3

The nature of sound

The material of music is sound—a physical phenomenon. As Varèse (1939) observed:

> When you listen to music do you ever stop to realize that you are being subjected to a physical phenomenon? . . . In order to anticipate the result, a composer must understand the mechanics of the instruments and must know just as much as possible about acoustics.

The phenomenon of sound requires the dimensions of time and space, but also a vibrating medium. Unlike light, which penetrates a vacuum, sound is mechanical energy. That is, it needs a medium that can be vibrated—typically air in the Earth's troposphere.

Understanding the nature of sound, its properties and physics, is of prime importance for composers today. The terrain of music is sound, and the more one knows the terrain, the better one can navigate within it.

The science of acoustics deals with the physical properties of sound. Acoustics has theoretical, experimental, and practical sides. Acoustical theory describes the physics of wave mechanics by means of mathematical models (Morse 1981). Experimental acoustics is concerned with the development of new acoustic devices—microphones and loudspeakers, for example—and overlaps with electrical engineering. Architectural acousticians design auditoria, concert halls, and recording studios. A branch of practical acoustics focuses on the pernicious problem of noise pollution.

Sound touches the body and penetrates rapidly to the brain in the form of electrochemical signals. Once it reaches our awareness, we inevitably dissect it to decode its messages. Thus, in contrast to acoustics, psychoacoustics explores the impact of sound on human beings—our bodily and psychological responses to it. It overlaps with the science of hearing, which is tied to anatomy and the neuroscience of the auditory system (Avanzini et al. 2003). In the rest of this chapter, we examine both acoustic and psychoacoustic phenomena.

While some of the material in this chapter is basic, other parts disclose new facts or perspectives. Advanced readers might quickly scan both this chapter and chapter 4 on sound materials.

What is sound?

> Sound is an alternation in pressure, particle displacement, or particle velocity propagated in an elastic material.
>
> —HARRY F. OLSON (1957)

Sound results from vibration in a material medium. Vibration occurs when mechanical energy interacts with the medium. In acoustical terminology, the energy is referred to as the *excitation*, and the vibration that it induces is the *response* or *resonance*. For example, a cellist vibrates a taut string by drawing a bow across it. This grating sound is amplified and filtered by the resonances of the cello body.

We hear sound because we live on the bottom of a vast ocean of air,[16] a vibrational medium. In scientific parlance, the word "sound" refers not only to phenomena in air responsible for the sensation of hearing but also "whatever else is governed by analogous physical principles" (Pierce 1994). Sound can be defined in a general sense as mechanical radiant energy that is transmitted by pressure waves in a material medium. Thus, besides the airborne frequencies that our ears perceive, one can speak of underwater sound, sound in solids, or structure-borne sound. Mechanical vibrations even take place on the inaudible atomic level, resulting in quantum units of sound energy called phonons. The term "acoustics" is likewise independent of air and human perception, and is distinguished from optics in that it involves mechanical—rather than electromagnetic—wave motion.

How does sound affect human beings?

Sound is heard by the ear but also felt by the body. As Stockhausen (1972) observed:

> Sound waves penetrate very deep into the molecular and atomic layers of our selves. Whenever we hear sounds, we are changed. We are no longer the same. . . . This is more the case when we hear organized sounds, sounds organized by a human being: music.

Airborne sounds enter the body via the ear. The ear is a miraculously sensitive organ connected to a complicated neural structure known as the *central auditory system*, extending deep into the brain. The outer ear takes in sound pressure waves and transduces them into mechanical vibrations in the middle ear. From this point, mechanical vibrations are transduced into liquid vibrations in the inner ear, and then into electrical impulses transmitted via nerves leading to the brain.

At dangerous intensities, sound affects the entire body (Miller 1978a, 1978b, 1978c). At the same time, experiments show that when sound waves are projected on a naked human body, it is almost entirely reflected; little energy penetrates the skin (Conti et al. 2003, 2004). This is because the unadorned human body is acoustically similar to a bag of water, which tends to reflect sound. When human beings don several layers of clothing, they become sound absorbers. Thus a thousand people dressed in several layers of clothing in a concert hall have a damping effect on the acoustics of the hall.

We do not need ears to sense sound vibrations. Mobile phones and other vibrating devices transmit sound energy directly to the body by the sense of touch, bypassing the auditory system.

At a basic level, the human body is an electrochemical system; chemical changes trigger electrical signals (Galambos 1962). For example, muscles are batteries, and muscle fatigue is literally a drop in electrical charge. The sensation of sound provokes immediate and major electrochemical changes in the paralimbic system, which includes the amygdala, hippocampus, and other brain structures associated with emotional responses (Brown et al. 2004; Molavi 2005). These in turn affect the hypothalamus and the pituitary gland—the brain's regulators, which emit hormonal secretions that affect the rest of the nervous system.

Hormones affect muscle tone, breathing rate and depth, blood pressure, and heart pulse, among other involuntary bodily functions. Meanwhile, the mind takes in conscious and unconscious sensations; it remembers what has happened and anticipates what is to come, that is, it constructs a narrative. Depending on the path of this narrative, intense emotional reactions can occur, covering the gamut from tears of joy to insufferable boredom and abject torment. Here is a universal principle: Every music, even the most coldly conceptual and unromantic in inspiration, triggers emotional reactions. The aesthetic implications of this

principle are profound: All music is perceived emotionally and romantically. This should not be a surprise, as Meyer (1956) observed:

> Thinking and feeling need not be viewed as polar opposites but as different manifestations of a single psychological process.

Or as Marvin Minsky (2006) observed, human beings never think purely rationally because our minds are always affected by assumptions and preferences (values and beliefs) and driven by goals (desires).

Sound waveforms

Let us now look at visualizations of sound from a scientific perspective. A direct method of visualizing sound waveforms is to draw them in the form of a graph of air pressure versus time (figure 3.1).

This is called a *time-domain* representation or *pressure graph.* When the curved line is descending, the air pressure is decreasing. When the curve is rising, the air pressure is increasing. The *amplitude* of the waveform is the amount

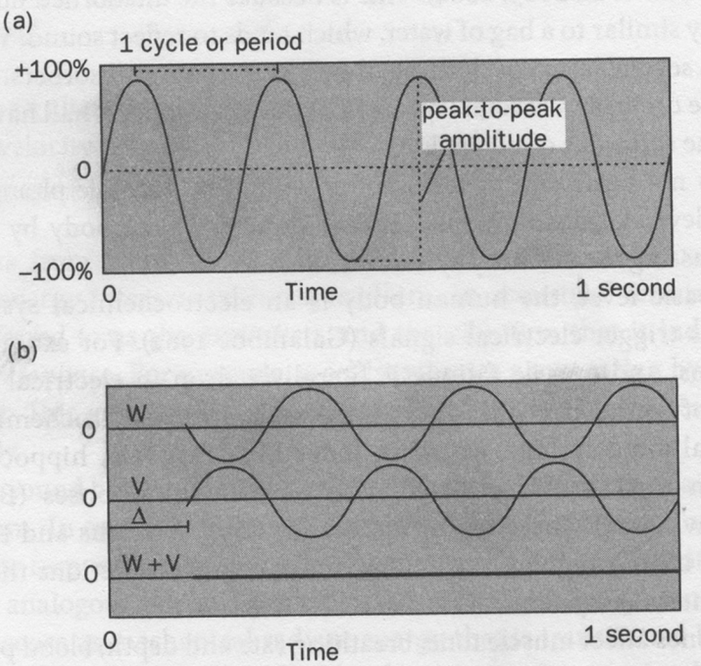

FIGURE 3.1 Time-domain representation of waveforms. (a) The distance between repetitions of a waveform is a cycle or a period. The peak-to-peak amplitude is the distance between the highest peak and the lowest trough. In this diagram, the amplitude scale is measured in percentage, but other scales, such as decibels, can be used. (b) This diagram shows two waveforms *W* and *V*, one of which is phase-shifted with respect to the other by the delay time Δ. The two waveforms are out of phase (i.e., not phase-aligned). Indeed, waveform *V* is phase-inverted with respect to *W*. If we were to add *W* and *V*, the result would be zero amplitude, shown in the third example. This is called *destructive interference.*

of air pressure change; we can measure amplitude as the vertical distance from the zero pressure point (in the middle) to the highest (or lowest) points of a given waveform segment.

Digital audio editing programs usually display pressure graphs, which plot the amplitude profile on a linear scale, that is, from 0% (silence) to ±100% (positive maximum and negative maximum). Some editors let users see the amplitude in terms of the numerical sample values. For example, for a 16-bit sound, the samples range from a high of +32767 to a low of −32768 (corresponding to 2^{16} different values), where one bit is reserved for indicating the positive or negative sign.

The positive and negative excursions of the waveform correspond to compression and rarefaction of air molecules when sound energy passes through air. Any vibrating source causes such effects. For example, a loudspeaker creates sound by moving a membrane back and forth according to changes in an electronic signal. As the loudspeaker pushes outward, the air pressure near the loudspeaker is raised (compression). When the loudspeaker pulls inward from its position at rest, the air pressure decreases (rarefaction).

PERIODICITY AND FREQUENCY

Figure 3.1a portrays a simple sinusoidal wave. A sine repeats at exactly constant intervals of time. Repeating waveforms are called *periodic*. If there is no discernible repetition pattern, it is called *aperiodic* or *noise*. In between the extremes of periodic and aperiodic is a vast domain of quasi-periodic tones.

The rate of repetition of a periodic sound is called its *fundamental frequency*, measured in cycles per second. The scientific term for cycles per second is *hertz* (abbreviated Hz) after the acoustician Heinrich Hertz. Logically, as the interval of the period increases, the frequency decreases, and vice versa. Specifically, the period is 1/frequency. Thus the period of a waveform at 100 Hz is 1/100th of a second.

