m. Find: (*a*) α; (*b*) β; (*c*) λ; (*d*) v_p; (*e*) Z_0.

Ans. 2.25 mNp/m; 2.50 rad/m; 2.51 m; 2×10^8 m/sec; $50.0 - j0.0350\ \Omega$

10.8 POWER TRANSMISSION AND THE USE OF DECIBELS IN LOSS CHARACTERIZATION

Having found the sinusoidal voltage and current in a lossy transmission line, we next evaluate the power transmitted over a specified distance as a function of voltage and current amplitudes. We start with the *instantaneous* power, given simply as the product of the real voltage and current. Consider the forward-propagating term in (49), where again, the amplitude, $V_0^+ = |V_0|$, is taken to be real. The current waveform will be similar, but will generally be shifted in phase. Both current and voltage attenuate according to the factor $e^{-\alpha z}$. The instantaneous power therefore becomes:

$$\mathcal{P}(z,t) = \mathcal{V}(z,t)\mathcal{I}(z,t) = |V_0||I_0|e^{-2\alpha z}\cos(\omega t - \beta z)\cos(\omega t - \beta z + \theta) \tag{57}$$

Usually, the *time-averaged* power, $\langle \mathcal{P} \rangle$, is of interest. We find this through:

$$\langle \mathcal{P} \rangle = \frac{1}{T}\int_0^T |V_0||I_0|e^{-2\alpha z}\cos(\omega t - \beta z)\cos(\omega t - \beta z + \theta)dt \tag{58}$$

where $T = 2\pi/\omega$ is the time period for one oscillation cycle. Using a trigonometric identity, the product of cosines in the integrand can be written as the sum of individual cosines at the sum and difference frequencies:

$$\langle \mathcal{P} \rangle = \frac{1}{T}\int_0^T \frac{1}{2}|V_0||I_0|e^{-2\alpha z}[\cos(2\omega t - 2\beta z + \theta) + \cos(\theta)]\,dt \tag{59}$$

The first cosine term integrates to zero, leaving the $\cos\theta$ term. The remaining integral easily evaluates as

$$\langle \mathcal{P} \rangle = \frac{1}{2}|V_0||I_0|e^{-2\alpha z}\cos\theta = \frac{1}{2}\frac{|V_0|^2}{|Z_0|}e^{-2\alpha z}\cos\theta \text{ [W]} \tag{60}$$

The same result can be obtained directly from the phasor voltage and current. We begin with these, expressed as

$$V_s(z) = V_0 e^{-\alpha z}e^{-j\beta z} \tag{61}$$

and

$$I_s(z) = I_0 e^{-\alpha z}e^{-j\beta z} = \frac{V_0}{Z_0}e^{-\alpha z}e^{-j\beta z} \tag{62}$$

where $Z_0 = |Z_0|e^{j\theta}$. We now note that the time-averaged power as expressed in (60) can be obtained from the phasor forms through:

$$\boxed{\langle \mathcal{P} \rangle = \frac{1}{2}\text{Re}\{V_s I_s^*\}} \tag{63}$$

where again, the asterisk (*) denotes the complex conjugate (applied in this case to the current phasor only). Using (61) and (62) in (63), it is found that

$$\begin{aligned}\langle \mathcal{P} \rangle &= \frac{1}{2}\text{Re}\left\{V_0 e^{-\alpha z}e^{-j\beta z}\frac{V_0^*}{|Z_0|e^{-j\theta}}e^{-\alpha z}e^{+j\beta z}\right\} \\ &= \frac{1}{2}\text{Re}\left\{\frac{V_0 V_0^*}{|Z_0|}e^{-2\alpha z}e^{j\theta}\right\} = \frac{1}{2}\frac{|V_0|^2}{|Z_0|}e^{-2\alpha z}\cos\theta\end{aligned} \tag{64}$$

which we note is identical to the time-integrated result in (60). Eq. (63) applies to any single-frequency wave.

An important result of the preceding exercise is that power attenuates as $e^{-2\alpha z}$, or

$$\boxed{\langle \mathcal{P}(z) \rangle = \langle \mathcal{P}(0) \rangle e^{-2\alpha z}} \tag{65}$$

Power drops at twice the exponential rate with distance as either voltage or current.

A convenient measure of power loss is in *decibel* units. This is based on expressing the power decrease as a power of 10. Specifically, we write

$$\frac{\langle \mathcal{P}(z) \rangle}{\langle \mathcal{P}(0) \rangle} = e^{-2\alpha z} = 10^{-\kappa \alpha z} \tag{66}$$

where the constant, κ, is to be determined. Setting $\alpha z = 1$, we find

$$e^{-2} = 10^{-\kappa} \quad \Rightarrow \quad \kappa = \log_{10}(e^2) = 0.869 \tag{67}$$

Now, by definition, the power loss in decibels (dB) is

$$\boxed{\text{Power loss (dB)} = 10 \log_{10} \left[\frac{\langle \mathcal{P}(0) \rangle}{\langle \mathcal{P}(z) \rangle} \right] = 8.69 \alpha z} \tag{68}$$

where we note that inverting the power ratio in the argument of the log function [as compared to the ratio in (66)] yields a positive number for the dB loss. Also, noting that $\langle \mathcal{P} \rangle \propto |V_0|^2$, we may write, equivalently:

$$\boxed{\text{Power loss (dB)} = 10 \log_{10} \left[\frac{\langle \mathcal{P}(0) \rangle}{\langle \mathcal{P}(z) \rangle} \right] = 20 \log_{10} \left[\frac{|V_0(0)|}{|V_0(z)|} \right]} \tag{69}$$

where $|V_0(z)| = |V_0(0)| e^{-\alpha z}$.

