1 Physical Concepts

This book is concerned with hearing, and what we hear is sound. Thus, both intuition and reason make it clear that a basic understanding of the nature of sound is prerequisite to an understanding of audition. The study of sound is acoustics. An understanding of **acoustics**, in turn, rests upon knowing several fundamental physical principles. This is so because acoustics is, after all, the physics of sound. We will therefore begin by reviewing a number of physical principles so that the following chapters can proceed without the constant need for the distracting insertions of basic definitions and concepts. The material in this chapter is intended to be a review of principles that were previously learned. Therefore, the review will be rapid and somewhat cursory, and the reader may wish to consult the American National Standard addressing acoustical terminology and a physics or acoustics textbook for a broader coverage of these topics (e.g., Pearce and David, 1958; van Bergeijk et al., 1960; Peterson and Gross, 1972; Beranek, 1986; Everest, 2000; Kinsler et al., 1999; Speaks, 1960; Rossing et al., 2002; Hewitt, 2005; Young and Freedman, 2007),[1] as well as the American National Standard addressing acoustical terminology (ANSI, 2004).

PHYSICAL QUANTITIES

Physical quantities may be thought of as being basic or derived, and as either scalars or vectors. The **basic quantities** of concern here are **time**, **length (distance)**, and **mass**. The **derived quantities** are the results of various combinations of the basic quantities (and other derived quantities), and include such phenomena as velocity, force, and work. If a quantity can be described completely in terms of *just* its *magnitude* (size), then it is a **scalar**. Length is a good example of a scalar. On the other hand, a quantity is a **vector** if it needs to be described by *both* its *magnitude* and its *direction*. For example, if a body moves 1 m from point x_1 to point x_2, then we say that it has been displaced. Here, the scalar quantity of length becomes the vector quantity of **displacement** when both magnitude and direction are involved. A derived quantity is a vector if any of its components is a vector. For example, force is a vector because it involves the components of mass (a scalar) and acceleration (a vector). The distinction between scalars and vectors is not just some esoteric concept. One must be able to distinguish between scalars and vectors because they are manipulated differently in calculations.

The basic quantities may be more or less appreciated in terms of one's personal experience and are expressed in terms of conventionally agreed upon units. These units are values that are measurable and repeatable. The unit of **time (t)** is the **second (s)**, the unit of **length (L)** is the **meter (m)**, and the unit of **mass (M)** is the **kilogram (kg)**. There is a common misconception that mass and weight are synonymous. This is actually untrue. Mass is related to the density of a body, which is the same for that body no matter where it is located. On the other hand, an object's weight is related to the force of gravity upon it so that weight changes as a function of gravitational attraction. It is a common knowledge that an object weighs more on the earth than it would on the moon, and that it weighs more at sea level than it would in a high-flying airplane. In each of these cases, the mass of the body is the same despite the fact that its weight is different.

A brief word is appropriate at this stage regarding the availability of several different systems of units. When we express length in meters and mass in kilograms, we are using the units of the *Système International d'Unités*, referred to as the **SI** or the **MKS system**. Here, MKS stands for *meters*, *kilograms*, and *seconds*. An alternative scheme using smaller metric units coexists with MKS, which is the **cgs system** (for *centimeters*, *grams*, and *seconds*), as does the English system of weights and measures. Table 1.1 presents a number of the major basic and derived physical quantities we will deal with, their units, and their conversion factors.[2]

Velocity (v) is the speed at which an object is moving and is derived from the basic quantities of displacement (which we have seen is a vector form of length) and time. On average, velocity is the distance traveled divided by the amount of time it takes to get from the starting point to the destination. Thus, if an object leaves point x_1 at time t_1 and arrives at x_2 at time t_2, then we can compute the average velocity as

$$v = \frac{(x_2 - x_1)}{(t_2 - t_1)}. \tag{1.1}$$

If we call $(x_2 - x_1)$ displacement (x) and $(t_2 - t_1)$ time (t), then, in general we have

$$v = \frac{x}{t}. \tag{1.2}$$

Because displacement (x) is measured in meters and time (t) in seconds, velocity is expressed in meters per second (m/s).

[1] Although no longer in print, the interested student may be able to find the classical books by Pearce and David (1958), van Bergeijk et al. (1960), and Peterson and Gross (1972) in some libraries.

[2] The student with a penchant for trivia will be delighted to know the following details: (1) The reference value for 1 kg of mass is that of a cylinder of platinum–iridium alloy kept in the International Bureau of Weights and Measures in France. (2) One second is the time needed to complete 9,192,631,700 cycles of the microwave radiation that causes a change between the two lowest energy states in a cesium atom. (3) One meter is 1,650,763.73 times the wavelength of orange-red light emitted by krypton-86 under certain conditions.

Table 1.1 Principal Physical Quantities

Quantity	Formula	SI (MKS) units	cgs units	Equivalent values
Time (t)	t	Second (s)	s	
Mass (M)	M	Kilogram (kg)	Gram (g)	1 kg = 1000 g
Displacement (x)	x	Meter (m)	Centimeter (cm)	1 m = 100 cm
Area (A)	A	m^2	cm^2	1 m^2 = 10^4 cm^2
Velocity (v)	v = x/t	m/s	cm/s	1 m/s = 100 cm/s
Acceleration (a)	a = v/t = x/t^2	m/s^2	cm/s^2	1 m/s^2 = 100 cm/s^2
Force (F)	F = Ma = Mv/t	Newton (N), kg·m/s^2	Dyne (d), g·cm/s^2	1 N = 10^5 d
Work (w)	w = Fx	Joule (J), N·m	erg, d·cm	1 J = 10^7 erg
Power (P)	P = w/t = Fx/t = Fv	Watt (W)	Watt (W)	1 W = 1 J/s = 10^7 erg/s
Intensity (I)	I = P/A	W/m^2	W/cm^2	Reference values: 10^{-12} W/m^2 or 10^{-16} W/cm^2
Pressure (p)	p = F/A	Pascal (Pa), N/m^2	Microbar (μbar) d/cm^2	Reference values: 2×10^{-5} N/m^2 (μPa) or 2×10^{-4} d/cm^2 (μbar)[a]

[a]The reference value for sound pressure in cgs units is often written as 0.0002 dynes/cm^2.

In contrast to **average velocity**, as just defined, **instantaneous velocity** is used when we are concerned with the speed of a moving body at a *specific moment* in time. Instantaneous velocity reflects the speed at some point in time when the displacement and time between that point and the next one approaches zero. Thus, students with a background in mathematics will recognize that instantaneous velocity is equal to the derivative of displacement with respect to time, or

$$v = \frac{dx}{dt}. \tag{1.3}$$

As common experience verifies, a fixed speed is rarely maintained over time. Rather, an object may speed up or slow down over time. Such a change of velocity over time is **acceleration (a)**. Suppose we are concerned with the average acceleration of a body moving between two points. The velocity of the body at the first point is v_1 and the time as it passes that point is t_1. Similarly, its velocity at the second point and the time when it passes this point are, respectively, v_2 and t_2. The **average acceleration** is the difference between these two velocities divided by the time interval involved:

$$a = \frac{(v_2 - v_1)}{(t_2 - t_1)} \tag{1.4}$$

or, in general:

$$a = \frac{v}{t}. \tag{1.5}$$

If we recall that velocity corresponds to displacement divided by time (Eq. 1.2), we can substitute x/t for v so that

$$a = \frac{\frac{x}{t}}{t} = \frac{x}{t^2}. \tag{1.6}$$

Therefore, acceleration is expressed in units of meters per second squared (m/s^2) or centimeters per second squared (cm/s^2).

The acceleration of a body at a given moment is called its **instantaneous acceleration**, which is the derivative of velocity with respect to time, or

$$a = \frac{dv}{dt}. \tag{1.7}$$

Recalling that velocity is the first derivative of displacement (Eq. 1.3), and substituting, we find that acceleration is the second derivative of displacement:

$$a = \frac{d^2x}{dt^2}. \tag{1.8}$$

Common experience and Newton's first law of motion tell us that if an object is not moving (is at rest), then it will tend to remain at rest, and that if an object is moving in some direction at a given speed, then it will tend to continue doing so. This phenomenon is **inertia**, which is the property of mass to continue doing what it is already doing. An outside influence is needed to make a stationary object move, or to change the speed or the direction of a moving object. That is, a **force (F)** is needed to overcome the body's inertia. Because a change in speed is acceleration, we may say that force is that which causes a mass to be accelerated, that is, to change its speed or direction. The amount of force is equal to the product of mass and acceleration (Newton's second law of motion):

$$F = Ma. \tag{1.9}$$

Recall that acceleration corresponds to velocity over time (Eq. 1.5). Substituting v/t for a (acceleration) reveals that force can also be defined in the form:

$$F = \frac{Mv}{t}, \tag{1.10}$$

where **Mv** is the property of **momentum**. Stated in this manner, force is equal to momentum over time.

Because force is the product of mass and acceleration, the amount of force is measured in kg·m/s^2. The unit of force is the **newton (N)**, which is the force needed to cause a 1-kg mass to

be accelerated by 1 kg·m/s^2 (i.e., 1 N = 1 kg·m/s^2). It would thus take a 2-N force to cause a 2-kg mass to be accelerated by 1 m/s^2, or a 1-kg mass to be accelerated by 2 kg·m/s^2. Similarly, the force required to accelerate a 6-kg mass by 3 m/s^2 would be 18 N. The unit of force in cgs units is the **dyne**, where 1 dyne = 1 g·cm/s^2 and 10^5 dynes = 1 N.

Actually, many forces tend to act upon a given body at the same time. Therefore, the force referred to in Eqs. 1.9 and 1.10 is actually the resultant or the net force, which is the net effect of all forces acting upon the object. The concept of net force is clarified by a few simple examples: If two forces are both pushing on a body in the same direction, then the net force would be the sum of these two forces. (For example, consider a force of 2 N that is pushing an object toward the north, and a second force of 5 N that is also pushing that object in the same direction. The net force would be 2 N + 5 N, or 7 N and the direction of acceleration would be to the north.) Alternatively, if two forces are pushing on the same body but in opposite directions, then the net force is the difference between the two, and the object will be accelerated in the direction of the greater force. (Suppose, for example, that a 2-N force is pushing an object toward the east and a 5-N force is simultaneously pushing it toward the west. Then the net force would be 5 N − 2 N, or 3 N, which would cause the body to accelerate toward the west.)

If two equal forces push in opposite directions, then the net force would be zero, in which case there would be no change in the motion of the object. This situation is called **equilibrium**. Thus, under conditions of equilibrium, if a body is already moving, it will continue in motion, and if it is already at rest, it will remain still. That is, of course, what Newton's first law of motion tells us.

Experience, however, tells us that a moving object in the real world tends to slow down and will eventually come to a halt. This occurs, for example, when a driver shifts to "neutral" and allows his car to coast on a level roadway. Is this a violation of the laws of physics? Clearly, the answer is no. The reason is that in the real world a moving body is constantly in contact with other objects or mediums. The sliding of one body against the other constitutes a force opposing the motion, called **friction** or **resistance**. For example, the coasting automobile is in contact with the surrounding air and the roadway; moreover, its internal parts are also moving one upon the other.

