McAdams, S., Depalle, P. & Clarke, E. (2004). Analyzing musical sound.
In E. Clarke & N. Cook (Eds.), *Empirical Musicology: Aims, Methods, Prospects* (pp. 157-196).
New York, NY: Oxford University Press.

CHAPTER 8

Analyzing Musical Sound

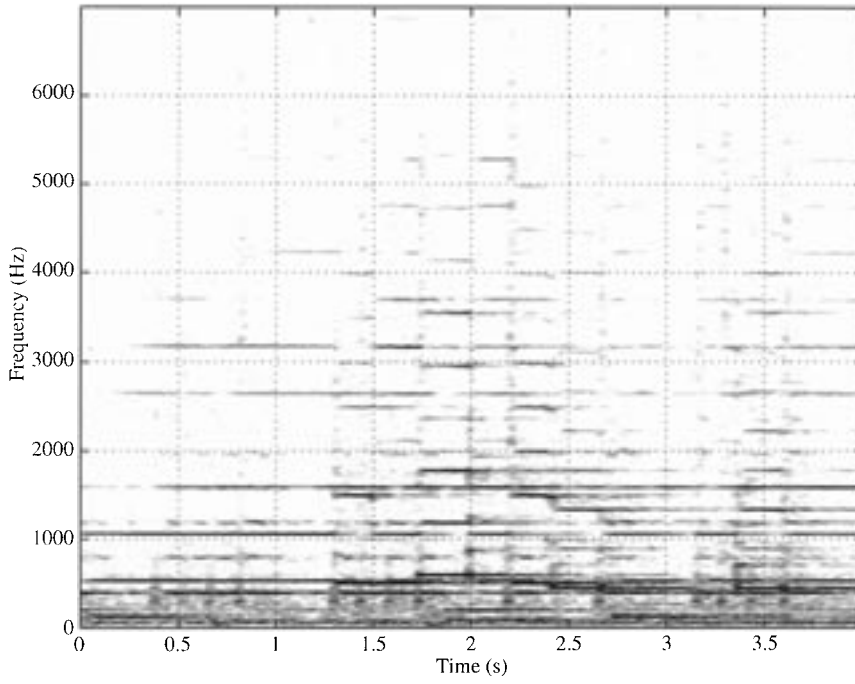
Stephen McAdams, Philippe Depalle and Eric Clarke

Introduction

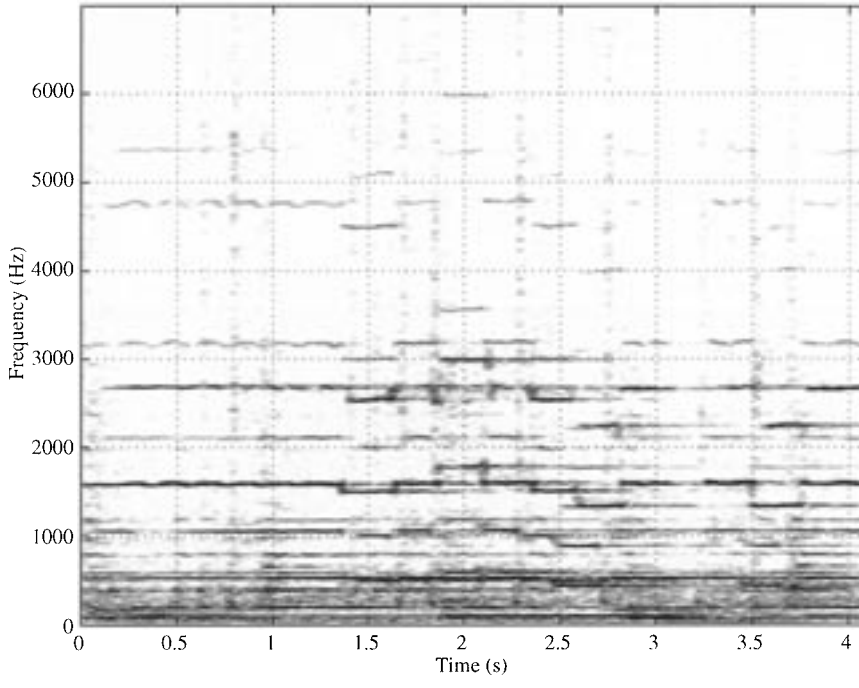
Musicologists have several starting points for their work, of which the two most prominent are text and sound documents (i.e., scores and recordings). One aim of this chapter is to show that there are important properties of sound that cannot be gleaned directly from the score but that may be inferred if the reader can bring to bear knowledge of the acoustic properties of sounds on the one hand, and of the processes by which they are perceptually organized on the other. Another aim is to provide the musicologist interested in the analysis of sound documents (music recorded from oral or improvising traditions, or electroacoustic works) with tools for the systematic analysis of unnotated—and in many cases, unnotatable—musics.

In order to get a sense of what this approach can bring to the study of musical objects, let us consider a few examples. Imagine Ravel's *Boléro*. This piece is structurally rather simple, alternating between two themes in a repetitive AABB form. However, the melodies are played successively by different instruments at the beginning, and by increasing numbers of instruments playing in parallel on different pitches as the piece progresses, finishing with a dramatic, full orchestral version. There is also a progressive crescendo from beginning to end, giving the piece a single, unified trajectory. It is not evident from the score that, if played in a particular way, the parallel instrumental melodies will fuse together into a single, new, composite timbre; and what might be called the "timbral trajectory" is also difficult to characterize from the score. What other representation might be useful in explaining, or simply describing, what happens perceptually?

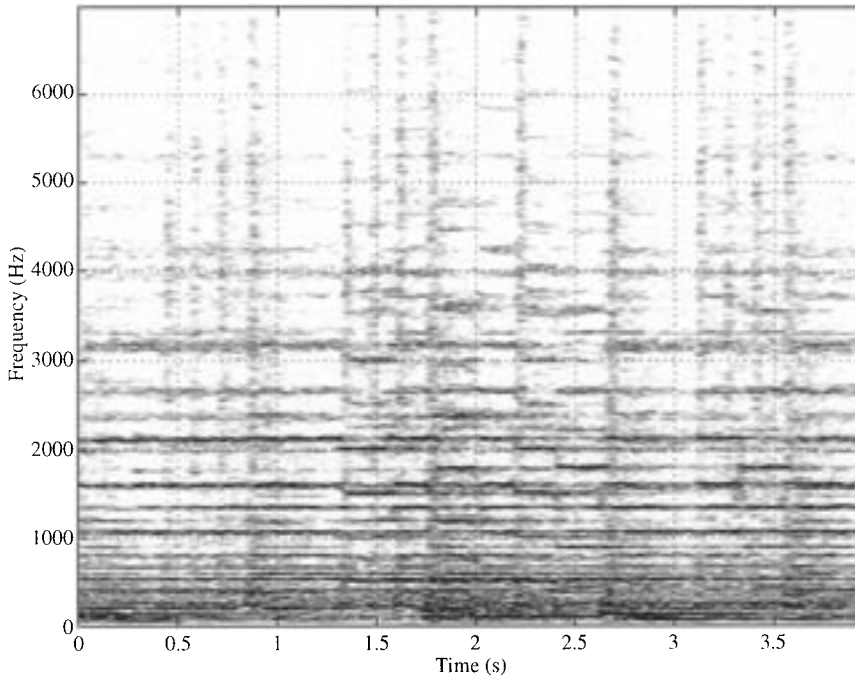
Figure 8.1 shows spectrographic representations (also called *spectrograms*) of the first 11 notes of the A melody from *Boléro*, in three orchestrations from different sections of the piece: (a) section 2, where it is played by one clarinet; (b) section 9, played in parallel intervals by a French horn, two piccolos, and celesta; and (c) section 14, played in parallel by most of the orchestra including the strings. We will come back to more detailed aspects of these representations later, but note that two kinds of structures are immediately visible: a series of horizontal lines that represent the frequencies of the instruments playing the melody, and a series of vertical bars that represent the rhythmic accompaniment. Note too that the density, intensity (represented by the blackness of the lines), and spectral extent (expansion toward the higher frequencies) can be seen to increase from section 2 through section 9 to section 14, reflecting the increasing number, dynamic level, and registral spread of instruments involved.


Frequency (Hz)
6000
5000
4000
3000
2000
1000
0
0
0.5
1
1.5
2
2.5
3
3.5
4
Time (s)

(a)


Frequency (Hz)
6000
5000
4000
3000
2000
1000
0
0
0.5
1
1.5
2
2.5
3
3.5
4
Time (s)

(b)



(c)

Figure 8.1 a. Spectrogram of the first 11 notes of the A melody from *Boléro* by Ravel (section 2). In this example, horizontal lines below 1,000 Hz represent notes, horizontal lines above 1,000 Hz their harmonic components. Percussive sounds appear as vertical bars (0.4 seconds, 0.8 seconds, 1.3 seconds, etc.). b. Spectrogram of the first 11 notes of the A melody from *Boléro* by Ravel (section 9). Notice the presence of instruments with higher pitches (higher frequencies). c. Spectrogram of the first 11 notes of the A melody from *Boléro* by Ravel (section 14). Notice the increase of intensity represented by increased blackness.

Now consider an example of electronic music produced with synthesizers: an excerpt from *Die Roboten*, by the electronic rock group Kraftwerk (Figure 8.2). First, note the relatively clean lines of the spectrographic representation, with little of the fuzziness found in the previous example. This is primarily due to the absence of the noise components and random fluctuations that are characteristic of natural sounds resulting from the complex onsets of notes, breath sounds, rattling snares, and the like. Several features of Figure 8.2 will be used in the following discussion, but it is interesting to note that most of the perceptual qualities of these sounds are not notatable in a score and can only be identified by concentrated listening or by visually examining acoustic analyses such as the spectrogram: for example, the opening sound, which at the beginning has many frequency components (horizontal lines extending from the bottom [low frequencies] to the top [high frequencies]), slowly dies down to a point (at about 1.4 seconds) where only the lower frequencies are

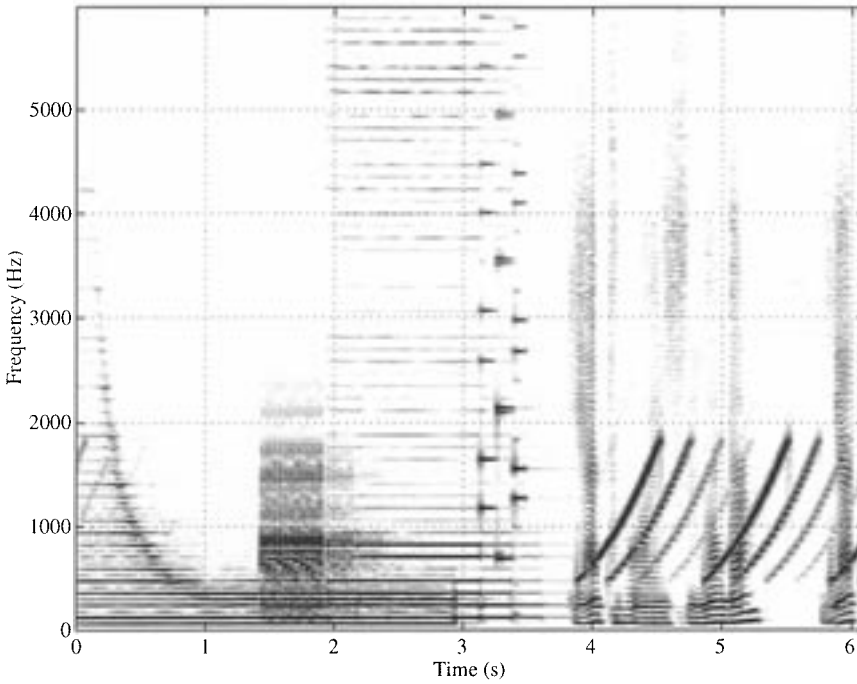


Figure 8.2. Spectrogram of an excerpt of *Die Roboten* by Kraftwerk. Notice the readability of the spectrographic representation that makes explicit continuous timbre, amplitude, and frequency variations.

present. This progressive filtering of the sound has a clear perceptual result that is directly discernible from this representation.

A spectrographic approach is also useful in the case of cultures in which music is transmitted by oral tradition rather than through notation. A telling example is the *Inanga chuchoté* from Burundi, in which the singer whispers (*chuchoter* is French for "to whisper") and accompanies himself on a low-pitched lute (Figure 8.3). This musical genre presents an interesting problem, in that the language of this people is tonal: the same syllable can have a different meaning with a rising or falling pitch contour. The fact that contour conveys meaning places a constraint on song production, since the melodic line must to some extent adhere to the pitch contour that corresponds to the intended meaning. But this is not possible with whispering, which has no specific pitch contour. The spectrogram reveals what is happening: the lute carries the melodic contour, reinforced by slight adjustments in the sound quality of the whispering (it is brighter when the pitch is higher, and duller when it is lower). There is a kind of perceptual fusion of the two sources, due to their temporal synchronization and spectral overlap, so that the pitch of the lute becomes "attached to" the voice.

In light of these examples, we can see how an approach involving acoustical analysis and interpretation based on perceptual principles can be useful in analyzing recorded sound. The spectrogram is just one possible means of representing

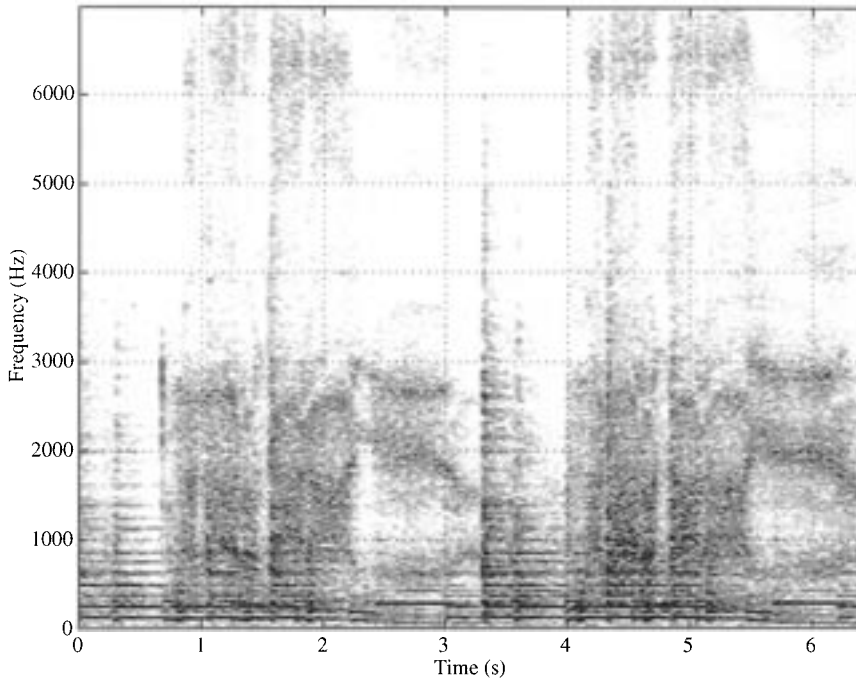


Figure 8.3. Spectrogram of an excerpt of *Inanga Chuchoté* from Burundi. Notice the movements of shaded zones, within the range 500 to 3,000 Hz, that represent timbral variations of the whispered sounds. The onsets of the lute notes and plosive consonants produced by the voice are indicated by the vertical lines in the representation.

sounds: others bring out different aspects of the sound, and it is characteristic of all such representations that different features can be brought out according to the settings that are used in creating them. (This is similar to the way in which the correct setting of a camera depends on what one wants to bring out in the photograph.) The goal of this chapter is therefore to introduce some of these ways of representing sound, and to provide a relatively nontechnical account of how such representations work and how they are to be interpreted. The chapter is organized in three main sections, the first dealing with basic characteristics and representations of sound, the second with acoustical analysis, and the third with perceptual analysis; the two analytical sections conclude with brief case studies taken from the literature.

