3 Cognitive Psychology and Music

Roger Shepard

3.1 Cognitive Psychology

What does cognitive psychology have to do with the perception of sound and music? There is a long chain of processes between the physical events going on in the world and the perceptual registration of those events by a human observer. The processes include the generation of energy by some external object or event, the transmission of the energy through the space between the event and the observer, the reception and processing of the energy by the observer's sensory receptors, and the transmission of signals to the brain, where still more processing takes place. Presumably, the end result is the formation of a representation in the brain of what is going on in the external world. The brain has been shaped by natural selection; only those organisms that were able to interpret correctly what goes on in the external world and to behave accordingly have survived to reproduce.

The way we experience all events in the world, including musical events, is the result of this process of interpretation in the brain. What is happening inside the eye on the surface of the retina, or on the basilar membrane in the ear, is of no significant interest whatsoever, except insofar as it provides information from which the brain is able to construct a representation of what is going on in the world. True, the signals from the receptors are generally the only source of information the brain has about what is actually going on in the external world, so it is important to understand the workings of the observer's eyes and ears. But what goes on in those sensory transducers has relatively little direct correspondence to the final representation experienced by the observer, which is the result of extensive further processing within the observer's brain.

Sensory psychophysicists and psychologists study what goes on in the sensory transducers, and the eye and ear appear fundamentally quite different in function and behavior. There are many things specific to a particular sensory organ, and they must be studied and discussed independently. In contrast, cognitive psychologists are principally interested in the final internal representation. If the

internal representation is to be useful, it must correspond to events in the real world. There is one world to be perceived, and all of the senses provide information to the observer about that world. Therefore, a confluence should emerge from the processing in the brain, regardless of whether the input is from the visual, auditory, or some other sensory modality. This chapter will point out some general principles of perception and cognition that, though similar for vision and audition, are directly relevant to the understanding of music and music perception.

Figure 3.1 demonstrates that internal representation can indeed be quite different from the physical stimulus on the retina. Two tables are depicted as if in different orientations in space, but stating that there are two tables already makes a cognitive interpretation. The figure actually consists only of a pattern of lines (or dots) on a two-dimensional surface. Still, humans tend to interpret the patterns of lines as three-dimensional objects, as two differently oriented tables with one larger than the other. If one were able to turn off the cognitive representation of "tables in space," one would see that the two parallelograms corresponding to the tabletops are of identical size and shape! Verify this with a ruler, or trace one parallelogram (tabletop) on a sheet of tracing paper and then slide it into congruence with the other. The fact that it is difficult to see the two tabletops abstractly as simple parallelograms, and thus to see them as the same size and shape, proves that the internal representation in the brain is quite different from the pattern present on the sensory surface (retina). We tend to represent the pattern of lines as objects in

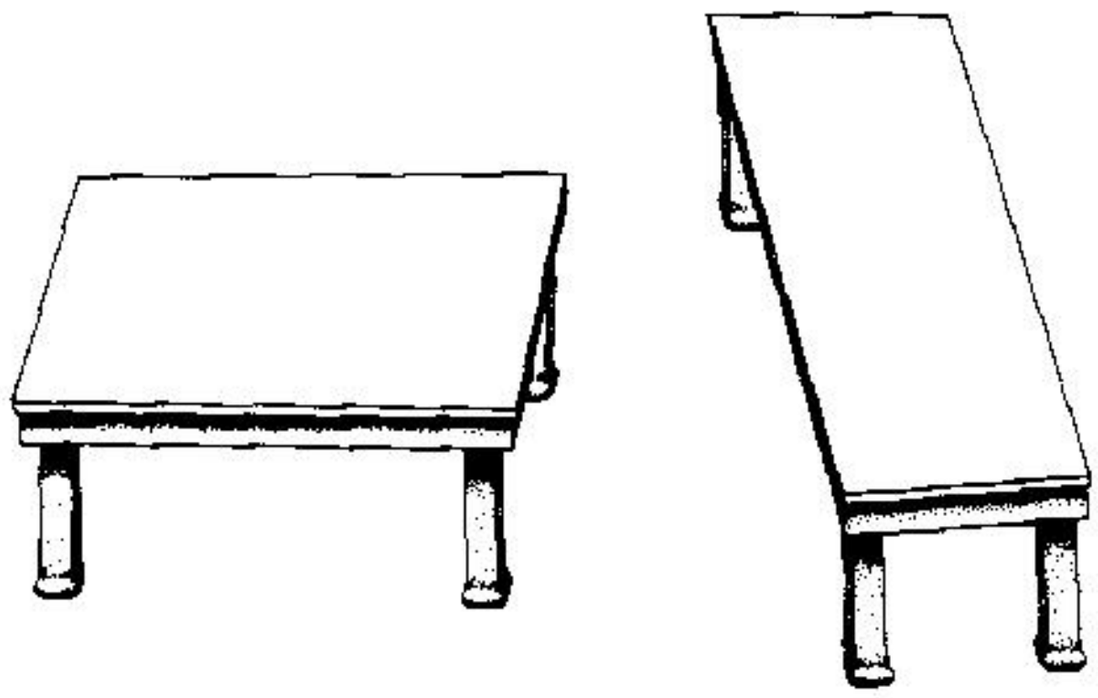

Figure 3.1 Things are sometimes different than they appear to be. Use a ruler to measure the tops of these two tables, specifically comparing the short ends and the long ends.

the external world because evolution has selected for such representation. The interpretation process in the brain has been shaped to be so automatic, swift, and efficient that it is virtually unconscious and outside of our control. As a result, we cannot suppress it even when we try.