The opposing force of friction depends on two factors. Differing amounts of friction occur depending upon what is sliding on what. The magnitude of friction between two given materials is called the **coefficient of friction**. Although the details of this quantity are beyond current interest, it is easily understood that the coefficient of friction is greater for "rough" materials than for "smooth" or "slick" ones.

The second factor affecting the force of friction is easily demonstrated by an experiment the reader can do by rubbing the palms of his hands back and forth on one another. First rub slowly and then rapidly. Not surprisingly, the rubbing will produce heat. The temperature rise is due to the conversion of the mechanical energy into heat as a result of the friction, and will be addressed again in another context. For the moment, we will accept the amount of heat as an indicator of the amount of friction. Note that the hands become hotter when they are rubbed together more rapidly. Thus, the amount of friction is due not only to the coefficient of friction (R) between the materials involved (here, the palms of the hands), but also to the velocity (v) of the motion. Stated as a formula, the force of friction (F) is thus

$$F = Rv. \tag{1.11}$$

A compressed spring will bounce back to its original shape once released. This property of a deformed object to return to its original form is called **elasticity**. The more elastic or stiff an object, the more readily it returns to its original form after being deformed. Suppose one is trying to compress a coil spring. It becomes increasingly more difficult to continue squeezing the spring as it becomes more and more compressed. Stated differently, the more the spring is being deformed, the more it opposes the applied force. The force that opposes the deformation of a spring-like material is called the **restoring force**.

As the example just cited suggests, the restoring force depends on two factors: the elastic modulus of the object's material and the degree to which the object is displaced. An **elastic modulus** is the ratio of stress to strain. **Stress** (s) is the ratio of the applied force (F) to the area (A) of an elastic object over which it is exerted, or

$$s = \frac{F}{A} \tag{1.12}$$

The resulting relative displacement or change in dimensions of the material subjected to the stress is called **strain**. Of particular interest is **Young's modulus**, which is the ratio of compressive stress to compressive strain. **Hooke's law** states that stress and strain are proportional within the elastic limits of the material, which is equivalent to stating that a material's elastic modulus is a constant within these limits. Thus, the restoring force (F) of an elastic material that opposes an applied force is

$$F = Sx \tag{1.13}$$

where S is the stiffness constant of the material and x is the amount of displacement.

The concept of "work" in physics is decidedly more specific than its general meaning in daily life. In the physical sense, **work (w)** is done when the application of a force to a body results in its displacement. The amount of work is therefore the product of the force applied and the resultant displacement, or

$$w = Fx \tag{1.14}$$

Thus, work can be accomplished only when there is displacement: If the displacement is zero, then the product of force and displacement will also be zero no matter how great the force. Work is quantified in newton-meters (N·m), and the unit of work is the **joule (J)**. Specifically, one joule (1 J) is equal to

1 N·m. In the cgs system, work is expressed in **ergs**, where 1 erg corresponds to 1 dyne-centimeter (1 d·cm).

The capability to do work is called **energy**. The energy of an object in motion is called **kinetic energy**, and the energy of a body at rest is its **potential energy**. **Total energy** is the body's kinetic energy plus its potential energy. Work corresponds to the change in the body's kinetic energy. The energy is not consumed, but rather is converted from one form to the other. Consider, for example, a pendulum that is swinging back and forth. Its kinetic energy is greatest when it is moving the fastest, which is when it passes through the midpoint of its swing. On the other hand, its potential energy is greatest at the instant that it reaches the extreme of its swing, when its speed is zero.

We are concerned not only with the amount of work, but also with how fast it is being accomplished. The rate at which work is done is **power (P)** and is equal to work divided by time,

$$P = \frac{w}{t} \tag{1.15}$$

in joules per second (J/s). The **watt (W)** is the unit of power, and 1 W is equal to 1 J/s. In the cgs system, the watt is equal to 10^7 ergs/s.

Recalling that w = Fx, Eq. 1.15 may be rewritten as

$$P = \frac{Fx}{t} \tag{1.16}$$

If we now substitute v for x/t (based on Eq. 1.2), we find that

$$P = Fv \tag{1.17}$$

Thus, power is equal to the product of force and velocity.

The amount of power per unit of area is called **intensity (I)**. In formal terms,

$$I = \frac{P}{A} \tag{1.18}$$

where I is intensity, P is power, and A is area. Therefore, intensity is measured in watts per square meter (W/m^2) in SI units, or in watts per square centimeter (W/cm^2) in cgs units. Because of the difference in the scale of the area units in the MKS and cgs systems, we find that $\mathbf{10^{-12}\ W/m^2}$ corresponds to $\mathbf{10^{-16}\ W/cm^2}$. This apparently peculiar choice of equivalent values is being provided because they represent the amount of intensity required to just barely hear a sound.

An understanding of intensity will be better appreciated if one considers the following. Using for the moment the common-knowledge idea of what sound is, imagine that a sound source is a tiny pulsating sphere. This *point source* of sound will produce a sound wave that will radiate outward in every direction so that the propagating wave may be conceived of as a sphere of ever-increasing size. Thus, as distance from the point source increases, the power of the sound will have to be divided over the ever-expanding surface. Suppose now that we measure how much power registers on a one-unit area of this surface at various distances from the source. As the overall size of the sphere is getting larger with distance from the source, so this one-unit sample must represent an ever-decreasing proportion of the total surface area. Therefore, less power "falls" onto the same area as the distance from the source increases. It follows that the magnitude of the sound appreciated by a listener would become less and less with increasing distance from a sound source.

The intensity of a sound decreases with distance from the source according to an orderly rule as long as there are no reflections, in which case a **free field** is said to exist. Under these conditions, increasing the distance (D) from a sound source causes the intensity to decrease to an amount equal to 1 over the square of the change in distance ($1/D^2$). This principle is known as the inverse-square law. In effect, the **inverse square law** says that doubling the distance from the sound source (e.g., from 1 to 2 m) causes the intensity to drop to $1/2^2$ or 1/4 of the original intensity. Similarly, tripling the distance causes the intensity to fall to $1/3^2$, or 1/9, of the prior value; four times the distance results in $1/4^2$, or 1/16, of the intensity; and a 10-fold increase in distance causes the intensity to fall $1/10^2$, or 1/100, of the starting value.

Just as power divided by area yields intensity, so force (F) divided by area yields a value called **pressure (p)**:

$$p = \frac{F}{A} \tag{1.19}$$

so that pressure is measured in N/m^2 or in dynes/cm^2. The unit of pressure is called the **pascal (Pa)**, where 1 Pa = 1 N/m^2. As for intensity, the softest audible sound can also be expressed in terms of its pressure, for which $\mathbf{2 \times 10^{-5}\ N/m^2}$ and $\mathbf{2 \times 10^{-4}}$ **dynes/cm²** are equivalent values.

DECIBEL NOTATION

The range of magnitudes we concern ourselves with in hearing is enormous. As we shall discuss in Chapter 9, the sound pressure of the loudest sound that we can tolerate is on the order of 10 million times greater than that of the softest audible sound. One can immediately imagine the cumbersome task that would be involved if we were to deal with such an immense range of numbers on a linear scale. The problems involved with and related to such a wide range of values make it desirable to transform the absolute physical magnitudes into another form, called **decibels (dB)**, which make the values both palatable and rationally meaningful.

One may conceive of the decibel as basically involving two characteristics, namely, ratios and logarithms. First, the value of a quantity is expressed in relation to some meaningful baseline value in the form of a ratio. Because it makes sense to use the softest sound one can hear as our baseline, we use the intensity or pressure of the softest audible sound as our reference value.

As introduced earlier, the **reference sound intensity** is 10^{-12} W/m^2, and the equivalent **reference sound pressure** is 2×10^{-5} N/m^2. Also, recall that the equivalent corresponding values in cgs units are 10^{-16} W/cm^2 for sound intensity and 2×10^{-4}

dynes/cm^2 for sound pressure. The appropriate reference value becomes the denominator of our ratio, and the absolute intensity or pressure of the sound in question becomes the numerator. Thus, instead of talking about a sound having an absolute intensity of 10^{-10} W/m^2, we express its intensity relatively in terms of how it relates to our reference, as the ratio:

$$\frac{(10^{-10}\,\mathrm{W/m^2})}{(10^{-12}\,\mathrm{W/m^2})},$$

which reduces to simply 10^2. This intensity ratio is then replaced with its common logarithm. The reason is that the linear distance between numbers having the same ratio relationship between them (say, 2:1) becomes wider when the absolute magnitudes of the numbers become larger. For example, the distance between the numbers in each of the following pairs increases appreciably as the size of the numbers becomes larger, even though they all involve the same 2:1 ratio: 1:2, 10:20, 100:200, and 1000:2000. The logarithmic conversion is used because equal ratios are represented as equal distances on a logarithmic scale.

The decibel is a relative entity. This means that the decibel in and of itself is a dimensionless quantity, and is meaningless without knowledge of the reference value, which constitutes the denominator of the ratio. Because of this, it is necessary to make the reference value explicit when the magnitude of a sound is expressed in decibel form. This is accomplished by stating that the magnitude of the sound is whatever number of decibels with respect to the reference quantity. Moreover, it is a common practice to add the word "level" to the original quantity when dealing with decibel values. Intensity expressed in decibels is called **intensity level (IL)**, and sound pressure in decibels is called **sound pressure level (SPL)**. The reference values indicated above are generally assumed when decibels are expressed as **dB IL** or **dB SPL**. For example, one might say that the intensity level of a sound is "50 dB *re*: 10^{-12} W/m^2" or "50 dB IL."

The general formula for the decibel is expressed in terms of power as

$$\mathrm{PL_{dB}} = 10 \cdot \log\left(\frac{\mathrm{P}}{\mathrm{P_0}}\right) \tag{1.20}$$

where P is the power of the sound being measured, P_0 is the reference power to which the former is being compared, and PL is the **power level**. Acoustical measurements are, however, typically made in terms of intensity or sound pressure. The applicable formula for decibels of intensity level is thus:

$$\mathrm{IL_{dB}} = 10 \cdot \log\left(\frac{\mathrm{I}}{\mathrm{I_0}}\right) \tag{1.21}$$

where I is the intensity (in W/m^2) of the sound in question, and I_0 is the reference intensity, or 10^{-12} W/m^2. Continuing with the example introduced above, where the value of I is 10^{-10} W/m^2, we thus find that

$$\begin{aligned}\mathrm{IL_{dB}} &= 10 \cdot \log\left(\frac{10^{-10}\,\mathrm{W/m^2}}{10^{-12}\,\mathrm{W/m^2}}\right)\\ &= 10 \cdot \log 10^2\\ &= 10 \times 2\\ &= 20\,\mathrm{dB} \quad re\colon 10^{-12}\,\mathrm{W/m^2}\end{aligned}$$

In other words, an *intensity* of 10^{-10} W/m^2 corresponds to an **intensity level** of 20 dB *re*: 10^{-12} W/m^2, or 20 dB IL.