Basic Characteristics and Representations of Sound

Sound is a wave that propagates between a source and a receiver through a medium. (The source can be an instrument, a whole orchestra or loudspeakers; the receiver can be the ears of a listener or a microphone; the medium is usually air.) It can also be considered as a signal that conveys information from an instrument or loudspeaker to the ears of a listener, who decodes the information by hearing the time

evolution of the acoustic wave, and recognizes instruments, notes played, a piece of music, a specific performer, a conductor, and so on. Using machines to analyze sound signals involves structuring the information in a way that is similar to what the ear does; such analyses usually provide symbolic information or—as in this chapter—graphical representations.

The analysis of a sound, then, starts with a microphone that captures variations in air pressure (produced by a flute, for example) and transduces them into an electrical signal. This signal can be represented as a mathematical function of *time*, and is therefore called a *temporal representation* of the sound. A graphical display of such a temporal representation is an intuitive way to begin to analyze it, and in the case of a solo flute note we might get the temporal representation in Figure 8.4, with time on the horizontal axis and the amplitude of the signal on the vertical axis. The figure reveals the way the sound starts (the attack), the sustained part with a slight oscillation of the level, and the release of the flute sound at the end. However, it fails to help us in determining the nature of the instrument and the note played, and in the case of more complex sounds—say an excerpt of an orchestral composition—very little is likely to emerge beyond a rough impression of dynamic level. That means that we need to find alternative representations based on mathematical transformations of the simple temporal representation. The most important of these transformations involve the idea of periodicity, since this is intimately linked to the perception of pitch—a primary feature of most musical sounds.

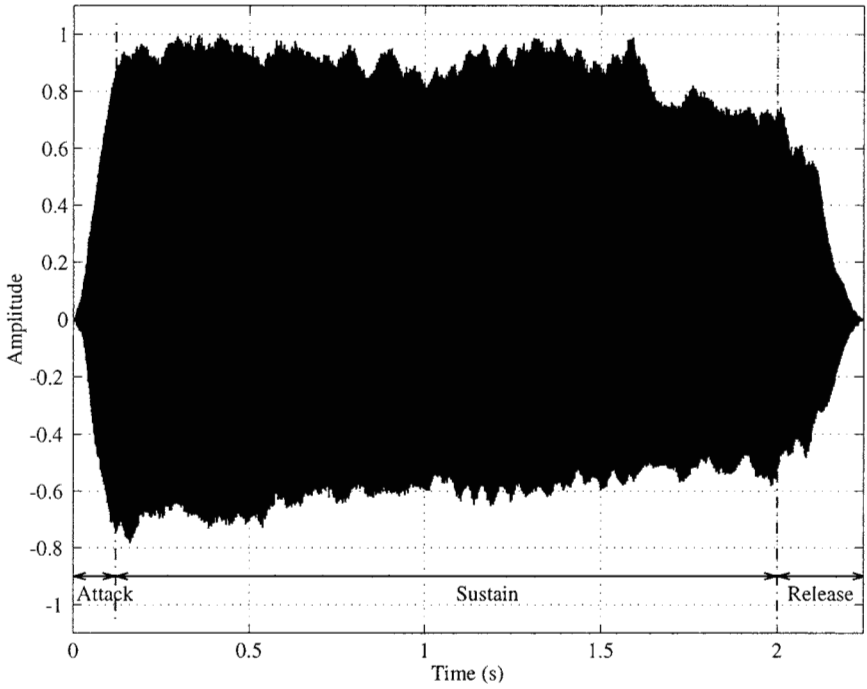


Figure 8.4. Temporal representation of a simple flute note.

Figure 8.5 is a simple temporal representation, like Figure 8.4, but the temporal profile is shown at a much higher level of magnification. We can now begin to see the repeated patterns that define periodic sounds: the term *period* refers to the duration of the cycle, and the number of times the cycle repeats itself per second is called the *frequency* (or fundamental frequency). Thus, the frequency is the reciprocal of the period, and its unit is the Hertz (Hz). It can be seen from Figure 5 that six periods are a little shorter than seven divisions of the time axis (which are hundredths of a second), so that the fundamental frequency is 87.2 Hz—which is the F at the bottom of the bass clef.

While frequency determines the pitch of the clarinet sound in Figure 8.5, the particular shape of the wave is related to factors that determine its timbral properties. How might it be possible to classify or model the range of different shapes that sound waves can take? A *spectral representation* (or spectrum) attempts to model sounds through the superimposition of any number of waves of different frequencies, with each individual wave taking the form of a "sinusoid": this is a function that endlessly oscillates at a given frequency, and which can be approximated by the sustained part of the sound of a struck tuning-fork. Figure 8.6.a shows a few sinusoidal oscillations at a frequency of 440 Hz (the standard tuning fork A). Now if, instead of showing amplitude against time as in Figure 8.6a, we were to show it against frequency, we would see a single vertical line corresponding to 440 Hz: this is shown in Figure 8.6b, a much more condensed and exhaustive representation of the signal by

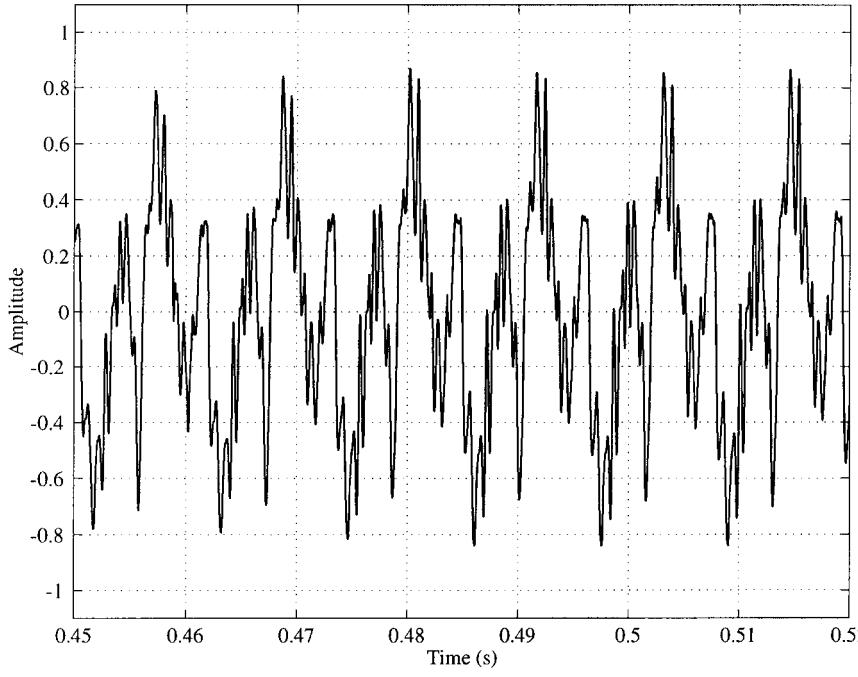
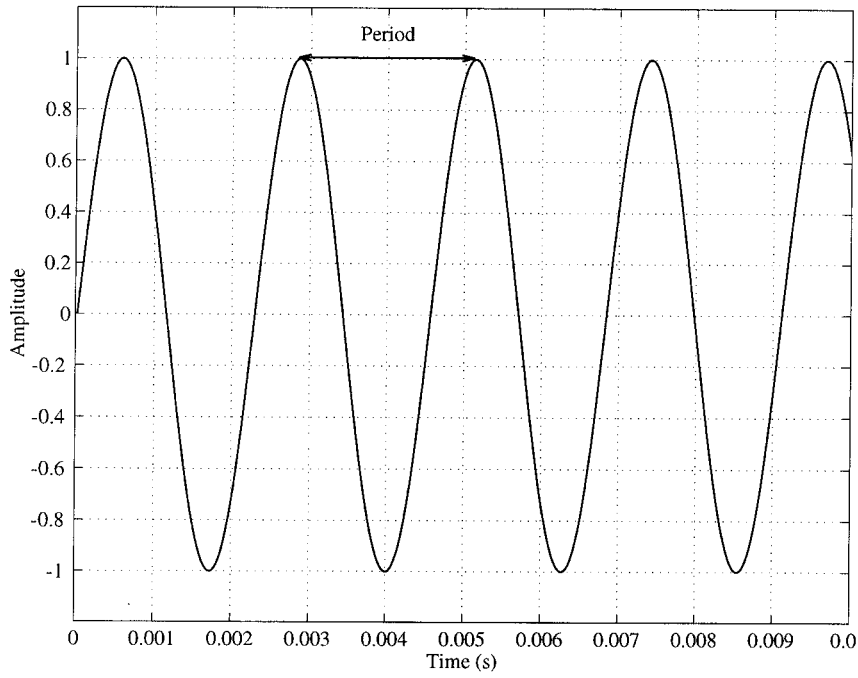
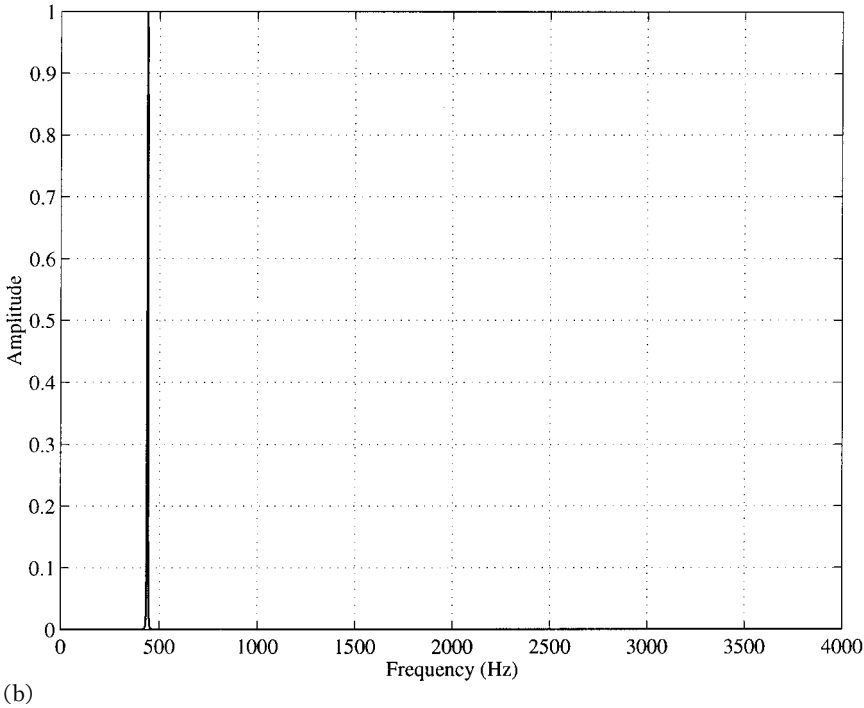


Figure 8.5. Simple periodic sound: six periods of a bass clarinet sound.



(a)



(b)

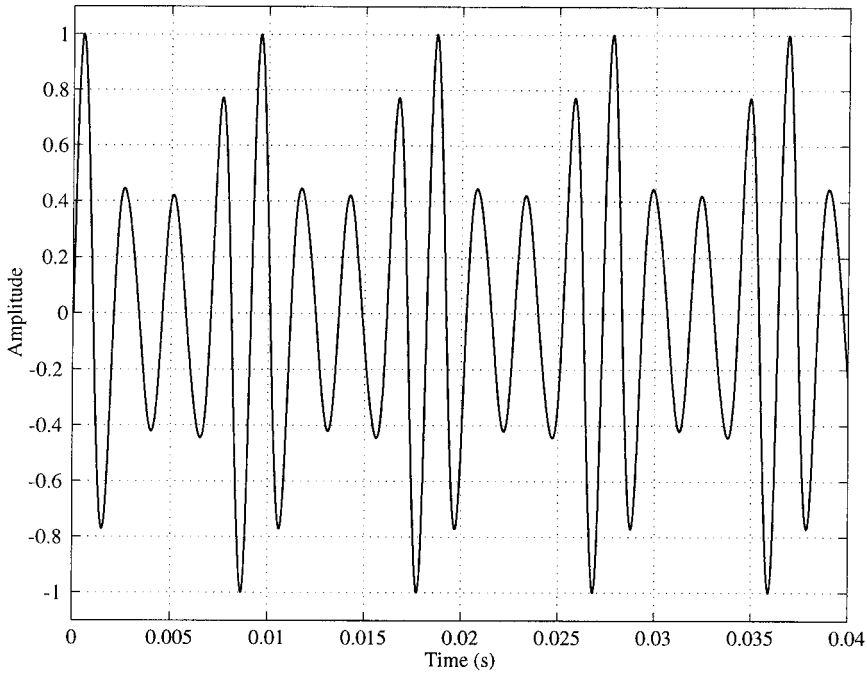
Figure 8.6 a. Sinusoidal sound (frequency = 440 Hz, amplitude = 1.0): temporal representation with a single period indicated. b. Sinusoidal sound (frequency = 440 Hz, amplitude = 1.0): spectral representation.

its *spectral content*, than by its temporal representation. The same representation can obviously show any number of different spectral components—different sinusoids—at different levels of amplitude: Figure 8.7a shows a temporal representation of a mix of three sinusoids at different amplitudes, and Figure 8.7b the resulting spectrum. While the number, frequency, or amplitude of the individual components are difficult to estimate from Figure 8.7a, are all immediately apparent from Figure 8.7b.