EXAMPLE 10.4

A 20-m length of transmission line is known to produce a 2.0-dB drop in power from end to end. (*a*) What fraction of the input power reaches the output? (*b*) What fraction of the input power reaches the midpoint of the line? (*c*) What exponential attenuation coefficient, α, does this represent?

Solution. (*a*) The power fraction will be

$$\frac{\langle \mathcal{P}(20) \rangle}{\langle \mathcal{P}(0) \rangle} = 10^{-0.2} = 0.63$$

(*b*) 2 dB in 20 m implies a loss rating of 0.2 dB/m. So, over a 10-m span, the loss is 1.0 dB. This represents the power fraction, $10^{-0.1} = 0.79$.

(*c*) The exponential attenuation coefficient is found through

$$\alpha = \frac{2.0 \text{ dB}}{(8.69 \text{ dB/Np})(20 \text{ m})} = 0.012 \text{ [Np/m]}$$

A final point addresses the question: Why use decibels? The most compelling reason is that when evaluating the accumulated loss for several lines and devices that are all end-to-end connected, the net loss in dB for the entire span is just the sum of the dB losses of the individual elements.

D10.2. Two transmission lines are to be joined end to end. Line 1 is 30 m long and is rated at 0.1 dB/m. Line 2 is 45 m long and is rated at 0.15 dB/m. The joint is not done well and imparts a 3-dB loss. What percentage of the input power reaches the output of the combination?

Ans. 5.3%

10.9 WAVE REFLECTION AT DISCONTINUITIES

The concept of wave reflection was introduced in Section 10.1. As implied there, the need for a reflected wave originates from the necessity to satisfy all voltage and current boundary conditions at the ends of transmission lines and at locations at which two dissimilar lines are connected to each other. The consequences of reflected waves are usually less than desirable, in that some of the power that was intended to be transmitted to a load, for example, reflects and propagates back to the source. Conditions for achieving *no* reflected waves are therefore important to understand.

The basic reflection problem is illustrated in Figure 10.5. In it, a transmission line of characteristic impedance Z_0 is terminated by a load having complex impedance, $Z_L = R_L + jX_L$. If the line is lossy, then we know that Z_0 will also be complex. For convenience, we assign coordinates such that the load is at location $z = 0$. Therefore, the line occupies the region $z < 0$. A voltage wave is presumed to be incident on the load, and is expressed in phasor form for all z:

$$V_i(z) = V_{0i} e^{-\alpha z} e^{-j\beta z} \tag{70a}$$

When the wave reaches the load, a reflected wave is generated that back-propagates:

$$V_r(z) = V_{0r} e^{+\alpha z} e^{+j\beta z} \tag{70b}$$

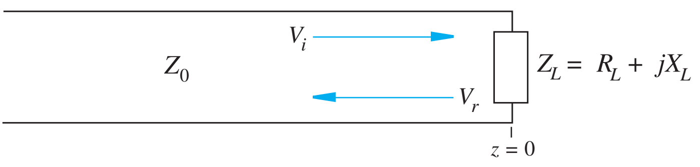

Figure 10.5 Voltage wave reflection from a complex load impedance.

The phasor voltage at the load is now the sum of the incident and reflected voltage phasors, evaluated at $z = 0$:

$$V_L = V_{0i} + V_{0r} \tag{71}$$

Additionally, the current through the load is the sum of the incident and reflected currents, also at $z = 0$:

$$I_L = I_{0i} + I_{0r} = \frac{1}{Z_0}[V_{0i} - V_{0r}] = \frac{V_L}{Z_L} = \frac{1}{Z_L}[V_{0i} + V_{0r}] \tag{72}$$

We can now solve for the ratio of the reflected voltage amplitude to the incident voltage amplitude, defined as the *reflection coefficient,* Γ:

$$\boxed{\Gamma \equiv \frac{V_{0r}}{V_{0i}} = \frac{Z_L - Z_0}{Z_L + Z_0} = |\Gamma|e^{j\phi_r}} \tag{73}$$

where we emphasize the complex nature of Γ—meaning that, in general, a reflected wave will experience a reduction in amplitude and a phase shift, relative to the incident wave.

Now, using (71) with (73), we may write

$$V_L = V_{0i} + \Gamma V_{0i} \tag{74}$$

from which we find the *transmission coefficient,* defined as the ratio of the load voltage amplitude to the incident voltage amplitude:

$$\boxed{\tau \equiv \frac{V_L}{V_{0i}} = 1 + \Gamma = \frac{2Z_L}{Z_0 + Z_L} = |\tau|e^{j\phi_t}} \tag{75}$$

A point that may at first cause some alarm is that if Γ is a positive real number, then $\tau > 1$; the voltage amplitude at the load is thus greater than the incident voltage. Although this would seem counterintuitive, it is not a problem because the load current will be lower than that in the incident wave. We will find that this always results in an average *power* at the load that is less than or equal to that in the incident wave. An additional point concerns the possibility of loss in the line. The incident wave amplitude that is used in (73) and (75) is always the amplitude that occurs *at the load*—after loss has occurred in propagating from the input.

Usually, the main objective in transmitting power to a load is to configure the line/load combination such that there is no reflection. The load therefore receives all the transmitted power. The condition for this is $\Gamma = 0$, which means that the load impedance must be equal to the line impedance. In such cases the load is said to be *matched* to the line (or vice versa). Various impedance-matching methods exist, many of which will be explored later in this chapter.