3.2 Unconscious Inference

Hermann von Helmholtz (born 1821) made more contributions to the understanding of hearing and vision than perhaps any other individual. In addition to his fundamental contributions to physics and to physiology, in cognitive psychology he is known for his formulation of the principle of *unconscious inference.* Figure 3.2 illustrates the principle of unconscious inference. Our perceptual machinery automatically makes the inference to three-dimensional objects on the basis of perceptual cues that are present in the two-dimensional pattern on the retina. Cues—particularly linear perspective—support the inference to the three-dimensional interpretation, but the inference is quite unconscious.

Many retinal cues enable us to construct a three-dimensional representation from purely two-dimensional representation input. Following are a few examples of these cues:

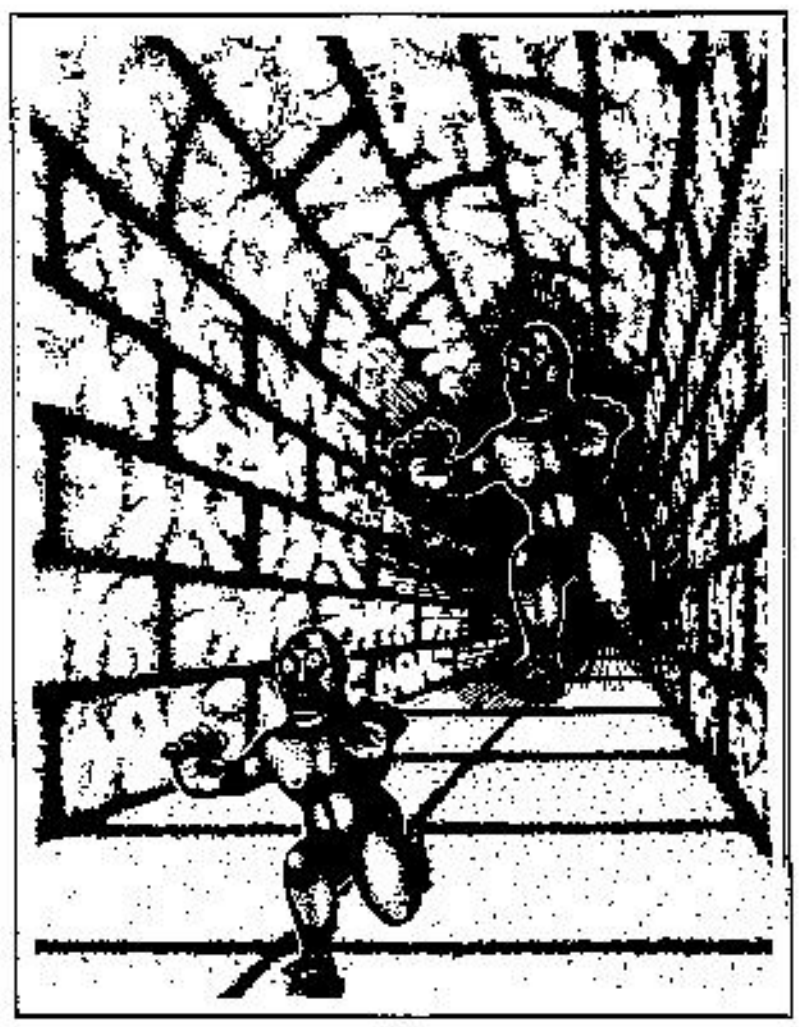

Figure 3.2 Unconscious inference is at work in this picture. Even though both "monsters" are exactly the same size (measure them with a ruler), the perspective placement makes the chaser look bigger than the one being chased.

Linear perspective. Converging lines in a two-dimensional drawing convey parallel lines and depth in three dimensions. This is evident in the rows of stones in figure 3.2.

Gradient of size. The elements of a uniform texture decrease in size as they approach the horizon. This is evident in figure 3.2, where the stone patterns get smaller in the receding tunnel.

Aerial perspective. Objects in the far distance appear lighter and blue (for the same reason that the sky appears light and blue).

Binocular parallax. Each of our two eyes receives a slightly different image, and from these the brain is able to make quite precise inferences about the relative distances of objects. This is particularly true for objects close to the observer.

Motion parallax. Movement on the part of the observer changes the images on each retina, and the differences between successive viewpoints is used to infer distances, just as in binocular parallax.

It is interesting to note that in general we have no notion of the cues that our brains are using. Experiments have shown that some of the cues can be missing (or intentionally removed); but as long as some subset of these cues is still available, the observer sees things in depth and can make accurate judgments about the relative distances and placements of objects. Even though the examples printed in this book are just two-dimensional drawings, the important thing to remember is that all images end up entering our retinas as two-dimensional images. We use unconscious inference to make sense of the real world just as we use it to interpret drawings, photographs, and movies.