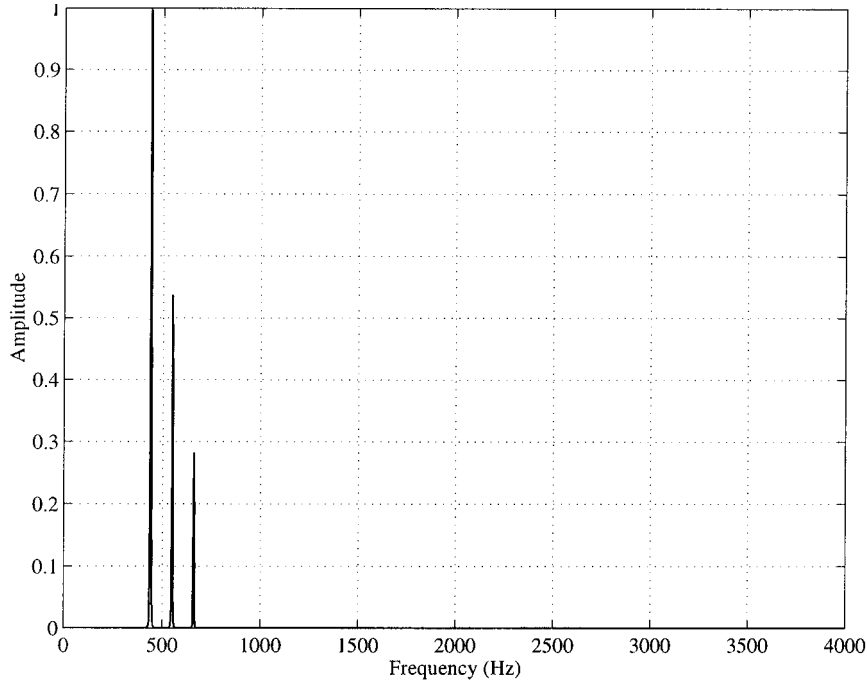
As there are only three sinusoids in Figures 8.7a and 8.7b, all in the central auditory range and at relatively similar dynamic levels, each will be heard as a separate pitch: in fact, since their frequencies are 440, 550, and 660 Hz, the percept will be an A major chord. A different set of sinusoids, by contrast, might produce the effect of a single pitch with a distinctive timbre. Because it represents only the acoustical qualities of the signal, not its perceptual correlates, the difference cannot be directly seen in a spectral representation.

The principle of decomposing a complex waveform into separate elements can be taken a good deal further than this. The mathematician Joseph Fourier demonstrated that a periodic signal, whatever the shape of its waveform, can always be analyzed into a set of harmonically related sinusoids ("harmonically," meaning that the frequencies of these sinusoids are multiples of the fundamental frequency—as, for instance, 440, 880, 1320 Hz, and so on are integer multiples of 440 Hz). The collection of these harmonics constitutes the *Fourier series* of the signal, and is an important property since it roughly corresponds to the way sounds are analyzed by the auditory system. In practice, then, analyzing a periodic or harmonic signal consists of determining the fundamental frequency and the amplitude of each harmonic component. Figure 8.8 compares waveforms and Fourier analyses of two simple signals often used in commercial synthesizers. (Note that in the Fourier representations in Figures 8.a2 and 8.b2 the frequency axis shows values as multiples of 10^4 Hz: thus "1" represents 10,000 Hz or 10 kHz.) The equidistant vertical lines represent the different harmonics of the fundamental frequency, which is again 440 Hz. Comparison of the sawtooth waveform (Figure 8.8a2) with a square wave (Figure 8.8b2) shows that the latter lacks even-numbered harmonics. Figure 8.8c, by contrast, shows the spectrum of the clarinet sound from Figure 8.5, which—while it exhibits vertical harmonic peaks—does not look as clean as the synthesized examples.

In addition to harmonic signals, there are *inharmonic* signals such as the sounds of bells (Figure 8.9a) or tympani (Figure 8.9b): these are not periodic, but can still be described as a series of superimposed sinusoids—although the sinusoids are no longer harmonically related. (They are therefore called *partials* rather than harmonics, and can take any frequency value.) Inharmonic sounds do not have a precise overall pitch, though they may have a pitch that corresponds to the dominant partial, or even several pitches. The bell sound in Figure 8.9a includes a series of near-harmonic components (a "fundamental frequency" at 103 Hz (G♯), a second harmonic at 206 Hz, a ninth one at 927 Hz, a thirteenth one at 1,339 Hz, and so on) but also other inharmonic components; these components give the sound a chord-like quality. The tympani spectrum in Figure 8.9b is similarly inharmonic, with some partials conforming to a nearly harmonic relationship with a fundamental frequency at 66 Hz (C). Indeed there are many sounds that we think of as clearly pitched but which are slightly inharmonic: Figure 8.9c shows the spectrum of a piano note



(a)



(b)

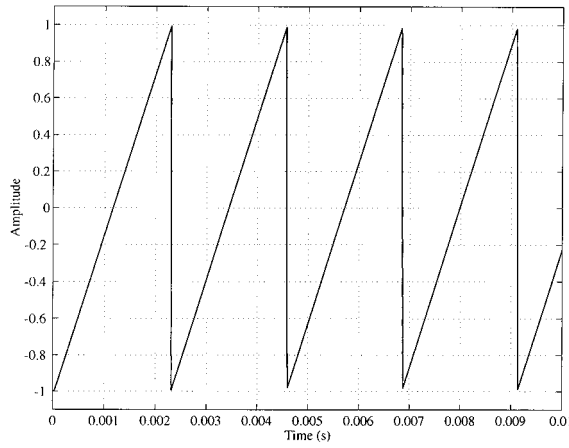
Figure 8.7 a. Mix of three sinusoids (frequencies are 440, 550, and 660 Hz; amplitudes are 1.0, 0.5, and 0.25, respectively): temporal representation. b. Mix of three sinusoids (frequencies are 440, 550, and 660 Hz; amplitudes are 1.0, 0.5, and 0.25, respectively): spectral representation.

whose partial components are near-harmonics of the fundamental frequency 831 Hz (g♯″), and one can see that the inharmonicity—represented by the difference between the positions of the actual partials and the theoretical positions shown by dashed lines—increases with frequency. (The cluster of low-frequency components, below the fundamental up to around 2,500 Hz, is produced not by the vibration of the piano strings, but by the soundboard.) Here the deviation from harmonicity at frequencies below about 5 kHz is sufficiently small that the listener perceives a precise pitch.

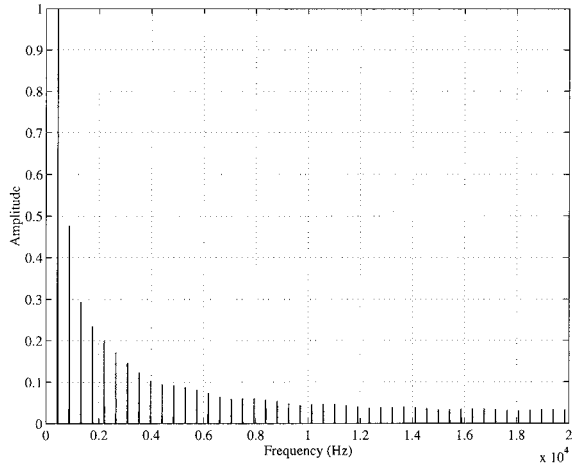
For the sound characterization to be complete, the category of *noisy* sounds has to be considered. By definition, these have a random temporal representation (such as the whispered sound between 2.2 and 3.2 seconds in Figure 8.3). Few instrumental sounds are completely noisy, but most of them include a certain amount of noise (the player's breath in the case of wind instruments, impact noise for percussion, and so on), and as we shall later see, this noisy part is nearly always very important for the perception of timbre. The sounds of wind and surf, replicated by electronic noise generators, are by contrast completely noisy. Such sounds are composed of a random mix of all possible frequencies, and their spectral representation is the statistical average of the spectral components. For example, white noise has a random waveform (Figure 8.10a). Although it has a spectrum that is theoretically flat, with all frequencies appearing at the same level, in practice the spectrum revealed by Fourier analysis is far from being perfectly flat (Figure 8.10b), and exhibits variations that are due to the lack of averaging of the random fluctuations in level of the different frequencies. These fluctuations can be reduced by taking the average of several spectra computed on successive time-limited samples of the noisy sound.

Now that the basic characteristics of sounds have been described, it is important to mention one major aspect of "natural" sound signals: their characteristics (frequency, amplitude, waveform, inharmonicity, or noise content) always vary over time. Sounds with perfectly stable characteristics (such as the sinusoids in Figure 8.6, or the stable low-frequency sound in *Die Roboten*, between 2 and 3 seconds in Figure 8.2) sound "unnatural" or "synthetic.." These time-varying characteristics are called *modulations* and can take various forms. Amplitude modulations range from uncontrolled random fluctuations, such as in the flute note in Figure 8.4, to the tremolo on the low sustained note in *Die Roboten* (between 1.5 and 2 seconds in Figure 8.2); frequency modulations can include vibrato (the undulating horizontal lines produced by the piccolo during the first 1.5 second of Figure 8.1b) or pitch glides (the upward sweeping "chirp" sound in *Die Roboten* between 3.8 and 4.5 seconds, Figure 8.2). Apart from their impact on the character of individual sounds, the presence and synchronization of modulations are very important for the perception of fused sounds, and will be discussed in the final section of this chapter.

This means that there is a significant element of approximation or idealization in all but the simplest representations of sound which we have been discussing. Most obviously, spectral representations (Fourier or otherwise) relate amplitude and frequency: they represent values averaged over a discrete temporal "window" or "frame," and therefore tell us nothing about changes in the sound during that period of time. (One could represent the changes by showing a series of spectral representations one after another, in the manner of an animation, but even then each frame



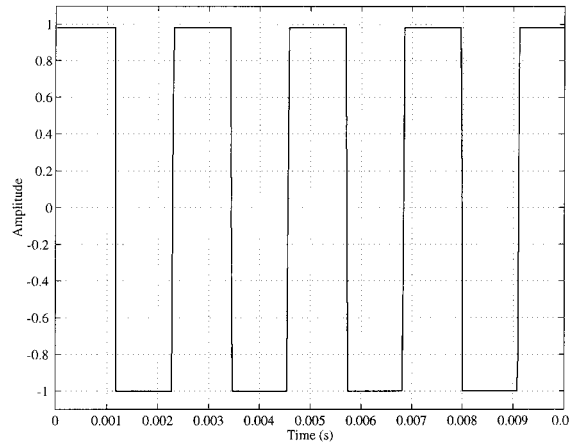
(a.1)



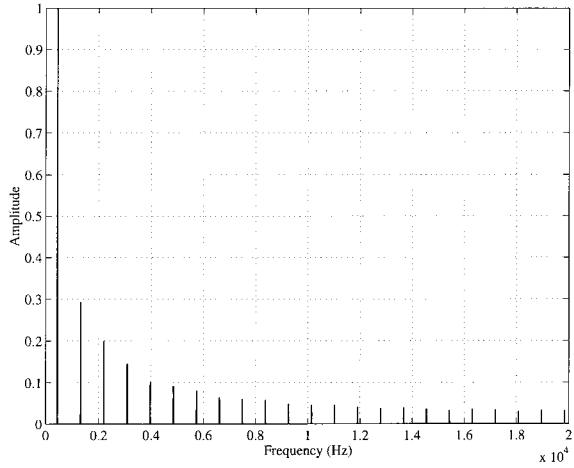
(a.2)

Figure 8.8 a.1. Sawtooth waveform. a.2. Spectrum (Fourier series) of a sawtooth waveform. b.1. Square waveform. b.2. Spectrum (Fourier series) of a square waveform. c. Spectrum (Fourier series) of a clarinet waveform.

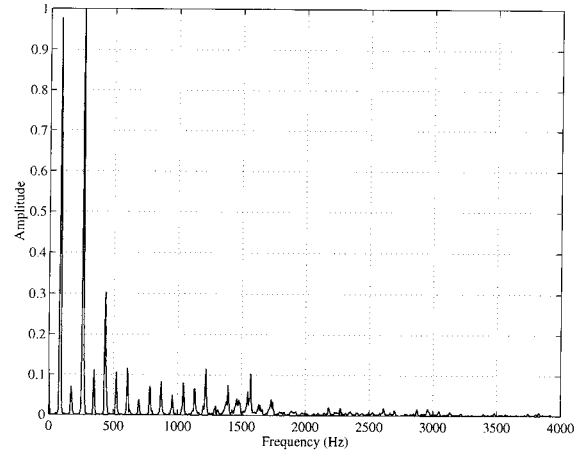
would represent the average over a given time span.) There is also a similar point in relation to periodicity. While from a mathematical point of view a periodic signal reproduces exactly the same cycle indefinitely, in the real world any sound has a beginning and an end: for this reason alone, musical sounds are not mathematically periodic. Moreover, they nearly always show slight differences from one period to the next (as can be seen from the waveform of the clarinet sound in Figure 8.5). In practice, a sound is perceived as having a definite pitch as soon as there is a sufficient degree of periodicity, not when it is mathematically periodic, and we therefore need an analytical tool that will identify degrees of periodicity on a local basis, show-


Amplitude
Time (s)

(b.1)


Amplitude
Frequency (Hz)
x 10^4

(b.2)


Amplitude
Frequency (Hz)

(c)

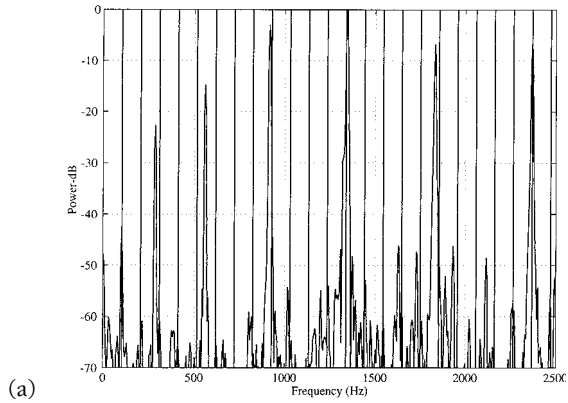
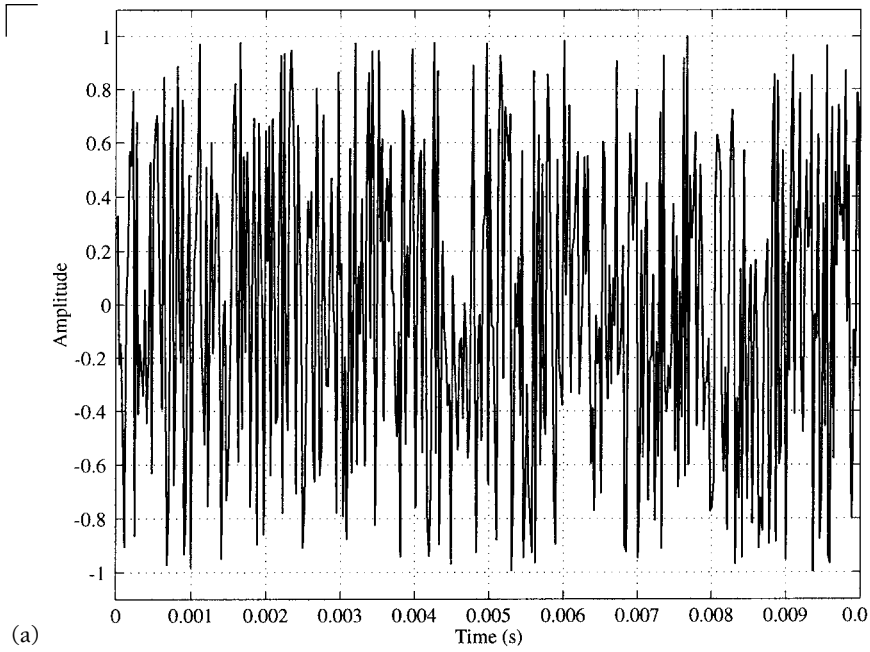
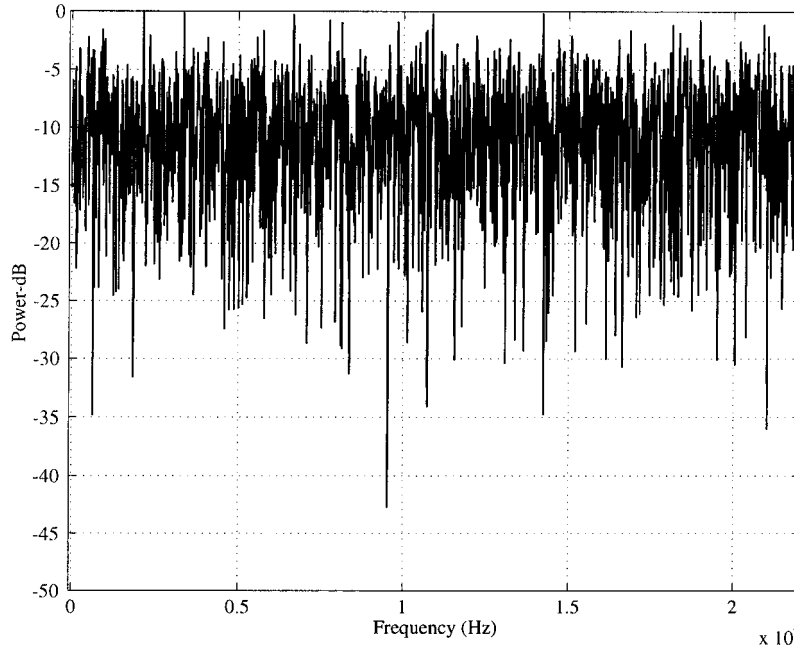