PHASE

The starting point of a periodic waveform on the *y* or amplitude axis is its *initial phase*. The cycle of periodic waveform repetition can be mapped to rotation around a circle, where one complete cycle is 360 degrees. For example, a sine wave starts at the amplitude point 0 and completes its cycle at 0. If we displace the starting point by 90 degrees (a quarter of a 360-degree cycle) then the sinusoidal wave starts at 1 on the amplitude axis. By convention, this is called a cosine wave. In effect, a cosine is equivalent to a sine wave that is *phase-shifted* by −90 degrees.

When identical waveforms start at the same initial phase, they are said to be *in phase* or *phase-aligned*. (It makes little sense to compare the phases of non-identical waveforms.) Conversely, when two waveforms start at different initial phases, they are said to be *out of phase*. In figure 3.1b, sine wave *W* starts at 0 on the amplitude axis. Notice that sine wave *V* starts after a half-cycle delay. Thus *V* is 180 degrees out of phase with respect to *W*. When two identical signals are

180 degrees out of phase, we say that they are *phase-inverted* with respect to one another. One could also say that *V* has *reversed polarity* with respect to *W*.

Notice the zero-valued waveform (*W* + *V*) in figure 3.1b. When summing two signals that are exactly out of phase, they cancel out each other.

Phase manipulations are behind a variety of audio transformations, including filtering and spatialization among others (Roads 1996). See Laitinen et al. (2013) for more on phase perception.

Spectra: the frequency domain

Many frequencies can superimpose in a waveform. A *frequency-domain* or *spectrum* representation shows the distribution of frequency energy in a sound. We can view a sound in the frequency domain after transforming it from the time domain to the frequency domain via *spectrum analysis* or *estimation*.

A working definition of spectrum is: a measure of the distribution of signal energy as a function of frequency. Such a definition may seem straightforward, but in practice, different analysis techniques measure properties that they each call "spectrum" with diverging results. Except for isolated test cases, the practice of spectrum analysis is not an exact science (see Marple 1987 for a thorough discussion). The results are typically an approximation of the actual energy. (Roads 1996 presents a variety of different methods of spectrum estimation.)

Individual frequency components of the spectrum are referred to as *partials*. *Harmonic partials* (or simply *harmonics*) are a special case. Harmonics are simple integer multiples of the fundamental frequency (2:1, 3:1, 4:1, etc.). Thus, assuming a fundamental or first harmonic of 440 Hz, its second harmonic is 880 Hz (2 × 440), its third harmonic is 1320 Hz (3 × 440), and so on. More generally, any frequency component can be called a partial, whether or not it is an integer multiple of a fundamental. Indeed, many complex sounds have many partials but no particular fundamental frequency.

The frequency content of a waveform can be displayed in myriad ways. A standard method is to plot each partial as a vertical line along an *x*-axis. This is called a *line spectrum*. The height of each line indicates the amplitude of each frequency component. The purest signal is a sine, which represents just one frequency component. If the segment being analyzed is exactly one period of the sine, then the spectrum shows a single line at the frequency corresponding to that period. Figure 3.2 depicts the time-domain and frequency-domain representations of several waveforms.

Another type of static plot is a *continuous spectrum*, which plots both inharmonic and harmonic energy. Figure 3.3 shows a typical continuous spectrum plot. The line and continuous spectral plots are static, timeless descriptions. They plot all the energy that occurs over a snapshot of time, but they do not indicate when this energy occurred within the snapshot. Since musical signals are non-stationary (i.e., constantly changing), these static views only describe a

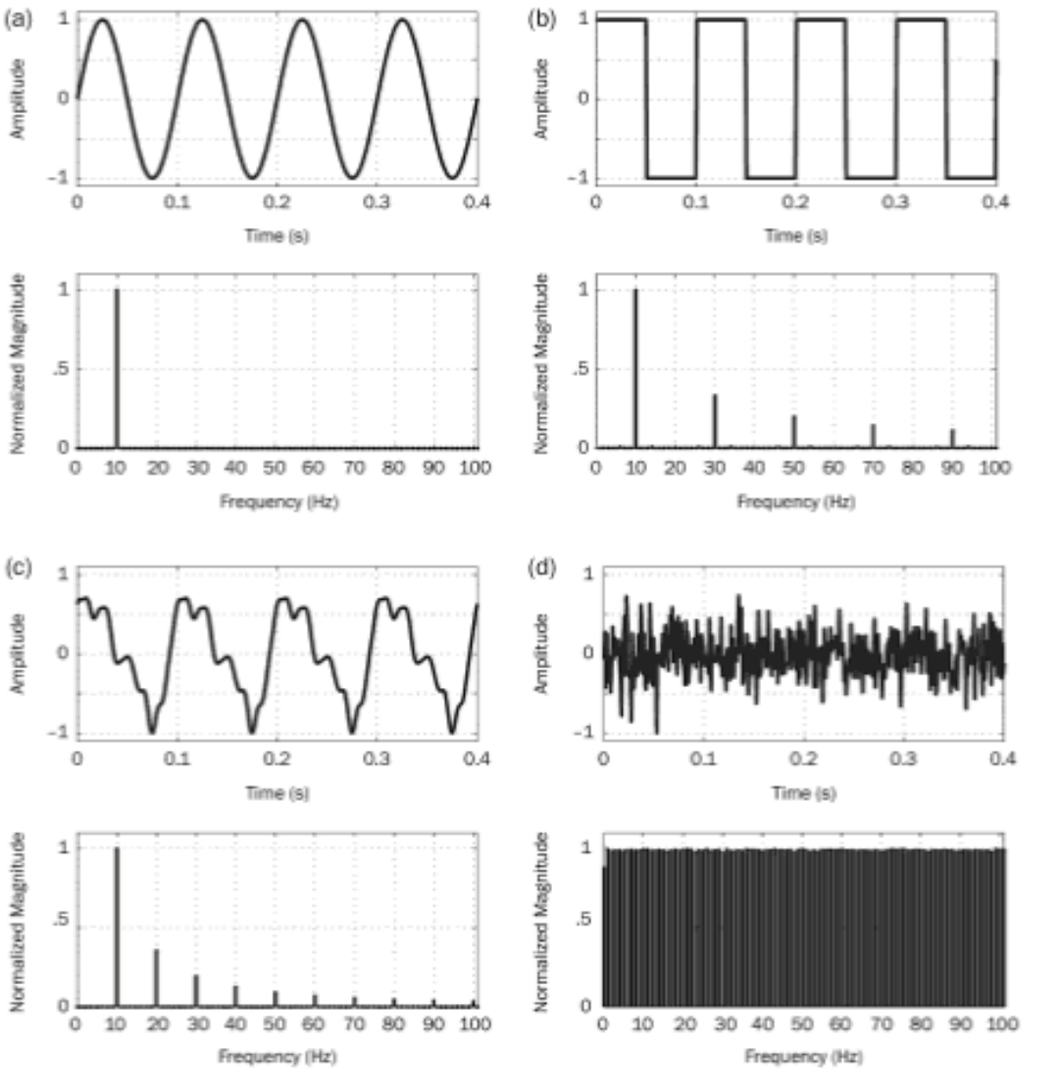

FIGURE 3.2 Time and frequency-domain representations of four audio signals. At the top is the waveform and underneath is its corresponding line spectrum. (a) 10 Hz sine, (b) 10 Hz square, (c) 10 Hz harmonic signal, (d) pink noise (notice the broad spectrum).

narrow window of time—usually less than a tenth of a second. Thus another type of plot is needed to capture variations of energy over time. One can design a time-varying visualization by analyzing sequential pieces of the signal, in the same way that a movie is nothing more than a series of snapshots.

A common time-varying plot is a *sonogram* (also called a *spectrogram*). Figure 3.4 shows a sonogram of speech, which plots frequency versus time. The darkness of the traces indicates the energy at a given frequency. The advantage of this display is that it provides a time-varying record of the sound—akin to a score—that can be studied in detail.

The sonogram is a venerable technique, with roots dating back to the "visible speech" of the 1930s (Dudley 1939). In digital form, it is implemented using the *Fast Fourier Transform* or FFT (Rabiner and Gold 1975; Allen and Rabiner 1977; Roads 1996).

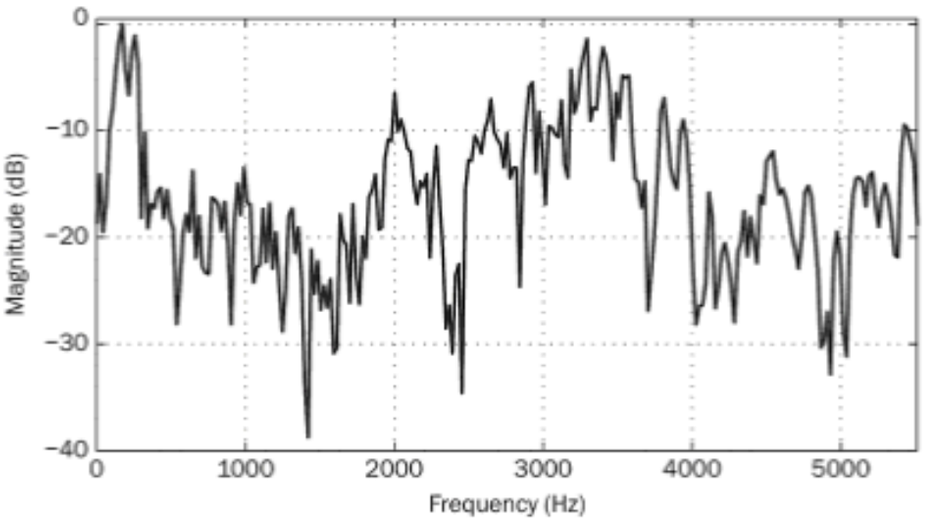

FIGURE 3.3 Continuous spectrum from 0 to 5.5 kHz of a speech sibilant noise, "shhh." The *x*-axis is frequency and the *y*-axis is magnitude. Notice the peak of energy between 3 and 4 kHz.