Sound intensity measurements are important and useful, and are preferred in certain situations. [See Rassmussen (1989) for a review of this topic.] However, most acoustical measurements involved in hearing are made in terms of sound pressure, and are thus expressed in decibels of **sound pressure level**. Here, we must be aware that intensity is proportional to pressure squared:

$$\mathrm{I} \propto \mathrm{p^2} \tag{1.22}$$

and

$$\mathrm{p} \propto \sqrt{\mathrm{I}} \tag{1.23}$$

As a result, converting the dB IL formula into the equivalent equation for dB SPL involves replacing the intensity values with the squares of the corresponding pressure values. Therefore

$$\mathrm{SPL_{dB}} = 10 \cdot \log\left(\frac{\mathrm{p^2}}{\mathrm{p_0^2}}\right) \tag{1.24}$$

where p is the measured sound pressure and p_0 is the reference sound pressure (2×10^{-5} N/m^2). This formula may be simplified to

$$\mathrm{SPL_{dB}} = 10 \cdot \log\left(\frac{\mathrm{p}}{\mathrm{p_0}}\right)^2 \tag{1.25}$$

Because the logarithm of a number squared corresponds to two times the logarithm of that number ($\log x = 2{\cdot}\log x$), the square may be removed to result in

$$\mathrm{SPL_{dB}} = 10 \cdot 2 \cdot \log\left(\frac{\mathrm{p}}{\mathrm{p_0}}\right) \tag{1.26}$$

Therefore, the simplified formula for decibels of SPL becomes

$$\mathrm{SPL_{dB}} = 20 \cdot \log\left(\frac{\mathrm{p}}{\mathrm{p_0}}\right) \tag{1.27}$$

where the value of 20 (instead of 10) is due to having removed the square from the earlier described version of the formula. (One *cannot* take the intensity ratio from the IL formula and simply insert it into the SPL formula, or vice versa. The square root of the intensity ratio yields the *corresponding* pressure ratio, which must be then placed into the SPL equation. Failure to use the proper terms will result in an erroneous doubling of the value in dB SPL.

By way of an example, a sound pressure of 2×10^{-4} N/m^2 corresponds to a SPL of 20 dB (*re*: 2×10^{-5} N/m^2), which may

be calculated as follows:

$$\begin{aligned} \mathrm{SPL_{dB}} &= 20 \cdot \log\left(\frac{2 \times 10^{-4}\,\mathrm{N/m^2}}{2 \times 10^{-5}\,\mathrm{N/m^2}}\right) \\ &= 20 \cdot \log 10^1 \\ &= 20 \times 1 \\ &= 20\,\mathrm{dB} \quad re\colon 10^{-5}\,\mathrm{N/m^2} \end{aligned}$$

What would happen if the intensity (or pressure) in question were the same as the reference intensity (or pressure)? In other words, what is the dB value of the reference itself? In terms of intensity, the answer to this question may be found by simply using 10^{-12} W/m^2 as both the numerator (I) and denominator (I_0) in the dB formula; thus

$$\mathrm{IL_{dB}} = 10 \cdot \log\left(\frac{10^{-12}\,\mathrm{W/m^2}}{10^{-12}\,\mathrm{W/m^2}}\right) \tag{1.28}$$

Because anything divided by itself equals 1, and the logarithm of 1 is 0, this equation reduces to:

$$\begin{aligned} \mathrm{IL_{dB}} &= 10 \cdot \log 1 \\ &= 10 \times 0 \\ &= 0\,\mathrm{dB} \quad re\colon 10^{-12}\,\mathrm{W/m^2} \end{aligned}$$

Hence, 0 dB IL is the intensity level of the reference intensity. Just as 0 dB IL indicates the intensity level of the reference intensity, so 0 dB SPL similarly implies that the measured sound pressure corresponds to that of the reference

$$\mathrm{SPL_{dB}} = 20 \cdot \log\left(\frac{2 \times 10^{-5}\,\mathrm{N/m^2}}{2 \times 10^{-5}\,\mathrm{N/m^2}}\right) \tag{1.29}$$

Just as we saw in the previous example, this equation is solved simply as follows:

$$\begin{aligned} \mathrm{SPL_{dB}} &= 20 \cdot \log 1 \\ &= 20 \times 0 \\ &= 0\,\mathrm{dB} \quad re\colon 10^{-5}\,\mathrm{N/m^2} \end{aligned}$$

In other words, 0 dB SPL indicates that the pressure of the sound in question corresponds to the reference sound pressure of 2×10^{-5} N/m^2. Notice that 0 dB does *not* mean "no sound." Rather, 0 dB implies that the quantity being measured is equal to the reference quantity. Negative decibel values indicate that the measured magnitude is smaller than the reference quantity.

Recall that sound intensity drops with distance from the sound source according to the inverse-square law. However, we want to know the effect of the inverse-square law in terms of *decibels of sound pressure level* because sound is usually expressed in these terms. To address this, we must first remember that pressure is proportional to the square root of intensity. Hence, pressure decreases according to the inverse of the distance change (1/D) instead of the inverse of the square of the distance change ($1/D^2$). In effect, the *inverse-square law* for *intensity* becomes an *inverse-distance law* when we are dealing with *pressure*. Let us assume a doubling as the distance change, because this is the most useful relationship. We can now calculate the size of the decrease in decibels between a point at some distance from the sound source (D_1, e.g., 1 m) and a point at twice the distance (D_2, e.g., 2 m) as follows:

$$\begin{aligned} \text{Level drop in SPL} &= 20 \cdot \log(D_2/D_1) \\ &= 20 \cdot \log(2/1) \\ &= 20 \cdot \log 2 \\ &= 20 \times 0.3 \\ &= 6\,\mathrm{dB} \end{aligned}$$

In other words, the inverse-square law causes the sound pressure level to decrease by 6 dB whenever the distance from the sound source is doubled. For example, if the sound pressure level is 60 dB at 1 m from the source, then it will be $60-6 = 54$ dB when the distance is doubled to 2 m, and $54-6 = 48$ dB when the distance is doubled again from 2 to 4 m.

HARMONIC MOTION AND SOUND

What is sound? It is convenient to answer this question with a formally stated sweeping generality. For example, one might say that sound is a form of vibration that propagates through a medium (such as air) in the form of a wave. Although this statement is correct and straightforward, it can also be uncomfortably vague and perplexing. This is so because it assumes a knowledge of definitions and concepts that are used in a very precise way, but which are familiar to most people only as "gut-level" generalities. As a result, we must address the underlying concepts and develop a functional vocabulary of physical terms that will not only make the general definition of sound meaningful, but will also allow the reader to appreciate its nature.

Vibration is the to-and-fro motion of a body, which could be anything from a guitar string to the floorboards under the family refrigerator, or a molecule of air. Moreover, the motion may have a very simple pattern as produced by a tuning fork, or an extremely complex one such as what one might hear at lunchtime in an elementary school cafeteria. Even though few sounds are as simple as that produced by a vibrating tuning fork, such an example provides what is needed to understand the nature of sound.

Figure 1.1 shows an artist's conceptualization of a vibrating tuning fork at different moments of its vibration pattern. The heavy arrow facing the prong to the reader's right in Fig. 1.1a represents the effect of applying an initial force to the fork, such as by striking it against a hard surface. The progression of the pictures in the figures from (a) through (e) represents the movements of the prongs as time proceeds from the moment that the outside force is applied.

Even though both prongs vibrate as mirror images of one another, it is convenient to consider just one of them for the time being. Figure 1.2 highlights the right prong's motion after being struck. Point C (center) is simply the position of the prong at rest. Upon being hit (as in Fig. 1.1a) the prong is pushed, as

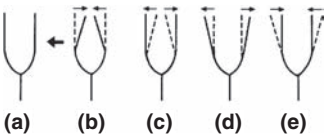

Figure 1.1 Striking a tuning fork (indicated by the heavy arrow) results in a pattern of movement that repeats itself over time. One complete cycle of these movements is represented from frames (a) through (e). Note that the two prongs move as mirror images of one another.

shown by arrow 1, to point L (left). The prong then bounces back (arrow 2), picking up speed along the way. Instead of stopping at the center (C), the rapidly moving prong overshoots this point. It now continues rightward (arrow 3), slowing down along the way until it comes to a halt at point R (right). It now reverses direction and begins moving leftward (arrow 4) at an ever-increasing speed so that it again overshoots the center. Now, again following arrow 1, the prong slows down until it reaches a halt at L, where it reverses direction and repeats the process.

The course of events just described is the result of applying a force to an object having the properties of elasticity and inertia (mass). The initial force to the tuning fork displaces the prong. Because the tuning fork possesses the property of elasticity, the deformation caused by the applied force is opposed by a restoring force in the opposite direction. In the case of the single prong in Fig. 1.2, the initial force toward the left is opposed by a restoring force toward the right. As the prong is pushed farther to the left, the magnitude of the restoring force increases relative to the initially applied force. As a result, the prong's movement is slowed down, brought to a halt at point L, and reversed in direction. Now, under the influence of its elasticity, the prong starts moving rightward. Here, we must consider the mass of the prong.

As the restoring force brings the prong back toward its resting position (C), the inertial force of its mass causes it to increase in speed, or accelerate. When the prong passes through the resting position, it is actually moving fastest. Here, inertia does not permit the moving mass (prong) to simply stop, so instead it

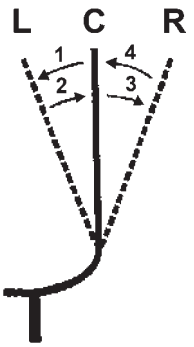

Figure 1.2 Movements toward the right (R) and left (L) of the center (C) resting position of a single tuning fork prong. The numbers and arrows are described in the text.

overshoots the center and continues its rightward movement under the force of its inertia. However, the prong's movement is now resulting in deformation of the metal again once it passes through the resting position. Elasticity therefore comes into play with the buildup of an opposing (now leftward) restoring force. As before, the restoring force eventually equals the applied (now inertial) force, thus halting the fork's displacement at point R and reversing the direction of its movement. Here, the course of events described above again comes into play (except that the direction is leftward), with the prong building up speed again and overshooting the center (C) position as a result of inertia. The process will continue over and over again until it dies out over time, seemingly "of its own accord."

Clearly, the dying out of the tuning fork's vibrations does not occur by some mystical influence. On the contrary, it is due to **resistance**. The vibrating prong is always in contact with the air around it. As a result, there will be **friction** between the vibrating metal and the surrounding air particles. The friction causes some of the mechanical energy involved in the movement of the tuning fork to be converted into heat. The energy that has been converted into heat by friction is no longer available to support the to-and-fro movements of the tuning fork. Hence, the oscillations die out, as continuing friction causes more and more of the energy to be converted into heat. This reduction in the size of the oscillations due to resistance is called **damping**.

The events and forces just described are summarized in Fig. 1.3, where the tuning fork's motion is represented by the curve. This curve represents the displacement to the right and left of the center (resting) position as the distance above and below the horizontal line, respectively. Horizontal distance from left to right represents the progression of time. The initial dotted line represents its initial displacement due to the applied force. The elastic restoring forces and inertial forces of the prong's mass are represented by arrows. Finally, damping is shown by the reduction in the displacement of the curve from center as time goes on.

The type of vibration just described is called **simple harmonic motion (SHM)** because the to-and-fro movements repeat themselves at the same rate over and over again. We will discuss the nature of SHM in greater detail below with respect to the motion of air particles in the sound wave.