Figure 8.9 a. Inharmonicity: spectrum of a bell sound. The series of vertical thin lines represent theoretical locations of a harmonic series of fundamental frequency 103 Hz. b. Inharmonicity: spectrum of a tympani sound. The series of vertical thin lines represent theoretical locations of a harmonic series of fundamental frequency 66 Hz. c. Inharmonicity: spectrum of a piano sound. The series of vertical dashed lines represents theoretical locations of a harmonic series of a fundamental frequency of 831 Hz.



(a)



(b)

Figure 8.10 a. Example of a noisy sound: white noise (temporal representation). b. Example of a noisy sound: white noise (spectral representation). Notice that the spectrum of a white noise is random and more or less flat on average.

ing how the spectrum changes over time. This is exactly what the spectrogram does, as illustrated in the introduction of this chapter—and this, again, is something to which we will return.

Acoustical Analysis of Sounds

A waveform display allows a user to analyze a sound quite intuitively by simply looking at a temporal representation of a sound. The representation can be created at different time scales, from the "microscopic" or short-term scale, to the "macroscopic" or long-term scale. A "microscopic" time scale, which in practice is usually on the order of a few periods, preserves the waveform shape, and allows a qualitative evaluation of the presence or absence of noise and its level, as well as the presence of strong, high-order harmonic components. Figures 8.5a, 8.6a, 8.7a, 8.8a.1, and 8.8b.1 are temporal representations of signals at a "microscopic" time scale: this also allows one to evaluate precisely the synchronization between acoustic events. By contrast, a "macroscopic" time scale makes visible long-term tendencies such as the global evolution of the sound level, amplitude changes in the course of a melodic line, or the way notes follow one another; an example is Figure 8.4 in which the temporal envelope of a flute sound can be discerned.

Sound signals are usually displayed on computer screens using sound editor programs; some examples are *ProTools*, *AudioSculpt, Peak, SpectroGramViewer*, and *Audacity*. However there is a problem when the number of samples to be represented on the screen becomes larger than the number of available pixels: as sounds are usually sampled at 44,100 Hz (the compact disc standard), and as the highest number of pixels on each line of current screens is usually less than 2000, a complete display is possible only for durations shorter than 50 milliseconds (msec). Beyond that, several sample values have to be averaged into one pixel value (in other words, the signal has to be smoothed), and this prevents precise investigation of long-term sound characteristics, limiting the direct use of "macroscopic" sound signal displays to the analysis of global temporal evolutions. Even in the case of these global evolutions, though, there is a perceptually more meaningful means of analysis: *temporal envelope estimation*.

As an example, Figure 8.11 displays the dynamic evolution of an excerpt from Ligeti's orchestral piece *Atmosphères*, starting with a slow crescendo/decrescendo over the first 20 seconds, followed by a fast crescendo to a median level, and a slow crescendo for the next 30 seconds, and so on. Such a temporal envelope, which is calculated automatically by most sound editing programs, is again based on a series of temporal windows or frames, the duration of which is normally between 10 and 200 msec, with the window sliding along the time axis to provide an estimate of the temporal envelope at set intervals. The only setting that you need to adjust in order to get a temporal envelope is the size of the window. The chosen size will inevitably be a compromise: it has to be large enough to smooth over the fine-grained oscillations of the signal, but at the same time small enough to preserve the shape of the attack, or of the other transient parts of the sound. The problem can be seen by comparing

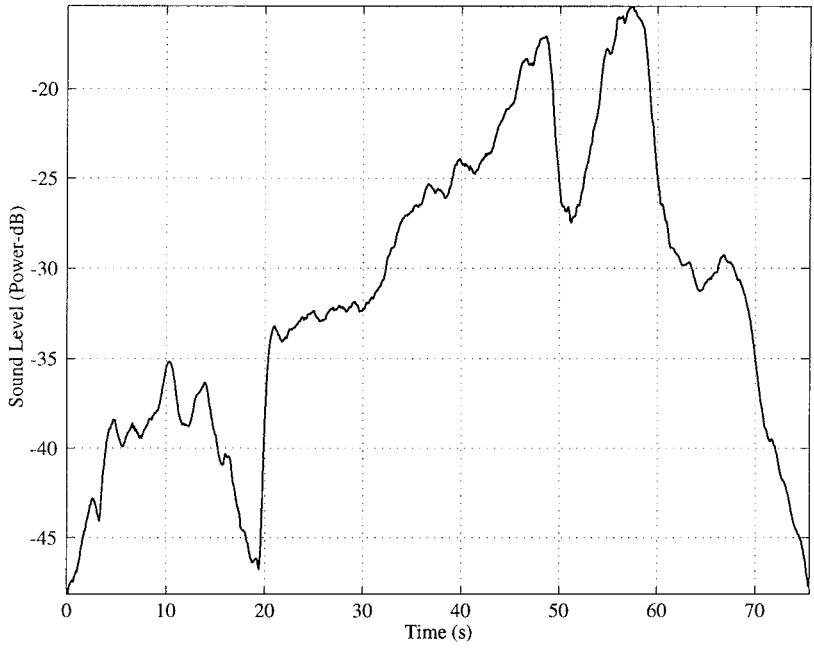


Figure 8.11. Temporal envelope estimation for an excerpt from Ligeti's *Atmosphères*.

three different temporal envelope estimations for the temporal representation shown in Figure 8.12a, the spectrogram of which was shown in Figure 8.2. In Figure 8.12b, too small a window (5.8 msec) has been chosen, and the profile exhibits spurious oscillations that are not perceived as changes in level; Figure 8.12c represents an appropriate value (23.2 msec), whereas in Figure 8.12d the window is too large (92.8 msec) and fast changes of level (particularly between 1.5 and 2 seconds) are smoothed out. The rule of thumb is to use more than the duration of the largest period contained in the original sound; in this way, musicologists who want to use a temporal envelope analysis need to set the window size in accordance with the particular music they are studying, as well as the particular aspects of the sound in which they are interested.

For many musicological purposes a spectral representation will be the best choice for sound signal analysis. There are, however, some practical problems associated with it: one is how to estimate the spectrum of nonperiodic sounds, while another is the relationship between the sampling rate, the sampling window, and the frequency of the sounds being studied. To be completely known, a nonperiodic signal has to be observed over its entire duration, unlike a periodic signal (where observation over one period is sufficient). This means that a Fourier series representation of a nonperiodic sound is not possible: the appropriate representation is instead a *Fourier transform*. A simple way to understand the extension of the Fourier series to the Fourier transform is to think of a nonperiodic signal as a periodic signal whose

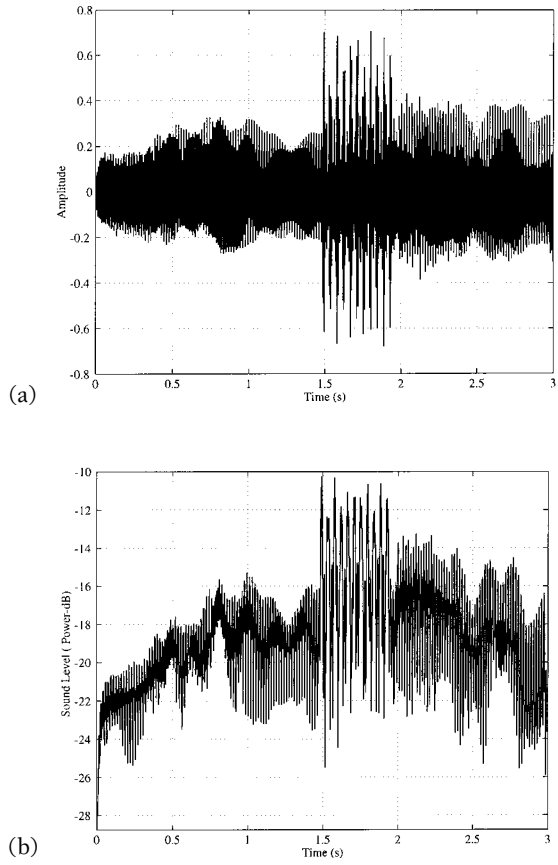
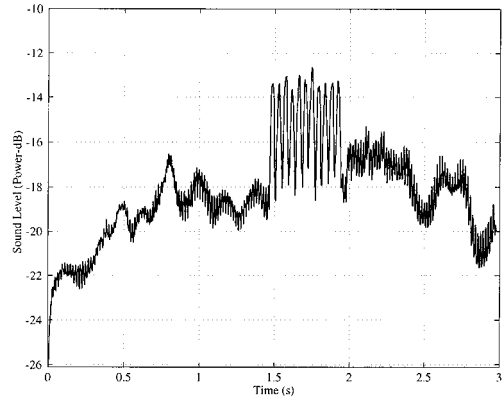


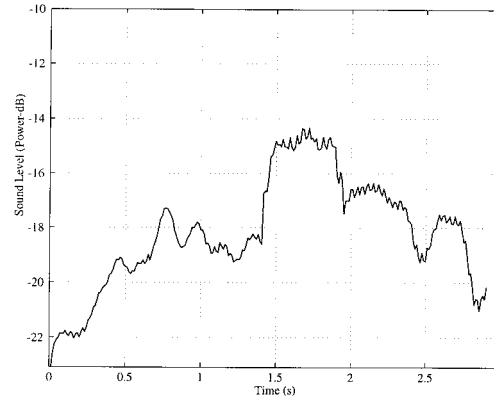
Figure 8.12 a. Temporal representation of the first three seconds of *Die Roboten* by Kraftwerk. b. Temporal envelope estimation of the sound signal displayed in Figure 8.12a. using a window size of 5.8 msec. c. Temporal envelope estimation of the sound signal displayed in (a). using a window size of 23.2 msec. d. Temporal envelope estimation of the sound signal displayed in Figure 12.a. using a window size of 92.8 msec.

period is infinite. An example of a nonperiodic signal and its Fourier transform is given in Figure 8.13. The sound signal is a damped sinusoid, that is a sinusoid whose amplitude decreases over time. The treatment of this kind of signal is significant, since many percussion instruments (including bells, timpani, pianos, and xylophones) produce a superposition of damped sinusoids.

The Fourier transform (Figure 8.13b) exhibits a maximum close to the oscillating frequency, but there is some power at other frequencies, particularly those close to the central frequency. The sharpness of the maximum, sometimes called a *for-*



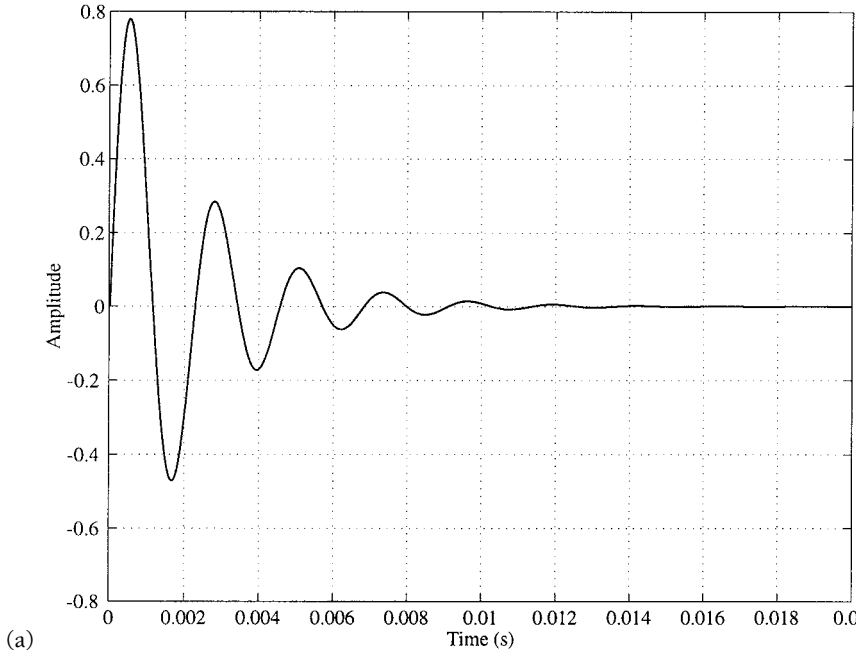
(c)



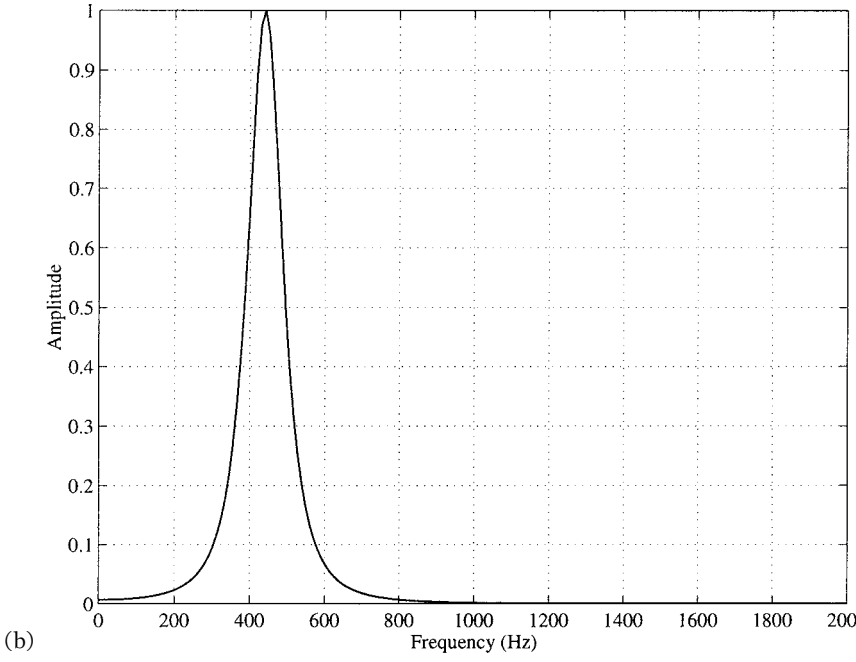
(d)

mant, varies in inverse proportion to the degree of damping. The damping makes the oscillating sinusoid nonperiodic and spreads its power out to other frequencies; the spectrum is therefore no longer a peak (as it is for a pure sinusoid) but a smoothed curve, whose *bandwidth* (the width of the curve) increases with the damping value.