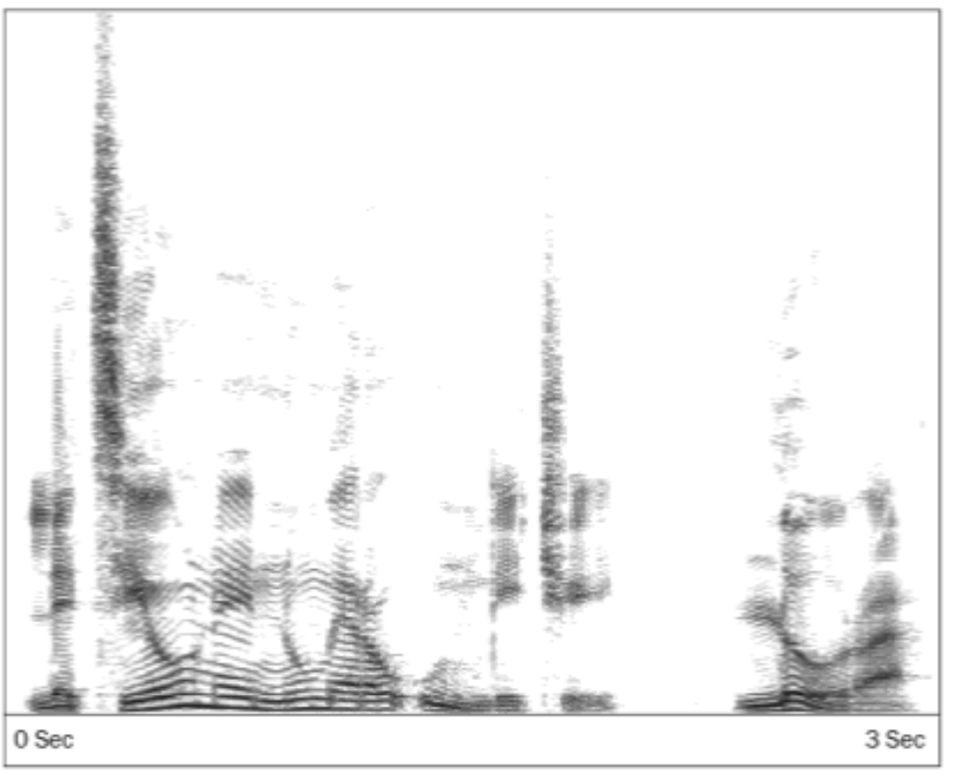

FIGURE 3.4 Sonogram projection of time-frequency energy of the Italian phrase "Lezione numero undice, l'ora" spoken by a male voice. The horizontal scale is time. The vertical scale represents frequency from 0 Hz to 12 kHz. Notice the wideband noise of the hard "z" in "Lezione" near the beginning.

Myriad alternatives to Fourier-based spectrum estimation exist (Roads 1996). Among them, one family is of particular interest, as they decompose a sound into a collection of *sound atoms* (analogous to grains). The resulting decomposition is called an *atomic time-frequency representation* of a sound. The first step in this method is to define a dictionary of all different types of atoms—long, short, high-frequency, low-frequency, etc. The next step is to analyze the sound by seeing if there is a match between the time-frequency energy in the sound and a given atom. The analysis looks at the sound and then searches through the dictionary, looking for an ideal match. If it finds one, it adds an atom to the atomic representation and subtracts that energy from the original signal. The process proceeds iteratively until all the important energy in the original is matched.

These techniques are called by various names, including *sparse approximations* and *dictionary-based pursuit* (Mallat and Zhang 1993; Mallat 1998; Sturm et al. 2009). They are called "sparse" because the atomic representation can closely approximate the original signal with a small number of atoms.

Dictionary-based pursuit is an analytical counterpart to granular synthesis (Roads 2001b). *Matching pursuit* (MP) decomposition is one sparse approximation technique. It offers a number of attractive properties, including excellent localization of time/frequency energy, customizable feature extraction, and malleability. This latter property means that the analysis data are robust under a variety of transformations. These transformations can easily be carried out in real time (Kling and Roads 2004; Sturm et al. 2008, 2009). Figure 3.5 is an analysis plot generated by the MP technique.

Sound magnitude

We all have an intuitive notion of sound level or magnitude. Even a small child understands the function of a volume knob. Dozens of terms have been devised by scientists and engineers to describe the magnitude of a sound. The following are among many:

- Peak-to-peak amplitude
- RMS amplitude
- Gain
- Sound energy
- Sound power
- Sound intensity
- Sound pressure level
- Loudness

From a scientific point of view, these are all different. In a scientific paper on acoustical measurements, a physicist should use precise and appropriate definitions. In discussing compositional issues, however, the extreme precision required by physicists is not always necessary. From a commonsense point of view, the terms listed above are all correlated and proportional to one another: A significant

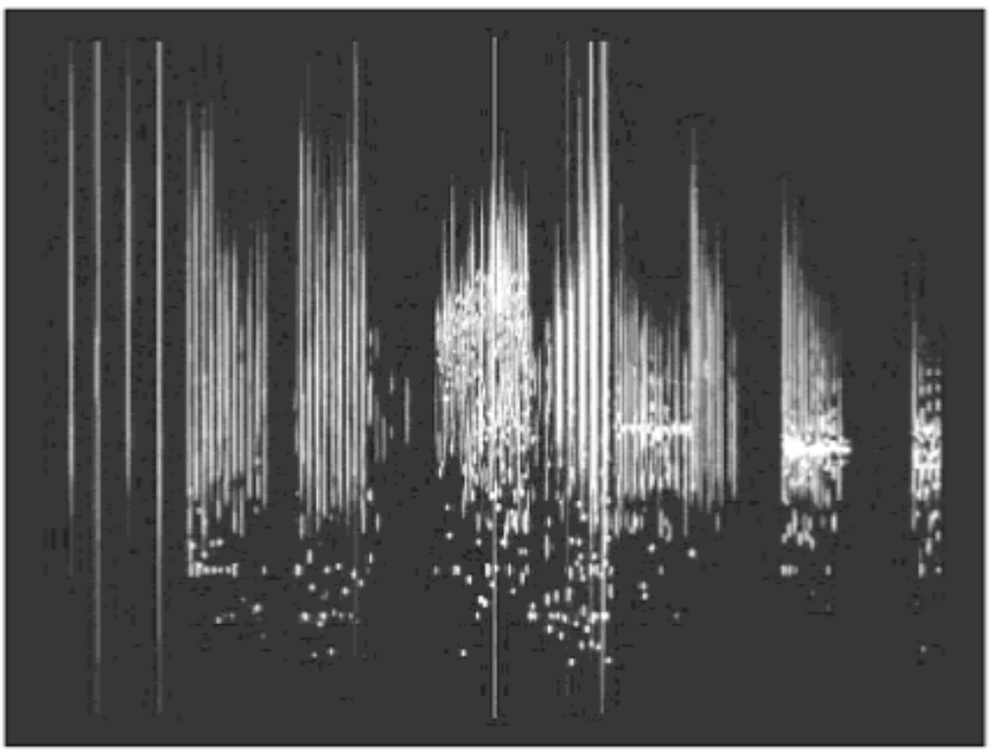

FIGURE 3.5 Frame of animation of *Pictor Alpha* (2003) by Curtis Roads. The display shows a white line in the center that signifies the now. To the left of center is the past and to the right is the future. Notice how noise around the "now" line takes the form of a cluster of atoms or grains. Animation by Garry Kling. From the DVD *POINT LINE CLOUD* (Asphodel 2004).

boost in one corresponds to a boost in all. Our ears are sharply attuned to sound level, so the concept of magnitude is physical and directly perceivable.

From a compositional point of view, the most useful terms are peak-to-peak and RMS amplitude (as seen in a sound editor), gain (a standard term for boosting or attenuating a sound), sound pressure level (what a sound level meter measures in the air), and loudness (perceived magnitude). Sound energy, power, and intensity are technical terms used by physicists to describe measures of sound magnitude in terms of the amount of work done (i.e., how much energy it takes to vibrate a medium).

Table 3.1 summarizes the formal definitions of these terms. The rest of this section explains the useful concept of *decibels* (dB).

DECIBELS

The ear is an extremely sensitive organ. Suppose that we are sitting three meters in front of a loudspeaker that is generating a sine tone at 1000 Hz, which we perceive as being very loud. Amazingly, one can reduce the power by a factor of one million and the tone is still audible. In a laboratory where all external sounds are eliminated, the reduction extends to a factor of more than one billion (Rossing 1990; Backus 1969).

TABLE 3.1
Units for measuring sound magnitude

Peak-to-peak amplitude	A measure of the peak-to-peak difference in waveform values expressed in percentage or dB (see figure 3.6). Useful for describing the magnitude of periodic waveforms in particular.
RMS amplitude	For complex signals such as noise, root mean squared (RMS) amplitude describes the average power of the waveform. RMS amplitude is the square root of the mean over time of the square of the vertical distance of the waveform from the rest position. (see figure 3.6).
Gain	Gain is a measure of the ratio of the input and the output amplitude (or power) of a process, usually measured in dB. A gain of greater than one is a boost, and a gain of less than one corresponds to attenuation.
Sound energy	A measure of work, sound energy is the ability to vibrate a medium, expressed in joules. A joule is a unit of energy corresponding to the work done by a force of one newton traveling through a distance of one meter. A newton is equal to the amount of force required to give a mass of one kilogram an acceleration of one meter per second squared.
Sound power	The rate at which work is done or energy is used. The standard unit of power is the watt, corresponding to one joule per second. One watt is the rate at which work is done when an object is moving at one meter per second against a force of one newton.
Sound intensity	Sound power per unit area, measured in watts per square meter.
Sound pressure level	Air pressure at a particular point, given in dB as a ratio of sound pressure to a reference sound pressure of 20 micropascals. A pascal is a unit of pressure equivalent to the force of one newton per square meter.
Loudness	A psychoacoustic measure based on queries of human subjects, measured in phons. One phon equals one dB SPL at one kHz.

Sound transports energy generated by the vibration of a source. The range of sound energy encompasses everything from the subsonic flutterings of a butterfly to massive explosions. A whisper produces only a few billionths of a watt. In contrast, a large rocket launch generates about 10 million watts of power.