The tuning fork serves as a sound source by transferring its vibration to the motion of the surrounding air particles (Fig. 1.4). (We will again concentrate on the activity to the right of the fork, remembering that a mirror image of this pattern occurs to the left.) The rightward motion of the tuning fork prong displaces air molecules to its right in the same direction as the prong's motion. These molecules are thus displaced to the right of their resting positions, thereby being forced closer and closer to the particles to their own right. In other words, the air pressure has been increased above its resting (ambient or atmospheric) pressure because the molecules are being compressed. This state is clearly identified by the term "**compression**." The amount of compression (increased air pressure) becomes greater as the

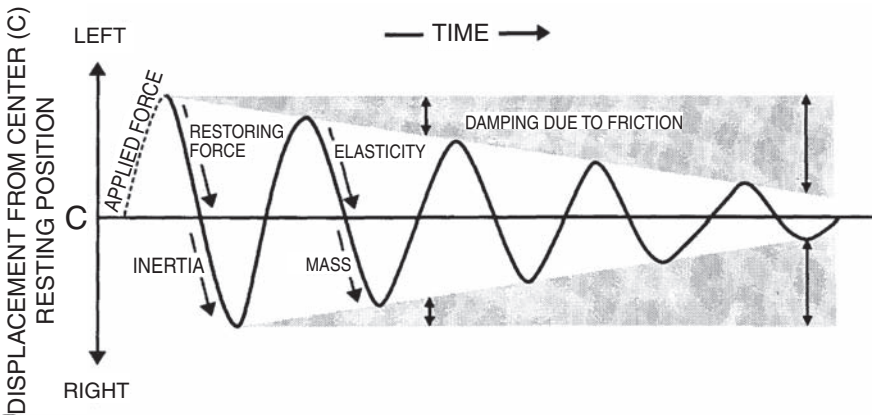

Figure 1.3 Conceptualized diagram graphing the to-and-fro movements of the tuning fork prong in Fig. 2. Vertical distance represents the displacement of the prong from its center (C) or resting position. The dotted line represents the initial displacement of the prong as a result of some applied force. Arrows indicate the effects of restoring forces due to the fork's elasticity, and the effects of inertia due to its mass. The damping effect due to resistance (or friction) is shown by the decreasing displacement of the curve as time progresses and is highlighted by the shaded triangles (and double-headed arrows) above and below the curve.

tuning fork continues displacing the air molecules rightward; it reaches a maximum positive pressure when the prong and air molecules attain their greatest rightward amplitude.

The prong will now reverse direction, overshoot its resting position, and then proceed to its extreme leftward position. The compressed air molecules will also reverse direction along with the prong. The reversal occurs because air is an elastic medium, so the rightwardly compressed particles undergo a leftward restoring force. The rebounding air molecules accelerate due to mass effects, overshoot their resting position, and continue to an extreme leftward position. The amount of compression decreases as the molecules travel leftward, and falls to zero at the moment when the molecules pass through their resting positions.

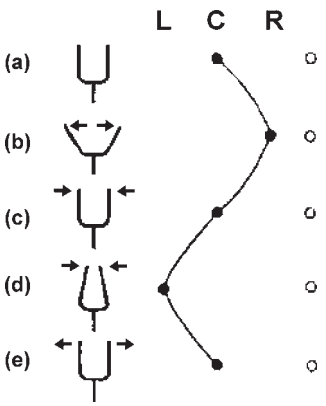

Figure 1.4 Transmittal of the vibratory pattern from a tuning fork to the surrounding air particles. Frames represent various phases of the tuning fork's vibratory cycle. In each frame, the filled circle represents an air particle next to the prong as well as its position, and the unfilled circle shows an air molecule adjacent to the first one. The latter particle is shown only in its resting position for illustrative purposes. Letters above the filled circle highlight the relative positions of the oscillating air particle [C, center (resting); L, leftward; R, rightward]. The line connecting the particle's positions going from frames (a) through (e) reveals a cycle of simple harmonic motion.

As the air molecules move left of their ambient positions, they are now at an increasingly greater distance from the molecules to their right than when they were in their resting positions. Consequently, the air pressure is reduced below atmospheric pressure. This state is the opposite of compression and is called **rarefaction**. The air particles are maximally rarefied so that the pressure is maximally negative when the molecules reach the leftmost position. Now, the restoring force yields a rightward movement of the air molecules, enhanced by the push of the tuning fork prong that has also reversed direction. The air molecules now accelerate rightward, overshoot their resting positions (when rarefaction and negative pressure are zero), and continue rightward. Hence, the SHM of the tuning fork has been transmitted to the surrounding air so that the air molecules are now also under SHM.

Consider now one of the air molecules set into SHM by the influence of the tuning fork. This air molecule will vibrate back and forth in the same direction as that of the vibrating prong. When this molecule moves rightward, it will cause a similar displacement of the particle to its own right. Thus, the SHM of the first air molecule is transmitted to the one next to it. The second one similarly initiates vibration of the one to its right, and so forth down the line.

In other words, each molecule moves to and fro around its own resting point, and causes successive molecules to vibrate back and forth around their own resting points, as shown schematically by the arrows marked "individual particles" in Fig. 1.5 Notice in the figure that each molecule stays in its own general location and moves to and fro about this average position, and that it is the vibratory pattern, which is transmitted.

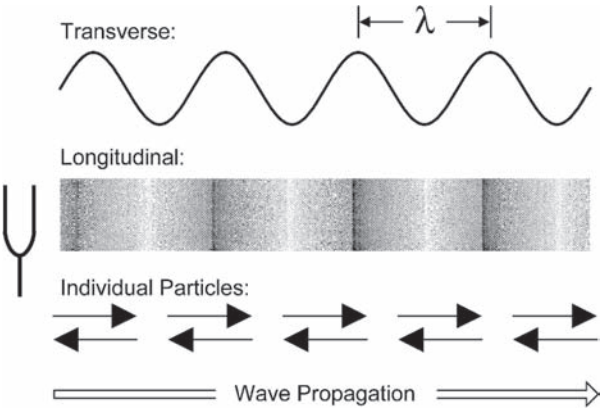

Figure 1.5 Transverse and longitudinal representations of a sinusoidal wave (illustrating points made in the text).

This propagation of vibratory motion from particle to particle constitutes the sound wave. This wave appears as alternating compressions and rarefactions radiating from the sound source as the particles transmit their motions outward and is represented in Fig. 1.5

The distance covered by one cycle of a propagating wave is called its **wavelength** (λ). If we begin where a given molecule is at the point of maximum positive displacement (compression), then the wavelength would be the distance to the next molecule, which is also at its point of maximum compression. This is the distance between any two successive positive peaks in the figure. (Needless to say, such a measurement would be equally correct if made between identical points on any two successive replications of the wave.) The wavelength of a sound is inversely proportional to its frequency, as follows:

$$\lambda = \frac{c}{f} \tag{1.30}$$

where f is frequency and c is a constant representing the speed of sound. (The speed of sound in air approximates 344 m/s at a temperature of 20°C.) Similarly, frequency can be derived if one knows the wavelength, as:

$$f = \frac{c}{\lambda} \tag{1.31}$$

Figure 1.5 reveals that the to-and-fro motions of each air molecule is in the same direction as that in which the overall wave is propagating. This kind of wave, which characterizes sound, is a **longitudinal wave**. In contrast to longitudinal waves, most people are more familiar with **transverse waves**, such as those that develop on the water's surface when a pebble is dropped into a still pool. The latter are called transverse waves because the water particles vibrate up and down around their resting positions at right angles (transverse) to the horizontal propagation of the surface waves out from the spot where the pebble hit the water.

Even though sound waves are longitudinal, it is more convenient to show them diagrammatically as though they were transverse, as in upper part of Fig. 1.5 Here, the dashed horizontal baseline represents the particle's resting position (ambient pressure), distance above the baseline denotes compression (positive pressure), and distance below the baseline shows rarefaction (negative pressure). The passage of time is represented by the distance from left to right. Beginning at the resting position, the air molecule is represented as having gone through one cycle (or complete repetition) of SHM at point 1, two cycles at point 2, three complete cycles at point 3, and four cycles at point 4.

The curves in Fig. 1.5 reveal that the **waveform** of SHM is a sinusoidal function and is thus called a **sinusoidal wave**, also known as a **sine wave** or a **sinusoid**. Figure 1.6 elucidates this concept and also indicates a number of the characteristics of sine waves. The center of the figure shows one complete cycle of SHM, going from points a through i. The circles around the sine wave correspond to the various points on the wave, as indicated by corresponding letters. Circle (a) corresponds to point a on the curve, which falls on the baseline. This point corresponds to the particle's resting position.

Circle (a) shows a horizontal radius (r) drawn from the center to the circumference on the right. Imagine as well a second radius (r′) that will rotate around the circle in a counterclockwise direction. The two radii are superimposed in circle (a) so that the angle between them is 0°. There is clearly no distance between these two superimposed lines. This situation corresponds to point a on the sine wave at the center of the figure. Hence, point a may be said to have an angle of 0°, and no displacement from the origin. This concept may appear quite vague at first, but it will become clear as the second radius (r′) rotates around the circle.

Let us assume that radius r′ is rotating counterclockwise at a *fixed speed*. When r′ has rotated 45°, it arrives in the position shown in circle (b). Here, r′ is at an angle of 45° to r. We will call this angle as the **phase angle** (θ), which simply reflects the degree of rotation around the circle, or the number of degrees into the sine wave at the corresponding point b. We now drop a vertical line from the point where r′ intersects the circle down to r. We label this line d, representing the vertical distance between r and the point where r′ intersects the circle. The length of this line corresponds to the displacement of point b from the baseline of the sine wave (dotted line at b). We now see that point b on the sine wave is 45° into the cycle of SHM, at which the displacement of the air particle from its resting position is represented by the height of the point above the baseline. It should now be clear that the sine wave is related to the degrees of rotation around a circle. The shape of the sine wave corresponds to the sine of θ as r′ rotates around the circle, which is simply equal to d/r′.

The positive peak of the sine wave at point c corresponds to circle (c), in which r′ has rotated to the straight up position. It is now at a 90° angle to r, and the distance (d) down to the

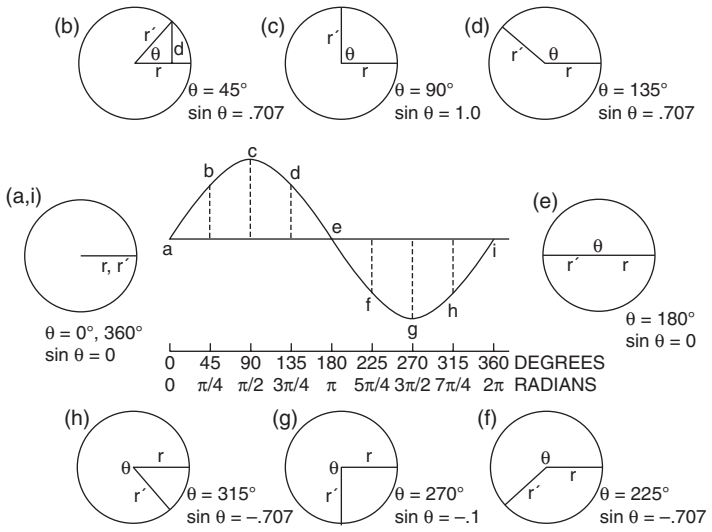

Figure 1.6 The nature of sinusoidal motion (see text).

horizontal radius (r) is the greatest. Here, we have completed a quarter of the wave and an arc equal to quarter the circumference of the circle. Notice now that further counterclockwise rotation of r′ results in decreasing the distance (d) down to the horizontal, as shown in circle (d), as well as by the displacement of point d from the baseline of the sine wave. Note also that θ is now 135°. Here, the air particle has reversed direction and is now moving back toward the resting position. When the particle reaches the resting position (point e), it is again at no displacement. The zero displacement condition is shown in circle (e) by the fact that r and r′ constitute a single horizontal line (diameter). Alternatively stated, r and r′ intersect the circle's circumference at points that are 180° apart. Here, we have completed half of the cycle of SHM, and the phase angle is 180° and the displacement from the baseline is again zero.

