Chapter 1

The Auditory Scene

Historical Difference between Auditory and Visual Perception

If you were to pick up a general textbook on perception written before 1965 and leaf through it, you would not find any great concern with the perceptual or ecological questions about audition. By a perceptual question I mean one that asks how our auditory systems could build a picture of the world around us through their sensitivity to sound, whereas by an ecological one I am referring to one that asks how our environment tends to create and shape the sound around us. (The two kinds of questions are related. Only by being aware of how the sound is created and shaped in the world can we know how to use it to derive the properties of the sound-producing events around us.)

Instead, you would find discussions of such basic auditory qualities as loudness and pitch. For each of these, the textbook might discuss the psychophysical question: which physical property of the sound gives rise to the perceptual quality that we experience? It might also consider the question of how the physiology of the ear and nervous system could respond to those properties of sound. The most perceptual of the topics that you might encounter would be concerned with how the sense of hearing can tell the listener where sounds are coming from. Under this heading, some consideration would be given to the role of audition in telling us about the world around us. For the most part, instead of arising from everyday life, the motivation of much of the research on audition seems to have its origins in the medical study of deafness, where the major concerns are the sensitivity of the auditory system to weak sounds, the growth in perceived intensity with increases in the energy of the signal, and the effects of exposure to noise.

The situation would be quite different in the treatment of vision. It is true that you would see a treatment of psychophysics and physiology, and indeed there would be some consideration of such deficits as colorblindness, but this would not be the whole story. You would also find discussions of higher-level principles of organization, such

as those responsible for the constancies. There would, for example, be a description of size constancy, the fact that we tend to see the size of an object as unchanged when it is at a different distance, despite the fact that the image that it projects on our retinas shrinks as it moves further away. Apparently some complex analysis by the brain takes into account clues other than retinal size in arriving at the perceived size of an object.

Why should there be such a difference? A proponent of the "great man" theory of history might argue that it was because the fathers of Gestalt psychology, who opened up the whole question of perceptual organization, had focused on vision and never quite got around to audition.

However, it is more likely that there is a deeper reason. We came to know about the puzzles of visual perception through the arts of drawing and painting. The desire for accurate portrayal led to an understanding of the cues for distance and certain facts about projective geometry. This was accompanied by the development of the physical analysis of projected images, and eventually the invention of the camera. Early on, the psychologist was faced with the discrepancy between what was on the photograph or canvas and what the person saw.

The earlier development of sophisticated thinking in the field of visual perception may also have been due to the fact that it was much easier to create a visual display with exactly specified properties than it was to shape sound in equally exact ways. If so, the present-day development of the computer analysis and synthesis of sound ought to greatly accelerate the study of auditory perception.

Of course there is another possibility that explains the slighting of audition in the textbook: Perhaps audition is really a much simpler sense and there are no important perceptual phenomena like the visual constancies to be discovered.

This is a notion that can be rejected. We can show that such complex phenomena as constancies exist in hearing, too. One example is timbre constancy. A friend's voice has the same perceived timbre in a quiet room as at a cocktail party. Yet at the party, the set of frequency components arising from that voice is mixed at the listener's ear with frequency components from other sources. The total spectrum of energy that reaches the ear may be quite different in different environments. To recognize the unique timbre of the voice we have to isolate the frequency components that are responsible for it from others that are present at the same time. A wrong choice of frequency components would change the perceived timbre of the voice. The fact that we can usually recognize the timbre implies that we regularly choose

the right components in different contexts. Just as in the case of the visual constancies, timbre constancy will have to be explained in terms of a complicated analysis by the brain, and not merely in terms of a simple registration of the input by the brain.

There are some practical reasons for trying to understand this constancy. There are engineers currently trying to design computers that can understand what a person is saying. However, in a noisy environment the speaker's voice comes mixed with other sounds. To a naive computer, each different sound that the voice comes mixed with makes it sound as if different words were being spoken or as if they were spoken by a different person. The machine cannot correct for the particular listening conditions as a human can. If the study of human audition were able to lay bare the principles that govern the human skill, there is some hope that a computer could be designed to mimic it.

The Problem of Scene Analysis

It is not entirely true that textbooks ignore complex perceptual phenomena in audition. However, they are often presented as an array of baffling illusions.[1] They seem more like disconnected fragments than a foundation for a theory of auditory perception. My purpose in this book is to try to see them as oblique glimpses of a general auditory process of organization that has evolved, in our auditory systems, to solve a problem that I will refer to as "auditory scene analysis."

Let me clarify what I mean by auditory scene analysis. The best way to begin is to ask ourselves what perception is for. Since Aristotle, many philosophers and psychologists have believed that perception is the process of using the information provided by our senses to form mental representations of the world around us. In using the word representations, we are implying the existence of a two-part system: one part forms the representations and another uses them to do such things as calculate appropriate plans and actions. The job of perception, then, is to take the sensory input and to derive a useful representation of reality from it.

An important part of building a representation is to decide which parts of the sensory stimulation are telling us about the same environmental object or event. Unless we put the right combination of sensory evidence together, we will not be able to recognize what is going on. A simple example is shown in the top line of figure 1.1. The pattern of letters is meaningful, but the meaning cannot be extracted because the letters are actually a mixture from two sentences, and

Figure 1.1
Top line: a string of letters that makes no sense because it is a mixture of two messages. Bottom line: the component messages are segregated by visual factors. (From Bregman 1981b.)

the two cannot be separated. However, if, as in the lower line of the figure, we give the eyes some assistance, the meaning becomes apparent.

This business of separating evidence has been faced in the design of computer systems for recognizing the objects in natural scenes or in drawings. Figure 1.2 shows a line drawing of some blocks.[2] We can imagine that the picture has been translated into a pattern in the memory of the computer by some process that need not concern us. We might think that once it was entered, all that we would have to do to enable the computer to decide which objects were present in the scene would be to supply it with a description of the shape of each possible one. But the problem is not as easy as all that. Before the machine could make any decision, it would have to be able to tell which parts of the picture represented parts of the same object. To our human eyes it appears that the regions labeled A and B are parts of a single block. This is not immediately obvious to a computer. In simple line drawings there is a rule that states that any white area totally surrounded by lines must depict a single surface. This rule implies that in figure 1.2 the whole of region A is part of a single surface. The reason for grouping region A with B is much more complex. The question of how it can be done can be set aside for the moment. The point of the example is that unless regions A and B are indeed considered part of a single object, the description that the computer will be able to construct will not be correct and the elongated shape formed out of A, B, and other regions will not be seen. It seems as though a preliminary step along the road to recognition would be to program the computer to do the equivalent of taking a set of crayons and coloring in, with the same color, all those regions that were parts of the same block. Then some subsequent recognition process could simply try to form a description of a single shape from each set in which the regions were the same color. This allocation of regions to objects is what is known to researchers in machine vision as the scene analysis problem.

There are similar problems in hearing. Take the case of a baby being spoken to by her mother. The baby starts to imitate her

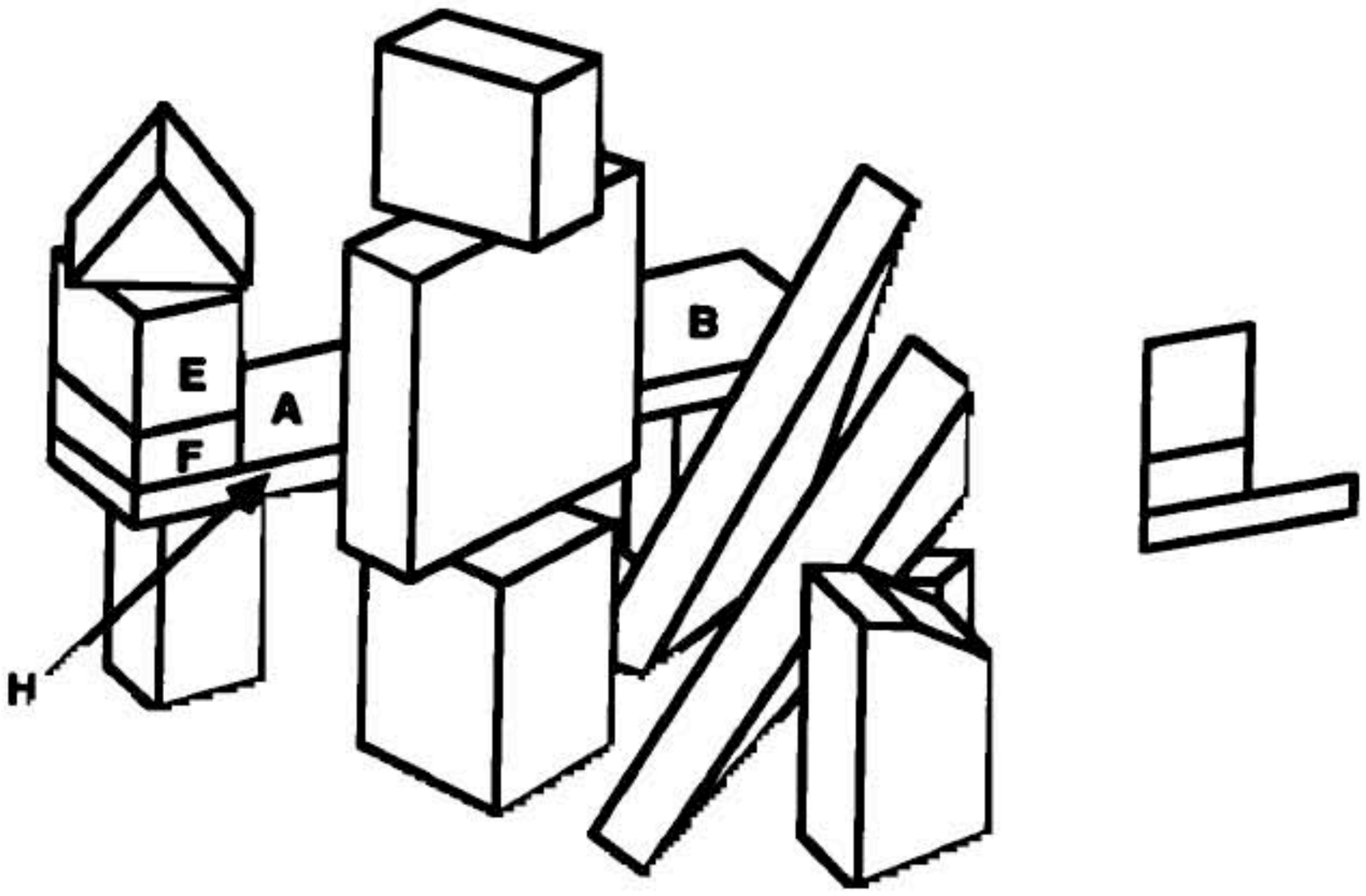

Figure 1.2
A line drawing of blocks for visual scene analysis. (After Guzman 1969.)

mother's voice. However, she does not insert into the imitation the squeaks of her cradle that have been occurring at the same time. Why not? A physical record of what she has heard would include them. Somehow she has been able to reject the squeak as not being part of the perceptual "object" formed by her mother's voice. In doing so, the infant has solved a scene analysis problem in audition.

It is important to emphasize again that the way that sensory inputs are grouped by our nervous systems determines the patterns that we perceive. In the case of the drawings of blocks, if areas E, F, and H were grouped as parts of the same object, we would see the L-shaped object shown at the right. The shape of the object formed by this grouping of areas is an emergent property, since it is not a property of any of the parts taken individually, but emerges only as a result of the grouping of the areas. Normally, in perception, emergent properties are accurate portrayals of the properties of the objects in our environment. However, if scene analysis processes fail, the emergent perceived shapes will not correspond to any environmental shapes. They will be entirely chimerical.

The difficulties that are involved in the scene analysis processes in audition often escape our notice. This example can make them more obvious. Imagine that you are on the edge of a lake and a friend challenges you to play a game. The game is this: Your friend digs two narrow channels up from the side of the lake. Each is a few feet long and a few inches wide and they are spaced a few feet apart. Halfway up each one, your friend stretches a handkerchief and fastens it to the sides of the channel. As waves reach the side of the lake they travel up the channels and cause the two handkerchiefs to go into motion. You

are allowed to look only at the handkerchiefs and from their motions to answer a series of questions: How many boats are there on the lake and where are they? Which is the most powerful one? Which one is closer? Is the wind blowing? Has any large object been dropped suddenly into the lake?

Solving this problem seems impossible, but it is a strict analogy to the problem faced by our auditory systems. The lake represents the lake of air that surrounds us. The two channels are our two ear canals, and the handkerchiefs are our ear drums. The only information that the auditory system has available to it, or ever will have, is the vibrations of these two ear drums. Yet it seems to be able to answer questions very like the ones that were asked by the side of the lake: How many people are talking? Which one is louder, or closer? Is there a machine humming in the background? We are not surprised when our sense of hearing succeeds in answering these questions any more than we are when our eye, looking at the handkerchiefs, fails.

The difficulty in the examples of the lake, the infant, the sequence of letters, and the block drawings is that the evidence arising from each distinct physical cause in the environment is compounded with the effects of the other ones when it reaches the sense organ. If correct perceptual representations of the world are to be formed, the evidence must be partitioned appropriately.

In vision, you can describe the problem of scene analysis in terms of the correct grouping of regions. Most people know that the retina of the eye acts something like a sensitive photographic film and that it records, in the form of neural impulses, the "image" that has been written onto it by the light. This image has regions. Therefore, it is possible to imagine some process that groups them. But what about the sense of hearing? What are the basic parts that must be grouped to make a sound?

Rather than considering this question in terms of a direct discussion of the auditory system, it will be simpler to introduce the topic by looking at a spectrogram, a widely used description of sound. Figure 1.3 shows one for the spoken word "shoe". The picture is rather like a sheet of music. Time proceeds from left to right, and the vertical dimension represents the physical dimension of frequency, which corresponds to our impression of the highness of the sound. The sound of a voice is complex. At any moment of time, the spectrogram shows more than one frequency. It does so because any complex sound can actually be viewed as a set of simultaneous frequency components. A steady pure tone, which is much simpler than a voice, would simply be shown as a horizontal line because at any moment it would have only one frequency.