The dB unit compresses these exponential variations into a smaller range by means of logarithms. It can be applied to myriad physical phenomena; however, the definition changes according to the phenomenon being measured. A standard unit in audio is dB SPL (*sound pressure level*). This compares a given SPL with a standard reference level. The logarithm (base 10) of this ratio is the level in decibels, hence:

$$\text{SPL in decibels} = 20 \log_{10} (W/W_0),$$

where W is the actual SPL of the signal being measured and W_0 is a standard reference level of 20 micropascals of pressure. This corresponds to the quietest sound that a human being can hear.

To calculate the dB value of a digital audio waveform, we compare a sample value to a reference level. For a 16-bit audio file, the sample values vary from –32768 to +32767, a range of 65536. Thus, for such a digital audio file, the dynamic range is

$$20 \log_{10} (65536/1) = 96.32 \text{ dB}.$$

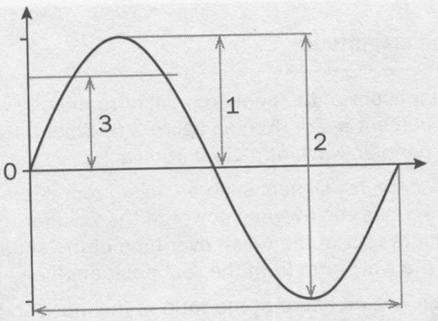

FIGURE 3.6 Measures of amplitude. (1) Peak amplitude. (2) Peak-to-peak amplitude. (3) RMS amplitude.

TABLE 3.2
Amplitude in percent versus decibels

100%	0 dB
70%	–3 dB
50%	–6 dB
25%	–12 dB
12.5%	–18 dB
6.25%	–24 dB
3.125%	–30 dB
1.562%	–36 dB
0.781%	–42 dB
0.39%	–48 dB
0.195%	–54 dB
0.097%	–60 dB
0.048%	–66 dB
0.024%	–72 dB
0.012%	–78 dB
0.006%	–84 dB
0.003%	–90 dB

Describing sound levels in terms of dB enables a wide range of amplitudes Table 3.2 shows how the decibel unit compresses large changes in percentage amplitude into relatively small changes in dB.[17]

As we move away from a sound source, its SPL diminishes according to the distance. Specifically, each doubling of distance decreases SPL by about 6 dB, which represents a 50% decrease in its amplitude. This is the famous *inverse square law*: Intensity diminishes as the square of the distance.

So far we have been talking in terms of amplitude and SPL. Another two concepts—*volume* or *loudness*—are intuitive. Technically, loudness refers to perceived or subjective intensity measured through psychoacoustic tests on human beings, not to sound pressure level measured by laboratory instruments. For example,

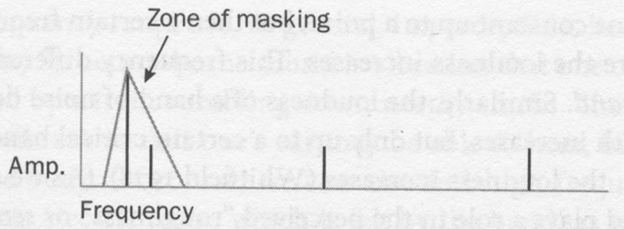

FIGURE 3.7 A loud low-frequency tone masks soft tones nearby. It does not mask tones outside the zone of masking.

the ear is especially sensitive to frequencies between 1000 Hz and 4000 Hz. Tones in this region sound louder than tones of equal intensity in other frequencies. Thus the measurement of loudness falls under the realm of psychoacoustics. In order to differentiate loudness level (a perceptual characteristic) from sound pressure level (a physical characteristic), a unit *phon* (rhymes with John) is used. For example, to sound equally loud (60 phons), a tone at about 30 Hz needs to be boosted 40 dB more than a 1000 Hz tone.

The ear is quite sensitive to sound intensity. A classical question of psychoacoustics is: What is the minimal intensity difference between two sounds, otherwise identical, that allows a listener to reliably report that one sound is louder than another? This is the *just noticeable difference* (JND). Interestingly, this is not a constant, but varies according to the frequency and intensity of the signal, and also from person to person. For example, in the ear's most sensitive range, at 70 dB between 1000 and 4000 Hz, the JND is less than 0.5 dB (Scharf 1978). However, the JND expands to 10 dB for low-frequency (35 Hz), low-intensity (30 dB) tones.

MASKING AND CRITICAL BANDS

The subject of *masking* inevitably arises in discussions of loudness perception. In its most basic form, masking describes a phenomenon wherein a low-level sound is obscured by a higher-level sound (figure 3.7).

For example, standing in a shower masks many sounds, such as someone speaking nearby. Masking is the process by which the threshold of inaudibility of one sound is raised by the presence of another sound. In this case, the voice, which would normally be perceived clearly, is reduced in apparent loudness. This effect is called *partial masking* because the masking signal does not completely eliminate the masked signal. Partial masking depends not only on the intensity of the masker, but also on the frequency of the masking signal relative to the frequency of the masked signal.

Human hearing can be considered as divided into a number of overlapping frequency bands. The interactions between sounds within adjacent bands can lead to a variety of frequency- and bandwidth-dependent loudness phenomena. For example, if we play two sine waves that are very close in frequency, the total loudness we perceive is less than the sum of the two loudnesses we would hear from the tones played separately. As we separate the tones in frequency, this

loudness remains constant up to a point, but then a certain frequency difference is reached where the loudness increases. This frequency difference corresponds to the *critical band*. Similarly, the loudness of a band of noise does not increase as the bandwidth increases, but only up to a certain critical bandwidth. Beyond this bandwidth, the loudness increases (Whitfield 1978). (As we see in chapter 7, the critical band plays a role in the perceived "roughness" or *sensory dissonance* of a pitch combination.)

PERCEPTUAL CODING AND DATA COMPRESSION

Taking advantage of the fact that the presence of one sound can partially or completely mask a second sound, *perceptual coding* techniques are designed for *data compression*—a large reduction in the amount of data needed to transmit an audio signal. These techniques estimate which frequency bands are being masked so that they can throw them away before transmission. Common data reduction schemes like MPEG-1 Audio Layer 3 (or MP3) and Advanced Audio Coding (AAC) are based on such methods, which are also called *lossy compression* schemes. A large research literature surrounds this topic.

So far we have discussed *simultaneous masking* effects. Two other types of masking effects are time-based: *forward* and *backward masking*. Consider a short sound that ends abruptly. The human auditory system continues to react for a short time (about a half second) after the sound ends (Zwislocki 1978). This "resonance" can blur our perception of the onset of a second sound. Indeed, when the time interval between impulses is less than about 50 ms, the ear no longer perceives them as separate impulses but collectively as continuous tones. Thus forward masking is strongly related to our perception of pitch.

Backward masking is a curious phenomenon. Basically, a loud click or noise coming less than 100 ms after another sound can obscure our perception of the earlier sound (Zwislocki 1978). The masking sound can disrupt the brain's ability to hear a preceding sound, hence the term backward).

It is interesting that both forward masking and backward masking occur in the visual domain as well. This would suggest that they are indicators of the limitations of the brain in handling events that are too closely spaced in time. We continue this discussion in the next section.

Zones of frequency and intensity

Some sounds we can hear; other sounds we cannot. The *audio* frequencies are perceptible to the ear. They span the range of about 20 Hz to 20 kHz, where the specific boundaries vary depending on age and the individual.

Low-frequency impulsive events are perceived as rhythms. These are the *infrasonic frequencies* in the range below about 20 Hz. The infectious beating rhythms of percussion instruments fall within this range. (Note that sine waves

in this same frequency range are, in general, imperceptible, because they have little bandwidth.) Structure-borne sound is vibration that one can feel, like the vibration caused by a train rumbling down nearby tracks. These are typically low-frequency vibrations that one's ear may be able to hear, but they are also felt through the body. Of course, we can also feel high frequencies, such as the buzzing of an electric razor, which is felt by the hand, as well as heard by the ear.

Ultrasound comprises the domain of high frequencies beyond the range of human audibility. The threshold of ultrasound varies according to the individual, their age, and the test conditions.

Some sounds are too soft to be perceived by the human ear, such as a caterpillar's delicate march across a leaf. The softest sounds we can hear stand at the *absolute threshold of hearing* (Zwikcker and Feldtkeller 1999). Below this is the zone of the *subabsolute* intensities, sounds too feeble to be perceived by the ear. This zone of faint sounds can sometimes be captured by a microphone and amplified into the realm of the audible to spectacular effect—a classic technique of musique concrète.

Other sounds are so loud that to perceive them directly is dangerous, since they are destructive to the human body. Very loud impulses can permanently damage the inner ear. Sustained exposure (typically in a noisy work environment) to sound levels above 85 dB induces permanent hearing loss. As the intensity increases, it takes less and less time to induce permanent loss. Around 130 dB, sound is not only heard but also felt as a painful pressure wave by the exposed tissues of the body (Pierce 1983). The loudest possible sound in air is about 194 dB, the point at which the nominal air pressure is reduced to from an average of 100,000 pascals to 0 pascals, creating a vacuum.

The dangerous zone of intensities extends into a range of highly destructive acoustic phenomena. The detonation of a high explosive device, for example, results in an intense acoustic shock wave. For lack of a better term, I call these the *perisonic* intensities (from the Latin "periculos" meaning dangerous). The *audible* intensities fall between these two ranges. Figure 3.8 depicts the zones of sound intensity and frequency. What I call the *alphazone* in the center is where the audio frequencies intersect with the audible intensities, enabling hearing. Notice that the α-zone is only a fraction of a larger range of sonic phenomena.

