Chapter 8

Summary and Conclusions: What We Do and Do Not Know about Auditory Scene Analysis

Summary of Previous Chapters

This chapter presents my conclusions about what we do and do not know about auditory scene analysis at present. The sequence of topics exactly follows that of the previous chapters, so if a point is not clear here, or if fuller details are wanted, the reader should have no difficulty in finding the corresponding material in the earlier discussion. In this chapter I present no references to the experiments on which the conclusions are based. The interested reader can look for them in the earlier chapters. Sometimes I have not undertaken to define a technical word here. In such cases, a rough definition can be found in the glossary at the end of the book.

Primitive Auditory Scene Analysis

The problem of scene analysis is this: Although we need to build separate mental descriptions of the different sound-producing events in our environments, the pattern of acoustic energy that is received by our ears is a mixture of the effects of the different events. It appears that our auditory systems solve the problem in two ways, by the use of primitive processes of auditory grouping and by governing the listening process by schemas that incorporate our knowledge of familiar sounds. This book has been mainly about the primitive process, although it has tried to discover whether the effects of the two kinds of processes can be distinguished from each other. I shall begin this summary by talking about the primitive process.

The primitive process of scene analysis seems to employ a strategy of first breaking down the incoming array of energy into a large number of separate analyses. These are local to particular moments of time and particular frequency regions in the acoustic spectrum. Each region is described in terms of its intensity, its fluctuation pattern, the direction of frequency transitions in it, an estimate of where the sound is coming from in space, and perhaps other features. After

these numerous separate analyses have been done, the auditory system has the problem of deciding how to group them so that each group has been derived from the same environmental event. The grouping has to be done in two dimensions (at least), across time and across the spectrum. I have called the temporal grouping "sequential integration" and have referred to the other one as "simultaneous integration." The chapters in this volume have been laid out according to this division. However, I have pointed out that the two forms of grouping often operate in conjunction to solve the problem.

Sequential Integration: Auditory Stream Segregation

Sequential integration is visible in a number of contexts, but a popular stimulus pattern for studying it has been the one that gives rise to auditory stream segregation. This occurs when a sequence of tones jumps rapidly up and down between different frequency regions. The simplest case is a rapid repeated alternation of a high and a low tone. If the alternation is fast enough and the frequency separation great enough, listeners will not experience a single stream of tones alternating in pitch, but will perceive two streams of tones, one consisting of repetitions of the lower tone and the other consisting of repetitions of the higher one. When two streams are heard, the listener has the impression of two different sources of sound, one high pitched and the other low, whose tones happen to occur at about the same time. In more complex patterns, where a number of higher tones of slightly different frequencies are interleaved with a number of lower tones, listeners will still hear two streams, but this time each stream will have a melodic pattern that is restricted to the tones of that stream.

One sort of stimulus that has been used is a short repeating loop of tones. The tones are chosen from two different frequency regions, one high and one low. The properties of the tones have been varied and the effects on stream segregation observed.

Another form of stimulus has been a tune or other short tonal pattern interleaved with distractor tones. The frequency relations between the distractor and the relevant tones have been varied. When the two sets of tones are in the same frequency region, the tune disappears into a pattern formed out of all the tones, but if they are in two nonoverlapping frequency regions, the tune is readily heard as a separate stream. This sort of stimulus has been used to study primitive segregation but is more suited for the study of schema-based segregation.

Auditory stream segregation has been known by musicians since the Baroque period, when it was used to produce the impression of

two lines of melody even though the instrument playing the sequence could produce only one note at a time. This could be done by having the instrument rapidly alternate between a higher and a lower melodic line.

Factors Influencing Stream Segregation The most important influences on the segregation of the streams are the rate of the tonal sequence and the frequency separation between the two subsets of tones that are interleaved. As the subsets are moved further apart in frequency, the segregation increases and it becomes harder and harder for a listener to hear the entire sequence as a single stream of sound. (A logarithmic scale seems to be one that reflects the segregability of the frequencies.) Furthermore the segregation increases when the sequence goes faster. As a consequence of these facts, the effects of frequency separation and time can be traded off against one another. As the frequency separation increases, the sequence must be slowed down if the listener is to be able to experience all the tones as part of a single, coherent stream of sound.

However, the effects of frequency separation and speed depend on what the listeners are trying to do. If they are trying to hear all the tones as part of a single sequence, the effects are as I have just described them. But if they are trying to focus their attention on the tones of just one of the streams, the effects of frequency separation and speed are different. The frequency separation of the high from the low tones need only exceed some small amount (a few semitones in the case of two alternating tones) before the target sequence can be followed by attention. Further increases in separation do not increase the segregation and the capacity to follow the selected stream is virtually unaffected by the speed of the sequence. Because of the difference in the effects when listeners are trying to hear coherence or segregation, I have proposed that two different factors are at work. One, primitive segregation, is affected by the rate and the frequency separation. The other, schema-based segregation that involves attention, is used for focusing on one of the streams. The attention of the listener can always be directed toward a narrow frequency range when required by the task. Beyond the minimum frequency separation that it requires so as not to confuse the notes of the stream that it is following with the tones of other streams, the attentional process is unaffected by the separation between the frequencies.