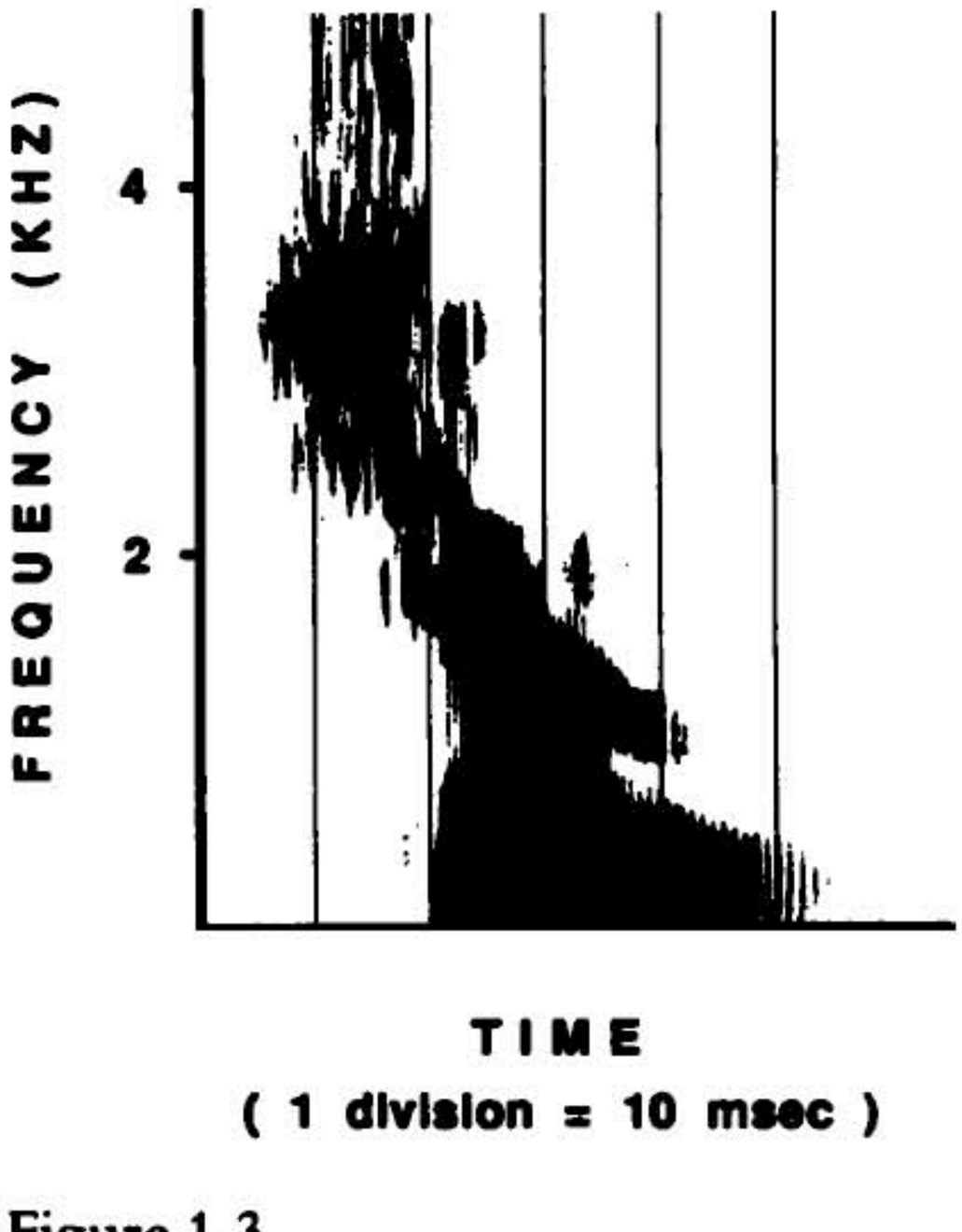

Figure 1.3
Spectrogram of the word "shoe" spoken in isolation.

Once we see that the sound can be made into a picture, we are tempted to believe that such a picture could be used by a computer to recognize speech sounds. Different classes of speech sounds, stop consonants such as "b" and fricatives such as "s" for example, have characteristically different appearances on the spectrogram. We ought to be able to equip the computer with a set of tests with which to examine such a picture and to determine whether the shape representing a particular speech sound is present in the image. This makes the problem sound much like the one faced by vision in recognizing the blocks in figure 1.2.

If a computer could solve the recognition problem by the use of a spectrogram, it would be very exciting news for researchers in human audition, because there is some reason to believe that the human auditory system provides the brain with a pattern of neural excitation that is very much like a spectrogram. Without going into too much detail, we can sketch this process as follows. As sound enters the ear, it eventually reaches a part called the inner ear where it affects an organ called the basilar membrane, a long coiled ribbon. Different frequency components in the incoming sound will cause different parts of this organ to vibrate most vigorously. It reacts most strongly to the lowest audible frequencies at one end, to the highest at the other, with an orderly progression from low to high in between. A different group of neurons connects with each location along the basilar membrane and is responsible for recording the vibration at that

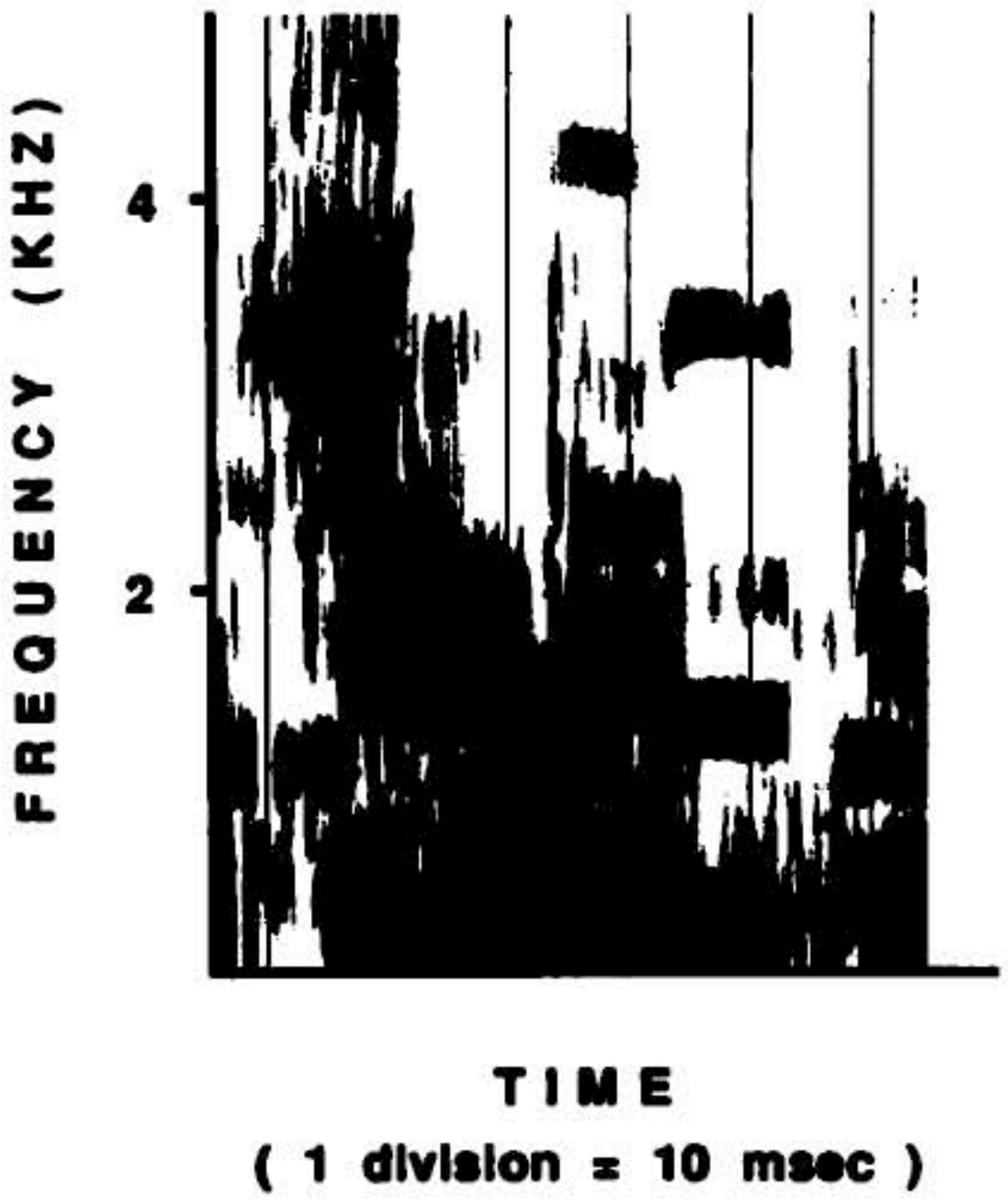

Figure 1.4
A spectrogram of a mixture of sounds (containing the word "shoe").

location (primarily). As the sound changes over time, different combinations of neural groups are activated. If we imagined the basilar membrane oriented vertically so that the neural groups responsive to the highest frequencies were at the top, and also imagined that each group was attached to a pen, with the pen active whenever a neural group was, the pens would write out a picture of the sound that looked like a spectrogram. So the brain has all the information that is visible in the spectrogram, and providing that it could store a record of this information for some brief period of time, it would have a neural spectrogram.

The account that I have just given hides a deep problem. The spectrographic record of most situations would not have the pristine purity of figure 1.3, which represents speech recorded in an absolutely quiet background. The real world is a great deal messier. A typical acoustic result is shown in figure 1.4. Here all the sounds are being mixed together in the listener's ear in exactly the same way that the waves of the lake, in our earlier example, were mixed in each of the channels that ran off it. The spectrogram for a mixture of sounds looks somewhat like a picture created by making a spectrogram of each of the individual sounds on a separate piece of transparent plastic, and then overlaying the individual spectrograms to create a composite. The spectrogram of the word shoe is actually one of the component spectrograms of the mixture.

Although the theorist has the privilege of building the composite up from the pictures of its components, the auditory system, or any machine trying to imitate it, would be presented only with the spectrogram of the mixture and would have to try to infer the set of pictures that was overlaid to produce it.

The recognizer would have to solve the following problems: How many sources have created the mixture? Is a particular discontinuity in the picture a change in one sound or an interruption by a second one? Should two dark regions, one above the other in the picture (in other words, occurring at the same time), be grouped as a single sound with a complex timbre or separated to represent two simultaneous sounds with simpler timbres? We can see that if we look at a spectrogram representing a slice of real life, we would see a complex pattern of streaks, any pair of which could have been caused by the same acoustic event or by different ones. A single streak could have been the summation of one, two, or even more parts of different sounds. Furthermore, the frequency components from one source could be interlaced with those of another one; just because one horizontal streak happens to be immediately above another, it does not mean that they both arose from the same sonic event.

We can see that just as in the visual problem of recognizing a picture of blocks, there is a serious need for regions to be grouped appropriately. Again, it would be convenient to be able to hand the spectrogram over to a machine that did the equivalent of taking a set of crayons and coloring in, with the same color, all the regions on the spectrogram that came from the same source. This "coloring problem" or "auditory scene analysis problem" is what the rest of this volume is about.

Objects Compared to Streams

It is also about the concept of "auditory streams." An auditory stream is our perceptual grouping of the parts of the neural spectrogram that go together. To see the reasons for bringing in this concept, it is necessary to consider the relations between the physical world and our mental representations of it. As we saw before, the goal of scene analysis is the recovery of separate descriptions of each separate thing in the environment. What are these things? In vision, we are focused on objects. Light is reflected off objects, bounces back and forth between them, and eventually some of it reaches our eyes. Our visual sense uses this light to form separate descriptions of the individual objects. These descriptions include the object's shape, size, distance, coloring, and so on.

Then what sort of information is conveyed by sound? Sound is created when things of various types happen. The wind blows, an animal scurries through a clearing, the fire burns, a person calls. Acoustic information, therefore, tells us about physical "happenings." Many happenings go on at the same time in the world, each one a distinct event. If we are to react to them as distinct, there has to be a level of mental description in which there are separate representations of the individual ones.

I refer to the perceptual unit that represents a single happening as an auditory stream. Why not just call it a sound? There are two reasons why the word stream is better. First of all a physical happening (and correspondingly its mental representation) can incorporate more than one sound, just as a visual object can have more than one region. A series of footsteps, for instance, can form a single experienced event, despite the fact that each footstep is a separate sound. A soprano singing with a piano accompaniment is also heard as a coherent happening, despite being composed of distinct sounds (notes). Furthermore, the singer and piano together form a perceptual entity—the "performance"—that is distinct from other sounds that are occurring. Therefore, our mental representations of acoustic events can be multifold in a way that the mere word "sound" does not suggest. By coining a new word, "stream", we are free to load it up with whatever theoretical properties seem appropriate.

A second reason for preferring the word "stream" is that the word "sound" refers indifferently to the physical sound in the world and to our mental experience of it. It is useful to reserve the word "stream" for a perceptual representation, and the phrase "acoustic event" or the word "sound" for the physical cause.

I view a stream as a computational stage on the way to the full description of an auditory event. The stream serves the purpose of clustering related qualities. By doing so, it acts as a center for our description of an acoustic event. By way of analogy, consider how we talk about visible things. In our verbal descriptions of what we see, we say that an *object* is red, or that it is moving fast, that it is near, or that it is dangerous. In other words, the notion of an object, understood whenever the word "it" occurs in the previous sentence, serves as a center around which our verbal descriptions are clustered. This is not just a convenience of language. The perceptual representation of an object serves the same purpose as the "it" in the sentence. We can observe this when we dream. When, for some reason, the ideas of angry and dog and green are pulled out from our memories, they tend to coalesce into a single entity and we experience an angry green

dog and not merely anger, greenness, and dogness taken separately. Although the combination of these qualities has never occurred in our experience, and therefore the individual qualities must have been dredged up from separate experiences, those qualities can be experienced visually only as properties of an *object*. It is this "belonging to an object" that holds them together.

The stream plays the same role in auditory mental experience as the object does in visual. When we want to talk about auditory units (the auditory counterparts of visual objects), we generally employ the word "sound". We say that a sound is high pitched or low, that it is rising or falling, that it is rough or smooth, and so on. Again I am convinced that this is not simply a trick of language, but an essential aspect of both our conceptual and our perceptual representations of the world. Properties have to belong to something. This becomes particularly important when there is more than one "something" in our experience. Suppose there are two acoustic sources of sound, one high and near and the other low and far. It is only because of the fact that nearness and highness are grouped as properties of one stream and farness and lowness as properties of the other that we can experience the uniqueness of the two individual sounds rather than a mush of four properties.

A critic of this argument might reply that the world itself groups the "high" with the "near" and the "low" with the "far". It is not necessary for us to do it. However, it is not sufficient that these clusters of properties be distinct in the physical happenings around us. They must also be assigned by our brains to distinct mental entities. In auditory experience, these entities are the things that I am calling streams. As with our visual experience of objects, our auditory streams are ways of putting the sensory information together. This going together has obvious implications for action. For example, if we assign the properties "far" and "lion roar" to one auditory stream and the properties "near" and "crackling fire" to another one, we might be inclined to behave differently than if the distance assignments had been reversed.

When people familiar with the English language read the phrase "The gray wagon was on the black road", they know immediately that it is the wagon that is gray, not the road. They know it because they can *parse* the sentence, using their knowledge of English syntax to determine the correct "belongingness" relations between the concepts. Similarly, when listeners create a mental representation of the auditory input, they too must employ rules about what goes with what. In some sense, they can be said to be parsing this input too.