Finally, the fractions of the incident wave *power* that are reflected and dissipated by the load need to be determined. The incident power is found from (64), where this time we position the load at $z = L$, with the line input at $z = 0$.

$$\langle \mathcal{P}_i \rangle = \frac{1}{2}\text{Re}\left\{ \frac{V_0 V_0^*}{|Z_0|} e^{-2\alpha L} e^{j\theta} \right\} = \frac{1}{2}\frac{|V_0|^2}{|Z_0|} e^{-2\alpha L} \cos\theta \tag{76a}$$

The reflected power is then found by substituting the reflected wave voltage into (76a), where the latter is obtained by multiplying the incident voltage by Γ:

$$\langle \mathcal{P}_r \rangle = \frac{1}{2}\mathrm{Re}\left\{ \frac{(\Gamma V_0)(\Gamma^* V_0^*)}{|Z_0|} e^{-2\alpha L} e^{j\theta} \right\} = \frac{1}{2}\frac{|\Gamma|^2 |V_0|^2}{|Z_0|} e^{-2\alpha L} \cos\theta \tag{76b}$$

The reflected power fraction at the load is now determined by the ratio of (76b) to (76a):

$$\boxed{\frac{\langle \mathcal{P}_r \rangle}{\langle \mathcal{P}_i \rangle} = \Gamma\Gamma^* = |\Gamma|^2} \tag{77a}$$

The fraction of the incident power that is transmitted into the load (or dissipated by it) is therefore

$$\boxed{\frac{\langle \mathcal{P}_t \rangle}{\langle \mathcal{P}_i \rangle} = 1 - |\Gamma|^2} \tag{77b}$$

WW Illustrations

The reader should be aware that the transmitted power fraction is *not* $|\tau|^2$, as one might be tempted to conclude.

In situations involving the connection of two semi-infinite transmission lines having different characteristic impedances, reflections will occur at the junction, with the second line being treated as the load. For a wave incident from line 1 (Z_{01}) to line 2 (Z_{02}), we find

$$\boxed{\Gamma = \frac{Z_{02} - Z_{01}}{Z_{02} + Z_{01}}} \tag{78}$$

The fraction of the power that propagates into the second line is then $1 - |\Gamma|^2$.

EXAMPLE 10.5

A 50-Ω lossless transmission line is terminated by a load impedance, $Z_L = 50 - j75\ \Omega$. If the incident power is 100 mW, find the power dissipated by the load.

Solution. The reflection coefficient is

$$\Gamma = \frac{Z_L - Z_0}{Z_L + Z_0} = \frac{50 - j75 - 50}{50 - j75 + 50} = 0.36 - j0.48 = 0.60e^{-j.93}$$

Then

$$\langle \mathcal{P}_t \rangle = (1 - |\Gamma|^2)\langle \mathcal{P}_i \rangle = [1 - (0.60)^2](100) = 64 \text{ mW}$$

EXAMPLE 10.6

Two lossy lines are to be joined end to end. The first line is 10 m long and has a loss rating of 0.20 dB/m. The second line is 15 m long and has a loss rating of 0.10 dB/m. The reflection coefficient at the junction (line 1 to line 2) is $\Gamma = 0.30$. The input power (to line 1) is 100 mW. (*a*) Determine the total loss of the combination in dB. (*b*) Determine the power transmitted to the output end of line 2.

Solution. (*a*) The dB loss of the joint is

$$L_j(\text{dB}) = 10\log_{10}\left(\frac{1}{1-|\Gamma|^2}\right) = 10\log_{10}\left(\frac{1}{1-0.09}\right) = 0.41 \text{ dB}$$

The total loss of the link in dB is now

$$L_t(\text{dB}) = (0.20)(10) + 0.41 + (0.10)(15) = 3.91 \text{ dB}$$

(*b*) The output power will be $P_{\text{out}} = 100 \times 10^{-0.391} = 41$ mW.

10.10 VOLTAGE STANDING WAVE RATIO

In many instances, characteristics of transmission line performance are amenable to measurement. Included in these are measurements of unknown load impedances, or input impedances of lines that are terminated by known or unknown load impedances. Such techniques rely on the ability to measure voltage amplitudes that occur as functions of position within a line, usually designed for this purpose. A typical apparatus consists of a *slotted line,* which is a lossless coaxial transmission line having a longitudinal gap in the outer conductor along its entire length. The line is positioned between the sinusoidal voltage source and the impedance that is to be measured. Through the gap in the slotted line, a voltage probe may be inserted to measure the voltage amplitude between the inner and outer conductors. As the probe is moved along the length of the line, the maximum and minimum voltage amplitudes are noted, and their ratio, known as the *voltage standing wave ratio,* or VSWR, is determined. The significance of this measurement and its utility form the subject of this section.

To understand the meaning of the voltage measurements, we consider a few special cases. First, if the slotted line is terminated by a matched impedance, then no reflected wave occurs; the probe will indicate the same voltage amplitude at every point. Of course, the instantaneous voltages that the probe samples will differ in phase by $\beta(z_2 - z_1)$ rad as the probe is moved from $z = z_1$ to $z = z_2$, but the system is insensitive to the phase of the field. The equal-amplitude voltages are characteristic of an unattenuated traveling wave.

Second, if the slotted line is terminated by an open or short circuit (or in general a purely imaginary load impedance), the total voltage in the line is a standing wave and, as was shown in Example 10.1, the voltage probe provides no output when it is located at the nodes; these occur periodically with half-wavelength spacing. As the probe position is changed, its output varies as $|\cos(\beta z + \phi)|$, where z is the distance from the load, and where the phase, ϕ, depends on the load impedance. For example, if the load is a short circuit, the requirement of zero voltage at the short leads to a null occurring there, and so the voltage in the line will vary as $|\sin(\beta z)|$ (where $\phi = \pm\pi/2$).

A more complicated situation arises when the reflected voltage is neither 0 nor 100 percent of the incident voltage. Some energy is absorbed by the load and some is reflected. The slotted line, therefore, supports a voltage that is composed of both a traveling wave and a standing wave. It is customary to describe this voltage as a standing wave, even though a traveling wave is also present. We will see that the voltage does not have zero amplitude at any point for all time, and the degree to which the voltage is divided between a traveling wave and a true standing wave is expressed by the ratio of the maximum amplitude found by the probe to the minimum amplitude (VSWR). This information, along with the positions of the voltage minima or maxima with respect to that of the load, enable one to determine the load impedance. The VSWR also provides a measure of the quality of the termination. Specifically, a perfectly matched load yields a VSWR of exactly 1. A totally reflecting load produces an infinite VSWR.