We know that the result of different speeds comes from the fact that certain temporal intervals have been affected. But we do not know exactly which ones. If the effect is based on bringing the tones from the same frequency range closer together in time, we would

expect that the time interval between the end of one tone and the beginning of another in the same range would be the important one. Yet some evidence suggests that it is the separation of the *onsets* of the two that is important. There is another question too. Which is important: the interval between successive tones in the same frequency range or that between successive ones in the up-and-down sequence? Since most studies use sequences in which all tones and intertone silences are the same length, it is impossible to tell which of these intervals is the critical one.

It appears that the stream-forming process behaves in a manner analogous to the Gestalt principle of grouping by proximity. The high tones tend to group with other high ones if brought close to them in time by the speeding up of the sequence.

When we talk about the proximity of *tones* or *sounds* in frequency or in time, we are implying that the stream of sound is composed of discrete units. But what about cases in which the sound is more continuous? Where are the units? It appears as though there is a unit-forming process that is sensitive to discontinuities in the sound, particularly to sudden rises in intensity, and that creates unit boundaries when such discontinuities occur. Units can occur at different time scales and smaller units can be embedded in larger ones. When the sequence is speeded up, the changes that signal the smaller units may be missed by the auditory system, and other changes, too gradual to form units at the slower speed, may now be sudden enough to control the formation of units.

The units, once formed by these processes, can form groups with other similar ones. Similarity is determined by analyses that are applied to the units once they are formed. For example, suppose there is a glide in frequency, bounded by a rise and a fall in intensity. Between these boundaries, the change in frequency may be measured by the auditory system and assigned to the unit as one of its properties. This frequency-gliding unit will prefer to group with other ones whose frequency change has the same slope and which are in the same frequency region.

One of the similarities that affects the grouping of tones is their location in space. Engineers working on the automatic segregation of concurrent sounds have used spatial separation as a uniquely powerful way of determining whether the sounds have come from the same physical event (usually a talker). Humans use spatial origin too, but do not assign such an overwhelming role to it. They can do quite well at segregating more than one stream of sound coming from a single point in space, for example, from a single loudspeaker.

Primitive scene analysis tends to group sounds that come from the same point in space and to segregate those that come from different places. As a consequence, if two sounds, different in frequency, are alternated between the ears, they will not form a single coherent stream. The frequency separation, the rate, and the spatial separation combine to influence the segregation. Spatial differences seem to have their strongest effects on segregation when they are combined with other differences between the sounds. Illusions can be created by setting up a competition between the tendency to group sounds by their frequency similarity and by their spatial similarity. An example is Diana Deutsch's scale illusion.

When speech is rapidly switched back and forth between the ears, it is hard to follow. One reason is that when the switch occurs, it produces a sudden rise in intensity in one ear and a sudden drop in the other. If the listener simply combined the two changes there would be no net change in the signal. But this is not what happens. The changes in the two ears are treated as separate events. As a result, false beginnings and ends of syllables are perceived and this hurts the recognition of the speech.

When clicks are alternated between the two ears, the click rate seems to be slower than when all the clicks come to the same ear. When listeners are asked to count the total sequence, they are worse at doing so when the clicks are alternated than when they are not. Both effects seem to be the results of spatially based stream segregation.

So far I have spoken about frequency differences between tones. Most readers translate this to "pitch differences." This is all right as long as we are dealing with pure tones whose pitches are in one-to-one correspondence with their frequencies. Yet we know that in complex tones, having many frequency components, this simple equivalence can break down.

In a complex tone, the perceived pitch depends on the auditory system's estimate of the fundamental frequency of the set of harmonics in the tone. Yet this fundamental itself need not actually be present. All that is required (to a first approximation) is that all the partials that are present should be multiples of this fundamental. We can therefore have a case in which a tone with a lower pitch can have partials whose average is higher in frequency that the average of those in a tone whose pitch is higher. (The tone with the higher harmonics will sound "brighter" even though it has a lower pitch.)

This allows us to ask a question that we could not ask about sequences of pure tones. Is it the difference in the fundamental frequency (pitch) of the tones or the difference in the average frequency of its

partials (brightness) that affects perceptual grouping? The answer is that they both do and that the effects are additive. A pure tone has a different spectral content than a complex tone; so even though the pitches of the two may be the same, the tones will tend to segregate from one another in rapid sequences. Another kind of grouping also occurs: A pure tone, instead of grouping with the entire complex tone that follows it, may group with one of the frequency components of the latter.