Timescales of sound and music

Music theory has long recognized a temporal hierarchy of structure in music compositions. Adopting the terminology of mathematical graph theory (Bobrow and Arfib 1974; Aldous and Wilson 2000), this hierarchy can be plotted as an inverted tree structure (figure 3.9). The topmost vertex or *root* represents the entire piece (the *global* level). The root splits into multiple arcs, which connect to vertices that represent substructures of the piece. These in turn split into further

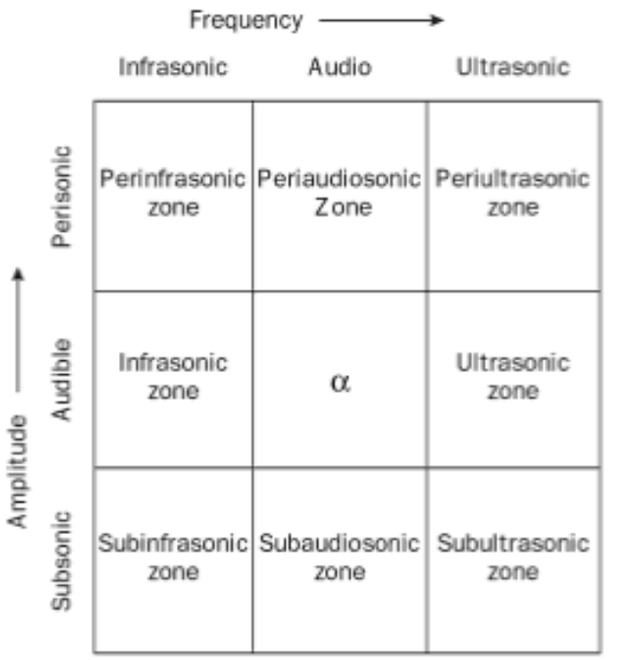

FIGURE 3.8 Zones of intensities and frequencies.

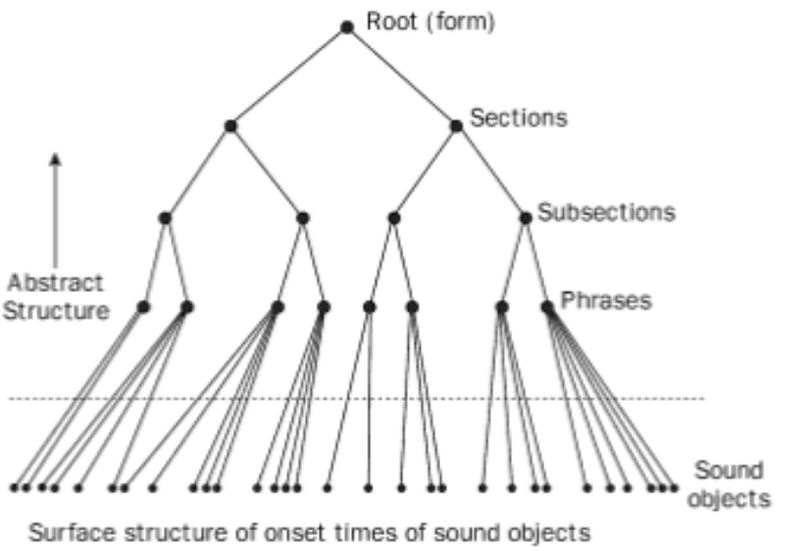

FIGURE 3.9 Idealized graph of hierarchical musical structure. This is a simple musical structure, typical of nursery rhymes.

substructures, ultimately arriving at the bottom or *terminal* layer of individual notes or sound objects (the *local* level).

This hierarchy, however, is incomplete. Above the level of an individual piece are the cultural time spans defining the oeuvre of a composer or a stylistic period. Beneath the level of the note lies another multilayered stratum, the microsonic hierarchy. Modern tools let us view and manipulate the microsonic layers, from which all acoustic phenomena emerge. Beyond these physical timescales,

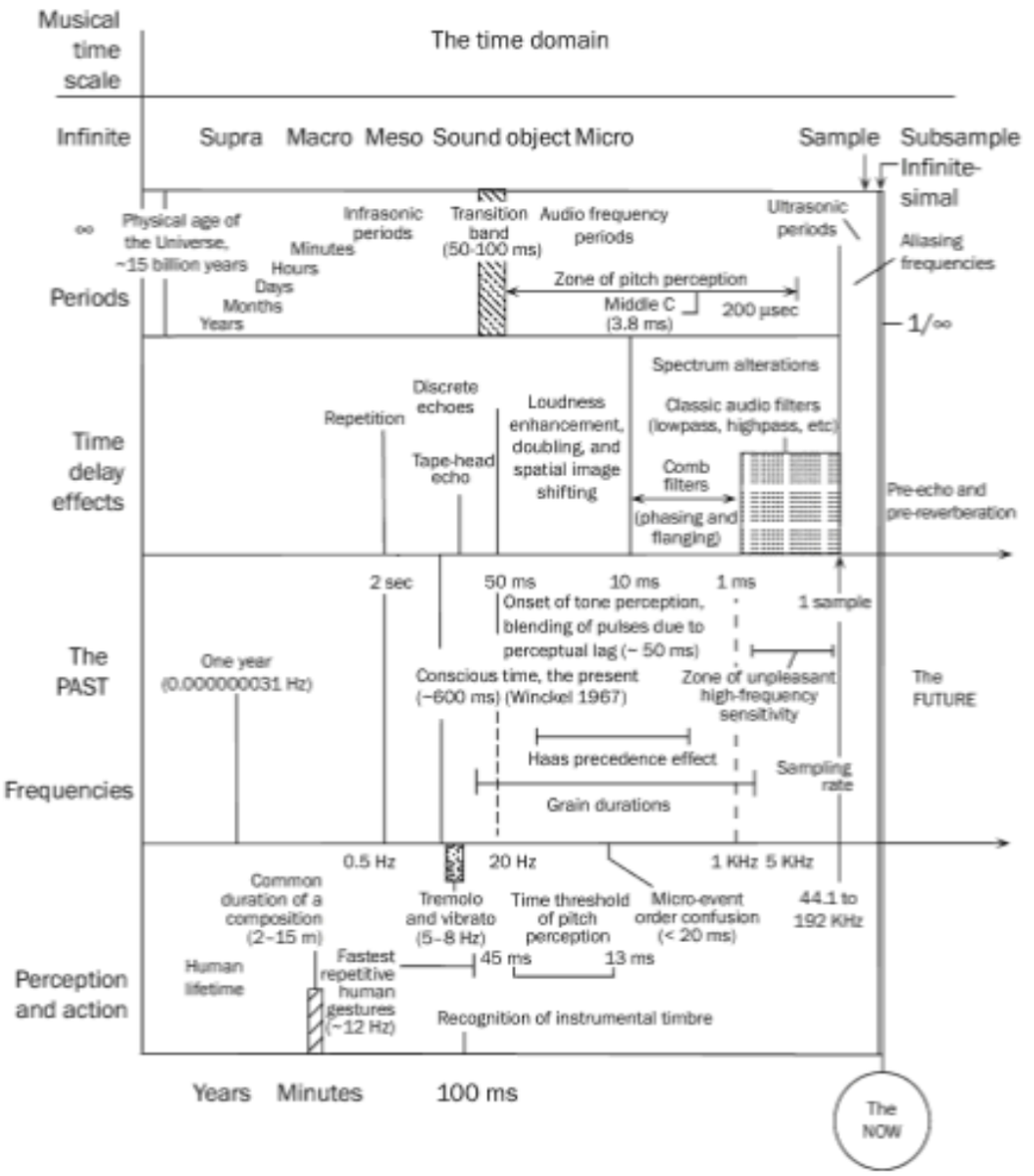

FIGURE 3.10 The time domain. Ranges are not drawn to scale.

mathematics defines two ideal temporal boundaries—the infinite and the infinitesimal, which appear in the theory of Fourier analysis.

Understanding the multiscale nature of sound is essential to the practice of composition today. Taking a comprehensive view, we distinguish nine timescales of music, as shown in figure 3.10.

The nine timescales are as follows:

1. *Infinite*—The ideal time span of mathematical durations such as the infinite sine waves of classical Fourier analysis. Harking back to a theological perspective, Messiaen (1994) called this *eternity*.
2. *Supra*—A timescale beyond that of an individual composition and extending into months, years, decades, and centuries. Musical cultures

are constructed out of supratemporal bricks: the eras of instruments, of styles, of musicians, and of composers.

3. *Macro*—The timescale of overall musical architecture or form, measured in minutes or hours, or in extreme cases, days (as in Wagner's *Ring* cycle, Japanese Kabuki rituals, etc.).
4. *Meso*—Divisions of form, groupings of sound objects into hierarchies of phrase structures of various sizes, measured in minutes or seconds. This "local" as opposed to "global" timescale is extremely important in composition. For it is most often on the meso level that the sequences, combinations, and transmutations that constitute musical ideas unfold. Local rhythmic patterns, as well as melodic, harmonic, and contrapuntal relations transpire at the meso layer, as do processes such as theme and variations, development, progression, and juxtaposition.
5. *Sound object*—A basic unit of musical structure, generalizing the traditional concept of note to include complex and mutating sound events on a timescale ranging from a fraction of a second to several seconds. Whereas notes are static and homogeneous (each note has a fixed pitch, duration, dynamic, and instrument name), sound objects can vary in time (they can mutate) and they are very heterogeneous. In electronic music, any sound from any source may serve a musical function. Thus the influence of the traditional intervallic pitch-duration grid (which assumes note homogeneity) is greatly diminished.
6. *Micro*—Sound particles on a timescale that extends down to the thresholds of auditory perception (measured in thousandths of a second or milliseconds). Thousands of microsonic particles such as grains and wavelets can serve as building blocks for a complex time-varying sound (Roads 2002).
7. *Sample*—The lowest level of digital audio systems: individual binary samples or numerical amplitude values, one following another at a fixed time interval. The period between samples is measured in millionths of a second (microseconds).
8. *Subsample*—Fluctuations on a timescale that are too brief to be properly recorded or perceived, measured in nano-, pico-, fempto-, atto-, zepto-, and yoctoseconds (1×10^{-24} seconds). The subsample timescale encompasses an enormous range of phenomena, from the perceptible to the imperceptible: aliased artifacts, ultrasounds, and the Planck interval (see Roads 2001b).
9. *Infinitesimal*—The ideal time span of mathematical durations such as the infinitely brief delta functions. One application of the delta function in signal processing is to tether the mathematical explanation of sampling (see Roads 2001b).