To derive the specific form of the total voltage, we begin with the forward and backward-propagating waves that occur within the slotted line. The load is positioned at $z = 0$, and so all positions within the slotted line occur at negative values of z. Taking the input wave amplitude as V_0, the total phasor voltage is

$$V_{sT}(z) = V_0 e^{-j\beta z} + \Gamma V_0 e^{j\beta z} \tag{79}$$

The line, being lossless, has real characteristic impedance, Z_0. The load impedance, Z_L, is in general complex, which leads to a complex reflection coefficient:

$$\boxed{\Gamma = \frac{Z_L - Z_0}{Z_L + Z_0} = |\Gamma| e^{j\phi}} \tag{80}$$

If the load is a short circuit ($Z_L = 0$), ϕ is equal to π; if Z_L is real and less than Z_0, ϕ is also equal to π; and if Z_L is real and greater than Z_0, ϕ is zero. Using (80), we may rewrite (79) in the form:

$$V_{sT}(z) = V_0 \left(e^{-j\beta z} + |\Gamma| e^{j(\beta z + \phi)}\right) = V_0 e^{j\phi/2} \left(e^{-j\beta z} e^{-j\phi/2} + |\Gamma| e^{j\beta z} e^{j\phi/2}\right) \tag{81}$$

To express (81) in a more useful form, we can apply the algebraic trick of adding and subtracting the term $V_0(1 - |\Gamma|)e^{-j\beta z}$:

$$V_{sT}(z) = V_0(1 - |\Gamma|)e^{-j\beta z} + V_0 |\Gamma| e^{j\phi/2} \left(e^{-j\beta z} e^{-j\phi/2} + e^{j\beta z} e^{j\phi/2}\right) \tag{82}$$

The last term in parentheses in (82) becomes a cosine, and we write

$$\boxed{V_{sT}(z) = V_0(1 - |\Gamma|)e^{-j\beta z} + 2V_0 |\Gamma| e^{j\phi/2} \cos(\beta z + \phi/2)} \tag{83}$$

The important characteristics of this result are most easily seen by converting it to real instantaneous form:

$$\mathcal{V}(z,t) = \text{Re}[V_{sT}(z)e^{j\omega t}] = \underbrace{V_0(1-|\Gamma|)\cos(\omega t - \beta z)}_{traveling\ wave} + \underbrace{2|\Gamma|V_0\cos(\beta z + \phi/2)\cos(\omega t + \phi/2)}_{standing\ wave} \tag{84}$$

Equation (84) is recognized as the sum of a traveling wave of amplitude $(1-|\Gamma|)V_0$ and a standing wave having amplitude $2|\Gamma|V_0$. We can visualize events as follows: The portion of the incident wave that reflects and back-propagates in the slotted line interferes with an equivalent portion of the incident wave to form a standing wave. The rest of the incident wave (which does not interfere) is the traveling wave part of (84). The maximum amplitude observed in the line is found where the amplitudes of the two terms in (84) add directly to give $(1+|\Gamma|)V_0$. The minimum amplitude is found where the standing wave achieves a null, leaving only the traveling wave amplitude of $(1-|\Gamma|)V_0$. The fact that the two terms in (84) combine in this way with the proper phasing is not immediately apparent, but the following arguments will show that this does occur.

To obtain the minimum and maximum voltage amplitudes, we may revisit the first part of Eq. (81):

$$V_{sT}(z) = V_0\left(e^{-j\beta z} + |\Gamma|e^{j(\beta z+\phi)}\right) \tag{85}$$

First, the minimum voltage amplitude is obtained when the two terms in (85) subtract directly (having a phase difference of π). This occurs at locations

$$z_{\text{min}} = -\frac{1}{2\beta}(\phi + (2m+1)\pi) \quad (m = 0, 1, 2, \ldots) \tag{86}$$

Note again that all positions within the slotted line occur at negative values of z. Substituting (86) into (85) leads to the minimum amplitude:

$$V_{sT}(z_{\text{min}}) = V_0(1-|\Gamma|) \tag{87}$$

The same result is obtained by substituting (86) into the real voltage, (84). This produces a null in the standing wave part, and we obtain

$$\mathcal{V}(z_{\text{min}},t) = \pm V_0(1-|\Gamma|)\sin(\omega t + \phi/2) \tag{88}$$

The voltage oscillates (through zero) in time, with amplitude $V_0(1-|\Gamma|)$. The plus and minus signs in (88) apply to even and odd values of m in (86), respectively.

Next, the maximum voltage amplitude is obtained when the two terms in (85) add in-phase. This will occur at locations given by

$$z_{\text{max}} = -\frac{1}{2\beta}(\phi + 2m\pi) \quad (m = 0, 1, 2, \ldots) \tag{89}$$

On substituting (89) into (85), we obtain

$$V_{sT}(z_{\max}) = V_0(1 + |\Gamma|) \tag{90}$$

As before, we may substitute (89) into the real instantaneous voltage (84). The effect is to produce a maximum in the standing wave part, which then adds in-phase to the running wave. The result is

$$\mathcal{V}(z_{\max}, t) = \pm V_0(1 + |\Gamma|)\cos(\omega t + \phi/2) \tag{91}$$

where the plus and minus signs apply to positive and negative values of m in (89), respectively. Again, the voltage oscillates through zero in time, with amplitude $V_0(1 + |\Gamma|)$.