As in the case of temporal envelope estimation, choosing the right duration as the basis for calculating the Fourier transform is a compromise: the duration needs to be small enough to maintain sufficient resolution between closely adjacent sinusoids, and yet not so large as to average out all of the temporal evolution of the sound's spectral characteristics. A good compromise is usually a duration of four to five times the period of the lowest frequency difference between the sinusoidal components of the sound. Figure 8.14 revisits the spectral analysis of the A major chord, made up of three sinusoidal sounds, that was shown in Figure 8.7: since the frequencies are 440, 550, and 660 Hz, the lowest frequency difference is 110 Hz, which (at a 44,100 Hz sampling rate) corresponds to approximately 400 samples. Figure 8.14 shows that a window size of 2,048 samples i.e., approximately 5 times 400 samples) successfully separates out the three components.



(a)



(b)

Figure 8.13 a. Nonperiodic signal: a damped sinusoid (frequency = 440 Hz). b. Nonperiodic signal: spectrum of a damped sinusoid (frequency = 440 Hz). Notice the continuous aspect of the spectrum.

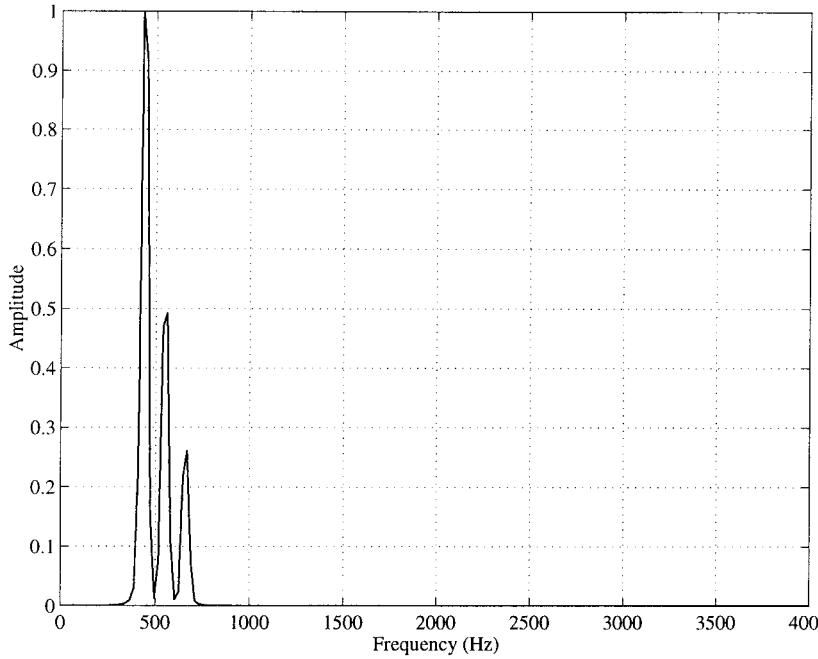
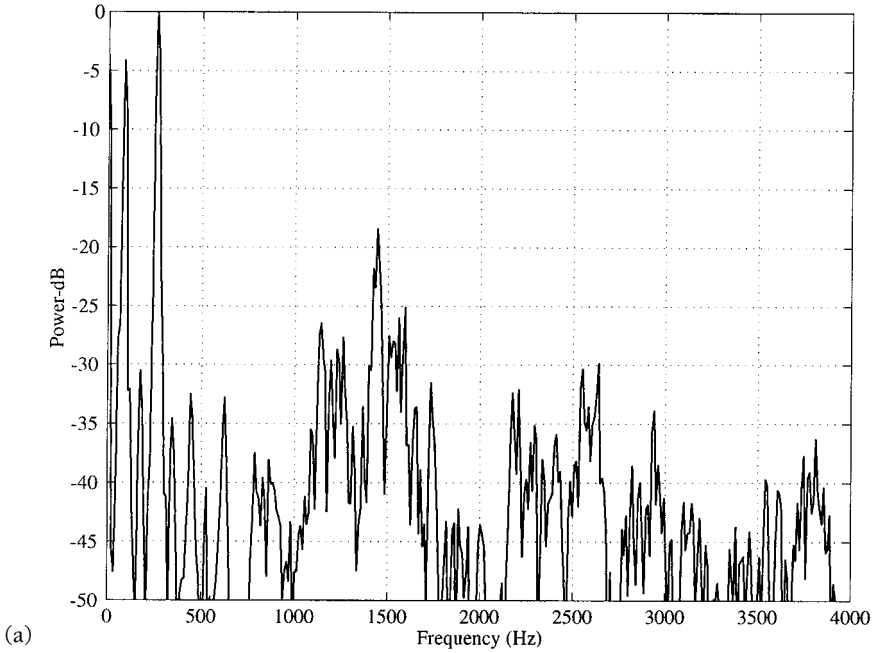


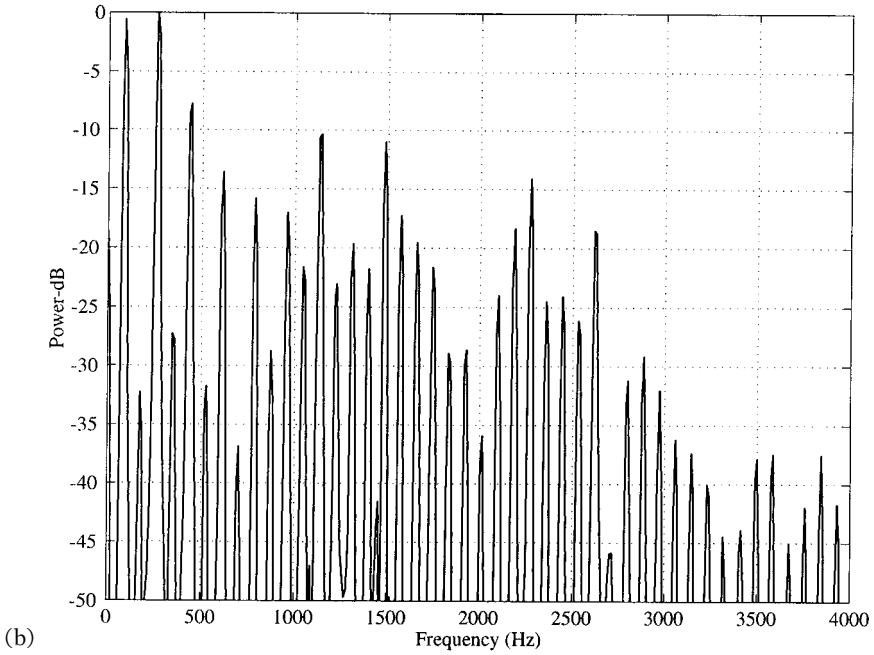
Figure 8.14. Time-limited Fourier analysis: spectrum of the major chord from Figure 8.7 estimated with a window size of 2,048 samples.

We are now ready to come full circle, establishing the link between all the concepts developed in this section and the spectrographic representation at the very beginning of the chapter. A spectrum, resulting from a Fourier transform performed over a finite duration, provides useful information about a sound signal only when the sound is known to be stable in time. However, as already mentioned, the characteristics of natural sounds always vary in time. A spectrum taken from a window located at the beginning of a sound (Figure 8.15a) is usually different from a spectrum taken from a window located in the middle of the sound (Figure 8.15b). In order to describe the temporal variations of the spectral properties of a sound, a simple idea is to compute a series of evenly spaced *local spectra*. This is achieved by computing a Fourier transform for each of a series of sliding windows taken from the signal. The time-step increment of the sliding window is usually a proportion of the window size, and a time-shift of an eighth of the window size or less ensures perfect tracking of the temporal evolution. Each Fourier transform represents an estimate of the spectral content of the signal at the time on which the window is centered, and a simple and efficient way to display this series of spectra is to create a time/frequency representation, with the darkness of the trace representing the amplitude of each frequency (Figure 8.16). This representation, which we have already encountered, is called a *spectrogram* (or sometimes a *sonogram*).

In summary, we have seen two different kinds of two-dimensional representation of sound, and a three dimensional representation. The two-dimensional repre-



(a)



(b)

Figure 8.15 a. Spectrum of a bass clarinet tone during the attack (window centered on 0.12 second, window size of 4,096 samples). b. Spectrum of a bass clarinet tone during the sustained part (window centered on 1 second, window size of 4,096 samples).

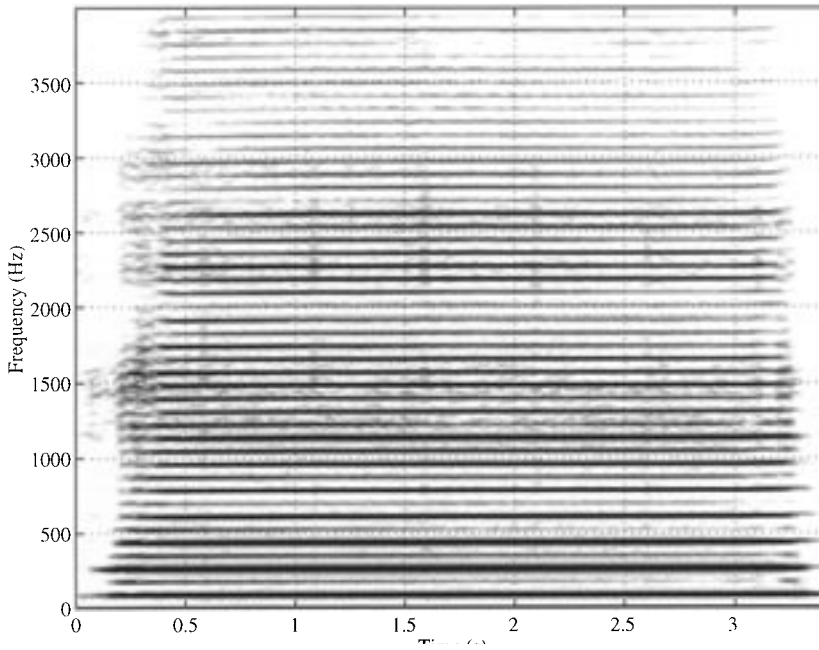


Figure 8.16. Spectrogram of a bass clarinet sound (window size of 4,096 samples, time-step increment of 512 samples).

sentations are temporal representations (amplitude against time), and spectral representations (amplitude against frequency); the three-dimensional representation—the spectrogram—shows frequency against time on the two axes, and intensity against time in the blackness of the trace.

Analytical Applications

The most sustained example of applying acoustical principles to musical analysis, and one which makes considerable use of spectrograms, is Robert Cogan's book *New Images of Musical Sound* (Cogan 1984). The central part of the book consists of a discussion and analysis of 17 "spectrum photos" (the equivalent of spectrograms) of music from a wide variety of traditions, including Gregorian chant, jazz, a movement of a Beethoven piano sonata, electroacoustic music, and Tibetan tantric chant; the examples range from half a minute to over 11 minutes in duration, with the majority around two to four minutes long. The spectrum photos show duration on the horizontal axis and frequency on the vertical axis, with intensity represented as the brightness of the trace (in other words, while intensity is represented by blackness against a white background in the spectrograms presented in this chapter, it is shown as whiteness against a black background in Cogan's photos). The spectrum photos were created using analog signal analysis equipment, with a camera used to take photographs of the cathode ray tube for successive sections of music; these were

then literally pasted together to create the resulting composite photos that appear in the book. Digital technology has made it possible to produce more flexible and finely graded representations of this kind far more easily, as the figures in this chapter demonstrate.

Cogan uses the spectrum photos to analyze and demonstrate a variety of different features of the music, the diversity of which is intended to show how many different features can be addressed through these means. His discussion of Billie Holiday's recording of "Strange Fruit," for example, focuses on the ways in which Holiday uses continuous pitch changes and the timbral effects of different vowel sounds to articulate semantic relationships in the text, and to expose its savage ironies: "Note bending is a motif that recurs with ever-increasing intensity. 'The gallant South' is immediately echoed with growing irony at 'sweet and fresh,' again bending to the voice's lowest depth. Then a string of increasingly bent phrases. . . . leads with gathering intensity to the explicit recall of the first stanza . . . " (Cogan 1984: 35)

By comparison, the discussion of Elliott Carter's *Etude III for Wind Quartet* focuses on the timbral changes that result both from instrumental entries and exits, and from continuous dynamic changes, rhythmic augmentation, and diminution. This analysis is unusual in using spectral representations (rather than temporal representations) to give more detail about the relative balance of different spectral components in the sound than is possible in the spectrograms used elsewhere in the book: because time is eliminated from the representation, Cogan presents a sequence of 18 spectral "snapshots" to demonstrate how the timbre evolves over the piece. Without getting involved in the detail of Cogan's analysis, a sense of what he claims such an analysis can achieve can be gathered from the following (Cogan 1984: 71–72):

> Spectrum analysis provides a tool whereby the important similarity of these passages—the initial one characterized by instrumental change and rhythmic diminution, the climactic one by dynamic change and rhythmic augmentation—can be discovered and shown. Remove the spectral features and the most critical formal links of the entire etude . . . disappear. Without spectral understanding, the link between the successive transformations—instrumental, rhythmic, and dynamic . . . would evaporate. . . . We noted at the beginning of this commentary that, in the light of earlier analytic methods, this etude could emerge only as incomprehensible, static, or both. It now, however, reveals itself to be a set of succinct, precise spectral formations whose roles and relationships, whether of identity or opposition, are clear at every instant.