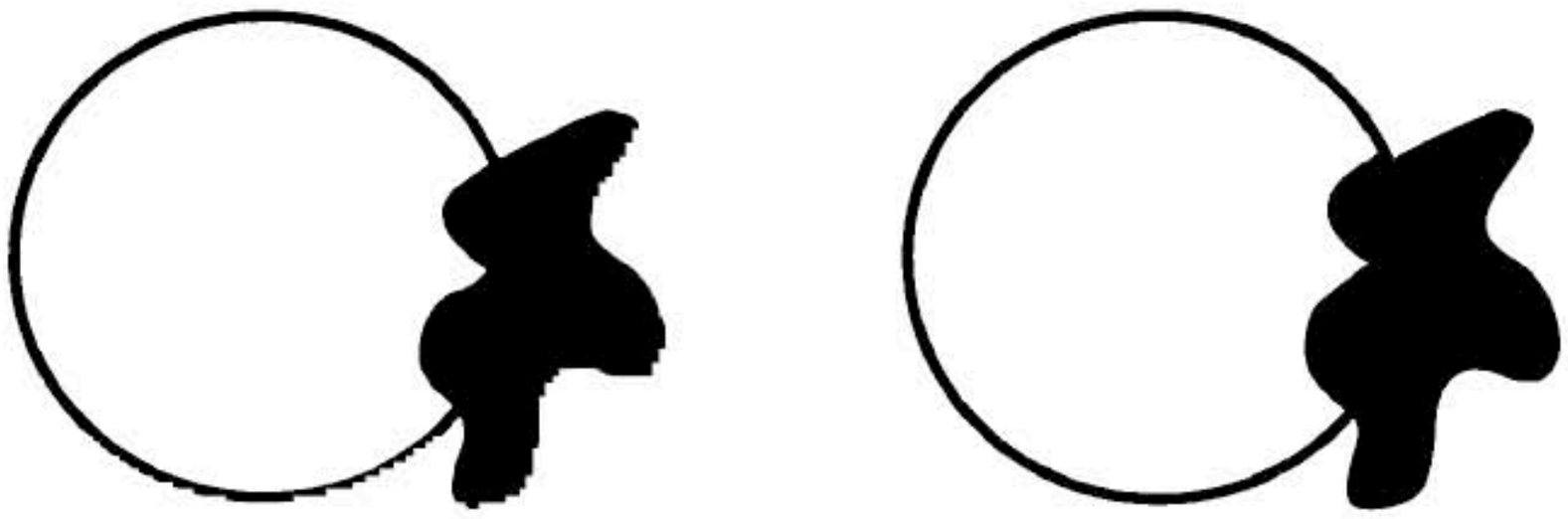

Figure 1.5
An example of "belongingness." The dark portion of the line seems to belong to the irregular form.

The Principle of Exclusive Allocation

Any system that attempts to build descriptions of a natural world scene must assign the perceptual qualities that it creates to one organization or another. The quality "loud" is assigned to the organization that represents the roar of the lion. The quality "far" is assigned as the distance of that same event. The Gestalt psychologists made this point by introducing the principle of belongingness. In describing the visual organization of drawings like the one in figure 1.5, they pointed out that the lines at which the drawn irregular figure overlaps the circle (shown as a dark line in part B of the figure) are generally seen as part of the irregular figure and not of the circle. That is, they *belong* to the irregular form. With an effort, we can see them as part of a circle; then they belong to the circle. In any mental representation of a drawing, a perceived line always belongs to some figure of which it forms a part. The belongingness may shift, for example, when we try to see the figure in a different way, but regardless of how we see it, it is always a property *of* something.

There is a second principle that I want to introduce here because it has a connection with the principle of belongingness. This is the principle of "exclusive allocation." It can be seen in an ambiguous visual figure such as the vase-faces illusion of the Gestalt psychologists. An example is shown in figure 1.6. We can interpret the figure as an outline of either a vase or two faces. The "exclusive allocation of evidence" describes how these interpretations affect the edge that separates the vase from a face. When we see the vase, that edge is allocated to the vase and defines its shape. When we see the face, the same edge is now allocated to the face. It is never allocated to both vase and face at the same time, but exclusively to one of them.

The exclusive allocation principle says that a sensory element should not be used in more than one description at a time. If the line is assigned to the vase, that assignment "uses up" the line so that its

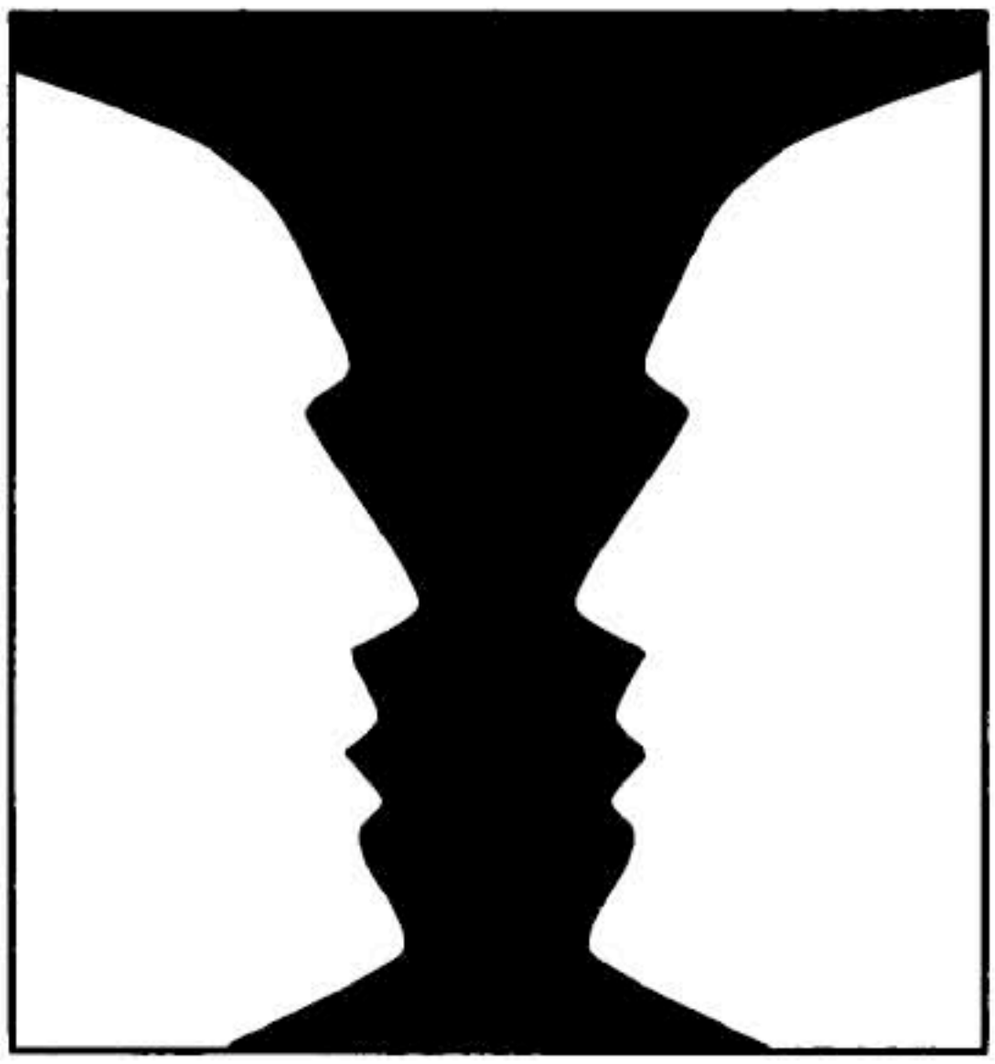

Figure 1.6
An ambiguous drawing in which either a vase at the center or two faces at the sides can be seen.

shape cannot contribute to the shape of another figure at the same time. We shall eventually see in chapter 7 that there are certain limits to this idea, but it holds true often enough that it is worth pointing it out as a separate principle. It is not identical to the principle of belongingness. The latter merely states that the line has to be seen as a property of a figure, but does not prevent it from being allocated to more than one at a time.

There is a certain ecological validity of the principle of exclusive allocation in vision. The term "ecological validity" means that it tends to give the right answers about how the visual image has probably originated in the external world. In the case of edges separating objects, there is a very low likelihood (except in jigsaw puzzles) that the touching edges of two objects will have the same shape exactly. Therefore the shape of the contour that separates our view of two objects probably tells us about the shape of only one of them—the nearer one. The decision as to which object the contour belongs to is determined by a number of cues that help the viewer to judge which object is closer.

Dividing evidence between distinct perceptual entities (visual objects or auditory streams) is useful because there really are distinct physical objects and events in the world that we humans inhabit. Therefore the evidence that is obtained by our senses really ought to be untangled and assigned to one or another of them.

Our initial example came from vision, but the arguments in audition are similar. For example, it is very unlikely that a sound will

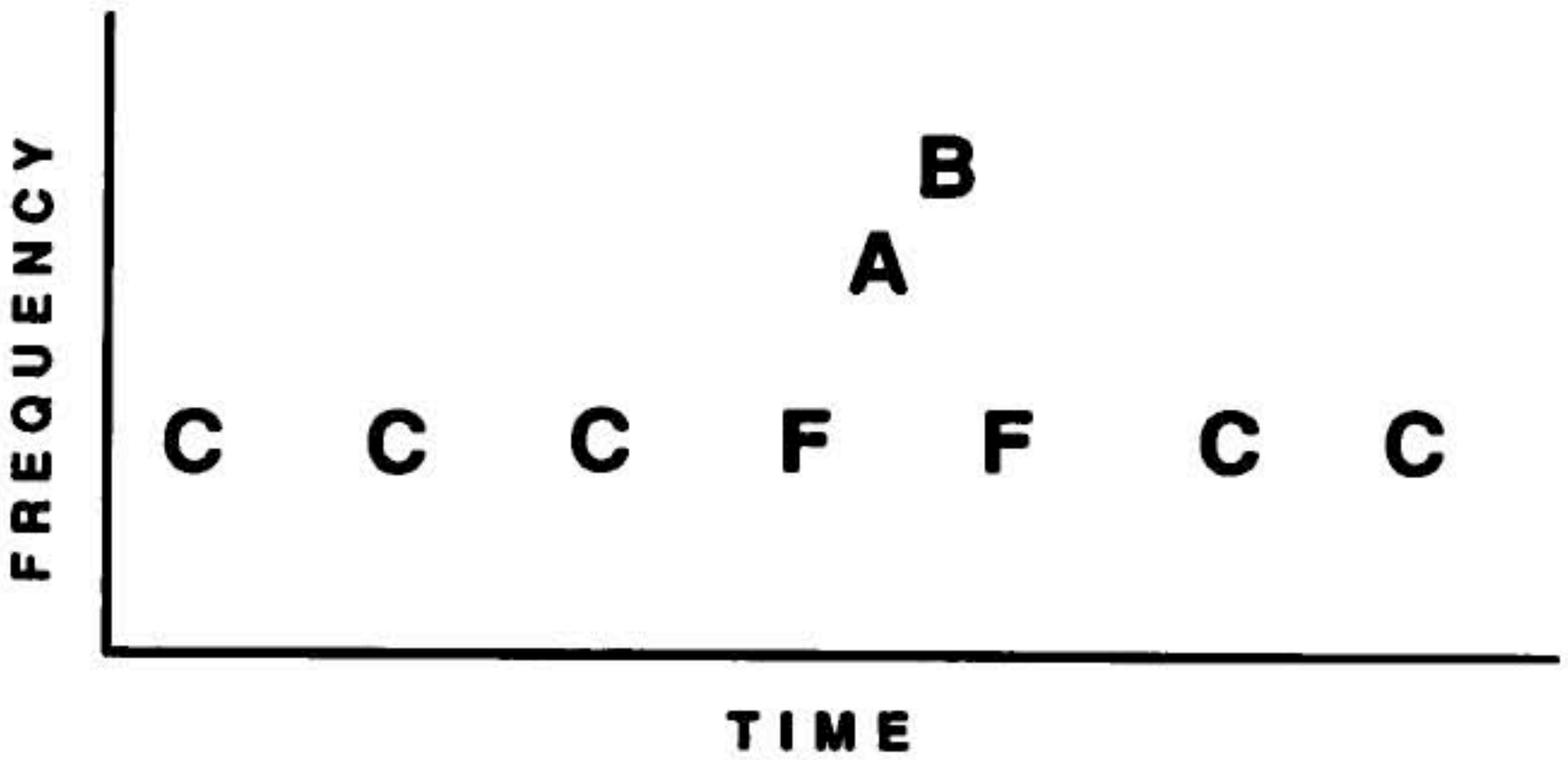

Figure 1.7
A tone sequence of the type used by Bregman and Rudnicky (1975).

terminate at exactly the moment that another begins. Therefore when the spectral composition of the incoming sensory data changes suddenly, the auditory system can conclude that only one sound in a mixture has gone on or off. We will see in chapter 3 that this conclusion can give rise to a search in the second sound for a continuation of the first one.

The strategy completes itself in the following way. Let us give the name A to the segment of sound that occurs prior to the change, and call the second part B. If spectral components are found in B that match the spectrum of A, they are considered to be the continuing parts of A. Accordingly, they can be subtracted out of B. This allows us a picture of the second sound free from the influence of the first. In chapter 3, this is called the "old-plus-new heuristic," and it is shown to be one of our most powerful tools in solving the scene analysis problem in audition. Here I want to point out that it is an example of the principle of exclusive allocation in which the allocation of the continuing spectral components to the first sound interferes with their being allocated to the second.

Another case of exclusive allocation is shown in an experiment by Bregman and Rudnicky, using the pattern of pure tones shown in figure 1.7.[3] In this figure the horizontal dimension represents time and the vertical one shows the frequency of the tones. The listener's task was to decide on the order of two target tones, A and B, embedded in the sequence. Were they in the order high-low or low-high? When A and B were presented alone, as an isolated pair of tones, this decision was very easy. However, when the two tones labeled F (for "flankers") were added to the pattern, the order of A and B became very hard to hear. Apparently when they were absorbed as

the middle elements of a larger pattern, FABF, the orders AB and BA lost their uniqueness.

This experiment was about the perceptual allocation of the F tones. As long as they were allocated to the same auditory stream as A and B, the order of A and B was hard to hear. However, Bregman and Rudnicky reasoned that if some principle of grouping were able to assign the F tones to a different perceptual stream, the order of A and B might become audible again. With this in mind, they introduced yet another group of tones, labeled C (for "captors") in figure 1.7. They varied the frequency of these C tones. When they were very low, much lower than the frequency of the F tones, the F tones grouped with the AB tones and the order of A and B was unclear to the listeners. However, when the C tones were brought up close to the frequency of the F tones, they captured them into a stream, CCCFFCC. One reason for this capturing is that tones tend to group perceptually with those that are nearest to them in frequency; a second is that the F tones were spaced so that they fell into a regular rhythmic pattern with the C tones. When the capturing occurred, the order of AB was heard more clearly because they were now in their own auditory stream that was separate from the CCCFFCC stream. The belongingness of the F tones had been altered, and the perceived auditory forms were changed.

Scene analysis, as I have described it, involves putting evidence together into a structure. Demonstrations of the perceptual systems acting in this way are seen in certain kinds of illusions where it appears that the correct features of the sensory input have been detected but have not been put together correctly. Two examples will make this clearer.