Note that a voltage maximum is located at the load ($z = 0$) if $\phi = 0$; moreover, $\phi = 0$ when Γ is real and positive. This occurs for real Z_L when $Z_L > Z_0$. Thus there is a voltage maximum at the load when the load impedance is greater than Z_0 and both impedances are real. With $\phi = 0$, maxima also occur at $z_{\max} = -m\pi/\beta = -m\lambda/2$. For a zero-load impedance, $\phi = \pi$, and the maxima are found at $z_{\max} = -\pi/(2\beta)$, $-3\pi/(2\beta)$, or $z_{\max} = -\lambda/4$, $-3\lambda/4$, and so forth.

The minima are separated by multiples of one half-wavelength (as are the maxima), and for a zero load impedance, the first minimum occurs when $-\beta z = 0$, or at the load. In general, a voltage minimum is found at $z = 0$ whenever $\phi = \pi$; this occurs if $Z_L < Z_0$ where Z_L is real. The general results are illustrated in Figure 10.6.

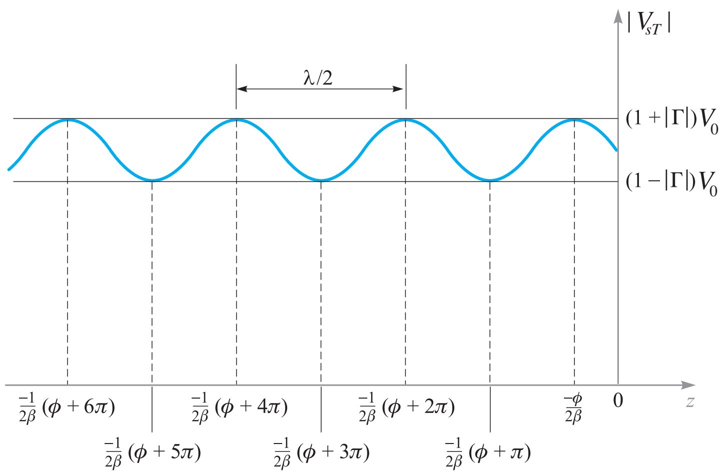

Figure 10.6 Plot of the magnitude of V_{sT} as found from Eq. (85) as a function of position, z, in front of the load (at $z = 0$). The reflection coefficient phase is ϕ, which leads to the indicated locations of maximum and minimum voltage amplitude, as found from Eqs. (86) and (89).

Finally, the voltage standing wave ratio is defined as:

$$\boxed{s \equiv \frac{V_{sT}(z_{\max})}{V_{sT}(z_{\min})} = \frac{1+|\Gamma|}{1-|\Gamma|}} \tag{92}$$

Since the absolute voltage amplitudes have divided out, our measured VSWR permits the immediate evaluation of $|\Gamma|$. The phase of Γ is then found by measuring the location of the first maximum or minimum with respect to the load, and then using (86) or (89) as appropriate. Once Γ is known, the load impedance can be found, assuming Z_0 is known.

WW Illustrations

D10.3. What voltage standing wave ratio results when $\Gamma = \pm 1/2$?

Ans. 3

EXAMPLE 10.7

Slotted line measurements yield a VSWR of 5, a 15-cm spacing between successive voltage maxima, and the first maximum at a distance of 7.5 cm in front of the load. Determine the load impedance, assuming a 50-Ω impedance for the slotted line.

Solution. The 15-cm spacing between maxima is $\lambda/2$, implying a wavelength of 30 cm. Because the slotted line is air-filled, the frequency is $f = c/\lambda = 1$ GHz. The first maximum at 7.5 cm is thus at a distance of $\lambda/4$ from the load, which means that a voltage minimum occurs at the load. Thus Γ will be real and negative. We use (92) to write

$$|\Gamma| = \frac{s-1}{s+1} = \frac{5-1}{5+1} = \frac{2}{3}$$

So

$$\Gamma = -\frac{2}{3} = \frac{Z_L - Z_0}{Z_L + Z_0}$$

which we solve for Z_L to obtain

$$Z_L = \frac{1}{5} Z_0 = \frac{50}{5} = 10\ \Omega$$

10.11 TRANSMISSION LINES OF FINITE LENGTH

A new type of problem emerges when considering the propagation of sinusoidal voltages on finite-length lines that have loads that are not impedance matched. In such cases, numerous reflections occur at the load and at the generator, setting up a multiwave bidirectional voltage distribution in the line. As always, the objective is to determine the net power transferred to the load in steady state, but we must now include the effect of the numerous forward- and backward-reflected waves.

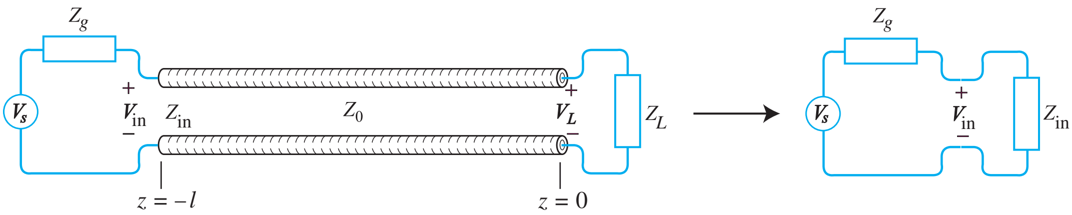

Figure 10.7 Finite-length transmission line configuration and its equivalent circuit.