Notice in the middle of figure 3.8, in the frequency column, a line indicating "Conscious time, the present (~600 ms)." This line marks off Winckel's (1967)

estimate of the "thickness of the present." The thickness extends to the line at the right indicating the physical NOW. This temporal interval constitutes an estimate of the accumulated lag time of the perceptual and cognitive mechanisms associated with hearing. Here is but one example of a disparity between *chronos* (physical time) and *tempus* (perceived time).

As a sound's duration passes from one timescale to another, it crosses perceptual boundaries. It seems to change quality. This is because human perception processes each timescale differently. Consider one period of a simple sinusoid wave transposed to several timescales (1 μsec, 1 ms, 1 sec, 1 minute, 1 hour). The waveform shape is identical, but one would have difficulty classifying auditory experiences of these waveforms in the same qualitative family.

In some cases, the borders between timescales are demarcated clearly, but ambiguous zones surround other boundaries. The ultrasonic threshold, for example, could be said to be different for each person, changing with age. Moreover, training and culture condition the perception of certain temporal phenomena. To notice a flat pitch or a dragging beat, for example, is to detect a temporal anomaly on a microscale that might not be noticed by everyone.

It is easy to distinguish the boundary separating the sample timescale from the subsample timescale. This boundary is the Nyquist frequency, or half the sampling frequency. However, the perceived effect of crossing this boundary is not always evident. Low-level aliased frequencies from the subsample time domain may mix unobtrusively with high frequencies in the sample time domain.

The border between other timescales is context-dependent. Between the sample and micro timescales, for example, is a region of transient events—too brief to evoke a sense of pitch but rich in timbral content. Between the micro and the object timescales is a stratum of brief events like short staccato notes. Another zone of ambiguity is the border between the sound object and meso levels, exemplified by an evolving texture. Consider a granular cloud lasting a minute or more; it is perceived as a unified entity but it may be constantly mutating. Sound art installations that are set up for weeks at a time blur the boundary between the macro and the supratemporal timescales. Indeed, an algorithmic composition system can spawn non-repeating musical patterns indefinitely, given a sufficiently large parameter space to explore (Collins 2002).

Timescales are interlinked, since the musical structure at each level comprises events on lower levels and is simultaneously subsumed by higher timescales. Hence, to operate on one level is to affect other levels. The nonlinear nature of musical structure means, however, that linear incremental changes in the parameters on one timescale cannot guarantee a linear perceptible effect on adjacent timescales. The most common example is a beating pulse at 1 Hz, which when sped up linearly, passes through several zones of perception (figure 3.8). At 20 Hz, it forms a continuous tone without identifiable pitch. Increasing the frequency of this tone, it turns into a pitched tone at about 40 Hz. However, the perception of pitch evaporates again at about 5 kHz. Increasing the frequency still further to 9 kHz, the tone becomes piercing to

the ear, while at 12.5 kHz, it gives a sense of transparency and air. The tone disappears from awareness altogether somewhere about 20 kHz, yet it is still registered by the brain considerably beyond this frequency threshold (Oohashi et al. 1991, 1993).

Sound phenomena on one timescale may travel by transposition to another timescale, but the voyage is not linear. Pertinent characteristics may not be maintained. In other words, the perceptual properties of a given timescale are not necessarily invariant across dilations and contractions. To cite an example, a melody loses all sense of pitch when transposed very high or very low. This inconsistency, of course, does not prevent us from applying such transpositions. It merely means that we must recognize that each timescale abides by its own rules.

The speed of sound

Sound propagates at different rates, depending on the medium of propagation. As the mathematician and physicist Leonard Euler wrote in his 1726 doctoral dissertation "De Sono":

> [As] both the density or weight and pressure of the air surrounding the earth are subject to various changes, the speed of sound is constantly changing also. Hence the maximum speed of sound will be [found] on the hottest days with a clear sky. . . . With the harshest cold and the fiercest storm, the speed of sound should be a minimum.

Sound traveling in a sparse medium of dry air propagates at about 331 meters (1100 feet) per second at 0 degrees centrigrade. In contrast, sound waves travel slowly in rubber—an absorbent medium—at speeds as low as 40 meters per second (Eargle 1995).

Sound travels much faster in water (1429 meters per second). The speed of sound in a bubbly medium is slowed, however, due to interference by the bubbles. Bubbles transport their own sounds. A gurgling brook releases bubbles of air, and as these air pockets contact the surface, they pop with a resonance according to the size of the bubble. Large bubbles make low-frequency sounds while the tiny bubbles in a glass of fine champagne emit a high fizzing sound.

Finally, sound travels fastest in solid metal media, such as a bar of steel, in which longitudinal sound waves travel over 5000 meters per second (Rossing 1990; Eargle 1995).

If the source producing the sound is moving toward or away from a listener, this affects what they hear. The *Doppler shift* effect is a continuous pitch bending that occurs when a moving sound source passes a stationary listener. If we are standing by the railroad tracks and a train approaches at high speed, we hear the pitch shift upward as it approaches and shift downward as it passes by. The pitch shift upward is attributable to the shortening of the wavefronts as the sound

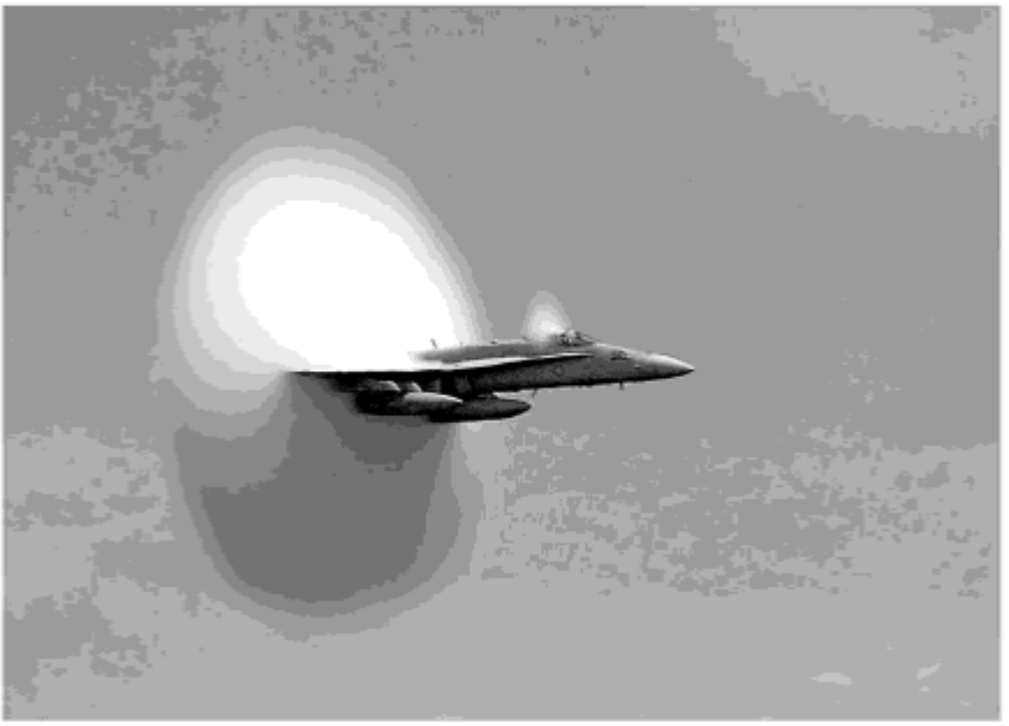

FIGURE 3.11 Boeing F-18 jet at Mach 1. A condensation cloud can appear around aircraft traveling at transonic velocities. (Photo: John Gay, www.navy.mil/navydata/images/hornetsb.jpg)

approaches, and the pitch shift downward is attributable to the corresponding lengthening of the wavefronts as it recedes into the distance.

MACH 1 AND BEYOND

The speed of sound, referred to as Mach 1, is approximately 1224 km/hour (761 mph) in air at sea level. When a sound-emitting source moves at *transonic* velocity (Mach 1), it is effectively aligned with its own sonic shock wave (figure 3.11). When it passes overhead, one hears a very loud impulse that represents its sonic history (Dowling and Williams 1983).

A curious effect occurs at exactly twice the speed of sound: Mach 2. The sound plays in reverse! As the source is ahead of its sound wave, the most recent sounds are heard first, and the earlier sounds arrive later. To the people in the airplane, the noise is heard normally.[18]

Sounds at extreme sonic velocities are destructive. Explosives can generate powerful transonic shock wave air currents traveling at up to 8000 meters per second, corresponding to Mach 24. These destroy everything in their path.

Shape, direction, and size of sounds

Sounds sculpt space. They form individual shapes of different sizes radiating in specific directions. This section examines these spatial attributes.

SOUND SHAPE

Sound waves in air tend to radiate spherically in three dimensions from a source. The direction and shape of the sound waves are variously called the *dispersion pattern*, *direction pattern*, or *radiation pattern*, and they can vary depending on the frequency band and the source. The dispersion pattern of a loudspeaker is usually fixed, whereas the dispersion pattern of an acoustic instrument varies as a function of frequency (figure 3.12).

A special-purpose *superdirectional* loudspeaker projects a narrowly focused sound beam—less than 50 cm in diameter at a distance of 2 meters (Holosonics 2010). A typical application is a museum installation, where listeners standing under a sound beam can hear a narrative description of a work of art, while others standing nearby cannot. (For more on superdirectional sound beams, see chapter 10.)