The first is in vision. Treisman and Schmidt carried out an experiment in which a row of symbols was flashed briefly in a tachistoscope.[4] There were three colored letters flanked by two black digits. The viewers were asked to first report what the digits were and then to report on the letters. Their reports of the digits were generally correct, but the properties of the letters were often scrambled. A subject might report a red O and a green X, when actually a green O and a red X had been presented. These combinations of features often seemed to the viewers to be their actual experiences rather than merely guesses based on partially registered features of the display. The experimenters argued that this showed that the human mind cannot consciously experience disembodied features and must assign them to perceived objects. That is, the mind obeys the principle of belongingness.

The second example comes from audition. In 1974, Diana Deutsch reported an interesting illusion that could be created when tones were sent to both ears of a listener over headphones. The listener was presented with a continuously repeating alternation of two events. Event A was a low tone presented to the left ear, accompanied by a high tone presented to the right ear. Event B was just the reverse: a low tone to the right ear together with a high tone to the left. The high and low tones were pure sine wave tones spaced exactly an octave apart. Because events A and B alternated, each ear was presented with a sequence of high and low tones. Another way to express it is that while both the high and low tones tones bounced back and forth between the ears, the high and low were always in opposite ears.

However the experience of many listeners did not resemble this description. Instead they heard a single sound bouncing back and forth between the ears. Furthermore, the perceived tone alternated between sounding high pitched and sounding low as it bounced from side to side. The only way this illusion could be explained was to argue that the listeners were assuming the existence of a single tone, deriving two different descriptions of it from two different types of perceptual analyses, and then putting the two descriptions together incorrectly. Apparently they derived the fact that the tone was changing in frequency by monitoring the changes in a single ear (usually the right). However, they derived the *position* of the assumed single sound by tracking the position of the higher tone. Therefore, they might report hearing a low tone on the left at the point in time at which, in actuality, a high tone had been presented on the left. Here we see an example of pitch and location assigned in the wrong combination to the representation of a sound. Therefore, this can be classified as a misassignment illusion just as Treisman and Schmidt's visual illusion was.

The question of why this illusion occurs can be set aside for the moment. What is important is that the illusion suggests that an assignment process is taking place, and this supports the idea that perception is a process of building descriptions. Only by being built could they be built incorrectly.

These illusions show that there are some similarities in how visual and auditory experiences are organized. A thoughtful discussion of the similarities and differences between vision and audition can be found in a paper by Bela Julesz and Ira Hirsh.[5] There is no shortage of parallels in audition to visual processes of organization. This chapter cannot afford the space to mention many examples, but it can at least discuss two of them, the streaming phenomenon and the continuity illusion.

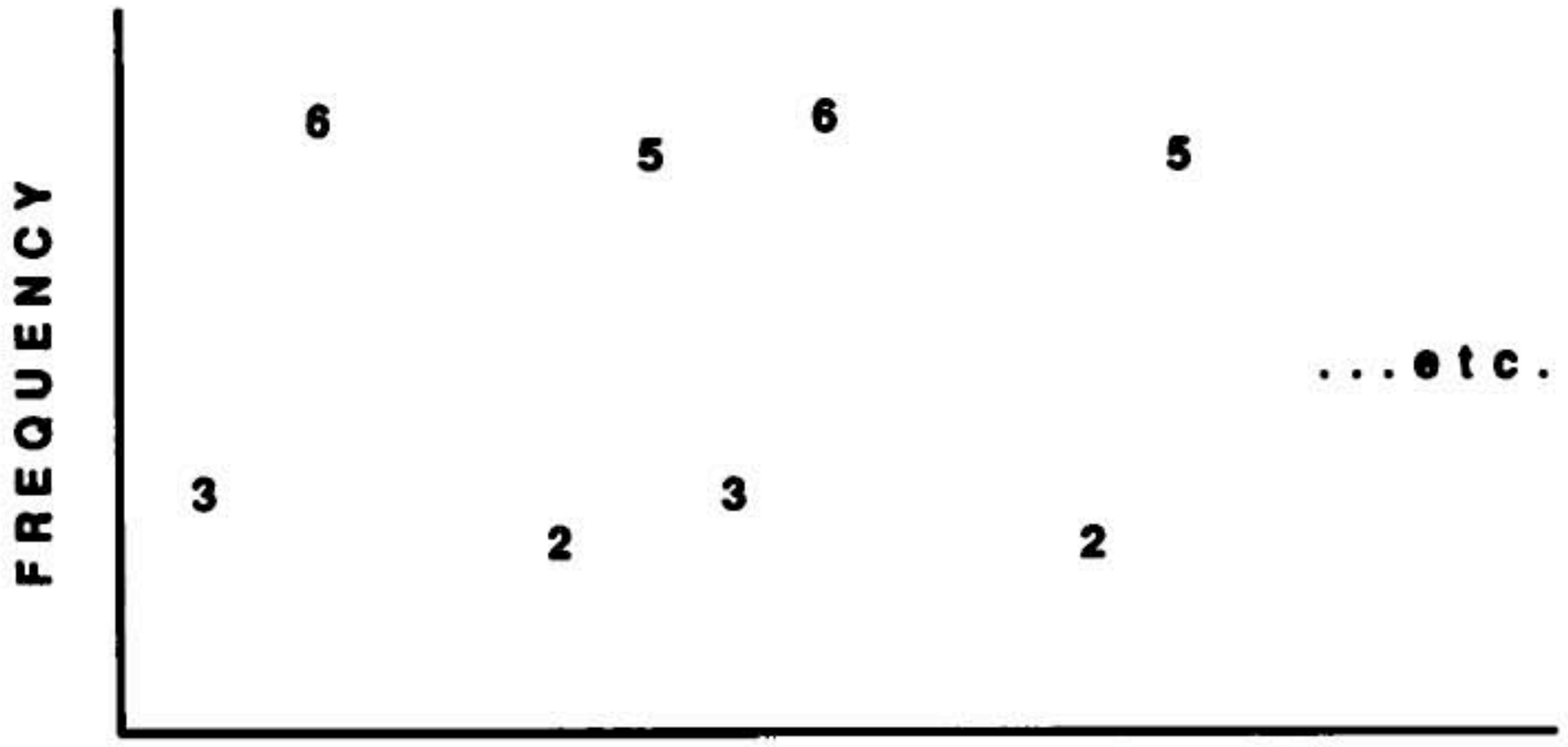

Figure 1.8
A repeating cycle of six tones, of the type used by Bregman and Campbell (1971).

Two Comparisons of Scene Analysis in Vision and Audition

Auditory Streaming and Apparent Motion

One auditory phenomenon with a direct parallel in vision is the auditory streaming effect. This is the phenomenon that originally got me interested in auditory organization. The effect occurred when listeners were presented with an endlessly repeating loop of tape on which were recorded a sequence of six different tones, three high ones and three low ones. The high ones were at least one and a half octaves above the low ones. High and low tones alternated. If tones are given numbers according to their pitches with 1 as the lowest and 6 as the highest the tones were arranged in the sequence 142536. The six tones, shown in figure 1.8, formed a repeating loop that was cycled over and over.

When the cycle of tones was presented very slowly the listeners heard the sequence of high and low tones in the order in which they occurred on the tape. However, as it was made faster, a strange perceptual effect became stronger and stronger and was extremely compelling when there was only one-tenth of a second between the onsets of consecutive tones. When the effect occurred, the listeners did not actually hear the tones in the correct order, 142536. Instead, they heard two streams of tones, one containing a repeating cycle of the three low pitched tones, 1–2–3– (where dashes indicate silences) and the other containing the three high ones (–4–5–6). The single sequence of tones seemed to have broken up perceptually into two parallel sequences, as if two different instruments were playing different, but interwoven parts. Furthermore it was impossible for the listeners to focus their attention on both streams at the same time.

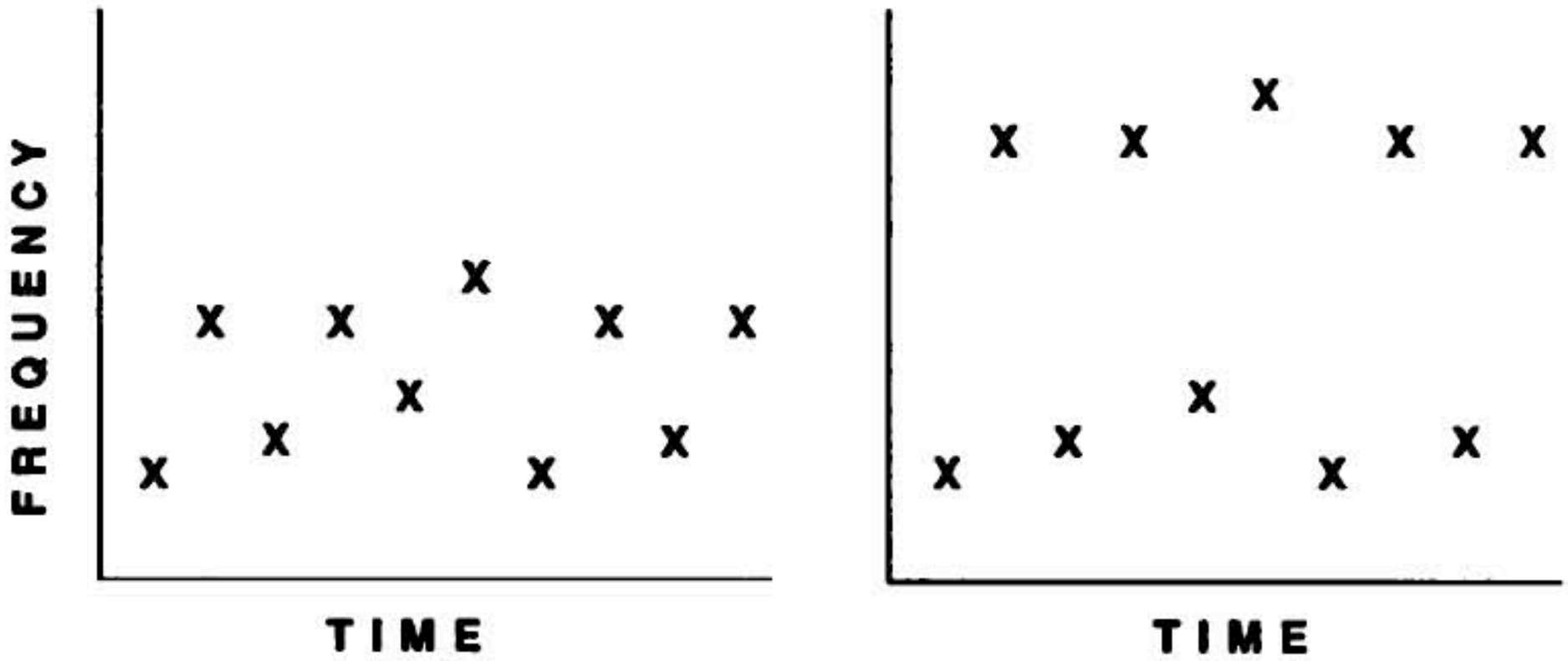

Figure 1.9
Stream segregation is stronger when the frequency separation between high and low tones is greater, as shown on the right.

When they focused on one of the streams, the other was heard as a vague background. As a consequence, while the listeners could easily judge the order of the high tones taken alone, or of the low ones taken alone, they could not put this information together to report the order of the six tones in the loop. Many listeners actually reported that the high tones all preceded the low ones, or vice versa, although this was never the case.

Other research has shown that the phenomenon of stream segregation obeys some fairly simple laws. If there are two sets of tones, one of them high in frequency and the other low, and the order of the two sets is shuffled together in the sequence (not necessarily a strict alternation of high and low), the degree of perceptual segregation of the high tones from the low ones will depend on the frequency separation of the two sets. Therefore if the two conditions shown in figure 1.9 are compared, the one on the right will show greater perceptual segregation into two streams. An interesting point is that visually, looking at figure 1.9, the perception of two distinct groups is also stronger on the right.

There is another important fact about stream segregation: the faster the sequence is presented, the greater is the perceptual segregation of high and low tones. Again there is a visual analogy, as shown in figure 1.10. We see the pattern in the right panel, in which there is a contraction of time (the same as an increase in speed), as more tightly grouped into two groups than the left panel is.

Gestalt Grouping Explanation

In the visual analogies, the grouping is predictable from the Gestalt psychologists' proximity principle, which states roughly that the

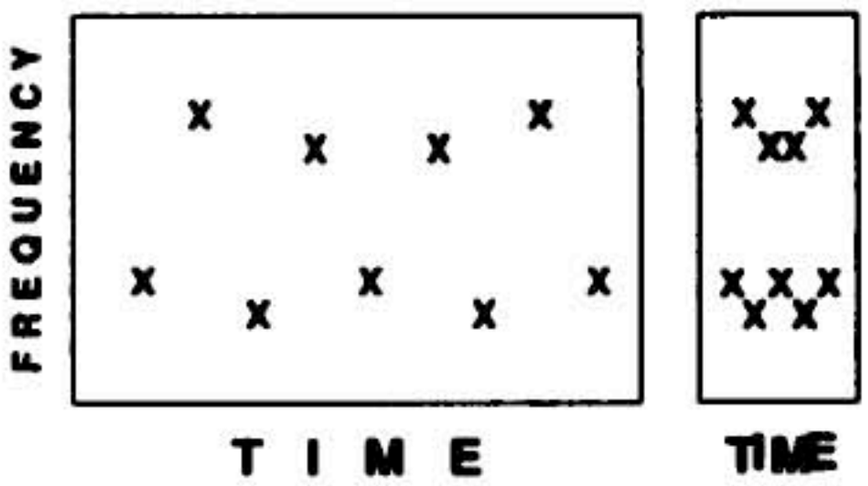

Figure 1.10
Stream segregation is higher at higher speeds, as shown on the right.

closer the visual elements in a set are to one another, the more strongly we tend to group them perceptually. The Gestalt psychologists thought of this grouping as if the perceptual elements—for example, the tones in figure 1.9—were attracting one another like miniature planets in space with the result that they tended to form clusters in our experience. If the analogy to audition is a valid one, this suggests that the spatial dimension of distance in vision has two analogies in audition. One is separation in time, and the other is separation in frequency. Both, according to this analogy, are distances, and Gestalt principles that involve distance should be valid for them.

The Gestalt principles of grouping were evolved by a group of German psychologists in the early part of this century to explain why elements in visual experience seemed highly connected to one another despite the fact that the incoming light rays, pressure energy, sound waves, and so on stimulated discrete sensory receptors such as the ones found in the retina of the eye. The word Gestalt means "pattern" and the theory described how the brain created mental patterns by forming connections between the elements of sensory input. We cannot go into much detail here about this subtle and philosophically sophisticated theory. However, we can examine a few of the observations that they made about the grouping of sensory elements. They are illustrated in the present discussion by means of the set of diagrams shown in figure 1.11.

Distinct visible elements will be grouped to form coherent perceptual organizations if they fulfill certain conditions. The first is similarity. In the first part of the figure, the black and white blobs can be seen as different subgroups because of the similarity of color within each group and the contrast between groups. Similarly, in audition we find that sounds of similar timbres will group together so that the successive sounds of the oboe will segregate from those of the harp, even when they are playing in the same register.