The use of the term *inference* does not imply that the cognitive processes of interpretation are mere probabilistic guesses, although situations do occur in which the number of cues is reduced to the point where unconscious inference may become a random guesslike process. James Gibson, a perceptual psychologist at Cornell University, emphasized that under most circumstances (when there is good illumination, we are free to move about with both eyes open, and our spatial perception is completely accurate and certain), the information is sufficient to construct an accurate representation of the disposition of objects in space. Gibson referred to this as *direct perception*, as contrasted to *unconscious inference*. The two can be reconciled by the fact that complex computation must go on to process the information coming into the sensory systems, and most of that computation goes on unconsciously. The information is integrated

in order to give very precise information about what is going on in the world, not random guesses based on fragmentary information.

3.3 Size and Loudness Constancy

Objects in the world are, in general, of constant size; but the image of an object on the retina expands and contracts as the object moves closer and farther away. What has been important for us and for our ancestors has been the ability to perceive objects as they are, independent of their distance from us. This is known as *size constancy*. Figure 3.3 demonstrates this principle.

In the auditory domain, *loudness constancy* is a direct analog of size constancy. If an instrument emitting a sound of constant output is moved farther away, the intensity that reaches a listener decreases. This is because the wave fronts emanating from the instrument are spherical in shape, and the surface area of a sphere increases with the square of the radius. The energy from the instrument is uniformly distributed over this spherical surface, and hence the intensity reaching the listener decreases with the square of the distance from the instrument to the listener. Not surprisingly then, if the amplitude of a sound is decreased, the sound may seem to come from farther away. But we could alternatively experience the source as decreasing in intensity without moving farther away. Similarly, a visually perceived balloon from which air is escaping may

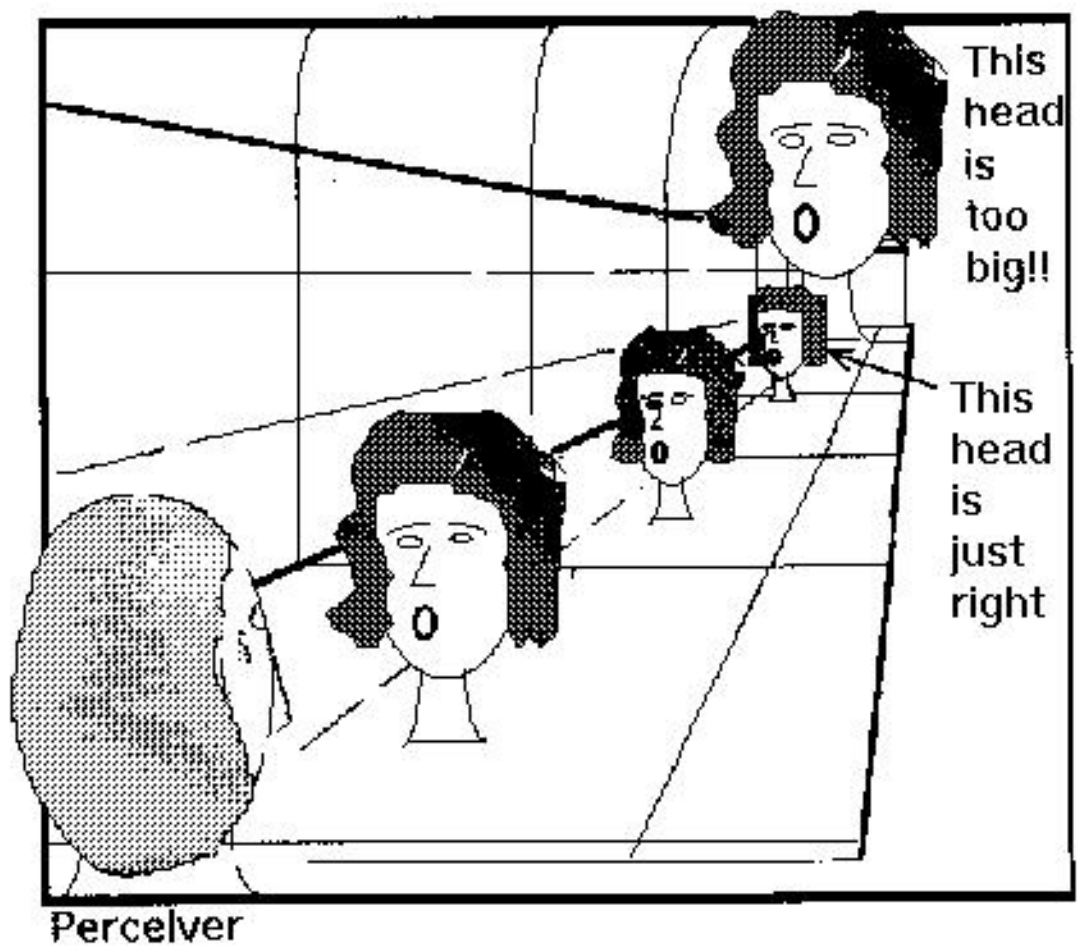

Figure 3.3 Size constancy. The head closest to the perceiver is the same physical size on the page as the "too-big" head farthest from the perceiver.

appear to be receding into the distance or simply shrinking in size. Other cues besides size or loudness may determine whether the change in the external world is in the size or the intensity of the source, or in its distance from the observer.

The intensity of a musical source can be decreased by playing the instrument more softly. There are accompanying changes in timbre, however, that are different from a simple decrease in amplitude. The higher-frequency components of the sound tend to increase and decrease with the effort exerted by the musician, an amount that is not proportional to the lower components of the spectrum. Thus, spectral balance as well as overall amplitude provides cues to the intensity versus distance of a source.