Two further examples of the way in which spectral analysis can be used for musicological purposes are provided by Peter Johnson's (1999) discussion of two performances of the aria "Erbarme Dich" from Bach's *St. Matthew Passion*, and David Brackett's (2000) analysis of a track by Elvis Costello. The subject of Johnson's paper is a wide-ranging discussion of the relationship between performance and listening, with Bach's aria as a focal example viewed from aesthetic and more concretely analytical perspectives. Central to Johnson's argument is his insistence that the impact of the *sound* of a performance on listeners' experience is consistently underestimated

by commentators and analysts, and an important part of the paper is thus devoted to a detailed consideration of the acoustical characteristics of two performances of the aria. Johnson focuses primarily on differences in the frequency domain, highlighting distinctions in the use of vibrato, timbre, and intonation in the first eight bars of the aria as taken from recordings directed by John Eliot Gardiner and Karl Richter. On the basis of both spectral and temporal representations of the sound characteristics of the first few bars of the instrumental opening of the aria, obtained using the signal processing and plotting software in the Matlab program, Johnson demonstrates how Gardiner's recording features a more transparent timbre, much less vibrato in the solo violin part, and a more fluctuating amplitude profile, with a consistent tendency for the amplitude to drop away at group and phrase endings; Richter's recording, by contrast, demonstrates a more constant vibrato and amplitude level, a thicker timbre, and the use of expressive intonation (a flattening of the mediant note).

Johnson acknowledges that

> much of what is shown by spectrographic analysis is little more than a visual analogue of what we have already recognized and perceived through listening. Nonetheless, acoustic analysis reinforces the experiential claims of the listening musician, namely that (1) performance can significantly determine the properties of the experience itself, and (2) the listening experience is not wholly private: hearing is not entirely "subjective" in the sense of a strictly unverifiable or purely solipsistic mode of perception. . . . Finally, acoustic analysis is a powerful medium for the education of the ear and as a diagnostic tool for the conscientious performer, the didactic possibilities of which have barely begun to be exploited. (Johnson 1999: 83–84)

In fact Johnson uses the acoustic characteristics of the two recordings to argue that the two interpretations offer distinctly different musical and theological perspectives. Richter's recording, he claims, conveys a sense of reverence and authority (in relation both to Bach and the biblical narrative) in its even lines, thick textures, constant amplitude, tempo and vibrato, and solemnly "depressed" expressive intonation. Gardiner's, by contrast, is more enigmatic, using a faster and more flexible tempo and a more transparent sound to conjoin the secular connotations of dance with the seriousness of the biblical text; Johnson describes it as a "rediscovery in later 20th century Bach performance practice of the physical, the kinesthetic, not (here) as licentiousness but as a medium through which even a Passion can find new (or old) meanings." (Johnson 1999: 99)

Brackett's (2000) use of spectrum photos is more cursory and restricted, but worth considering because of the comparative rarity of this approach in the study of popular music—perhaps surprisingly, given that it is a non-score-based tradition in which acoustic characteristics (such as timbre, texture and space) are acknowledged to be of particular importance. Brackett's aim in his chapter on Elvis Costello's song "Pills and Soap" is to demonstrate various ways in which Costello maintains an elusive relationship with different musical traditions—particularly in his negotiations with art music. Brackett uses spectrum photos similar to Cogan's to make points about both the overall timbral shape of "Pills and Soap," and more detailed aspects

of word setting. An example of the latter is Brackett's demonstration (p. 187) that successive repetitions of the word "needle" in the song become increasingly timbrally bright and accented, as a way of drawing attention to the word and its narrative/semantic function. At a "middleground" level, he points out that vocal timbre (as well as pitch height) is used to give a sense of teleology to each verse, pushing the song forward. Finally (and this is where the connection with art music becomes more explicit), Brackett uses spectral information to support his claim that the song represents a particular kind of skirmish with Western "art" music. He shows how an increasingly oppositional relationship between high- and low-frequency timbral components characterizes the large-scale shape of the song, and argues that this

> is much more typical of pieces of Western art music than it is of almost any other form of music in the world, be it popular, "traditional," or non-Western "art" music. Examination of the photos in Robert Cogan's *New Images of Musical Sound* reveals a greater similarity between the spectrum photo of "Pills and Soap" and the photos of a Gregorian chant, a Beethoven piano sonata, the "Confutatis" from Mozart's *Requiem*, Debussy's "Nuages," and Varèse's *Hyperprism*, than between "Pills and Soap" and the Tibetan Tantric chant or Balinese shadow-play music. For that matter, the photo of "Pills and Soap" more closely resembles these pieces of art music than it does the photo for "Hey Good Lookin," the photo of which may reveal timbral contrast on a local level without that contrast contributing to a larger sense of teleological form. (Brackett 2000: 195)

Whether the argument that Brackett advances here stands up to scrutiny or not (there might be all kinds of reasons why "Pills and Soap" doesn't have a spectral shape that looks anything like Tibetan chant, Balinese shadow-play music, or another arbitrarily chosen popular song), the point that it makes is that the empirical evidence provided by spectral and temporal representations can furnish an important tool in a musicological enterprise—and that is what this chapter is intended to demonstrate.

The examples presented here, however, also illustrate some of the problems and pitfalls of using such information; it is very hard to find representational methods and analytical approaches that successfully reconcile detailed investigation with some sense of overall shape. Johnson's analysis, in focusing on the details of vibrato, timing, and intonation, doesn't go beyond bar 8 of the Bach aria; by contrast, Brackett's analysis of Costello, and many of Cogan's analyses, present spectrum photos at such a global level and with such inadequate resolution that some of the features and distinctions they discuss are all but invisible—and have to be taken on trust to more or less the same degree as if the authors were simply to tell the reader that the timbre gets brighter, or that there's a tiny articulation between phrases, or that there is an increasing accent on a word. In other words, there is a question about whether all the visual apparatus can really convince a reader of very much at all.

In part this is a purely technological matter, and the technology has certainly improved dramatically since the time of Cogan's book. But as shown by the much more sophisticated representations that Johnson uses, and as argued in this chapter, the problem is by no means solved by technological progress: there is still a real

problem in extracting the salient features from a data representation that contains a potentially overwhelming amount of information, only a tiny fraction of which may be relevant at any moment. The problem is testimony to the extraordinary analytical powers of the human auditory system: in the mass of detail that is presented in a "'close-up" view of the sound, the auditory system finds structure and distinctiveness. Some of the principles that account for this human capacity, and the ways in which they may contribute to musicological considerations, are the subject of the final section of this chapter.

Perceptual Analysis of Sounds

Music presents a challenge to the human auditory system, because it often contains several sources of sound (instruments, voices, electronics) whose behavior is coordinated in time. In order to make sense of this kind of musical material, the characteristics of the individual sounds, of concurrent combinations of them, and of sequences of them, must be identified by the auditory system. But to do this, the brain has to "decide" which bits of sound belong together, and which bits do not. As we will see, the grouping of sounds into perceptual units (events, streams, and textures) determines the perceived properties or attributes of these units. Thus, in considering the perceptual impact of the sounds represented in a score or a spectrogram, it is necessary to keep in mind a certain number of basic principles of perceptual processing.

Music played by several instruments presents a complex sound field to the human auditory system. The vibrations created by each instrument are propagated through the air to the listeners' ears, and combine with those of the other instruments as well as with the echoes and reverberations that result from reflections off walls, ceiling, furniture, and so on. What arrives at the ears is a very complex waveform indeed. To make matters worse, this composite signal is initially analyzed as a whole. The vibrations transmitted through the ear canal to the eardrum and then through the ossicles of the middle ear are finally processed biomechanically in the inner ear (the cochlea), such that different frequency regions of the incoming signal stimulate different sets of auditory nerve fibers. This is the aural equivalent of the spectral analysis described in the first main section of this chapter; one might consider the activity in the auditory nerve fibers over time as a kind of neural spectrogram. So if several instruments have closely related frequencies of vibration in their waveforms, they will collectively stimulate the same fibers: that is, they will be mixed together in the sequence of neural spikes that travel along that fiber to the brain. As we shall see, this would be the case for the different instruments playing the *Boléro* melody in parallel in a close approximation to a harmonic series.

It should be noted, however, that the different frequencies are still represented in the time intervals between successive nerve spikes, since the time structure of the spike train is closely related to the acoustic waveform. Furthermore, a sound from a single musical instrument is composed of several different frequencies (see the bass clarinet example in Figure 8.16) and thus stimulates many different sets of fibers; that is, it is analyzed into separate components distributed across the array of audi-

tory nerve fibers. The problem that this presents to the brain is to aggregate the separate bits that come from the same source, and to segregate the information that comes from distinct sources. Furthermore, the sequence of events coming from the same sound source must be linked together over time, in order to follow a melody played by a given instrument. Let us consider a few examples of the kinds of problem that this poses.

In some polyphonic music (such as Bach's orchestral suites or Ligeti's *Wind Quintet*), the intention of the composer is to create counterpoint, the success of which clearly depends on achieving segregation of the different instruments (Wright and Bregman 1987): what must be done to ensure that the instruments don't fuse together? In other polyphonic music, however (Ravel's *Boléro*, Ligeti's *Atmosphères*), the composer may seek a blending of different instruments and this would depend on achieving fusion or textural integration of the instruments: what must the musicians do to maximize the fusion and how can this be evaluated objectively? Finally, in some instrumental music an impression of two or more "voices" can be created from a monophonic source (such as in Telemann's recorder music or Bach's cello suites), or a single melodic line may be composed across several timbrally distinct instruments (as in Webern's *Six Pieces for Large Orchestra*, op. 6): what determines melodic continuity over time, and how might the integration or fragmentation be predicted from the score or for a given performance? For all these questions, the most important issue is how the perceptual result can be characterized from representations of the music (scores for notated music, acoustic representations for recorded or synthesized music). Obviously one can simply listen and use an aurally based analytical approach, but this restricts the account to the analyst's own (perhaps idiosyncratic) perceptions; if the aim is to provide a more generalized interpretation, the solution is to use basic principles of auditory perception as tools for understanding the musical process.

Grouping processes determine the perception of unified musical events (notes or aggregates of notes forming a vertical sonority), of coherent streams of events (having the connectedness necessary to perceive melody and rhythm), and of more or less dense regions of events that give rise to a homogeneous texture. *Perceptual fusion* is a grouping of concurrent acoustic components into a single auditory event (a perceptual unit having a beginning and an end); the perception of musical attributes such as pitch, timbre, loudness, and spatial position depends on which acoustic components are grouped together. *Auditory stream integration* is a grouping of sequences of events into a coherent, connected form, and determines what is heard as melody and rhythm. Texture is a more difficult notion to define, and has been the object of very little perceptual research, but intuitively the perception of a homogenous musical texture requires a grouping of many events across pitch, timbre, and time into a kind of unitary structure, the textural quality of which depends on the relations among the events that are grouped together (certain works by Ligeti, Xenakis, and Penderecki come to mind, as do any number of electroacoustic works). Note that the main notion behind the word "grouping" is a kind of perceptual connectedness or association, called "binding" by neuroscientists. It seems clear that many levels of grouping can operate simultaneously, and that what is perceived depends to some extent on the kind of structure upon which a listener focuses. Since

a large amount of scientific research has been conducted on concurrent and sequential sound organization processes, we will consider these in more detail, before moving on to discuss the perception of the musical properties (spatial location, loudness, pitch, timbre) that emerge from the auditory images formed by the primary grouping process.

There are two main factors that determine the perceptual fusion of acoustic components into unified auditory events, or their segregation into separate events: *onset synchrony* and *harmonicity*. A number of other factors were originally thought by perception researchers to be involved in grouping, but are probably more implicated either in increasing the perceptual salience of an event (vibrato and tremolo), or in allowing a listener to focus on a given sound source in the presence of several others (spatial position; for reviews see McAdams 1984, Bregman 1990, 1993, Darwin and Carlyon 1995, Deutsch 1999). We will focus here on the grouping factors.

Acoustic components that start at the same time are unlikely to arise from different sound sources and so tend to be grouped together into a single event. Onset asynchronies between components on the order of as little as 30 msec are sufficient to give the impression of two sources and to allow listeners in some cases to identify the sounds as separate; to get a perspective on the accuracy necessary to produce synchrony within this very small time window, one might note that skilled professional musicians playing in trios (strings, winds, or recorders) have asynchronies in the range of 30 to 50 msec, giving a sense of playing together while allowing perceptual segregation of the instruments (Rasch 1988). If musicians play in perfect synchrony, by contrast, there is a greater tendency for their sounds to fuse together and for the identity of each instrument or voice to be lost. These phenomena can also be manipulated compositionally: Huron (1993) has shown by statistical analyses that the voice asynchronies used by Bach in his two-part inventions were greater than those used in his work as a whole, suggesting an intention on the part of the composer to maximize the separation of the voices in these works. If, on the other hand, voices in a polyphony are synchronous, what may result is a global timbre that comes from the fusion of the composite—though considerable precision is needed to achieve such a result.

The other main grouping principle is that sound components tend to be perceived as a single entity when they are related by a common fundamental period. This is particularly the case if, when the fundamental period changes, all of the components change in similar fashion, as would be the case in playing vibrato, or in a single instrument playing a legato melody. Forced vibrating systems such as blown air columns (wind instruments) and bowed strings create nearly perfect harmonic sounds, with a strongly fused quality and an unambiguous pitch—in contrast to the several audible pitches of some inharmonic, free-vibrating systems such as a struck gong or church bell. This harmonicity-based fusion principle has again been used intuitively by composers of polyphonic music: a statistical analysis of Bach's keyboard music by Huron (1991) showed that the composer avoided harmonic intervals in proportion to the degree to which they promote tonal fusion, thus helping to ensure voice independence.

An important perceptual principle is demonstrated through such fusion: if sounds are grouped together, the perceptual attributes that arise—such as a new

composite timbre—may be different from those of the individual constituent sounds, and may be difficult to imagine merely from looking at the score or even at a spectrogram. The principle that the perceived qualities of simultaneities depend on grouping led Wright and Bregman (1987) to examine the role of nonsimultaneous voice entries in the control of musical dissonance: they argued that the dissonant effect of an interval such as a major seventh is much reduced if the voices composing the interval do not enter synchronously, and similar results also apply to fusion based on harmonicity (see McAdams 1999). All this demonstrates the need to consider issues of sonority in the perceptual analysis of pitch structures.

As a concrete example, Ravel's *Bolléro* arguably represents an example of intended fusion. Up to bar 148, the main melody is played in succession by different instrumental soloists. But at this point it is played simultaneously by five voices on three types of instrument: French horn, celesta, and piccolos (Figure 8.1b); the basic melody is played by the French horn, and is transposed to the octave, 12th, double octave, and double octave plus a major third for the celesta (LH), piccolo (2), celesta (RH), and piccolo (1), respectively. Note that this forms a harmonic series and that these harmonic intervals are maintained since the transpositions are exact (so that the fundamental, octave, and double octave melodies are played in C major, the 12th melody in G major, and the double octave plus a third melody in E major). Ravel thus respects the harmonicity principle to the letter, and since all the melodies are also presented in strict synchrony, the resulting fusion—with the individual instrument identities subsumed into a single new composite timbre—depends only on accurate tuning and timing being maintained by the performers. This procedure is repeated by Ravel for various other instrumental combinations in the course of the piece, the consequent timbral evolution contributing to the global crescendo of the piece.