The second part of the figure shows grouping by a second factor, proximity, where the black blobs seem to fall into two separate clus-

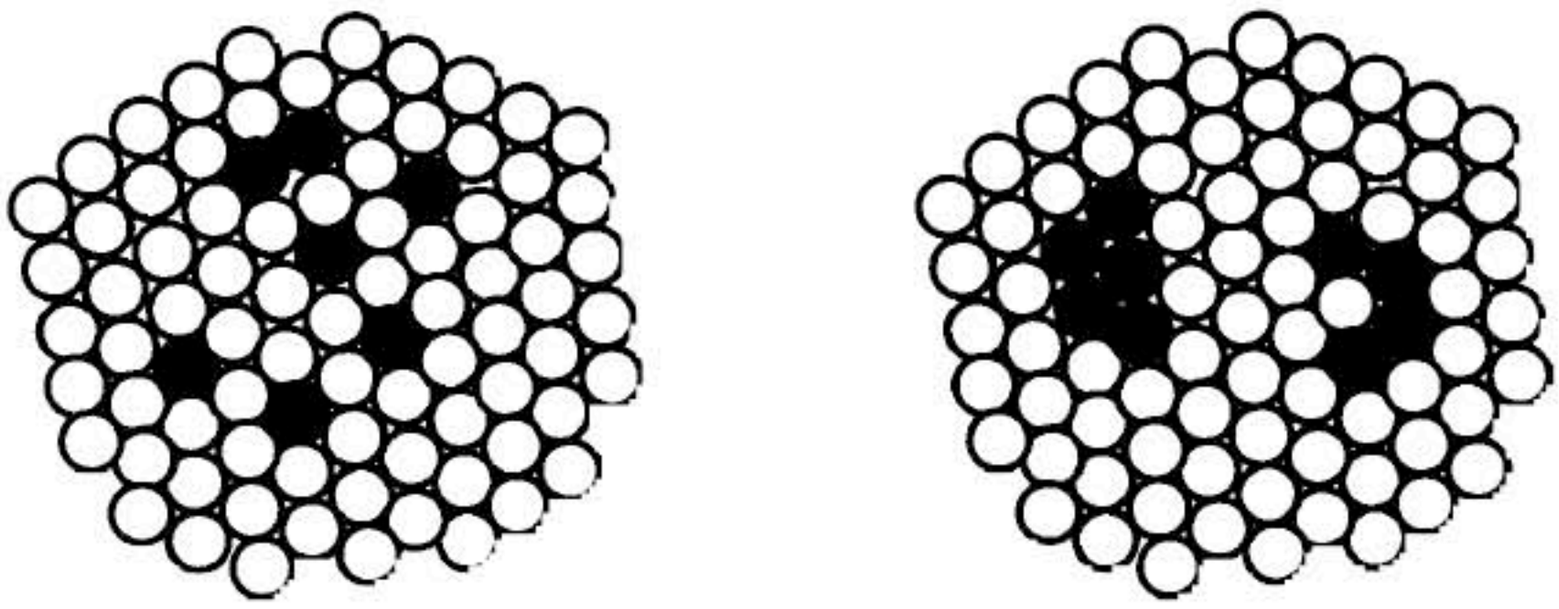

Figure 1.11
Illustration of the effects of the Gestalt principles of similarity and proximity on visual grouping.

ters because the members of one cluster are closer to other members of the same one than they are to the elements that form the other one. It would appear then that the example of stream segregation would follow directly from the Gestalt law of grouping by proximity. The high tones are closer to one another (in frequency) than they are to the low ones. As the high and low groups are moved further away from one another in frequency, the within-group attractions will become much stronger than the between-group attractions. Speeding the sequence up simply has the effect of moving things closer together on the time dimension. This attenuates the differences in time separations and therefore reduces the contribution of separations along the time dimension to the overall separation of the elements. In doing so, it exaggerates the effects of differences in the frequency dimension, since the latter become the dominant contributors to the total distance.

In both parts of figure 1.11, it is not just that the members of the same group go with one another well. The important thing is that they go with one another *better* than they go with members of the other group. The Gestalt theorists argued that there was always competition between the "forces of attraction" of elements for one another and that the perceptual organization that came out of this conflict would be a consequence of the distribution of forces across the whole perceptual "field," and not of the properties of individual parts taken in isolation.

The Gestalt psychologists' view was that the tendency to form perceptual organizations was innate and occurred automatically whenever we perceived anything. It was impossible, they claimed, to perceive sensory elements without their forming an organized

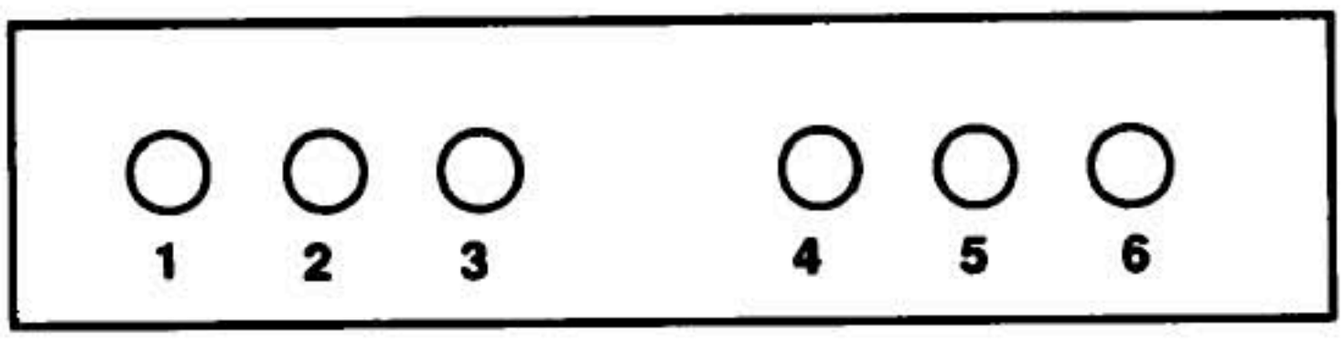

Figure 1.12
A visual display used to demonstrate visual motion segregation. Two groups of three lamps are arranged in a horizontal row.

whole. They argued that this organizing tendency was an automatic tendency of brain tissue.

Auditory Streaming versus Apparent Motion

We have been examining the phenomenon of auditory stream segregation as an example of how phenomena of auditory organization can exhibit the same complexities as are found in vision. This has led us to see interesting parallels in the principles that govern auditory stream segregation and visual grouping. But we have not yet discussed the most striking parallel, that between auditory stream segregation and the phenomenon of apparent motion in vision. Apparent motion is the perceptual effect that used to be very popular on the billboards of theatres, where the switching on and off of a series of electric light bulbs in sequence gave the experience of movement. In the laboratory it is usually created in a much simpler form. Two electric lamps, often seen as small white dots in an otherwise black room, are alternately switched on, each for a brief instant, so that a movement is seen that dances back and forth between the lights, always moving from the light that has just been flashed to the light that is currently being flashed. If the lamps are close together, it may seem that the light itself is moving back and forth. At greater distances the experience is just an impression of movement.

In 1915, Körte formulated a number of laws relating the duration, brightness, and spatial separation of the lamps to the strength of the impression of movement. Körte's third law stated that within certain ranges, if you want to increase the spatial separation between the lamps and still have a strong impression of motion, you had to slow down the alternation of flashes. It was almost as if the movement would not be able to keep up with the alternation of flashes if they were far separated in space unless the flashes were slowed down to compensate for their separation.

A more elaborate form of the apparent motion effect strongly resembles the streaming effect.[6] Instead of two lamps, there are six, arranged in a horizontal row as shown in figure 1.12. They are

arranged so that there is a wider gap between the left triplet of lights and the right triplet than there is between the lights within each triplet. If we label the lamps with the digits 1 to 6 from left to right, the order in which the lights are to be flashed can be expressed as the sequence 142536, repeated endlessly with no pause between repetitions. In this sequence there is an alternation between left-triplet and right-triplet flashes. At very low speeds, there is no apparent motion at all. The lights appear simply to go on and off in sequence. At a somewhat higher speed, the true sequence (142536) is seen as a form of irregular left-and-right motion between members of the two triplets. Then, as the speed is increased, the motion appears to split into two separate streams, one involving the leftmost three lamps and the other the rightmost three. The leftmost path of motion is 1–2–3 and the rightmost one is –4–5–6 (the dashes indicating the time periods in which the lights from the other stream are active). This segregation is exactly parallel to what happens in the auditory streaming effect. However, it is also directly explainable through Körte's third law.

This law simply states that as the speed increases, the distance between flashes must shrink if good motion is to be seen. Therefore, if we assume that potential motions between successive and non-successive flashes are competing with one another for dominance, and that we finally see the one that is most dominant, the results of our example follow directly. As we speed up the sequence there is an increased tendency for shorter movements to be favored by Körte's law so that the longer between-triplet motions are suppressed in favor of the stronger within-triplet motions.

I have set up the two examples, the streaming of tones and the splitting of apparent motion, in a parallel way so that the analogy can be directly seen. Horizontal position in space is made to correspond to the frequency of the tones, with time playing the role of the second dimension in both cases.

The success of Körte's law in explaining the visual case suggests that there is a parallel law in audition, with melodic motion taking the place of spatial motion.[7] This law would state that if you want to maintain the sense of melodic motion as the frequency separation between high and low tones increases, you must slow the sequence down. As with visual apparent motion it is as if the psychological mechanism responsible for the integration of auditory sequences could not keep up with rapid changes.

Scene-Analysis Explanation

However, Körte's law is not an accident of the construction of the human brain. In both visual motion and melodic motion, the laws of

grouping help to solve the scene analysis problem as the sensory input unfolds over time. In both domains, Körte's law is likely to group information appropriately. In vision it tends to group glimpses of a moving object with other glimpses of the same object rather than with those of different objects. This is important in a world where many objects can be moving at the same time and where parts of their trajectories can be hidden by closer objects such as trees. The law assumes that if a hidden object is moving a longer distance it takes it longer to get there. Hence the proportionality of distance and time that we find in the law.

The proportionality of frequency displacement and time that we observe in the streaming effect also has a value in scene analysis. What should the auditory system do if it hears a particular sound, A1, and then either a silence or an interruption by a loud sound of a different quality, and then a subsequent sound, A2, that resembles A1? Should it group A1 and A2 as coming from the same source? The auditory system assumes that the pitch of a sound tends to change continuously and therefore that the longer it has been since the sound was heard, the greater the change ought to have been. This has the effect that longer frequency jumps are tolerable only at longer time delays.

The experience of motion that we have when a succession of discrete events occurs is not a mere laboratory curiosity. When visual apparent motion is understood as a glimpse of a scene analysis process in action, new facts about it can be discovered. For example, it has been found that when the apparent movement seems to occur in depth, in a movement slanting away from the observer, the visual system allows more time for the object to move through the third dimension than it would have if it had appeared to be moving only in the horizontal plane.[8] This happens despite the fact that although a slanting-away motion would traverse more three-dimensional space, it produces the same displacement of an object's image as a horizontal motion does on the retina of an observer. Therefore Körte's law applies to real distance in the world and not to retinal distance, and therefore can best be understood as a sophisticated part of scene analysis.

Another example of a discovery that was guided by the assumption that the rules of apparent motion exist to group glimpses of real scenes was made by Michael Mills and myself.[9] We worked with an animation sequence in which a shape disappeared from one part of a drawing and appeared in another. This change was seen as motion only if the shape was seen as representing the outline of a "figure" both before and after the disappearance. If the observer was induced to see it as "ground" (the shape of an empty space between forms)

before it disappeared, and as "figure" (the shape of an actual figure) when it reappeared, the displacement was not seen as motion but as an appearance from nowhere of the figure.

Neither is the auditory streaming effect simply a laboratory curiosity. It is an oblique glimpse of a scene-analysis process doing the best it can in a situation in which the clues to the structure of the scene are very impoverished.

In general, all the Gestalt principles of grouping can be interpreted as rules for scene analysis. We can see this, for example, in the case of the principle of grouping by similarity. Consider the block-recognition problem shown earlier in figure 1.2 where the problem was to determine which areas of the drawing represented parts of the same block. Because this drawing is not very representative of the problem of scene analysis as we face it in everyday life, let us imagine it transformed into a real scene. In the natural world visible surfaces have brightness, color, and texture. It would be a good rule of thumb to prefer to group surfaces that were similar in appearance to one another on these dimensions. This would not always work, but if this principle were given a vote, along with a set of other rules of thumb, it is clear that it would contribute in a positive way to getting the right answer.

In the case of sound, the considerations are the same. If in a mixture of sounds we are able to detect moments of sound that strongly resemble one another, they should be grouped together as probably coming from the same happening. Furthermore, the closer in time two sounds that resemble each other occur, the more likely it is that they have originated with the same event. Both of these statements follow from the idea that events in the world tend to have some persistence. They do not change instantly or haphazardly. It seems likely that the auditory system, evolving as it has in such a world, has developed principles for "betting" on which parts of a sequence of sensory inputs have arisen from the same source. Such betting principles could take advantage of properties of sounds that had a reasonably high probability of indicating that the sounds had a common origin. Viewed from this perspective, the Gestalt principles are seen to be principles of scene analysis that will generally contribute to a correct decomposition of the mixture of effects that reaches our senses. I am not claiming that the auditory system "tries" to achieve this result, only that the processes have been selected by evolution because they did achieve them.

The argument that I have made does not imply that Gestalt theory is wrong. For the Gestaltists, the phenomena of perceptual grouping

arose from the fact that there were forces of attraction and segregation that operated in a perceptual field. This may indeed be the mechanism by which the grouping occurs. I am simply arguing that even if this is the form of the computation, the particular grouping force given to each property of the sensory input and the way in which the grouping forces are allowed to interact have been determined (through evolution) to be ones that will tend to contribute to the successful solution of the scene analysis problem.

Closure and Belongingness

Our senses of vision and audition, living in the same world, often face similar problems. So we should not be surprised if we often find them using similar approaches to overcome those problems. We have seen how the two systems sometimes deal with fragmented views of a sequence of events by connecting them in plausible ways. Another strong similarity between the sense modalities can be seen in the phenomenon of "perceived continuity." This is a phenomenon that is sometimes said to be an example of "perceptual closure."

The tendency to close certain "strong" perceptual forms such as circles was observed by the Gestalt psychologists. An example might be the drawing shown on the left side of figure 1.5 in which we are likely to see a circle partly obscured by an irregular form. The circle, though its outer edge is incomplete in the picture, is not seen as incomplete but as continuing on behind the other form. In other words, the circle has closed perceptually.

It is commonly said that the Gestalt principle of closure is concerned with completing forms with gaps in them. But if it did that, we would not be able to see any forms with gaps in them, which would be ridiculous. The principle is really one for completing *evidence* with gaps in it.

The Gestalt psychologists argued that closure would occur in an interrupted form if the contour was "strong" or "good" at the point of interruption. This would be true when the contours of the form continued smoothly on both sides of the interruption so that a smooth continuation could be perceived. Presumably laws of similarity would also hold so that if the regions on two sides of an interruption were the same brightness, for instance, they would be more likely to be seen as a single one continuing behind the interruption.