THE SIZE OF SOUNDS

Sounds have a specific physical size as well as shape. A quiet sound is physically petite. One has to put one's ear close to it because the body of air it perturbs is tiny. Other sound waves are gigantic, such as the Krakatoa explosion of August 1883. It was heard 4800 kilometers away, and the pressure wave traveled around the earth for 127 hours (Miller 1935). The shape of the sound is partly determined by the dispersion pattern of its source and also by the architecture of the space in

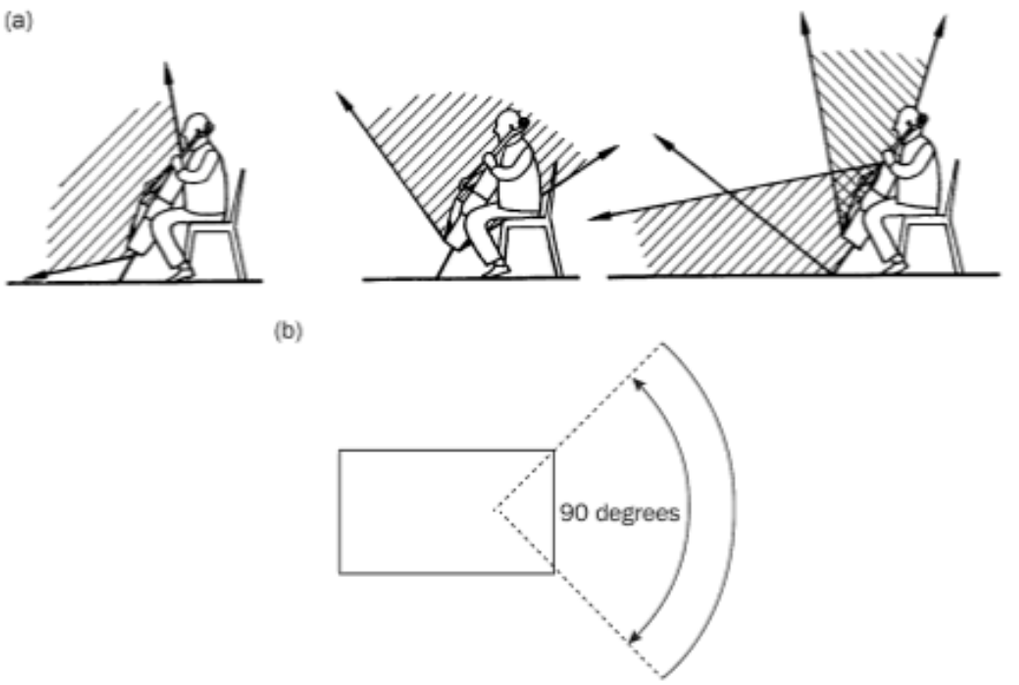

FIGURE 3.12 Comparison of dispersion pattern of a cello and a loudspeaker. (a) The dispersion pattern of a cello varies according to frequency (after Dickreiter 1989). (b) View of a loudspeaker from above. A typical loudspeaker dispersion pattern is relatively consistent for all frequencies.

which it is projected. For example, striking a triangle produces a spherical shape wave, while a sound beam emitted by a superdirectional loudspeaker is narrow. The reflections from a spherical sound wave tend to produce generalized reverberation, while a narrow sound beam will reflect like a beam of light, with echoes heard only in certain locations.

We can think of a sound wave as "unfolding" longitudinally from its source. Since sound takes time to travel in the air, we can measure the physical distance it takes for one period at a given frequency to unfold. This distance is called the *wavelength*. Since sound travels at 331 meters/second in air, the wavelength of a 331 Hz sine tone is one meter.

For loudspeaker listening, a room that is long (i.e., the distance between the loudspeakers and the rear wall is large) is preferable to a short room, since bass frequencies need space to unfold without distortion. For example, a deep bass tone at 68 Hz needs four meters to take form. In small rooms, parts of bass waveforms are reflecting back and colliding with other bass waveforms. There may be deep bass, but room-induced resonances (standing waves) can distort the tones produced by the loudspeakers.

PERCEPTION OF THE RATE OF SOUND EVENTS

In the world of acoustic music, the rate of successive sound events is limited by human performance. A virtuoso pianist can only play at maximum about 12 events per second. A skilled drummer can multiply this rate through the technique of the double-stroke or triple-stroke roll, in which every stroke produces two or three bounces, resulting in rolls that sound almost continuous due to the forward masking effect discussed in a previous section.

Not only is our muscular performance constrained, so is our perception. When events fly by too quickly, they "blur" in our brain. As Milton Babbitt (1964) observed:

> The constant self-question of the composer of the past, "Does what I have written exceed the capacities of the performer?" is now replaced by "Does what I have produced exceed the perceptual capacities of the trained listener?

Consider the effect of tempo on perception. I conducted an experiment with a MIDI score file of Mozart's familiar K .331 *Sonata for Piano Number 11*, the *Rondo Alla Turca: Allegretto* movement (ca. 1783). In a sequencer, one can change the playback tempo without altering the pitches. Here is what I observed when I steadily increased the playback tempo:

1 × tempo (the movement has a duration of 3 min, 45 sec): normal performance
2 × tempo: It is no longer playable by a human musician, but the ear follows it readily as an uptempo version of K .331.
3 × tempo: Certain melodic details are lost, but the identity of the piece is still unmistakable.

4 × tempo: Some melodic passages morph into glissandi, as if the pianist is skimming a finger up and down the keyboard.
6 × tempo (37 seconds duration): The piece loses structural coherence; certain parts sound like arbitrary glissandi.
12 × tempo (18 seconds duration): The sound degenerates into an unrecognizable swirl of grains.

The potential for uptempo performance was recognized in the early days of computer music. It led to a new genre of *machine music* characterized by superhuman speed and rhythmic precision. *Sonatina for CDC 3600* (1966) by Arthur Roberts is a classic example (on a recording accompanying Von Foerster and Beauchamp 1969).

Sound example 3.1. Excerpt of *Sonatina for CDC 3600* (1966) by Arthur Roberts.

Machine music found an echo in the manic sequenced electronica of the early 2000s, for example, *District Line II* (2003) by Squarepusher (Jenkinson 2003).

Sound example 3.2. Excerpt of *District Line II* (2003) by Squarepusher.

Psychoacousticians have studied a variety of perceptual effects that occur in the region between individuated event streams and continuous flows. These are mainly focused on processes of *fission* where alternating tones appear to be part of separate lines or streams, versus *fusion* in which they appear to be part of the same line or stream (Bregman 1978; Deutsch 1982). As in other aspects of perception, the laws of Gestalt psychology strongly influence how we segregate or group phenomena together. For example, the rule of "common fate" says that we tend to group together phenomena that change at the same time in the same way. (See chapter 6 for more on Gestalt grouping mechanisms.)

The perceptual threshold between individual events and a continuous flow is of great interest from an aesthetic point of view. For example, when discrete melodies are sped up, they lose their melodic quality and morph into continuous timbres. When rhythms are sped up, they morph into tones. When modulations like tremolo and vibrato are sped up, they morph into complex spectra. Streaming around the thresholds of this zone of morphosis—where discrete events turn into continuous tones—is intrinsically fascinating. It challenges our ability to keep pace with the flow. An example is Mario Davidovsky's *Synchronisms Number 6 for Piano and Electronic Sounds* (1970), which features glittering melodic strands; at one point, 14 tones flash by in just 1.14 seconds, each tone lasting 70 to 80 ms.

Sound example 3.3. Excerpt of *Synchronisms Number 6 for Piano and Electronic Sounds* by Mario Davidovsky.

A similar event rate appears at the end of my *Eleventh vortex* (2001), in which a burst of 40 sounds occurs in less than four seconds.

Sound example 3.4. Excerpt of *Eleventh vortex* by Curtis Roads. 

Timbre

> A compound color is produced by the admixture of two or more simple ones, and an assemblage of tones, such as we obtain when the fundamental tone and the harmonics of a string sound together, is called by the Germans *Klang*. May we not employ the English word clang to denote the same thing, and thus give the term a precise scientific meaning akin to its popular one? And may we not, like Helmholtz, add the word color or tint, to denote the character of the clang, using the term clang-tint as the equivalent of *Klangfarbe*?
>
> —JOHN TYNDALL (1875)

> Music theorists have directed little attention towards the compositional control of timbre. The primary emphasis has been on harmony and counterpoint. The reason for this probably lies in the fact that most acoustical instruments provide for very accurate control over pitch but provide little in the way of compositionally specifiable manipulation of timbre. With the potential of electroacoustic instruments the situation is quite different. Indeed one can now think in terms of providing accurate specifications for, by way of example, sequences of notes that change timbre one after another.
>
> —DAVID WESSEL (1979)

> In this day and age, the timbre of electroacoustic music could be expected to be the result of the compositional process as a whole, be it algorithmic or not.
>
> —CLARENCE BARLOW (2006)

As discussed in the preface, the standard definition of the term "timbre" is inadequate. The American National Standards Institute (1999) defines it as "an attribute of auditory sensation" that enables a listener to distinguish two sounds having the same loudness and pitch. This definition describes timbre as a perceptual phenomenon, and not an attribute of a physical sound. Despite this, everyone has an intuitive sense of timbre as a descriptive attribute of a sound (e.g., a gong sound, muted trumpet, voice-like sound, toy piano, etc.). Moreover, the spectrum of most instrumental sounds changes when they are played at different pitches, loudnesses, and durations, so even one instrument has many timbres.

Everyone agrees that timbre is a "multidimensional property," but there is no general scientific agreement about what these properties are or how to measure them.

Recognizing that the vast range of sound material opened up by musique concrète was largely undefined and unclassified, Pierre Schaeffer made a pioneering attempt to describe the correlates of timbre in his *Traité* (1966, 1976, 1977; Chion 2009) and the disc plus booklet *Solfège de l'object sonore* (Schaeffer, Reibel, and Ferreyra 1967). Although the vocabulary he developed is idiosyncratic, there is no question that Schaeffer made many discoveries on the nature of sound color.