In our normal surroundings, there are surfaces around us that reflect sound, causing echoes or reverberation. In general we have little direct awareness of the reflected sound reaching us via these paths, but we use the information in these reflected waves to make unconscious inferences about the surroundings and sound sources within those surroundings. The reflections tell us, for example, that we are in a room of a certain size and composition, and give us a sense of the space. We receive a signal from a sound source within the room, then some time later we receive signals via the reflected paths. If a sound source is close, the direct sound is relatively intense, and the reflected sounds occur at decreased intensity and later in time. If the sound source moves away, the direct sound decreases, but the reflected sound remains roughly constant in intensity. The time difference between the arrival of the direct and the reflected sounds also decreases as the source recedes. By unconscious inference, the intensity ratio of direct to reflected sound, and the time delay between the direct and reflected sound, are used, along with other cues, to determine the distance and intensity of a source.

3.4 Spatial and Temporal Inversion

Some of the correlations in the world are so common that we have developed special machinery for their interpretation. If a familiar pattern is transformed in some way, even though all of the information is retained intact, then that pattern will not be interpreted in the same way by a human observer because our machinery is "wired" to interpret the information only in its usually encountered form. Consider the simple transformation of rotation. Figure 3.4 shows a number of presentations of the same face. Because we are attuned to

perceiving faces in their usual upright orientation, the upper and lower rows of shapes shown in the figure are perceived as being of two different faces rather than as one face in two orientations. We tend to make the interpretation that is consistent with a standard face, in which the eyes are on the top and the mouth is on the bottom. Developmental studies have shown that up to a certain age, children are equally skilled at interpreting faces either right-side up or upside down, but with increasing age the skill at interpreting faces right-side up continues to increase after the ability to interpret inverted faces levels off. Eventually the right-side up exposure becomes so great that the perception dominates. We develop an impressive ability to recognize and to interpret the expressions of right-side up faces—an ability not yet matched by machine—but this ability does not generalize to upside-down faces, with which we have had much less practice.

An analog of this spatial inversion in the visual domain is a temporal reversal in the auditory domain. In normal surroundings, we receive direct and reflected sound. We generally do not hear the reflected echoes and reverberation as such, but make the unconscious inference that we are hearing the source in a certain type of space, where the impression of that space is determined by the character of the reflected signals.

It is curious that the addition of walls and boundaries, essentially limiting space, gives the sense of spaciousness in audition. In a purely anechoic room (a specially constructed space that minimizes

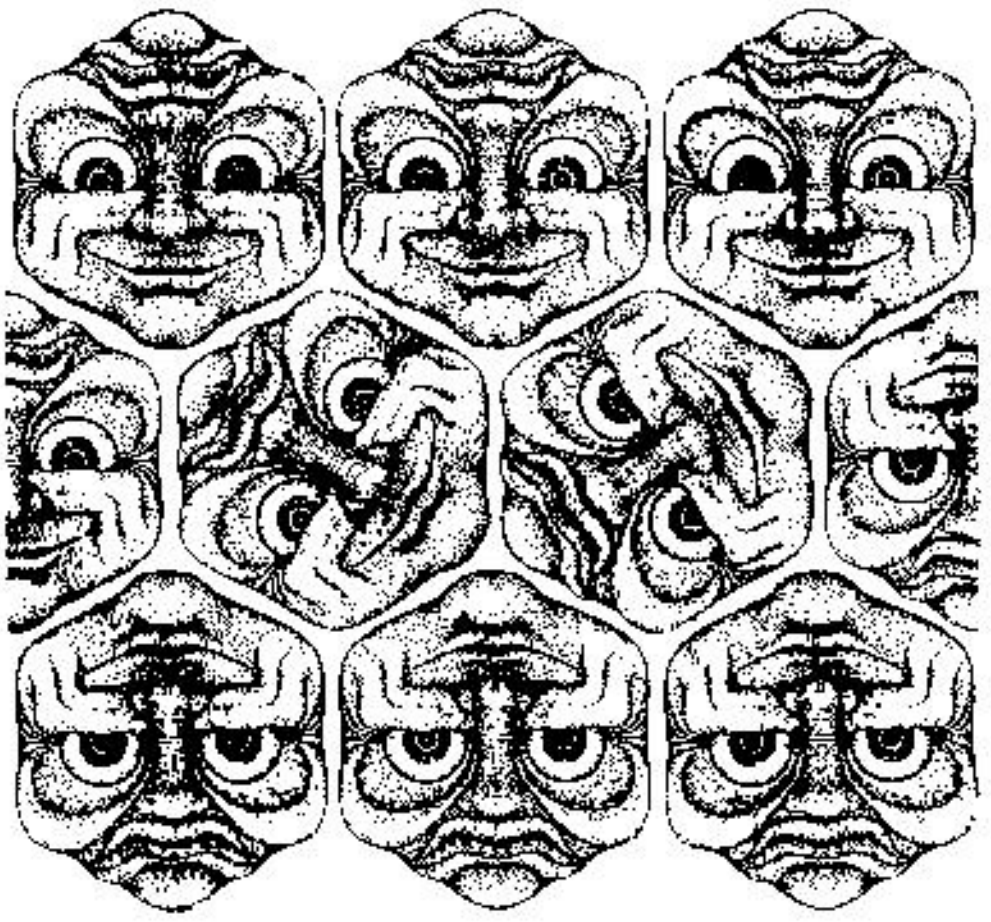

Figure 3.4 Turn this page over and you still will see faces right-side up. After infancy, we become more tuned to seeing faces right-side up, and thus must try hard to see the "frowning" faces as being upside down.

reflections from the walls, floor, and ceiling) we get no reverberation, and thus no sense of space. In vision, too, if an observer were in space with no objects around, there would be no sense of the space. Gibson pointed out that we do not perceive space but, rather, objects in space. In audition, we need surfaces to give us the sense of the space they define.