Like the perceptual grouping of discrete events, closure can also be seen as a scene-analysis principle. This can be illustrated with figure 1.13 which shows a number of fragments that are really parts of a familiar object or objects. The fragments were obtained by taking the

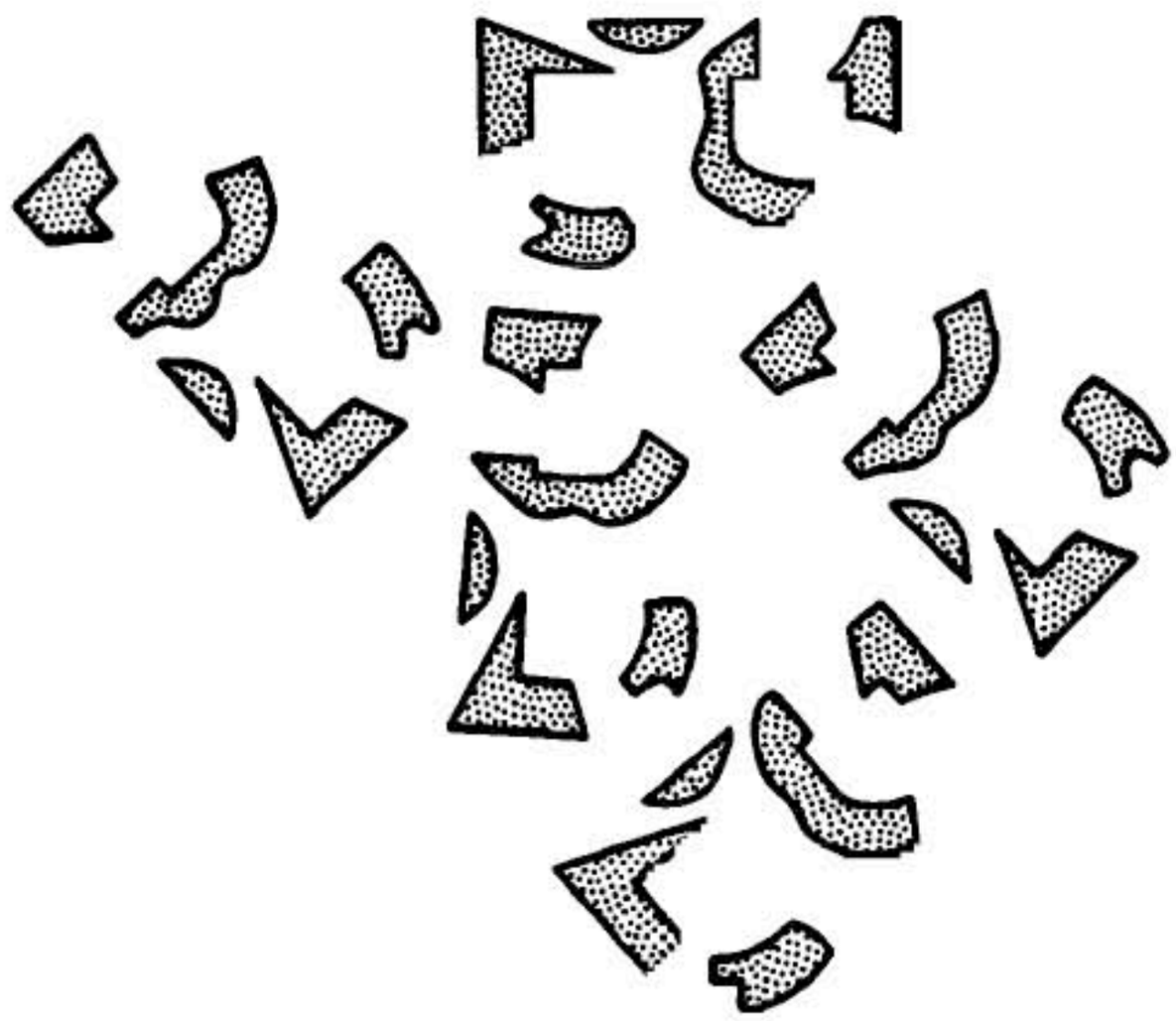

Figure 1.13
Fragments do not organize themselves strongly when there is no information for occlusion. (From Bregman 1981b.)

familiar display and laying an irregularly shaped mask over it. Then the parts that were underneath the mask were eliminated, leaving visible only those parts that had not been covered by it.

Why do the fragments not close up perceptually in this figure? A plausible Gestalt answer might be that the forces of closure are not strong enough. The contours of the fragments might not be similar enough or in good continuation with one another. However, it is easy to show that these are not the basic reasons for the lack of closure. The problem in this figure is that the visual system does not know where the evidence is incomplete. Look at what happens when the picture is shown with the mask present as in figure 1.14. The visual system quickly joins the fragments without the observer having to think about it. The Gestalt principle of closure has suddenly come alive in the presence of the mask.

What information could the mask be providing? It tells the eye two things. It explains which contours have been produced by the shape of the fragments themselves as contrasted with those that have been produced by the shape of the mask that is covering them. It also provides information about occlusion (which spaces between fragments were created by the fact that the mask occluded our view of the underneath shape). These spaces should be ignored and treated as missing evidence, not as actual spaces. The continuity among the contours of the fragments of a particular B undoubtedly contributes to their grouping, but this continuity becomes effective only in the presence of occlusion information.

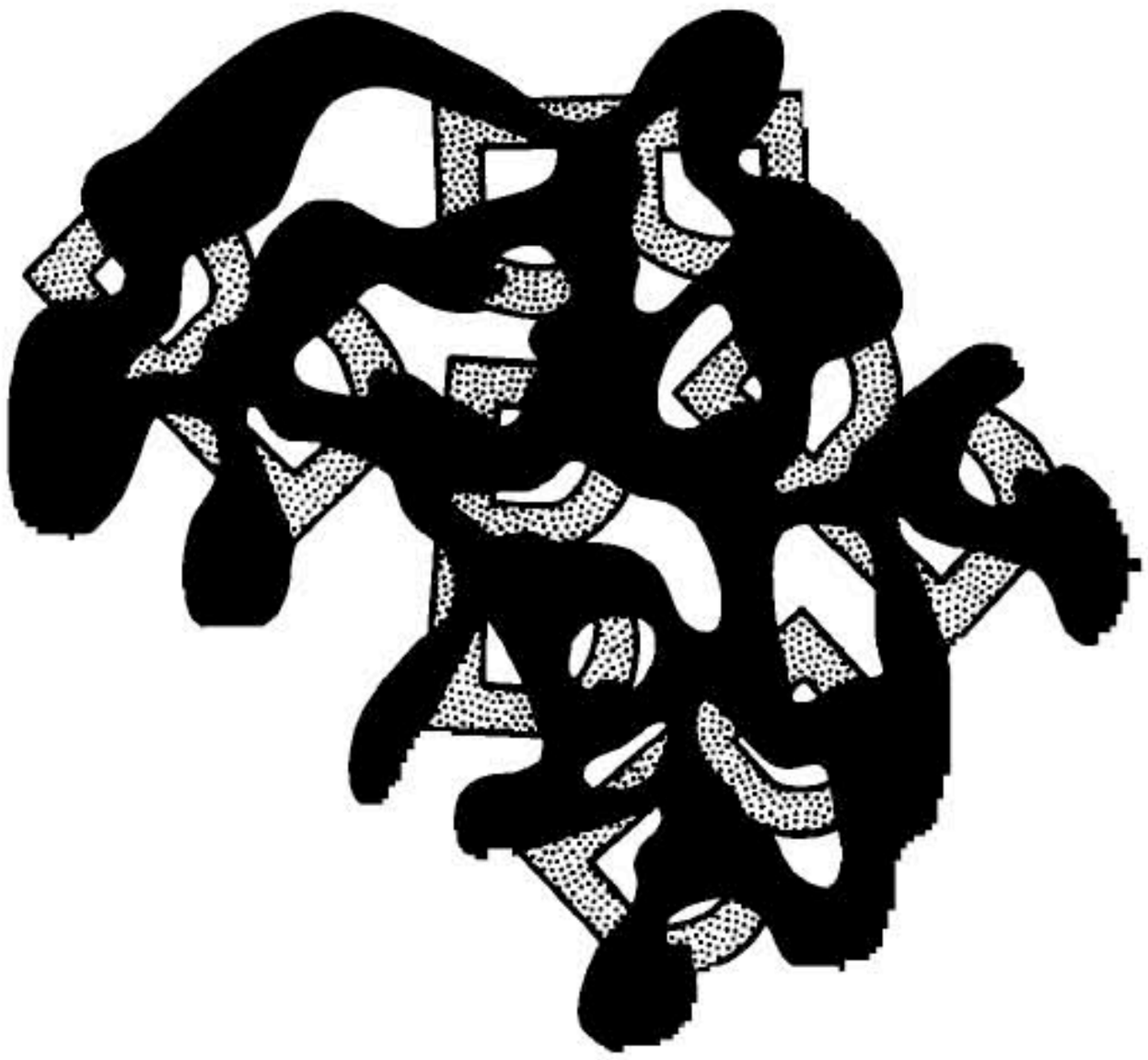

Figure 1.14
The same fragments shown earlier except that information for occlusion has been added, causing the fragments on the boundaries of the occluding form to be grouped. (From Bregman 1981b.)

The conclusion to be reached is this: the closure mechanism is really a way of dealing with missing evidence. But before our perceptual systems are willing to employ it, they first have to be shown that some evidence is missing. This explains how we can see figures with actual gaps in them; we have no reason to believe that the missing parts are merely being hidden. Figures 1.13 and 1.14 indicate that Gestalt principles are just oblique glimpses of a process of scene analysis that looks as much like an evidence-processing system as like the simple grouping-by-attraction system described by Gestalt psychology.

There is evidence that principles of grouping act in an equally subtle way in audition. There is a problem in hearing that is much like the problem of occlusion in seeing. This is the phenomenon of masking. Masking occurs when a loud sound covers up or drowns out a softer one. Despite the masking, if the softer sound is longer, and can be heard both before and after a brief burst of the louder one, it can be heard to continue behind the louder one just as B's were seen as continuing behind the occluding blob in figure 1.14, and as the circle seemed to continue behind the occluding form in the example of figure 1.5. What is more, even if the softer sound is *physically removed* during the brief loud sound, it is still heard as continuing through the interruption.

Figure 1.15
Tonal glides of the type used by Dannenbring (1976). Left: the stimulus with gaps. Right: the stimulus when the gaps are filled with noise.

This illusion has many names, but I will refer to it as the illusion of continuity. It occurs with a wide range of sounds. An example is shown in figure 1.15 where an alternately rising and falling pure-tone glide is periodically interrupted by a short loud burst of broad-band noise (like the noise between stations on a radio). When the glide is broken at certain places but no masking sound is present during the breaks, as in the left panel, the ear hears a series of rising and falling glides, but does not put them together as a single sound any more than the eye puts together the fragments of figure 1.13. However, if the masking noise is introduced in the gaps so as to exactly cover the silent spaces, as in the right panel, the ear hears the glide as one continuous rising and falling sound passing right through the interrupting noise. The integration of the continuous glide pattern resembles the mental synthesis of B's in figure 1.14. They are both effortless and automatic.

Again you could see the auditory effect as an example of the Gestalt principle of closure. However another way of looking at it may be more profitable. Richard Warren has interpreted it as resulting from an auditory mechanism that compensates for masking.[10] He has shown that the illusion can be obtained only when the interrupting noise would have masked the signal if it had really been there. The interrupting noise must be loud enough and have the right frequency components to do so. Putting that in the context of this chapter, we see that the illusion is another oblique glance of the auditory scene-analysis process in action.

We have seen how two types of explanation, one deriving from Gestalt psychology and the other derived from considerations of scene analysis, have been applicable to both the streaming and continuity effects. They differ in style. The Gestalt explanation sees the principles of grouping as phenomena in themselves, a self-sufficient

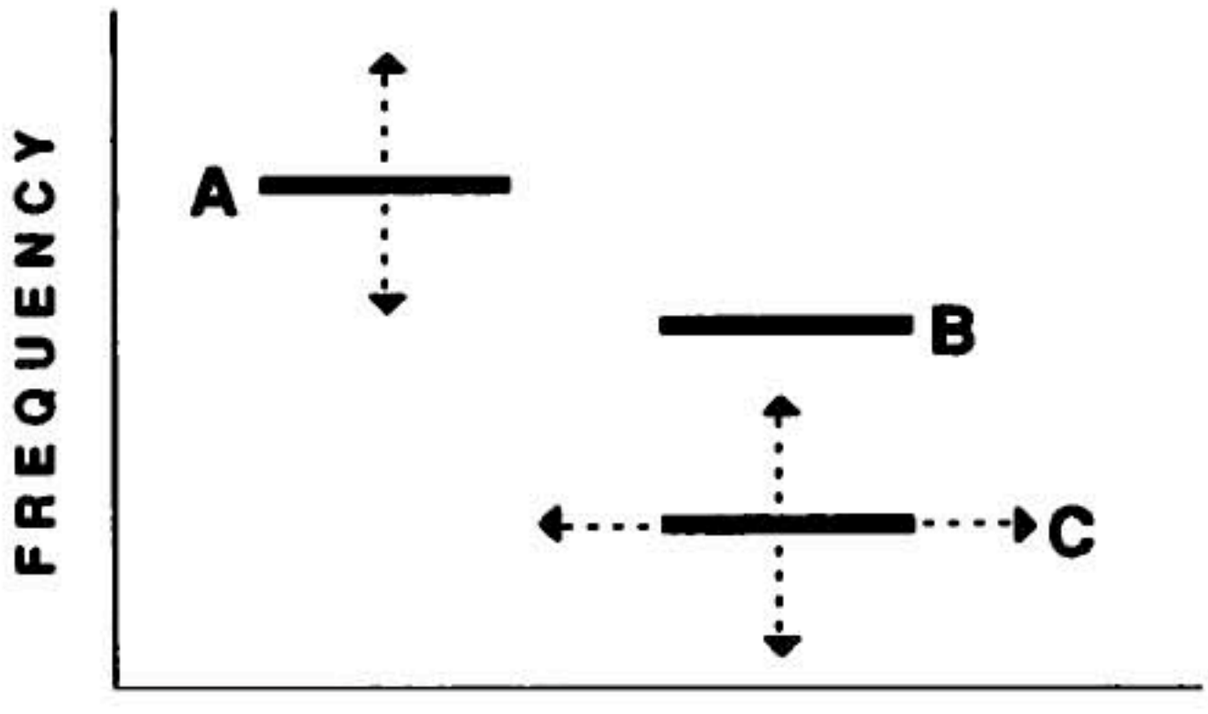

Figure 1.16
Stimulus used by Bregman and Pinker (1978). A, B, and C are pure tone components.

system whose business it is to organize things. The scene-analysis approach relates the process more to the environment, or, more particularly, to the problem that the environment poses to the perceiver as he or she (or it) tries to build descriptions of environmental situations.