Most scientific research on timbre has focused on traditional instrument and vocal tones. A classic example is the research of John Grey, who made a three-dimensional map of the perceived timbre space for different instrumental tones (Grey 1975, 1978). Tones that sound similar were close together in this space, while dissimilar tones were far apart. Wessel (1979) devised a scheme for navigating timbre space by means of additive synthesis. Only a few heroic attempts have been made to classify the vast universe of sound outside the territory of pitched acoustic instruments (Schaeffer 1977; Schaeffer, Reibel, and Ferreyra 1967).[19]

Numerous attributes of sounds inform timbre perception. These include the amplitude envelope of a sound (especially the attack shape), undulations due to vibrato and tremolo, perceived loudness, duration, and the time-varying spectrum envelope (the distribution of frequency energy over time) (Schaeffer 1966, 1976, 1977; Risset 1991; McAdams and Bregman 1979; McAdams 1987; Gordon and Grey 1977).

Certain tones, such as simple sawtooth waveforms, have a spectrum envelope that is *monotonic* (i.e., attenuating or rolling off linearly with increasing frequency; Mathews 1999). Other tones, such as spoken vowels, exhibit several sharp peaks called *formants* in their spectrum envelopes (figure 3.13). The formants move around as we speak to create the various phonemes of speech. For synthetic tones, we can choose to make the formant frequencies either independent of or dependent on pitch. In the latter case, the waveshape tends to be constant, and only the pitch period changes. For speech-like tones, the formants are in the range of 250 Hz to 4 kHz (Cook 1999).

MPEG-7 TIMBRAL DESCRIPTORS

Until the development of the MPEG-7 timbral descriptors in 2001, timbre was a vaguely defined territory described by numerous and incompatible maps. The MPEG-7 multimedia content description standard changed this situation. It provided a standard set of mathematically defined terms to describe a number of important aspects of timbre. Thus it represents a significant advance in timbral description.

A media file that conforms to the MPEG-7 format contains *metadata* (i.e., descriptors of the contents of the file). One category of descriptors built into this

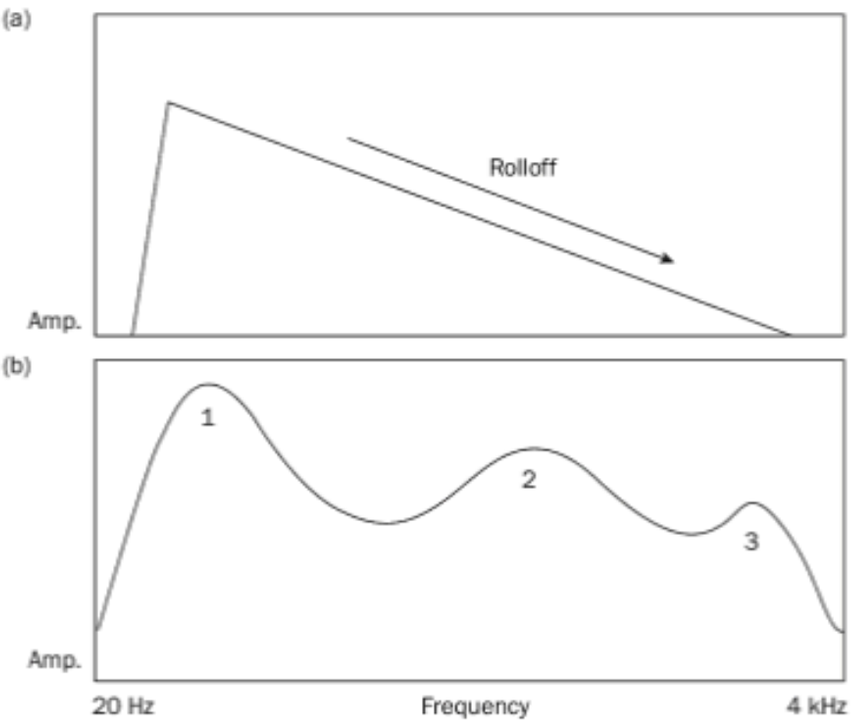

FIGURE 3.13 Spectrum envelopes. (a) Spectrum envelope with monotonic high-frequency rolloff. (b) Spectrum envelope with three labeled formants.

standard concerns timbre. Table 3.3 is a list of these descriptors, with capsule definitions adapted from Martinez (2004).

Although the definitions are described in capsule form here, in the standard, they are precisely defined mathematically. Indeed, my student Daniel Mintz (2007) developed a software synthesizer that could analyze a sound and synthesize similar sounds based on its MPEG-7 descriptors. For software resources in support of MPEG-7, see Casey (2010).

The descriptors form a solid scientific beginning for a taxonomy of timbre. It is only a beginning, however, because MPEG-7 is not a complete account of timbre. Its focus is specifically on harmonic, coherent, sustained sounds, and non-sustained percussive sounds.

Descriptors for the vast world of noises remain to be developed. Here the granular paradigm should prove useful, particularly since variations in granular density are characteristic of many noises.

FEATURE VECTORS

In parallel with the MPEG-7 standard, both academic and industrial researchers have developed schemes to analyze music (including its timbral aspects) for applications like *music information retrieval* (MIR) for categorization, recognition, retrieval, and recommendation of music (Wold et al. 1996). A typical MIR system takes a popular song and analyzes it according to dozens or even hundreds of quantified *feature vectors* (some similar to the MPEG-7 timbre

TABLE 3.3

MPEG-7 timbral descriptors

Audio waveform	The audio waveform envelope (minimum and maximum), typically for display purposes.
Audio power	Describes the temporally smoothed instantaneous power, which is useful as a quick summary of a signal, and in conjunction with the power spectrum.
Log-frequency power spectrum	Logarithmic-frequency spectrum, spaced by a power-of-two divisor or multiple of an octave.
Audio spectral envelope	A vector that describes the short-term power spectrum of an audio signal. It may be used to display a spectrogram, to synthesize a crude "auralization" of the data, or as a general-purpose descriptor for search and comparison.
Audio spectral centroid	Describes the center of gravity of the log-frequency power spectrum. A general indicator of "brightness," this is a concise description of the shape of the power spectrum, indicating whether the spectral content of a signal is dominated by high or low frequencies.
Audio spectral spread	Describes the second moment of the log-frequency power spectrum, indicating whether the power spectrum is centered near the spectral centroid, or spread out over the spectrum. This can help distinguish between pure-tone and noise-like sounds.
Audio spectral flatness	Describes the flatness properties of the spectrum of an audio signal for each of a number of frequency bands. When this vector indicates a high deviation from a flat spectral shape for a given band, it may signal the presence of tonal components.
Fundamental frequency	Describes the fundamental frequency of an audio signal. The representation of this descriptor allows for a confidence measure in recognition of the fact that the various extraction methods, commonly called "pitch-tracking," are not perfectly accurate, and in recognition of the fact that there may be sections of a signal (e.g., noise) for which no fundamental frequency may be extracted. Applies chiefly to periodic or quasi-periodic signals.
Harmonicity	Represents the distinction between sounds with a harmonic spectrum (e.g., musical tones or voiced speech [e.g., vowels]), sounds with an inharmonic spectrum (e.g., metallic or bell-like sounds), and sounds with a non-harmonic spectrum (e.g., noise, unvoiced speech [e.g., fricatives like "f"], or dense mixtures of instruments).
Log attack time	Characterizes the attack of a sound, the time it takes for the signal to rise from silence to the maximum amplitude. This feature signifies the difference between a sudden and a smooth sound.
Temporal centroid	Characterizes the signal envelope, representing where in time the energy of a signal is focused. This descriptor may, for example, distinguish between a decaying piano note and a sustained organ note when the lengths and the attacks of the two notes are identical.
Harmonic spectral centroid	The amplitude-weighted mean of the harmonic peaks of the spectrum. As such, it is very similar to the audio spectral centroid, but specialized for use in distinguishing musical instrument timbres. It is has a high correlation with the perceptual feature of the "sharpness" of a sound.
Harmonic spectral deviation	Indicates the spectral deviation of log-amplitude components from a global spectral envelope.

(continued)

TABLE 3.3 Continued

Harmonic spectral spread	Describes the amplitude-weighted standard deviation of the harmonic peaks of the spectrum, normalized by the instantaneous harmonic spectral centroid.
Harmonic spectral variation	The normalized correlation between the amplitude of the harmonic peaks between two subsequent time slices of the signal.
Audio spectrum basis (ASB)	A tool for indexing audio media using statistical methods. The ASB is a low-dimensional (data-reduced) projection of a high-dimensional spectral space consisting of a series of potentially time-varying and/or statistically independent basis functions derived from a normalized power spectrum. See Casey (2001), Casey et al. (2001), and Casey (2010) for details.
Audio spectrum projection (ASP)	A tool for indexing audio media using statistical methods. Used together with the ASB descriptor, the ASP represents low-dimensional features of a spectrum and is used to segregate different sources (e.g., instruments) in an audio document. See Casey (2001), Casey et al. (2001), and Casey (2010) for details.
Silent segment	Indicates a silent segment; can aid further segmentation of the audio stream or hint not to process a segment.

descriptors) that characterize many attributes of the sound. These can include low-level signal properties (e.g., zero-crossing rate, bandwidth, spectral centroid, signal energy), mel-frequency cepstral coefficients (often used in speech recognition), psychoacoustic features (e.g., roughness, loudness, sharpness), and auditory models. Layered on top of these can be myriad analyzers of pitch, event onset, instrumentation, melody, harmony, rhythmic organization, formal musical structure, genre, mood, artist, and so on (Tzanetakis and Cook 2002).

Conclusion

This chapter examined both the physical and psychoacoustical properties of sound materials: time domain waveforms, the time-frequency domain, sound magnitude, zones of frequency and intensity, timescales of sound, the speed of sound, the size and shape of sounds, the perception of the rate of sound events, and timbre. Understanding the physical nature of sound and its psychophysical impact is central to the practice of electronic music. Acquiring this knowledge is a gradual process, requiring book study (concepts, terminology, physical laws), as well as much listening.

Advances in technology have fostered enormous progress in tools for sound analysis, synthesis, and transformation. As composers become more knowledgeable, their use of these tools should become more sophisticated. Yet with each new tool, layers of psychoacoustic phenomena remain inexplicable. As one researcher observed:

> It is the immense difference between the physical acoustic signal on the one hand and the perceptual-cognitive world on the other hand that has