The ears can hear the direction of the source by comparing the differences between the arrival times and intensities at the two ears. The ears can similarly process differences in times and amplitudes of reflected sounds, and infer the source locations implied by those reflected sounds. In this way, we auditorally identify a sound source and a number of *virtual sources,* or copies of the sound source in virtual locations that lie *outside* the space actually enclosed by the walls. Figure 3.5 shows a sound source, an ear, and a few reflected sound paths. Only the first reflections (those that reflect from only one wall in going from the source to the listener) are shown, but there are many important second, third, and so on reflections. Figure 3.6 shows the same sound paths as direct paths from virtual sources. It is clear why reverberation gives the sense of space, with virtual sources distributed over a large space outside the room. The same sense of space can be experienced visually in a room (such as a barbershop or restaurant) with large mirrors on opposite walls.

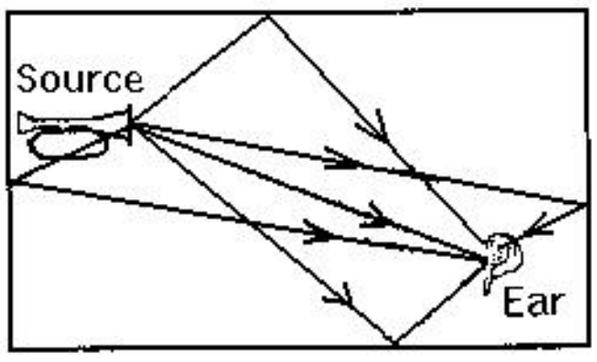

Figure 3.5 Many reflected acoustic paths in a room.

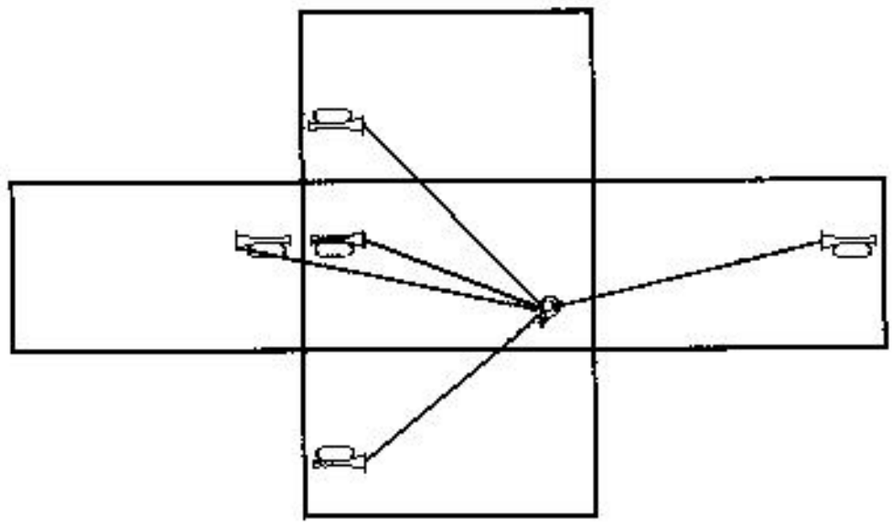

Figure 3.6 The same paths, shown as direct paths from "virtual sources."

There is a fundamental time asymmetry in the reception of direct and reflected sounds. All reflected sounds reach the listener *after* the direct signal. This is a manifestation of the second law of thermodynamics: in the absence of external energy input, order tends to go over into disorder. The direct sound may be orderly, but the randomly timed reflected copies of that sound appear to become random, with a momentary impulse decaying into white noise over time. Our auditory processing machinery evolved to process echoes and reverberation that follow a direct signal; it is ill-equipped to deal with an artificially produced case in which the echoes precede the signal.

Similarly, a resonant object, when struck, typically produces a sound that decays exponentially with time. Because of its unfamiliarity, a note or chord struck on a piano (with the damper lifted) sounds quite odd when played backward on a tape recorder. The sound suggests an organ more than a piano, slowly building up to an abrupt termination that gives no percussive impression. Moreover, if the individual notes of the chord are struck in rapid succession rather than simultaneously, their order is much more difficult to determine in the temporally reversed case than in the normal, forward presentation.

A sound of two hands clapping in a room is quite natural. In a normal-size room, the listener will hardly notice the reverberation; but in larger rooms it becomes noticeable, and the listener may think of the sounds only as indicating a large room, not a long reverberation. Completely unnatural sounds be can be created by mechanical manipulations using tape or a digital computer. A sound can be reversed, then reverberation added, then the resultant sound can be reversed. This generates a sound where the reverberation precedes the sound, but the sound itself still progresses forward. Speech processed in this fashion becomes extremely difficult to understand. This is because we are used to processing speech sounds in reverberant environments but are completely unfamiliar with an environment that would cause reverberation to come before a sound.