Sequential versus Spectral Organization

Perceptual Decomposition of Complex Sounds

We have looked at two laboratory phenomena in audition that show the activity of the scene-analysis process: the streaming effect and the illusory continuation of one sound behind another. There is a third phenomenon that deserves to be mentioned in this introductory chapter. It is introduced here not to demonstrate a parallel between vision and audition, but to show another dimension of the grouping problem. This is the perceptual decomposition of simultaneous sounds. It can be illustrated through an experiment by Bregman and Pinker.[11]

The sounds used in this experiment are shown in figure 1.16. They consist of a repeating cycle formed by a pure tone A alternating with a complex tone that has two pure-tone components, B and C. This is inherently an ambiguous event. For example, it could be created by giving an audio oscillator to each of two people. The oscillator given to one of them puts out the pure tone A, while the one given to the other puts out the complex tone BC. The two persons are asked to play their oscillators in rapid alternation. If this were the way the sound had been created, the correct perceptual analysis would be to hear a pure tone alternating with a rich-sounding complex tone. This,

however, is only one possibility for the origin of the sound. The second is that we have given out oscillators, as before, to two persons. This time, however, both of the oscillators can put out only pure tones. One person is told to sound his instrument twice on each cycle to make the tones A and B, whereas the other is told to play his tone only once on each cycle to make the tone C. He is told to synchronize his C tone with the B tone of his partner. If our auditory systems were to correctly represent the true causes of the sound in this second case, we should hear two streams: one consisting of the repetitions of tones A and B, accompanied by a second that contains only the repetitions of tone C. In this way of hearing the sequence, there should be no rich tone BC because the richness is an accidental by-product of the mixture of two signals. If the auditory system is built to hear the properties of meaningful events rather than of the accidental by-products of mixtures, it should discard the latter.

The experiment showed that it was possible to hear the sequence in either way, depending on two factors. The first was the frequency proximity of tones A and B. The closer they were to one another in frequency, the greater the likelihood of hearing A and B as forming a single stream separate from C. Apparently the auditory system uses the proximity of a succession of frequencies, much as it does in the case of the streaming phenomenon, as evidence that they are from a common source. The second factor was the synchrony of tones B and C. If their onsets and offsets were synchronized, they tended to be fused and heard as a single complex sound BC, which was heard as alternating with A. Furthermore, the effects of the BC synchrony were competitive with the effects of the AB frequency proximity. It was as if A and C were competing to see which one would get to group with B. If the synchrony of C with B was reduced, B would be more likely to group with A, unless, of course, the AB connection was made weaker by moving A further away in frequency from B.

Horizontal and Vertical Processes of Organization

There is a distinction that ought to be made now because it follows directly from the Bregman-Pinker experiment and because it is responsible for the structure of the later chapters. This is the distinction between the processes of sequential and spectral integration.

The process of putting A and B together into a stream can be referred to as sequential integration. This is the kind of integration that forms the melodic component of music. It is the process that connects events that have arisen at different times from the same source. It uses the changes in the spectrum and the speed of such changes as major

clues to the correct grouping. The sequential process is what is involved in the streaming effect that was discussed earlier.

The fusing of B with C into a single sound is what will be referred to as simultaneous integration or, in special contexts, as spectral integration, a term borrowed from James Cutting.[12] It is this process that takes acoustic inputs that occur at the same time, but at different places in the spectrum or in space, and treats them as properties of a single sound. It is responsible for the fact that we can interpret a single spectrum of sound as arising from the mixture of two or more sound sources, with the timbre of each one being computed from just those spectral components that have been allocated to that source. This happens, for example, when we hear two singers, one singing "ee" and the other "ah", on different pitches. Despite the fact that all we have is a single spectrum, with the harmonics from the two voices intermixed, we can clearly hear the two vowels. Since a vowel sound is a sort of timbre, this example shows that we can extract two timbres at the same time from a single signal.

If we turn back to the mixed spectrogram shown in figure 1.4, we see that in order to put together the streaks of darkness belonging to the same acoustic source, the same two kinds of grouping are necessary: (1) putting together events that follow one another in time (sequential grouping) and (2) integrating components that occur at the same time in different parts of the spectrum (simultaneous grouping). Musicians speak of a horizontal and a vertical dimension in written music. By horizontal, they refer to the groupings across the page that are seen as melody. By vertical, they refer to the simultaneous events that form chords and harmony. These are the same two dimensions as the ones called sequential and simultaneous.

It is useful to distinguish these two aspects of organization because they are controlled by different acoustic factors. Of course they interact, too, but that can be described separately. Therefore, chapter 2 discusses the sequential aspect of organization and chapter 3 deals with the fusion of simultaneous auditory components into a single sound.

Types of Explanation of These Phenomena

In the following chapters, I will attempt to show how the auditory system goes about trying to solve the scene-analysis problem. The presentation will include a number of approaches. First, it will try to show what acoustic information the auditory system uses to solve this problem. An example of this would be the fact that the synchronous onset of two frequency components is taken as evidence that

they are parts of the same sound. Second, it will show the perceptual effects that the grouping process has. For instance, if the two components are not allocated to the same stream, then properties that involve their combination, such as the timbre of the mixture, will tend not to be perceived. Third, it will point to a few general properties of the perceptual system that does these things, such as the fact that the groupings are competitive; for example, two components that, in the absence of any other components, might have been placed into the same stream can be captured into separate streams by other components with which they fit better.

This volume can be thought of as an attempt to build up a functional description of how the auditory system solves certain types of problems. It is possible to arrive at an appreciation of the problems of audition by taking the attitude of a researcher in artificial intelligence faced with the problem of trying to replicate human auditory functioning. Such a person, required to duplicate a skill such as segregating individual sound sources from mixtures, would first analyze the information available to the listener. What clues are available in the acoustic signal itself that could indicate which components arose from the same source? How can the environment alter the clues in ways that make them subject to error? What would be the best way to combine these clues if some of them are subject to error?

Whereas the following chapters maintain this attitude, they also deal with a large body of empirical data and try to keep the speculation within the bounds of what is supportable by that evidence. There are certain things that they do not do. They do not attempt to offer physiological explanations or proposals about explicit computational mechanisms. Their approach can best be viewed as an attempt to lay some constraints on theories of these two types.

Although the story is informal, it is interesting to take a moment to see how it is related to more developed theoretical positions. I will consider its relation to concepts drawn from computer modeling, syntactic theory, Gestalt psychology, and physiological explanation.

The computer modeling approach has contributed an important idea that will be used in the coming chapters. This is the notion of a heuristic. The idea was evolved in the process of designing computer programs to solve difficult problems for which no mathematical solution was known. The approach taken by the designers was to employ heuristics, which are defined as procedures that are not guaranteed to solve the problem, but are likely to lead to a good solution. An example would be the use of heuristic tests by computer chess programs to determine whether a proposed move would lead to a good position

(e.g., to test whether the move would result in the computer controlling the center of the board or whether the move would lead to an exchange of pieces that favored the computer). Each move is evaluated by a number of such heuristics. No one of them can guarantee success, but if there are a large number, each with some basis in the structure of the game of chess, a move that satisfies most of them will probably be a good one. Furthermore, if each of the heuristic evaluation processes has a chance to vote for or against the move, the program will be less likely to be tricked than it would be if it based its move on only one or two criteria, no matter how good they were.

I believe that the perceptual systems work in similar ways. Having evolved in a world of mixtures, humans have developed heuristic mechanisms capable of decomposing them. Because the conditions under which decomposition must be done are extremely variable, no single method is guaranteed to succeed. Therefore a number of heuristic criteria must be used to decide how to group the acoustic evidence. These criteria are allowed to combine their effects in a process very much like voting. No one factor will necessarily vote correctly, but if there are many of them, competing with or reinforcing one another, the right description of the input should generally emerge. If they all vote in the same way, the resulting percept is stable and unambiguous. When they are faced with artificial signals, set up in the laboratory, in which one heuristic is made to vote for integration and another for segregation, the resulting experiences can be unstable and ambiguous.

My use of the word "heuristic" does not imply a computer-like procedure that involves a long sequence of steps, extended over time. We have to bear in mind that the decisions of the auditory system are carried out in very short periods of time. I use the word heuristic in its functional sense only, as a process that contributes to the solution of a problem.

Whereas the perceptual phenomena that we examined earlier are the province of psychologists, the problem of how people build mental descriptions is a topic that has been looked at by linguists too. As a result, they have provided us with a metaphor for understanding auditory scene analysis. This metaphor, "deep structure," derives from the study of the syntactic structure of sentences.

One of the basic problems in syntax is how to describe the rules that allow the speaker to impose a meaning on a sentence by adding, subtracting, or rearranging elements in the sentence. For example, in English one of these rules imposes the form of a question on a sentence by placing the auxiliary verb at the beginning of the sentence.

Thus, the active sentence "He has gone there" is expressed in a question as "Has he gone there?" The difficulty that occurs when a language loads a sentence with meanings is that when a large number of form-shaping rules are piled on top of one another, it becomes difficult to untangle them and to appreciate the contribution of each of them to the final product. Somehow all speakers of English come to be able to do this, but the learning takes some time. In the 1960s, Noam Chomsky introduced the notion of the "deep structure" of a sentence, a description of a sentence that separately and explicitly described all the underlying syntactic forms and displayed their interrelationships. When a theorist, or a listener, starts with a given sentence and builds a description of its syntax, this is called "parsing" the sentence. It was argued by psychologists who were inspired by Chomsky's approach that in the course of understanding a sentence, the hearer parses a sentence and builds a deep structure for it.

We can talk about perception in a very similar way. Just as a spoken sentence imposes an extraordinary decoding problem upon the listener, so does a nonlinguistic sensory input. Whenever we experience an event, the sensory impression is always the result of an elaborate composition of physical influences. If we look at a four-inch-square area of a table top, for example, the local properties of this area have been affected by many factors: the table's shininess, the variations in its surface color, the unevenness of its surface, the shadow of a nearby object, the color of the light source, the slant of the surface of the table relative to our eyes, and perhaps many more. These factors are all simultaneously *shaping* the sensory information; they are not simply inserted side by side. The shininess is not at one place in our visual image, the surface color at another, and so on. Neither can they be extracted from the sense data independently of one another.

The same thing happens in audition. If we look at any one-tenth-second slice of figure 1.4, the information shown in that slice represents a composition of influences. The spectrum may have been shaped by voices and by other simultaneous sounds. Somehow, if we are able to understand the events that have shaped it, we are succeeding, as in sentence comprehension, in developing a mental description that displays the simple causative factors and their interrelationships in an explicit way.

There is a provocative similarity among the three examples—the syntactical, the visual, and the auditory. In all three cases, the perceivers are faced with a complex *shaping* of the sensory input by the effects of various simple features, and they must recover those features from their effects. Transposing the linguist's vocabulary to the field of perception, one might say that the job of the perceiver is to parse the

sensory input and arrive at its deep structure. In some sense the perceiver has to build up a description of the regularities in the world that have shaped the evidence of our senses. Such regularities would include the fact that there are solid objects with their own shapes and colors (in vision) and sounds with their own timbres and pitches (in audition).

Although the approach of this book is not physiological, it is important to see its relation to physiological explanation. We can take as an example the physiological explanations that have been offered for the streaming effect of figure 1.8. It has been proposed that the segregation into two streams occurs because a neural mechanism responsible for tracking changes in pitch has temporarily become less effective.[13] This interpretation is supported by the results of experiments that show that the segregation becomes stronger with longer repetitions of the cycle of tones. Presumably the detector for change has become habituated in the same manner as other feature detectors are thought to. This view of the stream segregation phenomenon sees it as a breakdown. This seems to be in serious conflict with the scene-analysis view presented earlier, in which stream segregation was seen as an accomplishment. So which is it to be, breakdown or accomplishment?

We do not know whether or not this physiological explanation is correct (the claim will be examined in chapter 3). But even if it is, its truth may not affect the scene analysis explanation of streaming. To demonstrate why, it is necessary to again appeal to an argument based on evolution. Every physiological mechanism that develops must stand the test of the winnowing process imposed by natural selection. However, the survival of an individual mechanism will often depend not just on what it does in isolation, but on the success of the larger functional system of which it forms a part.

Because of the indirect way in which the individual physiological mechanism contributes to the successful accomplishments displayed by the larger system, it is possible that what looks like a breakdown when seen at the single-mechanism level is actually contributing to an accomplishment at the system level. To take a homespun example, consider the case of a pitfall trap. When the top of the trap, covered with branches and leaves, "breaks down" and the animal falls through into the hole, we can see that the physical breakdown (of the trap cover) represents a functional success (of the entrapment). The breakdown and the achievement are at different levels of abstraction. By analogy, it would not be contradictory to assert that the streaming effect represented both the breakdown of a physiological mechanism

and the accomplishment of scene analysis. This example illustrates how indirect the relation can be between function and physiology.

Scene-Analysis View Prevents Missing of Vision-Audition Differences

It was argued in the earlier discussion that Gestalt explanations had to be supplemented by ones based on scene analysis because the latter might lead us to new phenomena, such as the role of the occluding mask in perceptual closure. There is another difference between the two approaches. Because the Gestalt theorists saw the principles of organization as following from general properties of neural tissue they focused on similarities between the senses rather than on differences. The laws of grouping were stated in a general way, in terms of adjectives (such as "proximity" or "similarity") that could apply equally well to different sense modalities. This has had both useful and harmful effects. On the positive side it has promoted the discovery of the similar way in which perceptual organization works in different sense modalities. For example, the similarities between apparent movement and auditory streaming have become apparent. However, an exclusive focus on the common Gestalt principles, neglecting the unique scene-analysis problems that each sense must solve, is likely to neglect differences between them and cause us to miss some excellent opportunities to study special problems in audition that make themselves evident once we consider the dissimilarities between the senses. The way to get at them is to consider the differences in the way in which information about the properties of the world that we care about are carried in sound and in light. The fact that certain Gestalt principles actually are shared between the senses could be thought of as existing because they are appropriate methods for scene analysis in both domains.

As an example of the way that the scene-analysis approach can reveal important differences between the senses, let us go through the exercise of considering the roles of direct energy, reflected energy, and their mixture in the two senses.

Differences in the Ecology of Vision and Audition

There is a crucial difference in the way that humans use acoustic and light energy to obtain information about the world. This has to do with the dissimilarities in the ecology of light and sound. In audition humans, unlike their relatives the bats, make use primarily of the sound-emitting rather than the sound-reflecting properties of things. They use their eyes to determine the shape and size of a car on the road by the way in which its surfaces reflect the light of the sun, but

use their ears to determine the intensity of the crash by receiving the energy that is emitted when this event occurs. The shape reflects energy; the crash creates it. For humans, sound serves to supplement vision by supplying information about the nature of events, defining the "energetics" of a situation.

There is another difference that is very much related to this one: sounds go around corners. Low-frequency sound bends around an obstruction while higher frequency sound bounces around it. This makes it possible for us to have a distant early warning system. The reader might be tempted to object that light too goes around corners. Although it does not bend around, in the way that low-frequency sound does, it often gets around by reflection; in effect, it bounces around the corner. But notice what a difference this bouncing makes in how we can use the light. Although the bounced-around light provides illumination that allows us to see the shapes of things on our own side of the corner, unless it has been bounced by means of mirrors it has lost the shape information that it picked up when it reflected off the objects on the opposite side. Sound is used differently. We use it to discover the time and frequency pattern of the source, not its spatial shape, and much of this information is retained even when it bends or bounces around the corner.