Figure 10.7 shows the basic problem. The line, assumed to be lossless, has characteristic impedance Z_0 and is of length l. The sinusoidal voltage source at frequency ω provides phasor voltage V_s. Associated with the souce is a complex internal impedance, Z_g, as shown. The load impedance, Z_L, is also assumed to be complex and is located at $z = 0$. The line thus exists along the negative z axis. The easiest method of approaching the problem is not to attempt to analyze every reflection individually, but rather to recognize that in steady state, there will exist one net forward wave and one net backward wave, representing the superposition of all waves that are incident on the load and all waves that are reflected from it. We may thus write the total voltage in the line as

$$V_{sT}(z) = V_0^+ e^{-j\beta z} + V_0^- e^{j\beta z} \tag{93}$$

in which V_0^+ and V_0^- are complex amplitudes, composed respectively of the sum of all individual forward and backward wave amplitudes and phases. In a similar way, we may write the total current in the line:

$$I_{sT}(z) = I_0^+ e^{-j\beta z} + I_0^- e^{j\beta z} \tag{94}$$

We now define the *wave impedance,* $Z_w(z)$, as the ratio of the total phasor voltage to the total phasor current. Using (93) and (94), this becomes:

$$\boxed{Z_w(z) \equiv \frac{V_{sT}(z)}{I_{sT}(z)} = \frac{V_0^+ e^{-j\beta z} + V_0^- e^{j\beta z}}{I_0^+ e^{-j\beta z} + I_0^- e^{j\beta z}}} \tag{95}$$

We next use the relations $V_0^- = \Gamma V_0^+$, $I_0^+ = V_0^+/Z_0$, and $I_0^- = -V_0^-/Z_0$. Eq. (95) simplifies to

$$Z_w(z) = Z_0 \left[\frac{e^{-j\beta z} + \Gamma e^{j\beta z}}{e^{-j\beta z} - \Gamma e^{j\beta z}} \right] \tag{96}$$

Now, using the Euler identity, (32), and substituting $\Gamma = (Z_L - Z_0)/(Z_L + Z_0)$, Eq. (96) becomes

$$Z_w(z) = Z_0 \left[\frac{Z_L \cos(\beta z) - jZ_0 \sin(\beta z)}{Z_0 \cos(\beta z) - jZ_L \sin(\beta z)} \right] \tag{97}$$

The wave impedance at the line input is now found by evaluating (97) at $z = -l$, obtaining

$$\boxed{Z_{\text{in}} = Z_0 \left[\frac{Z_L \cos(\beta l) + jZ_0 \sin(\beta l)}{Z_0 \cos(\beta l) + jZ_L \sin(\beta l)} \right]} \tag{98}$$

This is the quantity that we need in order to create the equivalent circuit in Figure 10.7.

One special case is that in which the line length is a half-wavelength, or an integer multiple thereof. In that case,

$$\beta l = \frac{2\pi}{\lambda} \frac{m\lambda}{2} = m\pi \quad (m = 0, 1, 2, \ldots)$$

Using this result in (98), we find

$$\boxed{Z_{\text{in}}(l = m\lambda/2) = Z_L} \tag{99}$$

For a half-wave line, the equivalent circuit can be constructed simply by removing the line completely and placing the load impedance at the input. This simplification works, of course, provided the line length is indeed an integer multiple of a half-wavelength. Once the frequency begins to vary, the condition is no longer satisfied, and (98) must be used in its general form to find Z_{in}.

Another important special case is that in which the line length is an odd multiple of a quarter wavelength:

$$\beta l = \frac{2\pi}{\lambda}(2m+1)\frac{\lambda}{4} = (2m+1)\frac{\pi}{2} \quad (m = 0, 1, 2, \ldots)$$

Using this result in (98) leads to

$$\boxed{Z_{\text{in}}(l = \lambda/4) = \frac{Z_0^2}{Z_L}} \tag{100}$$

An immediate application of (100) is to the problem of joining two lines having different characteristic impedances. Suppose the impedances are (from left to right) Z_{01} and Z_{03}. At the joint, we may insert an additional line whose characteristic impedance is Z_{02} and whose length is $\lambda/4$. We thus have a sequence of joined lines whose impedances progress as Z_{01}, Z_{02}, and Z_{03}, in that order. A voltage wave is now incident from line 1 onto the joint between Z_{01} and Z_{02}. Now the effective load at the far end of line 2 is Z_{03}. The input impedance to line 2 at any frequency is now

$$Z_{\text{in}} = Z_{02} \frac{Z_{03} \cos\beta_2 l + jZ_{02} \sin\beta_2 l}{Z_{02} \cos\beta_2 l + jZ_{03} \sin\beta_2 l} \tag{101}$$

Then, since the length of line 2 is $\lambda/4$,

$$Z_{\text{in}}(\text{line } 2) = \frac{Z_{02}^2}{Z_{03}} \tag{102}$$

Reflections at the Z_{01}–Z_{02} interface will not occur if $Z_{\text{in}} = Z_{01}$. Therefore, we can match the junction (allowing complete transmission through the three-line sequence) if Z_{02} is chosen so that

$$\boxed{Z_{02} = \sqrt{Z_{01} Z_{03}}} \tag{103}$$

This technique is called *quarter-wave matching* and again is limited to the frequency (or narrow band of frequencies) such that $l \doteq (2m+1)\lambda/4$. We will encounter more examples of these techniques when we explore electromagnetic wave reflection in Chapter 12. Meanwhile, further examples that involve the use of the input impedance and the VSWR are presented in Section 10.12.

10.12 SOME TRANSMISSION LINE EXAMPLES

In this section, we apply many of the results that we obtained in the previous sections to several typical transmission line problems. We simplify our work by restricting our attention to the lossless line.

Let us begin by assuming a two-wire 300 Ω line ($Z_0 = 300$ Ω), such as the lead-in wire from the antenna to a television or FM receiver. The circuit is shown in Figure 10.8. The line is 2 m long, and the values of L and C are such that the velocity on the line is 2.5×10^8 m/s. We will terminate the line with a receiver having an input resistance of 300 Ω and represent the antenna by its Thevenin equivalent $Z = 300\ \Omega$ in series with $V_s = 60$ V at 100 MHz. This antenna voltage is larger by a factor of about 10^5 than it would be in a practical case, but it also provides simpler values to work with; in order to think practical thoughts, divide currents or voltages by 10^5, divide powers by 10^{10}, and leave impedances alone.