Continuing rotation of r′ places its intersection with the circumference in the lower left quadrant of the circle, as in circle (f). Now, θ is 225°, and the particle has overshot and is moving away from its resting position in the negative (rarefaction) direction. The vertical displacement from the baseline is now downward or negative, indicating rarefaction. The negative peak of the wave occurs at 270°, where displacement is maximum in the negative direction [point and circle (g)].

Circle (h) and point h show that the negative displacement has become smaller as the rotating radius passes 315° around the circle. The air particle has reversed direction again and is now moving toward its original position. At point i, the air particle has once again returned to its resting position, where displacement is again zero. This situation corresponds to having completed a 360° rotation so that r and r′ are once again superimposed. Thus, 360° corresponds to 0°, and circle (i) is one and the same with circle (a). We have now completed one full cycle.

Recall that r′ has been rotating at a fixed speed. It therefore follows that the number of degrees traversed in a given amount of time is determined by how fast r′ is moving. If one complete rotation takes 1 s, then 360° is covered each second. It clearly follows that if 360° takes 1 s, then 180° takes 0.5 s, 90° takes 0.25 s, 270° takes 0.75 s, etc. It should now be apparent that the phase angle reflects the elapsed time from the onset of rotation. Recall from Fig. 1.3 that the **waveform** shows how particle displacement varies as a function of time. We may also speak of the horizontal axis in terms of phase, or the equivalent of the number of degrees of rotation around a circle. Hence, the **phase** of the wave at each of the labeled points in Fig. 1.6 would be 0° at a, 45° at b, 90° at c, 135° at d, 180° at e, 225° at f, 270° at g, 315° at h, and 360° at i. With an appreciation of phase, it should be apparent that each set of otherwise identical waves in Fig. 1.7 differs with respect to phase: (a) wave 2 is offset from wave 1 by

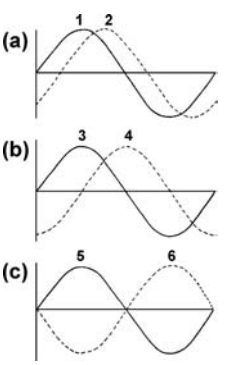

Figure 1.7 Pairs of sinusoidal waves of identical frequency differing in phase by (a) 45°, (b) 90°, and (c) 180°. The numbers serve only to identify the individual waves.

45°, (b) waves 3 and 4 are apart in phase by 90°, and (c) waves 5 and 6 are 180° out of phase.

We may now proceed to define a number of other fundamental aspects of sound waves. A cycle has already been defined as one complete repetition of the wave. Thus, four cycles of a sinusoidal wave were shown in Fig. 1.5 because it depicts four complete repetitions of the waveform. Because the waveform is repeated over time, this sound is said to be **periodic**. In contrast, a waveform that does not repeat itself over time would be called **aperiodic**.

The amount of time that it takes to complete one cycle is called its **period**, denoted by the symbol **t** (for time). For example, a periodic wave that repeats itself every millisecond is said to have a period of 1 ms, or $t = 1$ ms or 0.001 s. The periods of the waveforms considered in hearing science are overwhelmingly less than 1 s, typically in the milliseconds and even microseconds. However, there are instances when longer periods are encountered.

The number of times a waveform repeats itself per unit of time is its **frequency (f)**. The standard unit of time is the second; thus, frequency is the number of times a wave repeats itself in a second, or the number of **cycles per second (cps)**. By convention, the unit of cycles per second is the **hertz (Hz)**. Thus, a wave that is repeated 1000 times per second has a frequency of 1000 Hz, and the frequency of a wave that repeats at 2500 cycles per second is 2500 Hz.

If period is the time it takes to complete one cycle, and frequency is the number of cycles that occur each second, then it follows that period and frequency are intimately related. Consider a sine wave that is repeated 1000 times per second. By definition it has a frequency of 1000 Hz. Now, if exactly 1000 cycles take exactly 1 s, then each cycle must clearly have a duration of 1 ms, or 1/1000 s. Similarly, each cycle of a 250-Hz tone must last 1/250 s, or a period of 4 ms. Formally, then, frequency is the reciprocal of period, and period is the reciprocal of frequency:

$$f = \frac{1}{t} \tag{1.32}$$

and

$$t = \frac{1}{f} \tag{1.33}$$

It has already been noted that the oscillating air particle is moving back and forth around its resting or average position. In other words, the air particle's displacement changes over the course of each cycle. The magnitude of the air particle's displacement is called **amplitude**. Figure 1.8 illustrates a difference in the amplitude of a sinusoid, and contrasts this with a change in its frequency. In both frames of the figure, the tone represented by the finer curve has greater amplitude than the one portrayed by the heavier line. This is shown by the greater vertical distance from the baseline (amplitude) at any point along the horizontal axis (time). (Obviously, exceptions occur at those times when both curves have zero amplitudes.)

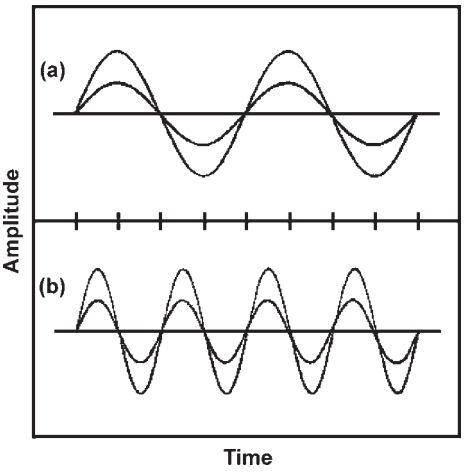

Figure 1.8 Within each frame (a and b), both sinusoidal waves have the same frequency, but the one depicted by the lighter curves has a greater amplitude than the one represented by the heavier curves. The curves in frame (b) have twice the frequency as those shown in frame (a).

At any given moment, the particle may be at its extreme positive or negative displacement from the resting position in one direction or the other, or it may be somewhere between these two extremes (including being at the resting position, where displacement is zero). Because each of these displacements is a momentary glimpse that holds true only for that instant, the magnitude of a signal at a given instant is aptly called its **instantaneous amplitude**.

Because the instantaneous amplitude changes from moment to moment, we also need to be able to describe the magnitude of a wave in more general terms. The overall displacement from the negative to positive peak yields the signal's **peak-to-peak amplitude**, while the magnitude from baseline to a peak is called the wave's **peak amplitude**. Of course, the actual magnitude is no more often at the peak than it is at any other phase of the sine wave. Thus, although peak amplitudes do have uses, we most often are interested in a kind of "average" amplitude that more reasonably reflects the magnitude of a wave throughout its cycles. The simple average of the sinusoid's positive and negative instantaneous amplitudes cannot be used because this number will always be equal to zero. The practical alternative is to use the **root-mean-square (rms) amplitude**. This value is generally and simply provided by measuring equipment, but it conceptually involves the following calculations: First, the values of all positive and negative displacements are squared so that all resulting values are positive numbers (and zero for those values that fall right on the resting position). Then the mean of all these values is obtained, and the rms value is finally obtained by taking the square root of this mean. The rms amplitude of a sinusoidal signal is numerically equal to 0.707 times the peak amplitude, or 0.354 times the peak-to-peak amplitude. Figure 1.9 illustrates the relationships among peak, peak-to-peak, and rms amplitudes.

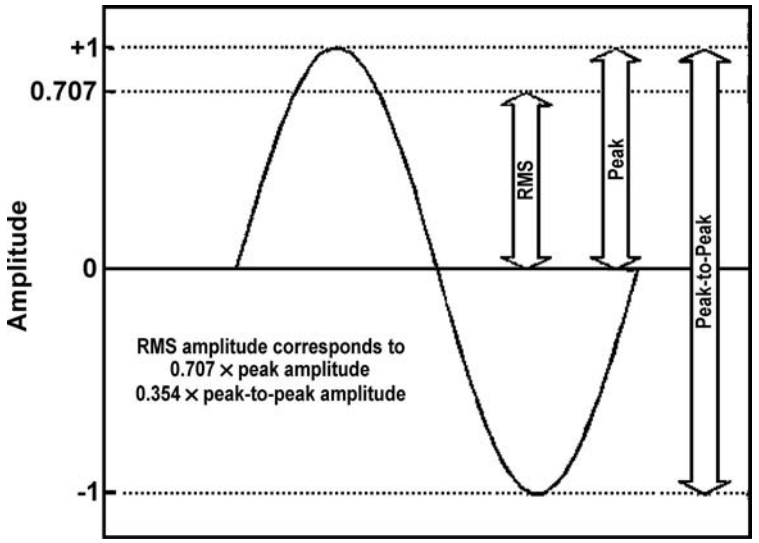

Figure 1.9 The relationships among the root-mean-square (rms), peak, and peak-to-peak amplitudes.

COMBINING WAVES

The sound associated with a sinusoidal wave is called a **pure tone**. Figure 1.10 shows what occurs when two sinusoids having the *same frequencies* and *amplitudes* are combined. In this case, the resulting wave will also be a pure tone, but the concepts illustrated in Fig. 1.10 reveal the principles that apply whenever waves are being combined. In Fig. 1.10a, the first and second sinusoids (labeled f_1 and f_2) are *in-phase* with each other. Here, the two waves are equal to one another in terms of (instantaneous) amplitude at every moment in time. The resulting wave (labeled $f_1 + f_2$) has twice the amplitude of the two components, but it is otherwise identical to them. This finding illustrates the central concept involved in combining waves: The amplitudes of the two waves being combined are *algebraically added to each other at every point along the horizontal (time) axis*. In the case of two identical, in-phase sinusoids, the resultant wave becomes twice as big at each point along the time axis and remains zero wherever the two waves have zero amplitudes. The latter occurs because the amplitudes of the two waves at the moments when they cross the baseline are zero; zero plus zero equals zero. For readily apparent reasons, the case shown in Fig. 1.10a is called **reinforcement**.

Figure 1.10b shows what happens when we combine two otherwise identical sinusoids that are *180° out of phase* with each other. This is, of course, the opposite of the relationship depicted in Fig. 1.10a. Here, wave f_1 is equal and opposite to wave f_2 at every moment in time. Algebraic addition under these circumstances causes the resulting amplitude to equal zero at all points along the horizontal (time) axis. Notice that the result $(f_1 + f_2)$ is complete **cancellation**.