3.5 Perceptual Completion

Another fundamental principle of perception is called *perceptual completion*. Sometimes we have incomplete information coming into our sensory systems. To infer what is going on, we have to do some amount of top-down processing in addition to the normal

bottom-up processing. We must complete the information to determine the most probable explanation for what is occurring in the real world that is consistent with the information presented to our senses. All of us can think of familiar examples of this from our own experience with camouflage, both in nature; with animals, insects, and birds, and in the artificial camouflage worn by humans. There are also many examples from the art world of the intentional use and manipulation of ambiguity and camouflage. Most famous perhaps are paintings by Bev Doolittle, such as her "Pintos on a Snowy Background," which depicts pinto horses against a snowy and rocky mountainside. Because of the patterning of the brown and white horse hair on the pintos, it is not easily distinguished from the background of brown rocks and white snow.

It is difficult to program a computer to correctly process ambiguous visual stimuli, because computers do not have the kind of real-world knowledge that humans have gained through evolution and learning. This knowledge allows us to make reasonable inferences about what is going on in the world, using only partial information. Figure 3.7 shows two (or more) objects, with one of the objects apparently covering part of the other object. The most probable explanation for the alignment of the objects is that the bar is one object that extends continuously under the disk. It is also possible that there are two shorter bars whose colors, alignments, and such just happen to coincide as shown in figure 3.8. But the simpler explanation is that it is a single bar. Research with young infants has shown that they, too, are sensitive to this type of environmental context, and if the disk of figure 3.8 is removed to reveal two bars, the infant registers surprise (as measured by breathing and heart rate increases). We will discuss more on early infant studies later in this chapter.

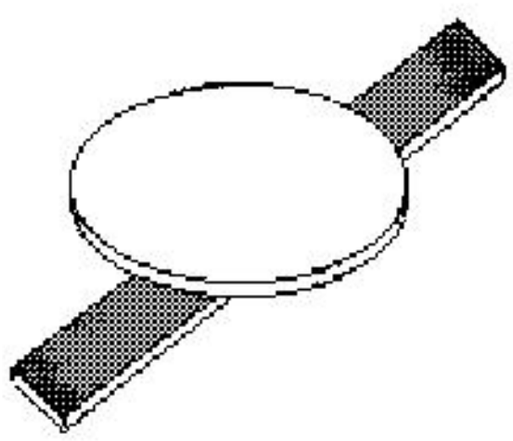

Figure 3.7 Continuation. We would normally assume one bar beneath the disk.

The Italian psychologist Gaetano Kanizsa has come up with a number of interesting examples of perceptual completion or *subjective contours.* Figure 3.9 demonstrates this phenomenon: it is difficult not to see a white triangle located at the center of the figure, although no such triangle actually exists. In the external world, the most probable cause for the improbable alignment of the objects is that a white triangle is lying on top of these objects, covering some and partially masking others.

This phenomenon can also be demonstrated in the auditory domain. Al Bregman has demonstrated this with sinusoidal tones that sweep up and down in frequency. These tones are interrupted with blank spaces, which cause quite obvious perceptual breaks. When the gaps are masked with bursts of white noise—just as the gaps in the inferred solid bar of figure 3.8 are masked by the disk—the listener makes the inference that the sinusoidal sweeps are continuous. The resulting perception is that a smoothly sweeping sinusoidal sound is occasionally covered up by noise bursts, not that the parts of the sinusoidal sound are actually replaced by bursts of noise, which is what is happening in the signal. The same thing can be done with music: the gaps sound like they are caused by a loose

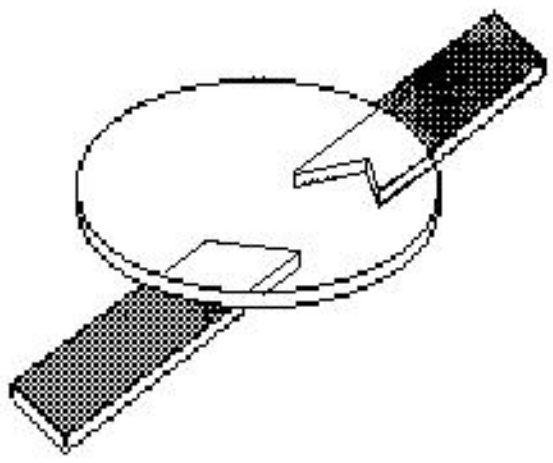

Figure 3.8 Another possible explanation of figure 3.7.