Because the load impedance is equal to the characteristic impedance, the line is matched; the reflection coefficient is zero, and the standing wave ratio is unity. For the given velocity and frequency, the wavelength on the line is $v/f = 2.5$ m, and the phase constant is $2\pi/\lambda = 0.8\pi$ rad/m; the attenuation constant is zero. The electrical length of the line is $\beta l = (0.8\pi)2$, or 1.6π rad. This length may also be expressed as 288°, or 0.8 wavelength.

The input impedance offered to the voltage source is 300 Ω, and since the internal impedance of the source is 300 Ω, the voltage at the input to the line is half of 60 V, or 30 V. The source is matched to the line and delivers the maximum available power

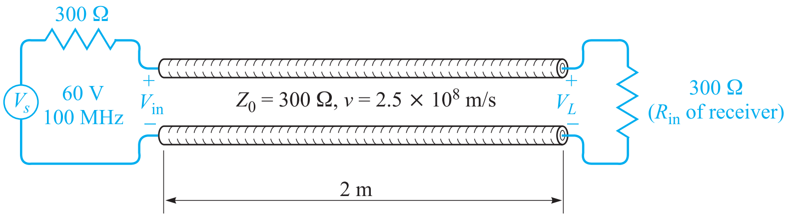

Figure 10.8 A transmission line that is matched at both ends produces no reflections and thus delivers maximum power to the load.

to the line. Because there is no reflection and no attenuation, the voltage at the load is 30 V, but it is delayed in phase by 1.6π rad. Thus,

$$V_{\text{in}} = 30\cos(2\pi 10^8 t)\ \text{V}$$

whereas

$$V_L = 30\cos(2\pi 10^8 t - 1.6\pi)\ \text{V}$$

The input current is

$$I_{\text{in}} = \frac{V_{\text{in}}}{300} = 0.1\cos(2\pi 10^8 t)\ \text{A}$$

while the load current is

$$I_L = 0.1\cos(2\pi 10^8 t - 1.6\pi)\ \text{A}$$

The average power delivered to the input of the line by the source must all be delivered to the load by the line,

$$P_{\text{in}} = P_L = \frac{1}{2} \times 30 \times 0.1 = 1.5\ \text{W}$$

Now let us connect a second receiver, also having an input resistance of 300 Ω, across the line in parallel with the first receiver. The load impedance is now 150 Ω, the reflection coefficient is

$$\Gamma = \frac{150 - 300}{150 + 300} = -\frac{1}{3}$$

and the standing wave ratio on the line is

$$s = \frac{1 + \frac{1}{3}}{1 - \frac{1}{3}} = 2$$

The input impedance is no longer 300 Ω, but is now

$$Z_{\text{in}} = Z_0 \frac{Z_L \cos\beta l + jZ_0 \sin\beta l}{Z_0 \cos\beta l + jZ_L \sin\beta l} = 300\,\frac{150\cos 288^\circ + j300\sin 288^\circ}{300\cos 288^\circ + j150\sin 288^\circ}$$
$$= 510\angle{-23.8^\circ} = 466 - j206\ \Omega$$

which is a capacitive impedance. Physically, this means that this length of line stores more energy in its electric field than in its magnetic field. The input current phasor is thus

$$I_{s,\text{in}} = \frac{60}{300 + 466 - j206} = 0.0756\angle 15.0^\circ\ \text{A}$$

and the power supplied to the line by the source is

$$P_{\text{in}} = \frac{1}{2} \times (0.0756)^2 \times 466 = 1.333\ \text{W}$$

Since there are no losses in the line, 1.333 W must also be delivered to the load. Note that this is less than the 1.50 W which we were able to deliver to a matched load; moreover, this power must divide equally between two receivers, and thus each receiver now receives only 0.667 W. Because the input impedance of each receiver is 300 Ω, the voltage across the receiver is easily found as

$$0.667 = \frac{1}{2}\frac{|V_{s,L}|^2}{300}$$

$$|V_{s,L}| = 20 \text{ V}$$

in comparison with the 30 V obtained across the single load.

Before we leave this example, let us ask ourselves several questions about the voltages on the transmission line. Where is the voltage a maximum and a minimum, and what are these values? Does the phase of the load voltage still differ from the input voltage by 288°? Presumably, if we can answer these questions for the voltage, we could do the same for the current.

Equation (89) serves to locate the voltage maxima at

$$z_{\max} = -\frac{1}{2\beta}(\phi + 2m\pi) \quad (m = 0, 1, 2, \ldots)$$

where $\Gamma = |\Gamma|e^{j\phi}$. Thus, with $\beta = 0.8\pi$ and $\phi = \pi$, we find

$$z_{\max} = -0.625 \quad \text{and} \quad -1.875 \text{ m}$$

while the minima are $\lambda/4$ distant from the maxima;

$$z_{\min} = 0 \quad \text{and} \quad -1.25 \text{ m}$$

and we find that the load voltage (at $z = 0$) is a voltage minimum. This, of course, verifies the general conclusion we reached earlier: a voltage minimum occurs at the load if $Z_L < Z_0$, and a voltage maximum occurs if $Z_L > Z_0$, where both impedances are pure resistances.

The minimum voltage on the line is thus the load voltage, 20 V; the maximum voltage must be 40 V, since the standing wave ratio is 2. The voltage at the input end of the line is

$$V_{s,\text{in}} = I_{s,\text{in}} Z_{\text{in}} = (0.0756\angle 15.0^\circ)(510\angle{-23.8^\circ}) = 38.5\angle{-8.8^\circ}$$

The input voltage is almost as large as the maximum voltage anywhere on the line because the line is about three-quarters of a wavelength long, a length which would place the voltage maximum at the input when $Z_L < Z_0$.