If the two otherwise identical sinusoids are out of phase by a value other than 180°, then the shape of the resulting wave will depend upon how their amplitudes compare at each moment in time. The two sinusoids in Fig. 1.10c are 90° out of phase. The result of algebraically adding their magnitudes on a point-by-point basis is shown by wave $f_1 + f_2$ below the two original waves. In general, combining two identical sinusoids having the same frequency that are out of phase (except 180° out of phase) results in a sinusoid with the same frequency, but that is different in its phase and amplitude.

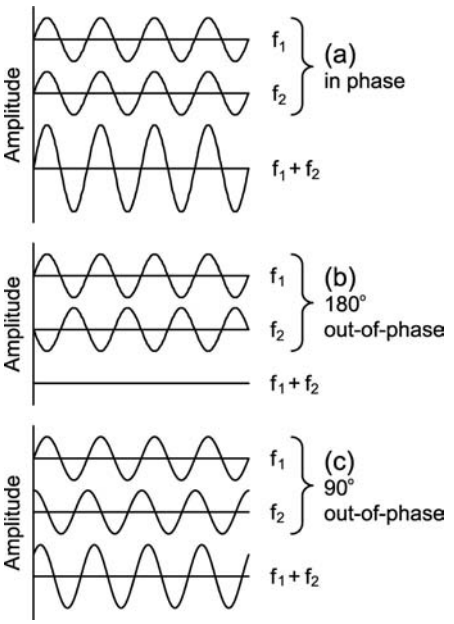

Figure 1.10 Combining sinusoids with equal frequencies and amplitudes that are (a) in phase, (b) 180° out of phase, and (c) 90° out of phase.

COMPLEX WAVES

Thus far, we have dealt only with the combination of sinusoids having the same frequency. What happens when we combined sinusoids that differ in frequency? When two or more pure tones are combined, the result is called a **complex** wave or a complex sound. The mechanism for combining waves of dissimilar frequencies is the same as what applies for those having the same frequency: *Any* two or more waves are combined by algebraically summing their instantaneous displacements on a point-by-point basis along the horizontal (time) axis, regardless of their individual frequencies and amplitudes or their phase relationships. However, the combination of waves having unequal frequencies will not yield a sinusoidal result. Instead, the result will depend upon the specifics of the sounds being combined.

Consider the three sinusoids at the top in Fig. 1.11, labeled f_1, f_2, and f_3. Note that the two cycles of f_1 are completed in the same time as four cycles of f_2 or six cycles of f_3. Thus, frequency of f_2 is *exactly* two times that of f_1, and the frequency of f_3 is *exactly* three times f_1. The actual frequencies of f_1, f_2, and f_3 could be any values meeting the described conditions; for example, 100, 200, and 300 Hz; 1000, 2000, and 3000 Hz, or 20, 40, and 60 Hz, etc. Because f_2 and f_3 are integral multiples of f_1, we say that they are **harmonics** of f_1. Hence, f_1, f_2, and f_3 constitute a harmonic series. The lowest frequency of this series is the **fundamental frequency**. Otherwise stated, harmonics are whole-number multiples of the fundamental frequency; the fundamental is the largest whole-number common denominator of its harmonics. Notice that the fundamental frequency (often written as $\mathbf{f}_0$) is also the first harmonic because its frequency is the value of the first harmonic, or $1 \times f_0$. Clearly, the harmonics are separated from one another by amounts equal to the fundamental frequency.

The lower three waves in Fig. 1.11 show what happens when f_1, f_2, and f_3 are combined in various ways. Notice that the combining of two or more sinusoidal waves differing in frequency generates a resultant wave that is no longer sinusoidal in character. Note, however, that the combined waveforms shown in this figure are still periodic. In other words, even though these combined waveforms are no longer sinusoidal, they still retain the characteristic of repeating themselves at regular intervals over time. Moreover, notice that all three waves ($f_1 + f_2$, $f_1 + f_3$, and $f_1 + f_2 + f_3$) repeat themselves with the same period as f_1, which is the lowest component in each case. These are examples of **complex periodic waves**, so called because (1) they are composed of more than one component and (2) they repeat themselves at regular time intervals. The lowest-frequency component of a complex periodic wave is its fundamental frequency. Hence, f_1 is the fundamental frequency of each of the complex periodic waves in Fig. 1.11 The period of the fundamental frequency constitutes the rate at which the complex periodic wave repeats itself. In other words, the time needed for one cycle of a complex periodic wave is the same as the period of its fundamental frequency.

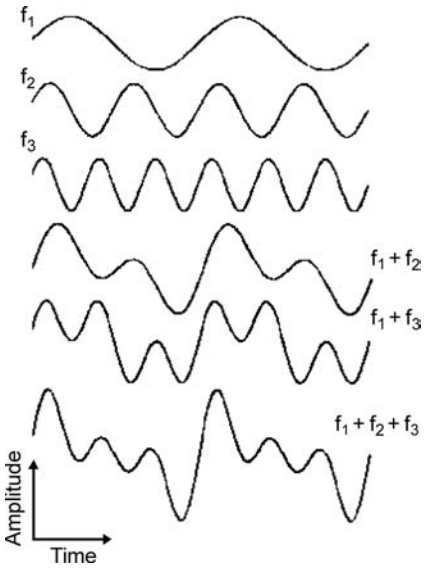

Figure 1.11 The in-phase addition of sinusoidal waves f_1, f_2, and f_3 into complex periodic waves, $f_1 + f_2$, $f_1 + f_3$, and $f_1 + f_2 + f_3$. The frequency of f_2 is twice that of f_1, and f_3 is three times the frequency of f_1. The frequency of f_1 is the fundamental frequency of each of the three complex periodic waves.

The example shown in Fig. 1.12 involves combining only odd harmonics of 1000 Hz (1000, 3000, 5000, and 7000 Hz) whose amplitudes become smaller with increasing frequency. The resulting complex periodic waveform becomes progressively squared off as the number of odd harmonics is increased, eventually resulting in the aptly named **square wave**. The complex periodic waveform at the bottom of the figure depicts the extent to which a square wave is approximated by the combination of the four odd harmonics shown above it.

The combination of components that are not harmonically related results in a complex waveform that does not repeat itself over time. Such sounds are thus called **aperiodic**. In the extreme case, consider a wave that is completely random. An artist's conceptualization of two separate glimpses of a random waveform is shown in Figs. 1.13 1.13a and 1.13b. The point of the two pictures is that the waveform is quite different from moment to moment. Over the long run, such a wave would contain all possible frequencies, and all of them would have the same average amplitudes. The sound described by such waveforms is often called **random noise** or **Gaussian noise**. Because all possible frequencies are equally represented, they are more commonly called **white noise** on analogy to white light. Abrupt sounds that are extremely short in duration must also be aperiodic because they are not repeated over time. Such sounds are called **transients**. The waveform of a transient is shown in Fig. 1.13c.

Because the **waveform** shows amplitude as a function of time, the frequency of a pure tone and the fundamental frequency of a complex periodic tone can be determined only indirectly by

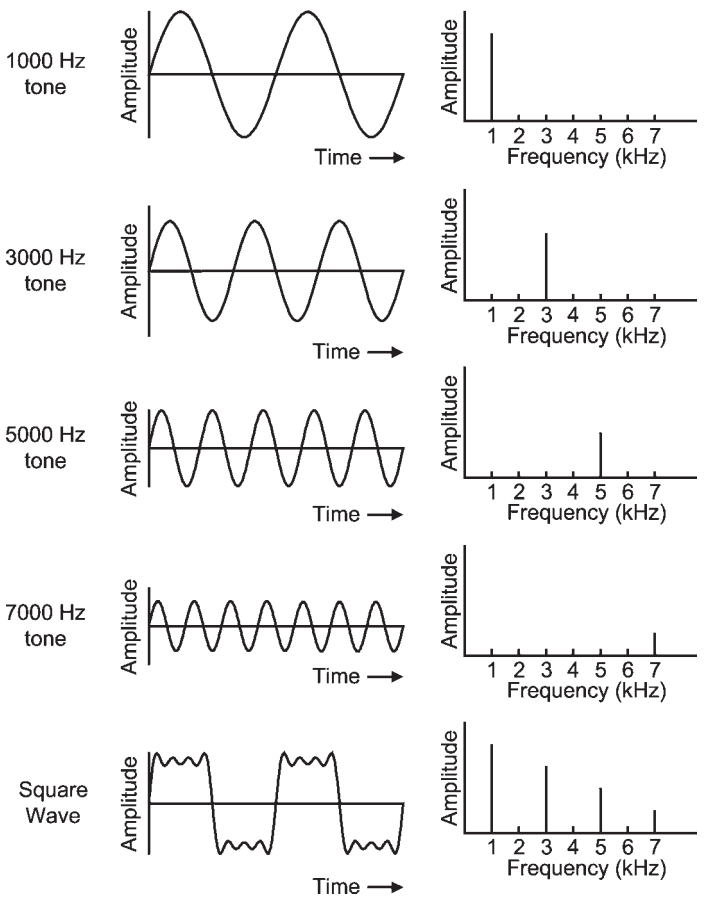

Figure 1.12 The addition of odd harmonics of 1000 Hz to produce a square wave. Waveforms (amplitude as a function of time) are shown in the left panels and corresponding spectra (amplitude as a function of frequency) are shown in the right panels.

examining such a representation, and then only if the time scale is explicit. Moreover, one cannot determine the frequency content of a complex sound by looking at its waveform. In fact, dramatically different waveforms result from the combination of the same component frequencies if their phase relationships are changed. Another means of presenting the material is therefore needed when one is primarily interested in information about frequency. This information is portrayed by the **spectrum,** which shows amplitude as a function of frequency. In effect, we are involved here with the issue of going between the time domain (shown by the waveform) and the frequency domain (shown by the spectrum). The underlying mathematical relationships are provided by **Fourier's theorem**, which basically says that a complex sound can be analyzed into its constituent sinusoidal components. The process by which one may break down the complex sound into its component parts is called **Fourier analysis**. Fourier analysis enables one to plot the spectrum of a complex sound.

The spectra of several periodic waves are shown in the right side of Fig. 1.12, and the spectrum of white noise is shown in Fig. 1.13d. The upper four spectra in Fig. 1.12 corresponds, respectively, to the waveforms of the sinusoids to their left. The top wave is that of a 1000-Hz tone. This information is shown on the associated spectrum as a single (discrete) vertical line drawn at the point along the abscissa corresponding to 1000 Hz. The height of the line indicates the amplitude of the wave. The second waveform in Fig. 1.12 is for a 3000-Hz tone that has a lower amplitude than does the 1000-Hz tone shown above it. The corresponding spectrum shows this as a single vertical line drawn at the 3000-Hz location along the abscissa. Similarly, the spectra of the 5000- and 7000-Hz tones are discrete vertical lines corresponding to their respective frequencies. Notice that the heights of the lines become successively smaller going from the spectrum of the 1000-Hz tone to that of the 7000-Hz tone, revealing that their amplitudes are progressively lower.