An inverse example can be found in the mixed instrumental and electroacoustic work *Archipelago* by Roger Reynolds, for ensemble and four-channel computer-generated tape. In the tape part, recordings of the musical materials used elsewhere in the work by different instruments were analyzed by computer and resynthesized with modifications. In particular, the even and odd harmonics were either processed together as in the original sound, giving a temporally extended resynthesis of the same instrument timbre, or processed separately with independent vibratos and spatial trajectories, resulting in a perceptual fission into two new sounds. Selecting only the odd harmonics of an instrument sound leaves the pitch the same, but makes the timbre more "hollow" sounding, moving in the direction of a clarinet sound (which has weak even-numbered harmonics in the lower part of its frequency spectrum); selecting only the even harmonics produces an octave jump in pitch, since a series of even harmonics is the same as a harmonic series an octave higher. The perceptual result is therefore two new sounds with pitches an octave apart and timbres that are also different compared to the original sound. An example from *Archipelago* is the split of an oboe sound (Figure 8.17), which resulted in a clarinetlike sound at the original pitch and a sopranolike sound an octave higher. When the vibrato patterns were made coherent again, the sound fused back into the original oboe.

Sequential sound organization concerns the integration of successive events into auditory streams and the segregation of streams that appear to come from different sources. In real-world settings, a stream generally constitutes a series of events

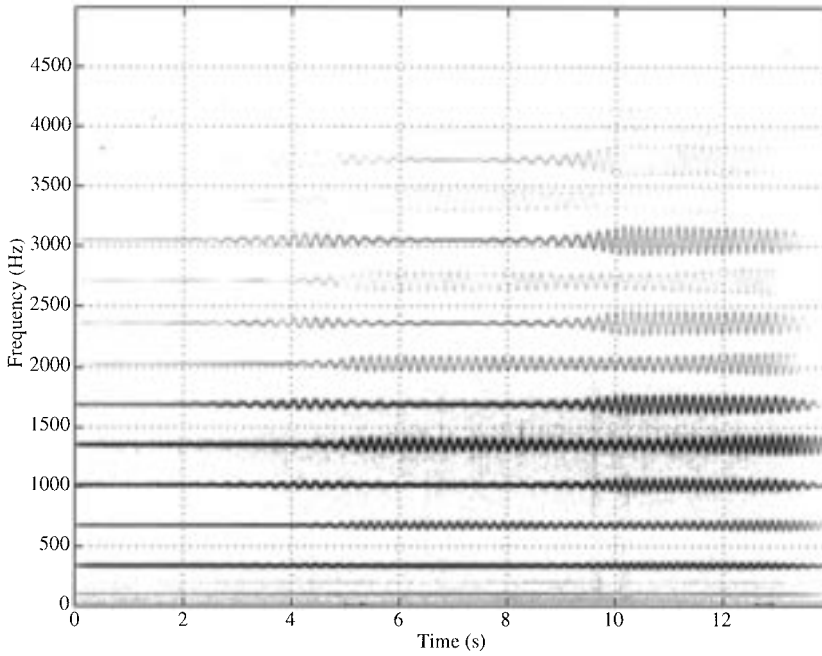


Figure 8.17. Splitting of an oboe sound in the Roger Reynolds's *Archipelago*. At around 3 seconds the odd-numbered harmonics start to have an independent vibrato which grows and then decays in strength. A similar pattern occurs on the even-numbered harmonics from about 5 seconds. Finally, each group swells in vibrato, but with independent vibration patterns at around 9 seconds.

emitted over time by a single source. As we will see, however, there are limits to what a listener can hear as an auditory stream, which does not always correspond to what real physical sources can actually do. So we can say that an auditory stream is a coherent "mental" representation of a succession of events. The main principle that affects this mental coherence is a trade-off between the temporal proximity of successive sound events and their relative similarity: the brain seems to prefer to connect a succession of sounds of similar quality which together create a perceptual continuity. Continuity, however, is relative since a given difference between successive events may be perceived as continuous at slow tempi, but will be split into different streams at fast tempi. The main parameters affecting continuity include spectrotemporal properties, sound level, and spatial position; continuous variation in all of these parameters gives a single stream, whereas rapid variation (particularly in all three together, as would often be the case for two independent sound sources playing at the same time) can induce the fission of a physical sequence of notes into two streams, one corresponding to the sequence emitted by each individual instrument.

In order to illustrate the basic principles, let us examine spectrotemporal continuity, which is affected by pitch and timbre change between successive notes. Melodies played by a single instrument with steps and small skips tend to be heard as unified, with easily detectable pitch and rhythmic intervals, while rapid jumps

across registers or between instruments may give rise to the perception of two or more melodies being played simultaneously, as illustrated in Figure 8.18 (an excerpt from a recorder piece by Telemann). In this case the perceived "melody" (i.e., the specific pattern of pitch and rhythmic intervals) corresponds to the relationships among the notes that have been grouped into a single stream or into multiple streams. Over the first six seconds of the excerpt shown, a listener will hear (and the spectrogram shows) a relatively slow ascending melody, a static pedal note, and a sequence of more rapid three note descending motifs. Because listeners often have great difficulty perceiving relations across streams, such as rhythmic intervals and even relative temporal order of events, there can be some surprising rhythmic results from apparently simple materials. The example in Figure 8.19 illustrates how two interleaved, isochronous rhythms played by separate xylophone players can produce a complex rhythmic pattern (note the irregular spacing of the sound events in the 250 to 1,000 Hz range in the spectrogram) due to the way the notes from the two players are combined perceptually into a single stream with unpredictable discontinuities in the melodic contour.

Once the acoustic waveform has been analyzed into separate source-related events, the auditory features of the events can be extracted. These musical qualities depend on various acoustic properties of the events, of which the most important are spatial location, loudness, pitch, and timbre. Each of these will be considered separately.

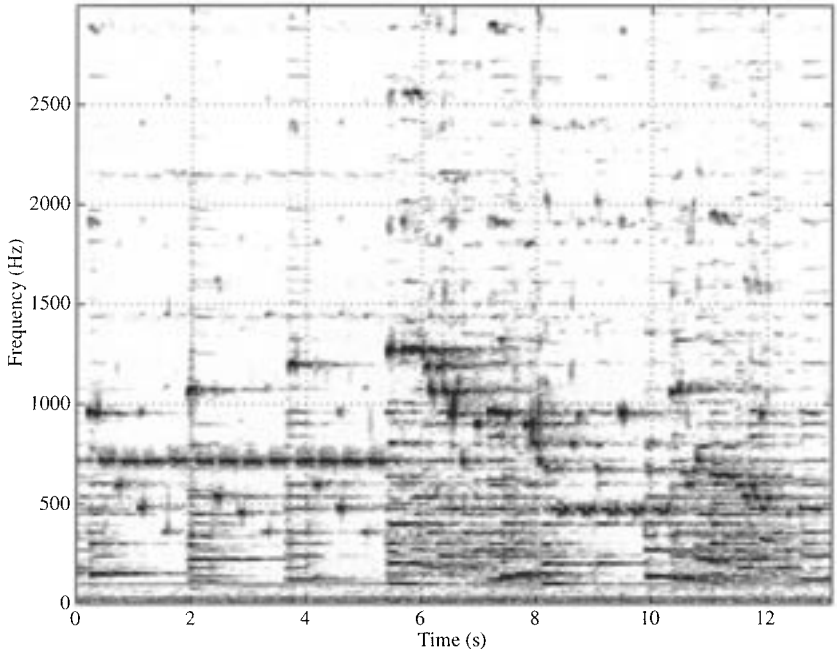


Figure 8.18. Spectrogram of an excerpt of a recorder piece by Telemann. The fundamental frequencies of the recorder notes lie in the range 400 to 1,500 Hz.

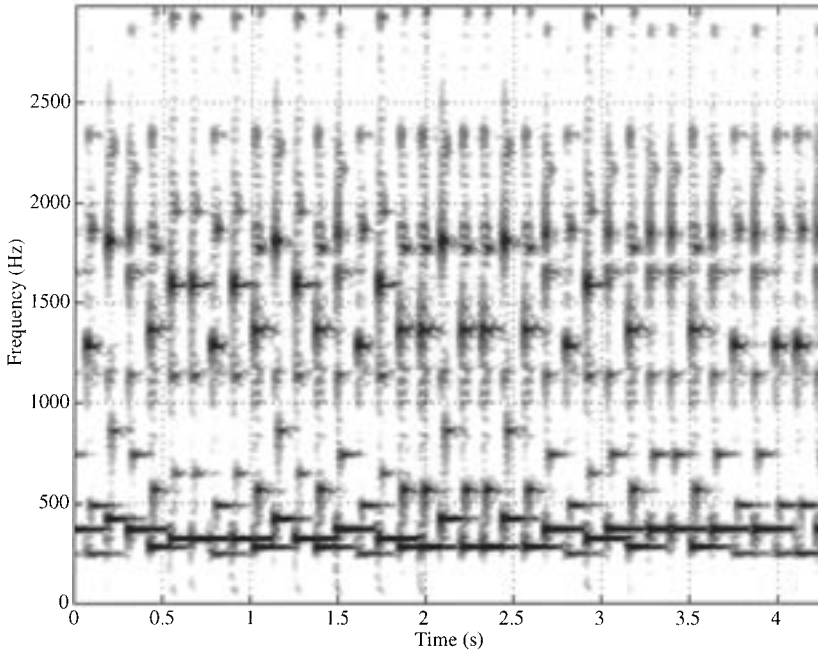


Figure 8.19. Spectrogram of a rhythm played by separate players on an African xylophone.

The *spatial location* of an event depends on several kinds of cues in the sound. In the first place, since we have two ears that are separated in space, the sound that arrives at the two ear drums depends on the position of the sound source relative to the listener's head, and is different for each ear: a sound coming from one side is both more intense and arrives earlier at the closer ear. Also the convoluted, irregular shape of the outer part of the ear (the pinna) creates position-dependent modifications of the sound entering the ear canal, and these are interpreted by the brain as cues for localization. Second, and more difficult to research, are the cues that allow us to infer the distance of the source (Grantham 1995). There are several possible acoustic cues for distance: one is the relative level, since level decreases as a function of distance, while another is the relative amount of reverberated sound in the environment as compared with the direct sound from the source (the ratio of reverberated to direct sound increases with distance). Finally, since higher frequencies are more easily absorbed and/or dispersed in the atmosphere than are lower frequencies, the spectral shape of the received signal can also contribute to the impression of distance. Such binaural, pinna, and distance cues are useful in virtual reality displays and in creating spatial effects in electroacoustic music.

For simple sounds *loudness* corresponds fairly directly to sound level; but for complex sounds, the global loudness results from a kind of summation of power across the whole frequency range. It is as if the brain adds together the activity in all of the auditory nerve fibers that are being stimulated by a musical sound to calcu-

late the total loudness. When several sounds are present at the same time and their frequency spectra overlap, a louder sound can cover up a softer sound either partially, making it even softer, or totally, making it inaudible: this process is called masking, and seems to be related to the neural activity of one sound swamping that of another. Masking may be partially responsible for the difficulty in hearing out inner voices when listening to polyphonies with three or more voices. Again, loudness is affected by duration: a very short staccato note (say around 50 msec) with the same physical intensity as a longer note (say around 500 msec) will sound softer. This seems to be because loudness accumulates over time, and the accumulation process takes time: for a long steady note, the perceived loudness levels out after about 200 msec. This principle is useful in instruments that produce sustained notes over whose intensity no control is possible, but whose duration can be controlled. For example, the production of agogic accents on the organ is obtained by playing certain notes slightly longer than their neighboring notes.

For any harmonic or periodic sound, the main *pitch* heard corresponds to the fundamental frequency, though this perceived pitch is the result of a perceptual synthesis of the acoustic information, rather than the analytic perception of the frequency component corresponding to the fundamental. (One can listen to a low-register instrument playing in the bottom of its tessitura over a small transistor radio and still hear the melody being played at the correct pitch, even though the spectrum of the signal shows that all of the lower-order harmonics are missing due to the very small size of the loudspeaker in the radio.) But many musical sound sources that are not purely harmonic (including carillon bells, tubular bells, and various percussion instruments) still give at least a vague impression of pitchedness; it seems that pitch perception is not an all-or-none affair, so that perceived pitch can be more or less strong or salient. For example, try singing a tune just by whispering and not using your vocal chords: you will find that you change the vowel you are singing to produce the pitch, which suggests that a noise sound with a prominent resonance peak can produce enough of a pitch percept to specify recognizable pitch relations between adjacent sound events. Similarly, the modern sound processing techniques used in electroacoustic and pop music can create spectral modifications of broadband noise (such as crowd or ocean sounds) with a regular series of peaks and dips in the spectrum: if the spacing between the centers of the noise peaks corresponds to a harmonic series, a weak pitch is heard, allowing musicians to "tune" noise sounds to more clearly pitched harmonic sounds.

Finally, *timbre* is a vague term that is used differently by different people and even according to the context. The "official" scientific definition is a nondefinition: the attribute of auditory sensation that distinguishes two sounds that are otherwise equal in terms of pitch, duration, and loudness, and that are presented under similar conditions (presumably taking into account room effects and so on). That leaves a lot of room for variation! Over the last 30 years, however, a new approach to timbre perception has been developed, which allows psychoacousticians to characterize more systematically what timbre is, rather than what it is not. Using special data analysis techniques, called multidimensional scaling (Plomp 1970, Grey 1977, McAdams et al. 1989), researchers have been able to identify a number of perceptual dimensions that constitute timbre, allowing a kind of deconstruction of this global category into

more precise elements. Attributes in terms of which timbres may be distinguished include the following:

- Spectral centroid (visible in a spectral representation: the relative weight of high and low-frequency parts of the spectrum, a higher centroid giving a "brighter" sound).
- Attack quality (visible in a temporal representation, and including the attack time and the presence of attack transients at the beginning of a sound).
- Smoothness of the spectral envelope (visible in a spectral representation: the clarinet has strong odd-numbered harmonics and weak even-numbered harmonics, giving a ragged spectral envelope).
- Degree of evolution of the spectral envelope over the course of a note (visible in a time-frequency representation: some instruments, like the clarinet, have a fairly steady envelope, whereas others have an envelope that opens up toward the high frequencies as the intensity of the note increases, as in the case of brass instruments).
- Roughness (visible in a temporal representation: smooth sounds have very little beating and fluctuation, whereas rough sounds are more grating and inherently dissonant).
- Noisiness/inharmonicity (visible in a spectral representation: nearly pure harmonic sounds, like blown and bowed instruments, can be distinguished from inharmonic sounds like tubular bells and steel drums, or from clearly noisy sounds such as those of crash cymbals and snare drums).