Finally, it is of interest to determine the load voltage in magnitude *and phase*. We begin with the total voltage in the line, using (93).

$$V_{sT} = \left(e^{-j\beta z} + \Gamma e^{j\beta z}\right) V_0^+ \tag{104}$$

We may use this expression to determine the voltage at any point on the line in terms of the voltage at any other point. Because we know the voltage at the input to the line, we let $z = -l$,

$$V_{s,\text{in}} = \left(e^{j\beta l} + \Gamma e^{-j\beta l}\right) V_0^+ \tag{105}$$

and solve for V_0^+,

$$V_0^+ = \frac{V_{s,\text{in}}}{e^{j\beta l} + \Gamma e^{-j\beta l}} = \frac{38.5\angle{-8.8^\circ}}{e^{j1.6\pi} - \frac{1}{3}e^{-j1.6\pi}} = 30.0\angle 72.0^\circ \text{ V}$$

We may now let $z = 0$ in (104) to find the load voltage,

$$V_{s,L} = (1 + \Gamma)V_0^+ = 20\angle 72^\circ = 20\angle{-288^\circ}$$

The amplitude agrees with our previous value. The presence of the reflected wave causes $V_{s,\text{in}}$ and $V_{s,L}$ to differ in phase by about -279° instead of -288°.

EXAMPLE 10.8

In order to provide a slightly more complicated example, let us now place a purely capacitive impedance of $-j300\,\Omega$ in parallel with the two 300 Ω receivers. We are to find the input impedance and the power delivered to each receiver.

Solution. The load impedance is now 150 Ω in parallel with $-j300\,\Omega$, or

$$Z_L = \frac{150(-j300)}{150 - j300} = \frac{-j300}{1 - j2} = 120 - j60\ \Omega$$

We first calculate the reflection coefficient and the VSWR:

$$\Gamma = \frac{120 - j60 - 300}{120 - j60 + 300} = \frac{-180 - j60}{420 - j60} = 0.447\angle{-153.4^\circ}$$

$$s = \frac{1 + 0.447}{1 - 0.447} = 2.62$$

Thus, the VSWR is higher and the mismatch is therefore worse. Let us next calculate the input impedance. The electrical length of the line is still 288°, so that

$$Z_{\text{in}} = 300\,\frac{(120 - j60)\cos 288^\circ + j300 \sin 288^\circ}{300 \cos 288^\circ + j(120 - j60) \sin 288^\circ} = 755 - j138.5\ \Omega$$

This leads to a source current of

$$I_{s,\text{in}} = \frac{V_{Th}}{Z_{Th} + Z_{\text{in}}} = \frac{60}{300 + 755 - j138.5} = 0.0564\angle 7.47^\circ\ \text{A}$$

Therefore, the average power delivered to the input of the line is $P_{\text{in}} = \frac{1}{2}(0.0564)^2(755) = 1.200$ W. Since the line is lossless, it follows that $P_L = 1.200$ W, and each receiver gets only 0.6 W.

EXAMPLE 10.9

As a final example, let us terminate our line with a purely capacitive impedance, $Z_L = -j300\,\Omega$. We seek the reflection coefficient, the VSWR, and the power delivered to the load.

Solution. Obviously, we cannot deliver any average power to the load since it is a pure reactance. As a consequence, the reflection coefficient is

$$\Gamma = \frac{-j300 - 300}{-j300 + 300} = -j1 = 1\angle{-90^\circ}$$

and the reflected wave is equal in amplitude to the incident wave. Hence, it should not surprise us to see that the VSWR is

$$s = \frac{1 + |-j1|}{1 - |-j1|} = \infty$$

and the input impedance is a pure reactance,

$$Z_{\text{in}} = 300\,\frac{-j300\cos 288^\circ + j300\sin 288^\circ}{300\cos 288^\circ + j(-j300)\sin 288^\circ} = j589$$

Thus, no average power can be delivered to the input impedance by the source, and therefore no average power can be delivered to the load.

Although we could continue to find numerous other facts and figures for these examples, much of the work may be done more easily for problems of this type by using graphical techniques. We encounter these in Section 10.13.

D10.4. A 50 W lossless line has a length of 0.4λ. The operating frequency is 300 MHz. A load $Z_L = 40 + j30\ \Omega$ is connected at $z = 0$, and the Thevenin-equivalent source at $z = -l$ is $12\angle 0^\circ$ V in series with $Z_{Th} = 50 + j0\ \Omega$. Find: (*a*) Γ; (*b*) s; (*c*) Z_{in}.

Ans. $0.333\angle 90^\circ$; 2.00; $25.5 + j5.90\ \Omega$

D10.5. For the transmission line of Problem D10.4, also find: (*a*) the phasor voltage at $z = -l$; (*b*) the phasor voltage at $z = 0$; (*c*) the average power delivered to Z_L.

Ans. $4.14\angle 8.58^\circ$ V; $6.32\angle -125.6^\circ$ V; 0.320 W

10.13 GRAPHICAL METHODS: THE SMITH CHART

Transmission line problems often involve manipulations with complex numbers, making the time and effort required for a solution several times greater than are needed for a similar sequence of operations on real numbers. One means of reducing the labor without seriously affecting the accuracy is by using transmission-line charts. Probably the most widely used one is the Smith chart.[3]

Basically, this diagram shows curves of constant resistance and constant reactance; these may represent either an input impedance or a load impedance. The latter, of course, is the input impedance of a zero-length line. An indication of location along the line is also provided, usually in terms of the fraction of a wavelength from a voltage maximum or minimum. Although they are not specifically shown on the chart, the standing-wave ratio and the magnitude and angle of the reflection coefficient are very

[3] P. H. Smith, "Transmission Line Calculator," *Electronics,* vol. 12, pp. 29–31, January 1939.