The lowest spectrum in Fig. 1.12 depicts the complex periodic wave produced by the combination of the four pure tones shown above it. It has four discrete vertical lines, one each at the 1000-, 3000-, 5000-, and 7000-Hz locations. This spectrum approximates that of a square wave. The spectrum of a square

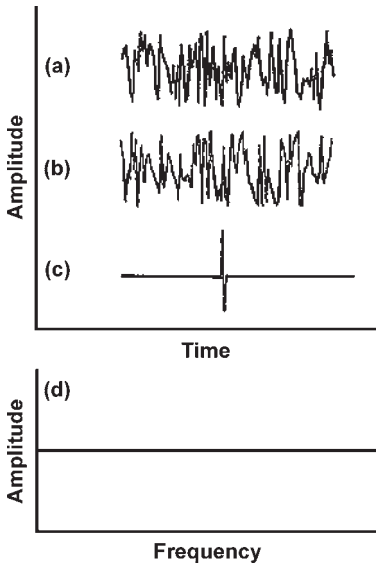

Figure 1.13 Artist's conceptualizations of the waveform of white noise as it might appear at two different times (a and b), and of a transient (c), along with the spectrum of white noise or a transient (d).

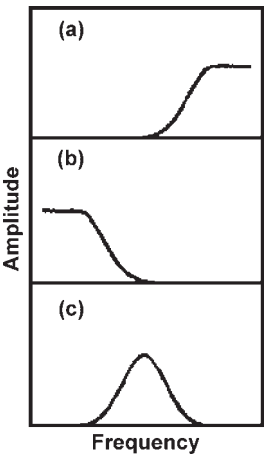

Figure 1.14 Continuous spectra of aperiodic sounds with (a) greater amplitude in the high frequencies, (b) greater amplitude in the lower frequencies, and (c) a concentration of energy within a given band of frequencies.

wave is composed of many discrete vertical lines, one each at the frequencies corresponding to odd multiples of its lowest (fundamental) component, with their heights decreasing as frequency increases.

To summarize, the spectrum of a periodic wave shows a vertical line at the frequency of each sinusoidal component of that wave, and the amplitude of each component is shown by the height of its corresponding line. Consequently, the spectrum of a periodic sound is referred to as a **discrete spectrum**. As should be apparent, the phase relationships among the various components are lost when a sound is represented by its spectrum.

Figure 1.13d shows the spectrum of white noise. Because white noise contains all conceivable frequencies, it would be a fruitless exercise to even try to draw individual (discrete) vertical lines at each of its component frequencies. The same point applies to the three spectra depicted in Fig. 1.14 shows the continuous spectra of aperiodic sounds that contain (1) greater amplitude in the higher frequencies, (2) greater amplitude in the lower frequencies, and (3) a concentration of energy within a particular range band (range) of frequencies.

FILTERS

The three spectra depicted in Fig. 1.14 may also be used to describe the manner in which a system transfers energy as a function of frequency. **Filters** are described according to the range of frequencies that they allow to *pass* as opposed to those that they *stop* or *reject*. Thus, Fig. 1.15a depicts a **high-pass filter** because the frequencies higher than a certain **cutoff frequency** are passed, whereas those below that cutoff frequency are stopped or rejected. On the other hand, Fig. 1.15b shows a **low-pass filter** because the frequencies lower than its cutoff frequency are passed whereas higher ones are rejected. A cutoff frequency is usually defined as the frequency where the power falls to half of its peak value. This location is called the **half-power point**. In decibels, the half-power point is 3 dB below that level of the peak, and is therefore also known as the **3-dB down point**.

Figure 1.15c illustrates a **band-pass filter** because the frequencies within the designated range are passed, whereas those below and above its lower and upper cutoff frequencies are rejected. A band-pass filter is usually described in terms of its **center frequency**, which is self-explanatory, and its **bandwidth**, which is how wide the filter is between its upper and lower cutoff frequencies. A filter that passes the frequencies above and below a certain band, but rejects the frequencies within that band is called a **band-reject filter** (Fig. 1.15d).

The sharpness with which a filter de-emphasizes the reject band is given by its **slope**, also know as its **rolloff**, **attenuation**, or **rejection rate**. The slope is usually expressed in *decibels per octave*. For example, a slope of 24 dB/octave means that the magnitude of the sound outside of the pass band is reduced at a rate of 24 dB for each doubling of frequency. Beginning at 1000 Hz, a 24-dB/octave rolloff rate would cause the signal to be reduced by 24 dB by 2000 Hz (an octave above 1000 Hz) and by an additional 24 dB by 4000 Hz (an octave above 2000 Hz). Besides using its slope, it is often convenient to describe the sharpness of tuning for a band-pass filter in terms of a value called **Q**, especially when comparing the characteristics of different filters. The Q of a filter is simply the ratio of its center frequency to its bandwidth.

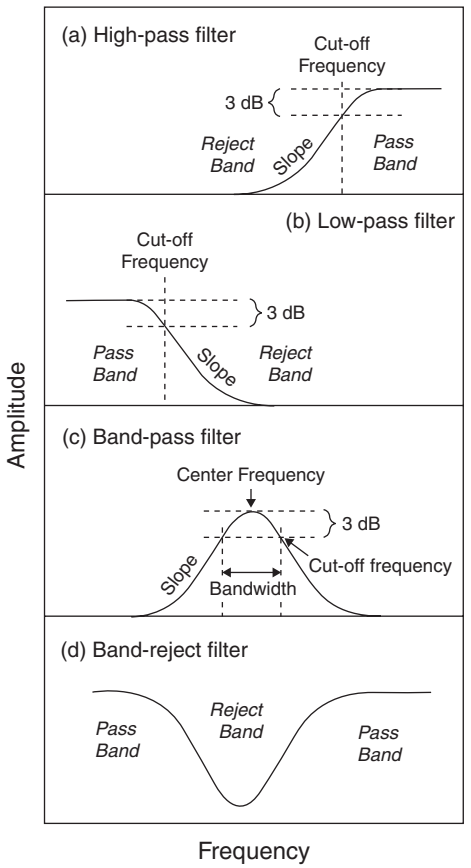

Figure 1.15 Examples of filters and some of their parameters: (a) high-pass filter, (b) low-pass filter, (c) band-pass filter, and (d) band-reject filter.

STANDING WAVES

Let us consider two situations. In the first situation, imagine the sound waves propagating rightward and leftward from a vibrating tuning fork placed in the center of a small room. These sound waves will hit and be reflected from the walls so that now there will also be reflected waves propagating in the opposite directions in addition to the original waves. (Considering for the moment only the right wall for simplicity, we can envision a rightward-going original wave and a leftward-going reflected wave.) The other situation involves plucking the center of a guitar string that is tied tautly at both ends. Here, the waves initiated by the pluck move outward toward each fixed end of the string, from which a reflected wave propagates in the opposite direction.

To reiterate, in both cases, just described, there are continuous original and continuous reflected waves moving toward one another. The reflected waves are equal in frequency to the original ones, and both the reflected and original waves are of course propagating at the same speed. Now, recall from prior discussion that two waves will interact with one another such that their instantaneous displacements add algebraically. Thus, the net displacement (of the air particles in the room or of the string) at any moment that occurs at any point (in the room or along the string) will be due to how the superimposed waves interact. It turns out that the resultant wave produced by this interaction constitutes a pattern that f_1, f_2, and f_3 actually stands still even though it is derived from component waves which themselves are propagating. Hence, the points of maximum displacement (peaks of the waves) and no displacement (baseline crossings of the waves) will always occur at fixed locations in the room or along the string. This phenomenon is quite descriptively called a **standing wave**.

Because the vibration pattern of the string is easily visualized, we will refer only to the string as example for the remainder of the discussion, although these points apply similarly to the room example as well. The locations of no (zero) displacement in the standing wave pattern are called **nodes**, and the places of maximum displacement are thus called **antinodes**. Even brief consideration will reveal that the displacement must be zero at the two ends of the string, where they are tied and thus cannot move. (This corresponds to the hard walls of the room, which prevent the air molecules from being displaced.) Hence, nodes must occur at the two ends of the string. It follows that if there is a node at each end of the string, then there must be an antinode at the center of the string, halfway between the two nodes. This notion should not be surprising, because we already know that the zero displacements (at 0° and 180° phase) and maximum displacements (at 90° and 270° phase) alternate for any cycle of a sinusoid.

This standing wave pattern is depicted in Fig. 1.16a. Some thought will confirm that the arrangement just described (a node at each end an antinode at the center) constitutes the

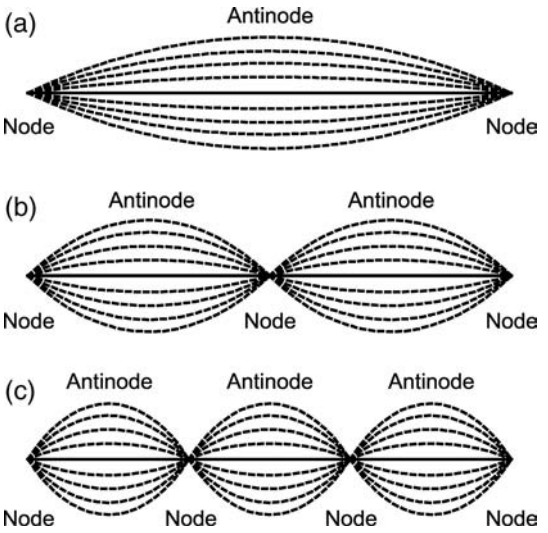

Figure 1.16 The (a) first, (b) second, and (c) third modes of a vibrating string.

longest possible standing wave pattern that can occur for any given string. We will call it the first **mode of vibration**. The figure also highlights the fact that this standing wave pattern comprises exactly one-half of a cycle (from 0° to 180°). Consequently, its length (L) corresponds to exactly one-half of a wavelength (λ), or $L = \lambda/2$. Its length therefore is equal to one-half of a wavelength (λ/2) of some frequency. This frequency, in turn, may be determined by applying the formula, $f = c/\lambda$ (Eq. 1.31). By substitution, the frequency of the string's first mode of vibration would be c/2L (where L is the length of the string and c is the appropriate speed of the wave for that string[3]). It should now be apparent that the first mode of vibration corresponds to the fundamental frequency.

The standing wave just described is not the only one that can occur for the string, but rather is the longest one. Other standing waves may develop as well as long as they meet the requirement that nodes occur at the two tied ends of the string. Several other examples are shown in Fig. 1.16, which reveals that each of these standing wave patterns must divide the string into parts that are exactly equal in length to one another. Thus, there will be standing wave patterns that divide the string into exact halves, thirds, fourths, fifths, etc. These are the second, third, fourth, fifth, etc., modes of vibration. In turn, they produce frequencies, which are exact multiples (harmonics) of the fundamental frequency.

Suppose we were to set the air inside of a tube into vibration by, for example, blowing across the open end of the tube. If we were to do this experiment for several tubes, we would find that the shorter tubes make higher-pitch sounds than do the longer ones. We would also find that the same tube would produce a higher pitch when it is open at both ends than when it is open at only one end. The frequency(ies) at which a body or medium vibrates is referred to as its **natural** or **resonant** frequency(ies).