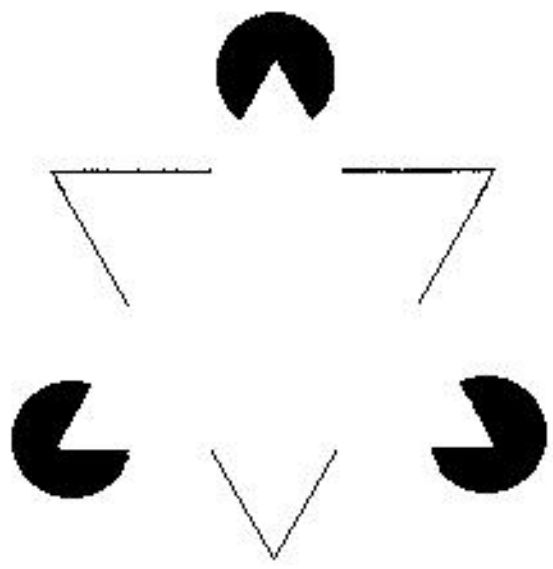

Figure 3.9 More continuation, and some symmetry.

connection in a circuit somewhere; but when the noise bursts fill in the gaps, the illusion is that the music continues throughout.

3.6 The Gestalt Grouping Principles

According to Max Wertheimer, one of the three principal founders of Gestalt psychology, Gestalt principles of grouping are used by the brain when parsing sensory input into objects in the world, especially when information is incomplete or missing altogether. Following are the Gestalt principles of grouping, which are all based on Helmholtz's concept of unconscious inference.

Proximity. Things that are located close together are likely to be grouped as being part of the same object. Figure 3.10a shows the principle of grouping by proximity.

Similarity. When objects are equally spaced, the ones that appear similar tend to be grouped as being related. If objects are similar in shape they are most probably related. (See figure 3.10b.)

Symmetry. Because random unrelated objects in the world are not expected to exhibit symmetry, it would be most improbable for unrelated objects to exhibit symmetric relationships. Figure 3.10c shows principles of both symmetry and similarity.

Good continuation. If objects are collinear, or arranged in such a way that it appears likely that they continue each other, they tend to be grouped perceptually. Figure 3.10d shows the principle of good continuation.

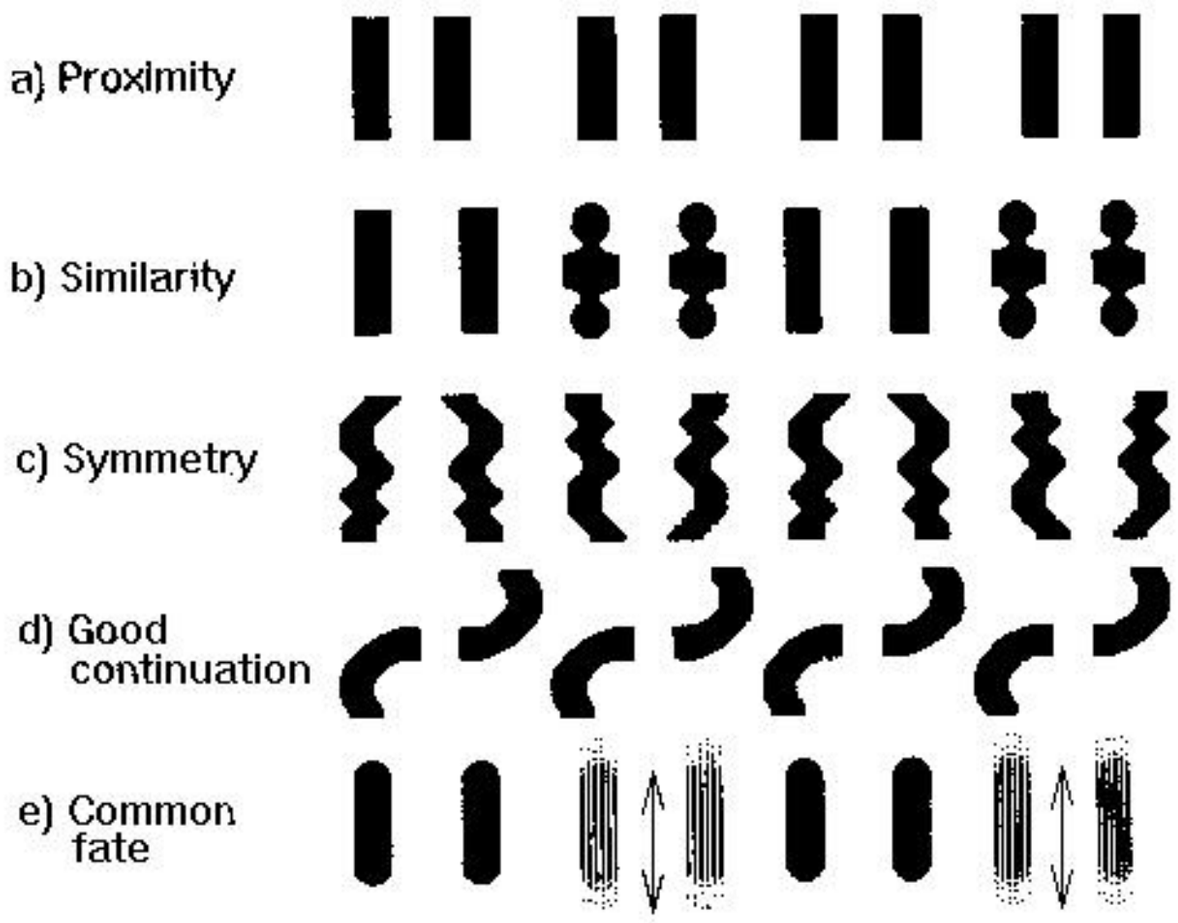

Figure 3.10 Gestalt grouping principles.