This way of using sound has the effect, however, of making acoustic events transparent; they do not occlude energy from what lies behind them. The auditory world is like the visual world would be if all objects were very, very transparent and glowed in sputters and starts by their own light, as well as reflecting the light of their neighbors. This would be a hard world for the visual system to deal with.

It is not true then that our auditory system is somehow more primitive simply because it does not deliver as detailed information about the shapes, sizes, and surface characteristics of objects. It simply has evolved a different function and lives in a different kind of world.

What of echoes? We never discuss echoes in light because its speed is so fast and the distances in a typical scene are so small that the echo arrives in synchrony with the original signal. Furthermore, in vision we are usually interested in the echoes, not the original signal, and certainly not in integrating the two into a single image. Light bounces around, reflecting off many objects in our environments, and eventually gets to our eyes with the imprint of the unoccluded objects still contained in it. Because the lens-and-retina system of the eye keeps this information in the same spatial order, it allows us access to the information about each form separately. Echoes are therefore very useful in specifying the shapes of objects in vision because the echoes

that come off different surfaces do not get mixed together on the way to our eye.

The case is otherwise in audition. Because our ears lack the lenses that could capture the spatial layout of the echoes from different surfaces, we are usually interested in the source of sound rather than in the shapes of objects that have reflected or absorbed it. The individual spatial origins of the parts of a reflected wave front are barely preserved at all for our ears. Therefore, when the sound bounces off other objects and these echoes mix with the original signal, they obscure the original properties of the sound. Although echoes are delayed copies and, as such, contain all the original structure of the sound, the mixing of the original and the echo creates problems in using this redundant structural information effectively.

The two senses also make different uses of the absorption of energy by the environment. The fact that different objects absorb light in different ways gives them their characteristic colors and brightnesses, but this differential absorption is not as valuable in hearing because our ears cannot separate the reflections from small individual objects. We do hear the "hardness" or "softness" of the entire room that we are in. This corresponds to the color information carried in light, but the acoustic information is about very large objects, whereas the information in light can be about very small ones.

In summary, we can see that the differences in how we use light and sound create different opportunities and difficulties for the two perceptual systems and that they probably have evolved specialized methods for dealing with them. This realization will be useful in chapter 7 when we begin to search for reasons for apparent violations of the principle of exclusive allocation of sensory evidence.

Primitive versus Schema-Based Stream Segregation

It seems reasonable to believe that the process of auditory scene analysis must be governed by both innate and learned constraints. In the chapters that follow, the effects of the unlearned constraints will be called "primitive segregation" and those of the learned ones will be called "schema-based segregation."

One reason for wanting to think that there are unlearned influences on segregation is the fact that there are certain constant properties of the environment that would have to be dealt with by every human everywhere. Different humans may face different languages, musics, and birds and animals that have their own particular cries. A desert certainly sounds different from a tropical forest. But certain essential physical facts remain constant. When a harmonically structured

sound changes over time, all the harmonics in it will tend to change together in frequency, in amplitude, and in direction, and to maintain a harmonic relationship. This is not true of just some particular environment but of broad classes of sounds in the world.

Such regularities can be used in reverse to infer the probable underlying structure of a mixture. When frequency components continue to maintain a harmonic relationship to one another despite changes in frequency, amplitude, and spatial origin, they will almost always have been caused by a coherent physical event. The later chapters show that the human auditory system makes use of such regularity in the sensory input. But is this innate? I think that it is. The internal organs of animals evolve to fit the requirements of certain constant factors in their environments. Why should their auditory systems not do likewise?

Roger Shepard has argued for a principle of "psychophysical complementarity," which states that the mental processes of animals have evolved to be complementary with the structure of the surrounding world.[14] For example, because the physical world allows an object to be rotated without changing its shape, the mind must have mechanisms for rotating its representations of objects without changing their shapes. The processes of auditory perception would fall under this principle of complementarity, the rules of auditory grouping being complementary with the redundancies that link the acoustic components that have arisen from the same source.

The Gestalt psychologists argued that the laws of perceptual organization were innate. They used two types of evidence to support their claim. One was the fact that the phenomenon of camouflage, which works by tricking the organizational processes into grouping parts of an object with parts of its surroundings, could be made to disguise even highly familiar shapes. Clearly, then, some general grouping rules were overriding learned knowledge about the shape of objects. The second was the fact that perceptual organization could be demonstrated with very young animals.

To the arguments offered by the Gestaltists can be added the following one: From an engineering point of view, it is generally easier to design a machine that can do some task directly than to design one that can *learn* to do it. We can design machines that can parse or generate fairly complex sentences, but there has been limited success in designing one that could learn grammatical rules from examples without any designed-in knowledge of the formal structure of those rules. By analogy, if you think of the physical world as having a "grammar" (the physical laws that are responsible for the sensory impressions that we receive), then each human must be equipped

either with mechanisms capable of learning about many of these laws from examples or with a mechanism whose genetic program has been developed once and for all by the species as a result of billions of parallel experiments over the course of history, where the lives of the members of the species and its ancestors represent the successes and the lives of countless extinct families the failures. To me, evolution seems more plausible than learning as a mechanism for acquiring at least a general capability to segregate sounds. Additional learning-based mechanisms could then refine the ability of the perceiver in more specific environments.

The innate influences on segregation should not be seen as being in opposition to principles of learning. The two must collaborate, the innate influences acting to "bootstrap" the learning process. In language, meaning is carried by words. Therefore if a child is to come to respond appropriately to utterances, it is necessary that the string be responded to in terms of the individual words that compose it. This is sometimes called the segmentation problem. Until you look at a spectrogram of continuous speech occurring in natural utterances, the task seems easy. However, on seeing the spectrogram, it becomes clear that the spaces that we insert into writing to mark the boundaries of words simply do not occur in speech. Even if sentences were written without spaces, adults could take advantage of prior knowledge to find the word boundaries. Because they already know the sequences of letters that make meaningful words, they could detect each such sequence and place tentative word boundaries on either side of it. But when infants respond to speech they have no such prior learning to fall back on. They would be able to make use only of innate constraints. I suspect a main factor used by infants to segment their first words is acoustic discontinuity. The baby may hear a word as a unit only when it is presented in isolation, that is, with silence (or much softer sound) both before and after it. This would be the result of an innate principle of boundary formation. If it were presented differently, for example, as part of a constant phrase, then the phrase and not the word would be treated as the unit. The acoustic continuity within a sample of speech and the discontinuities at its onset and termination would be available, even at the earliest stage of language acquisition, to label it as a single whole when it was heard in isolation. Once perceived as a whole, however, its properties could be learned. Then, after a few words were learned, recognition mechanisms could begin to help the segmentation process. The infant would now be able to use the beginnings and ends of these familiar patterns to establish boundaries for other words that might lie between them. We can

see in this example how an innate grouping rule could help a learning process to get started. (I am not suggesting that the establishing of acoustic boundaries at discontinuities is the only method that infants use to discover units, but I would be very surprised if it were not one of them.)

Another example of innate segregation that was given earlier concerned an infant trying to imitate an utterance by her mother. It was argued that the fact that the infant did not insert into her imitation the cradle's squeak that had occurred during her mother's speech displayed her capacity for auditory scene analysis. I am also proposing that this particular capacity is based on innately given constraints on organization.

There is much experimental evidence drawn from experiments on the vision of infants that supports the existence of innate constraints on perceptual organization. Corresponding experiments on auditory organization, however, are still in short supply.

One such study was carried out by Laurent Demany in Paris.[15] Young infants from $1\frac{1}{2}$ to $3\frac{1}{2}$ months of age were tested with sequences of tones. The method of habituation and dishabituation was used. This is a method that can be used with infants to discover whether they consider two types of auditory signals the same or different. At the beginning, a sound is played to the babies every time they look at a white spot on a screen in front of them. The sound acts as a reward and the babies repeatedly look at the white spot to get the interesting sound. After a number of repetitions of this "look and get rewarded" sequence, the novelty of the sound wears off and it loses its potency as a reward (the infants are said to have habituated to the sound). At this point the experimenter replaces the sound by a different one. If the newness of the sound restores its ability to act as a reward, we can conclude that the infants must consider it to be a different sound (in the language of the laboratory, they have become dishabituated), but if they continue ignoring it, they must consider it to be the same as the old one.

Using this method, Demany tried to discover whether infants would perceptually segregate high tones from low ones. The proof that they did so was indirect. The reasoning went as follows: Suppose that four tones, all with different pitches, are presented in a repeating cycle. Two are higher in pitch (H1 and H2) and two are lower (L1 and L2), and they are presented in the order H1,L1,H2,L2, ... If the high and low tones are segregated into different perceptual streams, the high stream will be heard as

H1–H2–H1–H2–H1–H2–. . .

and the low stream will be perceived as

L1–L2–L1–L2–L1–L2–. . .

(where the dashes represent brief within-stream silences). In each stream all that is heard is a pair of alternating tones.

Now consider what happens when the reverse order of tones is played, namely L2,H2,L1,H1, ... If the high tones segregate from the low ones, the high stream is heard as

H2–H1–H2–H1–H2–H1–. . .

and the low one as

L2–L1–L2–L1–L2–L1–. . . .

Again each stream is composed of two alternating tones. In fact, if the infant lost track of which one of the pair of tones started the sequence, the two streams would be considered to be exactly the same as they were with the original order of tones. Suppose, however, that the infant does not segregate the high from the low tones. In this case the forward and the backward orders of tones are quite different from one another and remain so even if the infant forgets which tone started the sequence.

To summarize, the segregated streams are quite similar for the forward and backward sequences whereas the unsegregated sequences are quite different. Using the habituation/dishabituation method, Demany tried to determine whether the infants considered the forward and backward sequences the same or different. The results showed that they were reacted to as being the same. This implied that stream segregation had occurred. In addition, Demany showed that this result was not due to the fact that the infants were incapable in general of distinguishing the order of tonal sequences. Pairs of sequences whose segregated substreams did not sound similar to an adult were not reacted to as being the same by infants. In general, the infant results paralleled those of adult perception and the older and younger infants did not differ in their reactions.

Undoubtedly more such research is required. After all, the infants were not newborns; they had had some weeks of exposure to the world of sound. But after this pioneering study, the burden of proof shifts to those who would argue that the basic patterns of auditory organization are learned. Unfortunately, working with very young infants is difficult and the amount of data collected per experiment is small.

The unlearned constraints on organization can clearly not be the only ones. We know that a trained musician, for example, can hear

the component sounds in a mixture that is impenetrable to the rest of us. I have also noticed that when researchers in my laboratory prepare studies on perceptual organization, they must listen to their own stimuli repeatedly. Gradually their intuitions about how easy it is to hear the stimulus in a particular way come to be less and less like the performance of the untrained listeners who are to serve as the subjects of the experiment.

Undoubtedly there are learned rules that affect the perceptual organization of sound. I shall refer to the effects of these rules as "schema-based integration" (a schema is a mental representation of some regularity in our experience). Schema-based analysis probably involves the learned control of attention and is very powerful indeed. The learning is based on the encounter of individuals with certain lawful patterns of their environments, speech and music being but two examples. Since different environments contain different languages, musics, speakers, animals, and so on, the schema-based stream segregation skills of different individuals will come to have strong differences, although they may have certain things in common. In later chapters, I will give some examples of the effects of schema-governed scene analysis in the fields of music and language, and will discuss a theory of sequential integration of sound, proposed by Mari Riess Jones, that is best understood as describing the influence of schemas on stream segregation.

Verification of the Theory

The theory presented in this volume proposes that there is an auditory stream-forming process that is responsible for a number of phenomena such as the streaming effect and the illusion of continuity, as well as for the everyday problems of grouping components correctly to hear that a car is approaching as we cross a street, or "hearing out" a voice or an instrument from a musical performance. This is not the type of theory that is likely to be accepted or rejected on the basis of one crucial experiment. Crucial experiments are rare in psychology in general. This is because the behavior that we observe in any psychological experiment is always the result of a large number of causal factors and is therefore interpretable in more than one way. When listeners participate in an experiment on stream segregation, they do not merely perceive; they must remember, choose, judge, and so on. Each experimental result is always affected by factors outside the theory, such as memory, attention, learning, and strategies for choosing one's answer. The theory must therefore be combined with extra assumptions to explain any particular outcome. Therefore it cannot easily be proven or falsified.

Theories of the type I am proposing do not perform their service by predicting the exact numerical values in experimental data. Rather they serve the role of guiding us among the infinite set of experiments that could be done and relationships between variables that could be studied. The notion of stream segregation serves to link a number of causes with a number of effects. Chapter 2, for example, will show how stream segregation is affected by the speed of the sequence, the frequency separation of sounds, the pitch separation of sounds, the spatial location of the sounds, and many other factors. In turn, the perceptual organization into separate streams influences a number of measurable effects, such as the ability to decide on the order of events, the tendency to hear rhythmic patterns within each segregated stream, and the inability to judge the order of events that are in different streams. Without the simplifying idea of a stream-forming process, we would be left with a large number of empirical relations between individual causal influences and measurable behaviors.

A theory of this type is substantiated by converging operations. This means that the concepts of "perceptual stream" and "scene-analysis process" will gain in plausibility if a large number of different kinds of experimental tasks yield results that are consistent with these ideas. With this in mind, in the remainder of this volume I will try to set out the pieces of evidence that fit together into a mutually supporting whole.

Summary

I started this chapter with a general introduction to the problems that would be considered in more detail in later chapters. I began with the claim that audition, no less than vision, must solve very complex problems in the interpretation of the incoming sensory stimulation. A central problem faced by audition was in dealing with mixtures of sounds. The sensory components that arise from distinct environmental events have to be segregated into separate perceptual representations. These representations (which I called streams) provide centers of description that connect sensory features so that the right combinations can serve as the basis for recognizing the environmental events. This was illustrated with three auditory phenomena, the streaming effect, the decomposition of complex tones (the ABC experiment), and perceptual closure through occluding sounds.

The explanation that I offered had two sides. It discussed both perceptual representations and the properties of the acoustic input that were used heuristically to do the segregation. I argued that one had to take the ecology of the world of sound into account in looking for the

methods that the auditory system might be using, and claimed that this could serve as a powerful supplement to the Gestalt theorist's strategy of looking for formal similarities in the activity of different senses. Finally I proposed that there were two kinds of constraints on the formation of perceptual representations, unlearned primitive ones and more sophisticated ones that existed in learned packages called schemas.

These theoretical ideas will be used in the remainder of the book to analyze the known evidence on auditory scene analysis. One chapter is devoted to looking at the problem of grouping auditory components sequentially. Another will look at the grouping of simultaneous sounds. There is a separate chapter that looks at the possible differences between primitive and schema-driven integration of sensory evidence. Eventually we look at the role of scene analysis in music and in speech. The reader will discover that there is actually an impressive body of data that can serve to constrain theorizing about auditory scene analysis.

The next chapters present a fairly detailed analysis of how auditory scene analysis is accomplished and analyze the supporting evidence. The reader who is interested only in the conclusions can turn to the final chapter for a summary of what we do and do not know at the present time.