In the case of a column of air vibrating in a tube open at both ends, the greatest pressure and the least particle displacement can occur in the center of the tube, while the greatest displacement and thus lowest pressure can occur at the two open ends (Fig. 1.17a). This is analogous to the vibration of the string. One may understand this in the sense that going from one end of the tube to the other involves going from a pressure node to an antinode to a node (or from displacement antinode to node to antinode), or 180° of a cycle. This pattern is related to the out-of-phase reflection of the wave at the two ends of the tube so that the pattern is duplicated when the length of the tube is covered twice. Hence, the lowest (fundamental) frequency capable of covering the tube exactly twice in one cycle must have a wavelength twice the length of the tube. Thus, the lowest

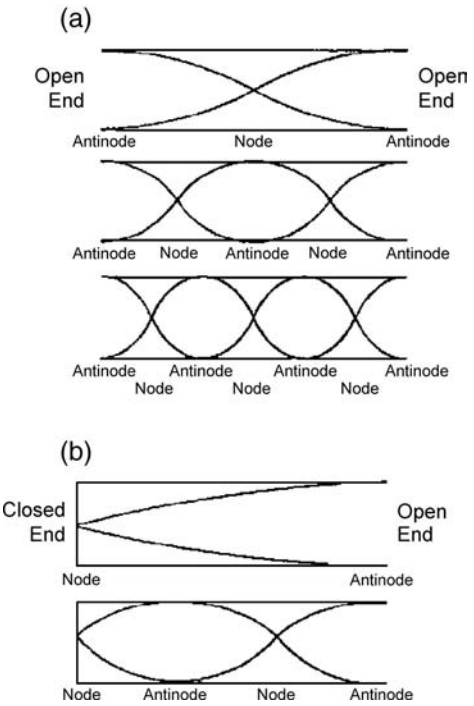

Figure 1.17 Standing waves in (a) a tube which is open at both ends (a half-wavelength resonator) and (b) a tube which is closed at one end and open at the other end (a quarter-wavelength resonator). *Source*: From Gelfand SA (2001). *Essentials of Audiology, Second Edition*. New York: Thieme Medical Publishers, with permission.

resonant frequency (f_1) of a tube open at both ends is the frequency whose wavelength is twice the length of the tube, or $f_1 = c/2L$. Harmonics will occur at each multiple of this fundamental frequency.

Air vibration in a tube closed at one end is most restricted at the closed end, where pressure must thus be greatest and displacement the least (Fig. 1.17b). (Reflections at the closed end occur without phase reversal.) Thus, in terms of displacement, there must be a node at the closed end and an antinode at the open end. This means that the length of the tube corresponds to a quarter of a wavelength so that the lowest resonant frequency (f_1) of a tube closed at one end and open at the other is the one whose wavelength is four times the length of the tube, or $f_1 = c/4L$. Since a node can occur at only one end, such a tube produces only the fundamental frequency and its odd harmonics (e.g., f_1, f_3, f_5, f_7, etc.).

IMPEDANCE

Impedance is the opposition to the flow of energy through a system. Some knowledge of impedance thus helps one to understand how a system transmits energy, and why it is more responsive to some frequencies than it is to others. We may generally define **impedance (Z)**, in **ohms**, as the ratio of force

[3] The speed of a wave along a vibrating string is not the same as for air. Instead, we would be dealing with the speed of a transverse wave along a string, which is the square root of the ratio of the string's tension (T) to its mass per unit area (M). Hence, the formula for the string's lowest resonant frequency actually would be $f = (1/2L)\sqrt{T/M}$.

to velocity:

$$Z = \frac{F}{v} \tag{1.34}$$

Therefore, the greater the amount of force needed to result in a given amount of velocity, the greater the impedance of the system.

We may also consider impedance in terms of its components. These are shown in the form of a mechanical representation of impedance in Fig. 1.18 Here, we see that impedance (Z) is the interaction between **resistance (R)** and two kinds of **reactance (X)**, known as **positive** or **mass reactance (X_m)** and **negative** or **stiffness reactance (X_s)**. These components are, respectively, related to friction, mass, and stiffness. In the figure, mass is represented by the block, and stiffness is provided by the spring. Friction is represented in the figure by the irregular surface across which the mass (represented by a solid block) is moved.

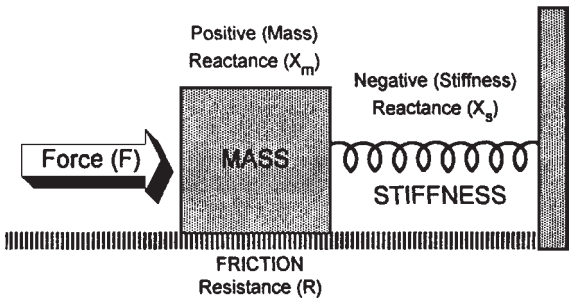

Figure 1.18 The components of impedance (Z) are friction or resistance (R), represented by the rough surface, mass (positive) reactance (X_m), represented by the block, and stiffness (negative) reactance (X_s), represented by the spring.

Let us imagine that a sinusoidal force (represented by the arrow) is being applied to the system depicted in the illustration. Friction causes a portion of the energy applied to the system to be converted into heat. This dissipation of energy into heat is termed resistance. Resistance is not related to frequency and occurs in phase with the applied force. In contrast, reactance is the storage (as opposed to the dissipation) of energy by the system. Mass reactance is, of course, associated with the mass of the system (the block in Fig. 1.18). Since mass is associated with the property of inertia, the application of a force causes the mass to accelerate according to the familiar formula F = Ma (where F is force, M is mass, and a is acceleration). If the force is applied sinusoidally, then the mass reactance is related to frequency as

$$X_m = M \cdot 2\pi f \tag{1.35}$$

where f is frequency. Thus, the magnitude of the mass reactance is directly proportional to frequency; that is, the higher the frequency, the greater the mass reactance. Since acceleration precedes force by a quarter-cycle, X_m will lead the applied force in phase by 90°. This is why X_m is termed positive reactance,

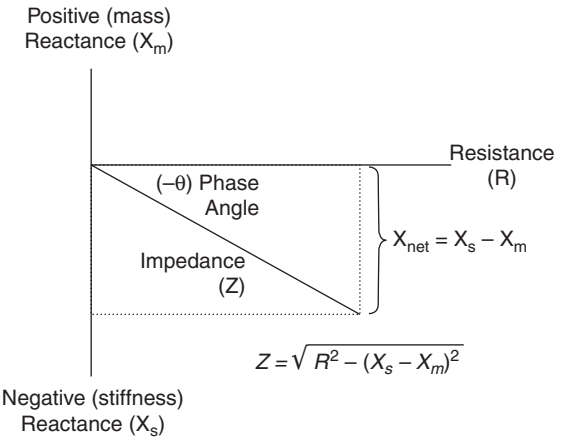

Figure 1.19 The relationships between impedance and its components.

and its value is shown in the positive direction on the y-axis in Fig. 1.19

Stiffness reactance is represented by the spring in Fig. 1.18 We will represent the stiffness as S. Applying a force compresses (displaces) the spring according to the formula F = Sx, where x is the amount of displacement. When the force is applied sinusoidally, then stiffness reactance is related to frequency as

$$X = \frac{S}{2\pi f} \tag{1.36}$$

In other words, the amount of stiffness reactance is *inversely proportional* to frequency; that is, stiffness reactance goes down as frequency goes up. Since displacement follows force by a quarter-cycle, X_s lags behind the applied force in phase by 90°. It is thus called negative reactance and is plotted downward on the y-axis in Fig. 1.19 It should be apparent at this point that X_m and X_s are 180° out of phase with each other.

Because stiffness and mass reactance are 180° out of phase, a system's net reactance is equal to the difference between them ($X_m - X_s$). This relationship is illustrated in Fig. 1.19 for the condition where X_s exceeds X_m, which is the case for lower frequencies in the normal ear (see Chap. 3). Notice that the impedance (Z) is a vector, which results from the interaction between the resistance (R) and the **net reactance (X_{net})**. The negative phase angle (−θ) in Fig. 1.19 shows that the net reactance is negative. The relationship among impedance, resistance, and reactance may now be summarized as follows:

$$Z = \sqrt{R^2 + (X_s - X_m)^2} \tag{1.37}$$

Looking at the effect of frequency, we find that

$$Z = \sqrt{R^2 + \left(\frac{S}{2\pi f} - M \cdot 2\pi f\right)^2} \tag{1.38}$$

The implication is that frequency counts. Because X_m is proportional to frequency, while X_s is inversely proportional to

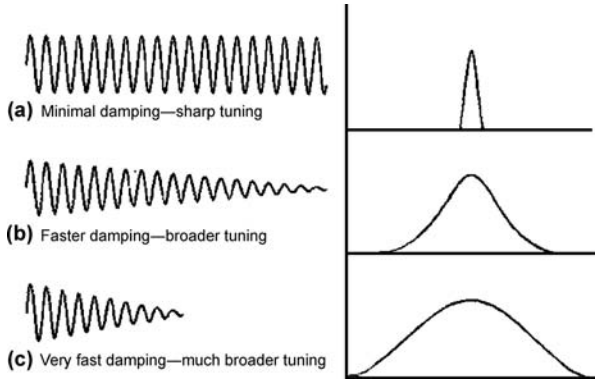

Figure 1.20 Artist's conceptualizations of the relationship between the amount of damping (*left panels*) and the sharpness of tuning around the resonant frequency (*right panels*).

frequency, they should be equal at some frequency. This is the system's **resonant frequency**, at which the reactance components cancel each other out, leaving only the resistance component.

The amount of resistance is associated with how rapidly damping occurs, and it determines the sharpness of the tuning around the resonant frequency. This relationship is illustrated in Fig. 1.20 The less the resistance (i.e., the slower the damping), the more narrowly tuned the resonance is; the more the resistance (i.e., the faster the damping), the broader the responsiveness of the system around the resonant frequency.

REFERENCES

American National Standards Institute, ANSI S1—1994 (R2004). 2004. *American National Standard Acoustical Terminology*. New York: ANSI.

Beranek LL. 1986. *Acoustics*. New York: American Institute of Physics.

Everest FA. 2000. *The Master Handbook of Acoustics, 4th ed*. New York: McGraw-Hill.

Gelfand SA. 2001. *Essentials of Audiology, 2nd ed*. New York: Thieme Medical Publishers.

Hewitt, PG 2005. *Conceptual Physics, 10th ed*. Boston, MA: Pearson Addison-Wesley.

Kinsler LE, Frey AR, Coopens AB, and Sanders JB. 1999. *Fundamentals of Acoustics, 4th ed*. New York, NY: Wiley.

Pearce JR, David EE Jr. 1958. *Man's World of Sound*. Garden City, New York, NY: Doubleday.

Peterson APG, Gross EE. 1972. *Handbook of Noise Measurement, 7th ed*. Concord, MA: General Radio.

Rassmussen G. Intensity—Its measurement and uses. *Sound Vibr* 1989; 23(3):12–21.

Rossing TD, Moore RF, Wheeler PA. 2002. *The Science of Sound, 3rd ed*. Boston, MA: Pearson Addison-Wesley.

Speaks CE. 1999. *Introduction to Sound: Acoustics for Hearing and Speech Sciences, 3rd ed*. San Diego, CA: Singular.

van Bergeijk WA, Pearce JR, David EE Jr. 1960. *Waves and the Ear*. Garden City, New York, NY: Doubleday.

Young HD, Freedman RA. 2007. *Sears and Zemansky's University Physics with Modern Physics, 12th ed*. Boston, MA: Pearson Addison-Wesley.