A greater understanding of the relative importance of these different "dimensions" of timbre may help musicologists develop systematic classification systems for musical instruments and even sound effects or electroacoustic sounds.

Analytical Application

As an example of the way in which psychoacoustical principles can be empirically applied to the analysis of pitch and timbre, Tsang (2002) uses a number of perceptually based approaches to analyze the structure of *Farben* (*Colors*)—the third of Schoenberg's *Five Orchestral Pieces, Op. 16*, which is celebrated for its innovative use of orchestral timbre. Taking principles developed by Parncutt (1989) for estimating the salience of individual pitches, and by Huron (2001) for explaining voice-leading in perceptual terms, Tsang discusses the perceptibility of the canonic structure of the opening section of *Farben*. By applying Parncutt's pitch salience algorithm (a formula used to calculate how noticeable any given pitch is in the context of other simultaneous pitches), Tsang concludes that "Schoenberg's choice of pitches ensures that relatively strong harmonic components often draw the listener's attention to the canonic voices that are moving or are about to move" (Tsang 2002: 29). Huron's perceptual principles relating to voice-leading, which take into account a wider variety of psychoacoustical considerations than pitch alone, partially support this conclusion, suggesting that Schoenberg tailored his choices of orchestration so as to bring out the canonic movement, but also indicate that other factors serve to disguise the canon.

At the end of his study, Tsang notes that the different attentional strategies listeners bring to bear on the music will inevitably result in different perceptual experiences, as will comparatively slight differences of interpretation on the part of conductors and orchestras—particularly in a piece that seems to place itself deliberately at the threshold of perceptual discriminability. These considerations suggest a high level of indeterminacy between what a perceptually informed analysis might suggest and what any particular listener may experience—an indeterminacy that would be damaging to a narrowly descriptive (let alone rigidly prescriptive) notion of the relationship between analysis and experience. But as many authors have pointed out (e.g., McAdams 1984, Cook 1990), to propose such a tight linkage is neither necessary nor even desirable.

A further example of an attempt to relate perceptual principles to musicological concerns is provided by Huron (2001). The goals of this ambitious paper are "to explain voice-leading practice by using perceptual principles, predominantly principles associated with the theory of auditory stream segregation . . .", and "to identify the goals of voice-leading, to show how following traditional voice-leading rules contributes to the achievement of these goals, and to propose a cognitive explanation for why the goals might be deemed worthwhile in the first place" (Huron, 2001: 2–3). As this makes clear, perceptual principles are being used here to address not only matters of compositional practice, but also aesthetic issues. The form of the paper is first to present a review of accepted rules of voice-leading for Western art music; second to identify a number of pertinent perceptual principles; third to see whether the rules of voice-leading can be derived from the perceptual principles; fourth to introduce a number of auxiliary perceptual principles which provide a perspective on different musical genres; and finally to consider the possible aesthetic motivations for the compositional practices that are commonly found in Western music and which do not always simply adhere to the perceptual principles that Huron identifies.

Huron makes use of six perceptual principles in the central part of the paper, each of which is supported with extensive empirical evidence from auditory and music perception research, and is shown to correspond to compositional practice often sampled over quite substantial bodies of musical repertoire (using Huron's *Humdrum* software—see chapter 6, this volume). To give some idea of what the perceptual principles are like, and how they are used to derive voice-leading rules, consider as an example the third perceptual principle, which Huron calls the "minimum masking principle": "In order to minimize auditory masking within some vertical sonority, approximately equivalent amounts of spectral energy should fall in each critical band. For typical complex harmonic tones, this generally means that simultaneously sounding notes should be more widely spaced as the register descends." (Huron 2001: 18)

In support of this principle, Huron assembles a considerable amount of evidence from well-established psychoacoustical research dating back to the 1960s showing that pitches falling within a certain range of one another (i.e., the "critical band," which roughly corresponds to the bandwidth of the auditory filters in the cochlea) both tend to obscure ("mask") one another, and interact to create a sense of instability or roughness (often referred to as "sensory dissonance"). When Huron

goes on to derive perceptually-based voice-leading rules, the "minimum masking principle" is used to motivate two rules, one traditional, and one which Huron calls "nontraditional"; "nontraditional" rules are those that seem to follow from perceptual principles, but are not acknowledged as explicit voice-leading rules in standard texts. The traditional rule is stated as follows (Huron, 2001: 33): "Chord Spacing Rule. *In general, chordal tones should be spaced with wider intervals between the lower voices.*" The nontraditional rule (ibid.) is: "Tessitura-Sensitive Spacing Rule. *It is more important to have large intervals separating the lower voices in the case of sonorities that are lower in overall pitch."*

And Huron refers to work showing that this rule, although not explicitly recognized in standard texts on voice-leading, is adhered to in musical practice. The five other perceptual principles work in a similar fashion, generating individually, and in combination with one another, a total of 22 voice-leading rules, of which nine are traditional. All of the 13 nontraditional rules are found to be supported by compositional practice, demonstrating rather convincingly how a perceptually-based approach can reveal implicit, but previously unrecognized, compositional principles.

Having derived the voice-leading principles, Huron introduces four additional perceptual principles which are used to address questions of musical genre. Again, to get a flavor of what is involved, consider just one of the four (Huron, 2001: 49): "Timbral Differentiation Principle. *If a composer intends to write music in which the parts have a high degree of perceptual independence, then each part should maintain a unique timbral character."*

The striking thing about this principle, as Huron points out, is the extent to which it is ignored in compositional practice. Although wind quintets and other small mixed chamber ensembles show significant differentiation, string quartets, brass ensembles, madrigal groups, and keyboards all make use of timbrally *un*differentiated textures for polyphonic purposes. Why is this? Huron suggests that there may be a number of factors. One is pragmatic: it may simply have been more difficult for composers to assemble heterogeneous instrumental groups, and so the goal of distinguishable polyphony was bracketed or abandoned in favor of practical possibility. A second reason may be the operation of a contrary aesthetic goal: Huron suggests that composers tend to prefer instrumental ensembles that show a high degree of "blend," and homogeneous instrumentation may be one way to achieve this. And a third reason may be balance: it is much harder to achieve an acceptable balance between voices in a very diverse instrumental group, and composers may have decided that this was a more important goal. It is interesting in this regard that Schoenberg's practice, from the middle period of his atonal style, of indicating instrumental parts as "Haupstimme" and "Nebenstimme" (main voice and subsidiary voice) was motivated by a concern that the correct balance between instrumental parts in his chamber and orchestral works might not be attained; given the dramatic explosion of writing for mixed chamber ensembles in the twentieth century, influenced strongly by the ensemble that Schoenberg used in *Pierrot Lunaire*, this does seem to be a recognition of the balance problem that Huron identifies. Equally, the striking way in which Webern uses timbral differentiation to cut across the serial structure in the first movement of the *Symphony*, Op. 21, provides support (through counterevidence) for the timbral differentiation principle: timbral identity and dif-

ferentiation disguise the serial structure, instead superimposing a different structure that is articulated by timbre itself—a timbral palindrome, in fact.

The Webern example already demonstrates the complex interaction between compositional and aesthetic goals on the one hand, and perceptual principles on the other—and it is this subject that the final part of Huron's paper addresses. Huron proposes that achieving perceptual clarity in perceptually challenging contexts (e.g., finding hidden objects in visual arrays, as exploited in many children's puzzles) is an intrinsically pleasurable and rewarding process, and that this is one way to understand why voice-leading rules and compositional practice both conform to, and flout, perceptual principles. If all music simply adhered to perceptual imperatives, then there would be little motivation to move beyond the most straightforward monophony. But social considerations (the need, or desire, to develop musical styles in which groups of singers or instrumentalists with different pitch ranges, timbral qualities or dynamic characteristics can sing and play together), aesthetic goals and a whole range of other factors have resulted in the historical development of an enormous variety of textures and styles. One strand within this, Huron suggests, is the possibility that some multipart music is organized deliberately to challenge the listener's perceptual capacities—precisely because of the pleasure that can be gained from successfully resolving these complex textures.

> Early Renaissance polyphonists discovered [that] . . . by challenging the listener's auditory parsing abilities, the potential for a pleasing effect could be heightened. However, this heightened pleasure would be possible only if listeners could indeed successfully parse the more complex auditory scenes. Having increased the perceptual challenge, composers would need to take care in providing adequate streaming cues. Following the rules of voice-leading and limiting the density of parts . . . might be essential aids for listeners. (Huron, 2001: 57)

What is exciting about this discussion is that it brings together perceptual principles based on extensive empirical support, aesthetic considerations, and a rather different perspective on music history in a way that manages to avoid the potential pitfalls of a perceptual determinism. Huron's final paragraph is significant for the care with which it recognizes that perceptual principles act neither as the arbiters of musical value, nor as constraints on future creativity. Noting that his interpretation of the aesthetic origins of voice-leading "should in no way be construed as evidence for the superiority of polyphonic music over other types of music," he goes on as follows (Huron 2001: 58):

> In the first instance, different genres might manifest different perceptual goals that evoke aesthetic pleasure in other ways. Nor should we expect pleasure to be limited only to perceptual phenomena. As we have already emphasized, the construction of a musical work may be influenced by innumerable goals, from social and cultural goals to financial and idiomatic goals. . . . The identification of perceptual mechanisms need not hamstring musical creativity. On the contrary, it may be that the overt identification of the operative perceptual principles may spur creative endeavors to generate unprecedented musical genres.

Conclusion

In this chapter, we have tried to show how acoustical and perceptual analyses can supplement and "animate" an analytical understanding drawn from scores and recordings in important ways. The representations and analytical methods considered here can be a significant key to a more *sound*-based understanding of music than the score-reading approach has traditionally encouraged, and in this way attributes of musical structure and process that remain hidden from view in the score, and that often pass by too rapidly and with too much complexity in performance or recording, can be brought to light and given appropriate consideration. But as we have seen, there are still problems to be overcome: as the figures in this chapter demonstrate, representations of sound often contain large amounts of information, with the result that it can be difficult to strike an effective balance between analyzing musically appropriate stretches of material and risking information overload on the one hand, and focusing on a frustratingly tiny fragment of music on the other. Through the analysis of musical sound is still in early stages of development, and more powerful summarizing tools will no doubt be developed in the future, but in the long run there may be no easy solution to a problem which is a testimony to the exceptional power of human perception.

As the example from Tsang (2002) has already made clear, the kind of approach discussed in this chapter is necessarily generic, and unable to *explain* individual listening experiences—even if it brings new tools with which to *illustrate* those experiences. It is, after all, based on a "culture-free" approach in which the salience and impact of events, for example, is based solely on their acoustical and perceptual properties and not on their cultural resonances or semiotic significance. A well-established finding in the psychology of perception, sometimes referred to as the "cocktail party phenomenon," demonstrates that when people are attending to multiple sound sources, a source that has special significance for them (such as their name, or an emotionally charged word) will catch their attention even when it competes with other sound sources that may be considerably more salient (louder, nearer, timbrally more prominent). Thus the analyst is in no measure freed by this wealth of empirical data from either the responsibility or the opportunity to explore, and try to explain, why music might be heard or understood in particular ways. Nonetheless, acoustical and perceptual analyses can usefully complement more culturally oriented approaches by providing a rich source of information on which the latter might be based, and in terms of which they are certainly grounded. As with any empirical approach, the value of such an outlook is not in the data that it may accumulate but in the way in which data rub up against theory—formal or informal—in ways that may be supportive and confirmatory, or uncomfortable and mind changing.

References

Brackett, D. (2000). *Interpreting Popular Music.* Berkeley: University of California Press.

Bregman, A. S. (1990). *Auditory Scene Analysis: The Perceptual Organization of Sound.* Cambridge, Mass.: MIT Press.

Bregman, A. S. (1993). "Auditory scene analysis: Hearing in complex environments," in S. McAdams and E. Bigand (eds.), *Thinking in Sound: The Cognitive Psychology of Human Audition*. Oxford: Oxford University Press, 10–36.

Cogan, R. (1984). *New Images of Musical Sound.* Cambridge, Mass.: Harvard University Press.

Cook, N. (1990). *Music, Imagination, and Culture*. Oxford: Clarendon Press.

Darwin, C. J., and Carlyon, R. P. (1995). "Auditory grouping," in B. C. J. Moore (ed.), *Hearing.* San Diego, Calif.: Academic Press, 387–424.

Deutsch, D. (1999). "Grouping mechanisms in music," in D. Deutsch (ed.), *The Psychology of Music,* 2nd ed. San Diego, Calif.: Academic Press, 299–348.

Grantham, D. W. (1995). "Spatial hearing and related phenomena," in B. C. J. Moore (ed.), *Hearing.* San Diego, Calif.: Academic Press, 297–345.

Grey, J. M. (1977). "Multidimensional perceptual scaling of musical timbres." *Journal of the Acoustical Society of America* 61: 1270–1277.

Huron, D. (1991). "Tonal consonance versus tonal fusion in polyphonic sonorities." *Music Perception* 9: 135–154.

Huron, D. (1993). "Note-onset asynchrony in J. S. Bach's two-part inventions." *Music Perception* 10: 435–444.

Huron, D. (2001) "Tone and voice: A derivation of the rules of voice-leading from perceptual principles." *Music Perception* 19: 1–64.

Johnson, P. (1999) "Performance and the listening experience: Bach's Erbarme" Dich," in N. Cook, P. Johnson, and H. Zender (eds.), *Theory into Practice: Composition, Performance and the Listening Experience*. Leuven (Louvain), Belgium: Leuven University Press, 55–101.

McAdams, S. (1984). "The auditory image: A metaphor for musical and psychological research on auditory organization," in W. R. Crozier and A. J. Chapman (eds.), *Cognitive Processes in the Perception of Art*. Amsterdam: North-Holland, 289–323.

McAdams, S. (1999). "Perspectives on the contribution of timbre to musical structure." *Computer Music Journal* 23: 96–113.

McAdams, S., Winsberg, S., Donnadieu, S., De Soete, G., and Krimphoff, J. (1995). "Perceptual scaling of synthesized musical timbres: Common dimensions, specificities, and latent subject classes." *Psychological Research* 58: 177–192.

Parncutt, R. (1989). *Harmony: A Psychoacoustical Approach.* Berlin: Springer-Verlag.

Plomp, R. (1970). "Timbre as a multidimensional attribute of complex tones," in R. Plomp and G. F. Smoorenburg (eds.), *Frequency Analysis and Periodicity Detection in Hearing*. Leiden: Sijthoff, 397–414.

Rasch, R. A. (1988). "Timing and synchronization in ensemble performance," in J. A. Sloboda (ed.), *Generative Processes in Music: The Psychology of Performance, Improvisation, and Composition*. Oxford: Oxford University Press, 71–90.

Tsang, L. (2002). "Towards a theory of timbre for music analysis." *Musicae Scientiae* 6: 23–52.

Wright, J. K., and Bregman, A. S. (1987). "Auditory stream segregation and the control of dissonance in polyphonic music." *Contemporary Music Review* 2/1: 63–92.