The principles of proximity, similarity, symmetry, and good continuation are considered *weak* principles of grouping, and are often used when the information is incomplete or noisy, or the perceiver has little to go on except the sensory input.

The principle of *common fate* (figure 3.10e) is much stronger. Common fate dictates that objects that move together are likely to be connected. In the world, it is extremely improbable that two things move in a perfectly correlated way unless they are in some way connected. For example, figure 3.11 shows a field of dots, and figure 3.12 shows another field of dots. If figure 3.12 is superimposed over figure 3.11 and moved back and forth, the face shape emerges from the random field of dots, made apparent by the fact that the dots that compose the face move together, and the others do not move.

Demonstrations of auditory common fate typically involve common onset time, common amplitude modulation, and common frequency modulation. One such example involves the grouping of partials and harmonics of a source: we are able to isolate the voice of a speaker or the musical line of a solo instrument in a complex auditory field. The task of isolating a sound source is essentially one of grouping the harmonics or partials that make up the sound; this is done by grouping those partials by the principle of common fate. The partials tend to move in ensemble, in both frequency and amplitude, and are thus recognized as being part of one object. Individual voices, even though they may be singing the same note, exhibit microfine deviations in pitch and amplitude that allow us to group the voices individually. This will be discussed in more detail in chapters 16 and 20.

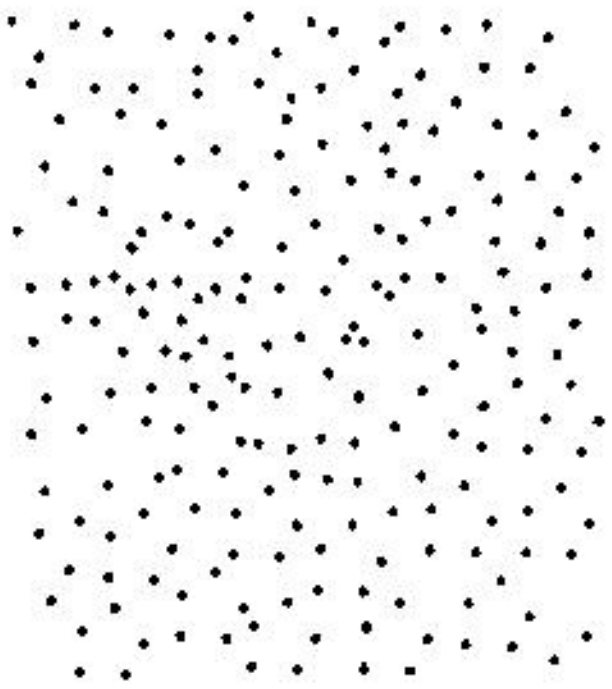

Figure 3.11 Common fate: some "random" dots. Photocopy figure 3.12 onto a transparency sheet, then lay it over figure 3.11. Slide the transparency slightly back and forth, and you will see a woman appear from the field of dots.

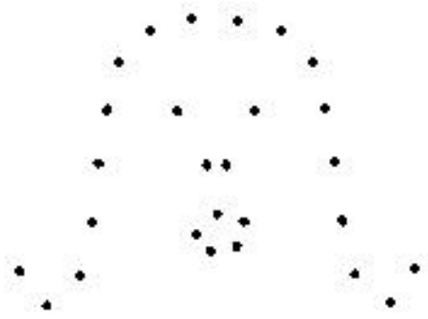

Figure 3.12 Some "random" dots. See figure 3.11.

Chowning's examples in chapter 20 will demonstrate grouping sound sources by common fate. One such demonstration involves a complex bell-like sound consisting of many inharmonic partials. The partials were computer-generated in such a way that they can be grouped into three sets of harmonic partials, each making up a female sung vowel spectrum. When the three voice sets are given a small amount of periodic and random pitch deviation (vibrato), the bell sound is transformed into the sound of three women singing. When the vibrato is removed, the three female voices merge again to form the original bell sound. This is another example of how common fate influences perception of sound.

As will be discussed in Chapter 16, there are styles of singing in which the vibrato is suppressed as much as possible. Such singing has quite a different effect than typical Western European singing; when the singers are successful in suppressing the vibrato to a sufficient extent, the chorus sound is replaced by a more instrumental timbre. The percept is not one of a group of singers but of a large, complex instrument.

The grouping principles discussed here are actually "wired into" our perceptual machinery. They do not have to be learned by trial and possibly fatal error, because they generally hold in the real world. For example, Elizabeth Spelky did work with early infant development and found that the principle of common fate is used by very young infants. She presented infants with displays of three-dimensional objects and moved some of them together. The infants registered surprise (measured physiologically) when they were shown that the objects that were moving together were not actually parts of the same object, but were artificially caused to move in synchrony. The infants were thus making an unconscious inference based on common fate and good continuation.

We have seen how the Gestalt grouping principle of common fate applies in both vision and audition. In later chapters we will explore

how some other Gestalt principles—those of similarity and proximity, for example—might apply to auditory stimuli, and in particular to musical events.

References

Bregman, A. S. (1990). *Auditory Scene Analysis: The Perceptual Organization of Sound.* Cambridge, Mass.: MIT Press. An excellent overview of many grouping principles in audio.

Flavell, J. H., and E. M. Markman, eds. (1983). *Cognitive Development.* New York: Wiley. A good book on general cognitive development, including early infant development.