Pitch and brightness are both one-dimensional properties of a spectrum of sound. However, when a spectrum has a number of peaks in it (as the spectrum of of a vowel does), there can be many ways in which one spectrum resembles another. Few studies of the stream segregation of sequences containing such complex sounds have been done and those that have been carried out have not looked analytically at the dimensions that the auditory system uses to summarize the various complex shapes of spectra.

There is another type of highness that can be heard in bands of filtered noise. When the noise burst contains higher-frequency components, it sounds higher. Differences in frequency content can cause sequences formed out of bursts of noise to segregate into high and low streams in the same way as sequences formed out of tones.

Timbre Timbre is another factor that affects the similarity of tones and hence their grouping into streams. The difficulty is that timbre is not a simple one-dimensional property of tones. It may not even be reducible to a small number of dimensions.

Probably one distinct dimension of timbre is brightness. We know that tones that are similar in their brightness will tend to be assigned to the same stream. Brightness is, roughly speaking, the mean frequency that is obtained when all frequency components of a sound are weighted according to their loudness. Bright tones have more of their energy concentrated in higher frequencies than dull tones do.

The difference in quality between noises and tones can also be thought of as a timbre difference. The frequency components of noises constantly change in amplitude and phase whereas those of tones are fairly constant. A rapid sequence can be created by alternating pure tones with noises whose center frequencies have been matched to those of the tones. The sequence will segregate into one formed out of the noises and another formed out of the tones. Much more research needs to be done on the effects of spectral noisiness in controlling the grouping of sounds.

I have suggested that the pattern of peaks and valleys in the spectra of sounds may affect their grouping. Another way of saying this is

that grouping is affected by the pattern of the intensities of various harmonics in the spectra of successive tones. However, we do not know how to compare the patterns in two successive tones whose fundamentals differ. We could consider two tones as having the same pattern of intensities if their harmonics had their peaks at exactly the same frequencies. In nature, this would probably mean that they had passed through the same set of resonators (the vocal tract of the same talker for example). On the other hand, we could also consider them the same if the corresponding harmonics were of proportional intensity. This would mean that if the fundamental frequency of the second tone were double that of the first, all the peaks in the spectrum would also be at double the frequency. In nature this would mean that the properties of the two vibrating bodies were similar (rather than the properties of the resonators that they had been passed through). The available evidence (and it is scanty) suggests that both forms of spectral similarity are used in auditory scene analysis to group successive tones.

Most of the research on the sequential grouping of sounds has used either tones or noise bursts that are presented in regular rhythmic patterns. Yet most of the sounds in the natural world are not like this. For example, natural sounds change their properties over time. Think of a voice, which has all kinds of sounds in it, or a plucked guitar string, which has a abrupt, noisy, and intense onset and dies out gradually to almost a pure tone. There has been almost no research on the sequential grouping of changing tones other than a few studies in the field of speech and studies on the grouping of pure-tone glides. The study of how changing tones are perceptually organized affords a rich opportunity for research. Among the factors that deserve study are the abruptness of the onset, the changes over time in the strengths of harmonics, the fluctuations in the overall intensity and pitch of the sound, and the granularity of the sound.

If you take each moment of each frequency component into account, you realize that sounds can differ from one another acoustically in an astonishingly large number of ways. Does the auditory system deal with this complexity by collapsing the differences into a small number of dimensions? We know that loudness, pitch, and possibly brightness are separable dimensions, but is there a limited number of other dimensions of timbre? A demonstration that we understood some set of dimensions would be our ability to construct metameric timbres. In vision, metameric colors are ones that look identical despite the fact that their spectral content is different. By analogy, metameric timbres would sound the same (in some respect) despite obvious acoustic differences. For our purposes in the study of

scene analysis, it would not be so important that they sounded exactly alike as that their grouping tendencies were identical. That is, the substitution of one for the other would leave the groupings the same in patterns in which they appeared.

It has been argued that amplitude differences between sounds will control their grouping. Loud sounds will tend to group with other loud ones and soft ones with soft. However, the results obtained with tones differing in loudness makes me question whether the grouping based on loudness is the result of primitive scene analysis or of some schema-governed process of selection. The resolution of this issue may be as follows. Tones that differ only in loudness may not have a tendency to segregate from one another, but when there are also other differences between the sounds, the loudness differences may strengthen the segregation. Again we really do not know and research is required to resolve the question.

When you listen to a repeating alternation of high and low tones that are sufficiently far apart in frequency, at first you can follow the alternation of tones as a single stream, but eventually the sequence seems to split into two separate streams, one high and the other low. This shows that there is a cumulative effect of the alternation between frequency ranges. The tendency for primitive stream segregation to subdivide the sequence builds up for at least 4 seconds and takes at least 4 seconds to go away after the sequence stops. I have interpreted this sluggishness of the changes as having a useful function. It prevents the system from oscillating wildly among different ways of organizing the auditory scene. Once some interpretation of a number of sonic sources has occurred it does not disappear instantly just because some source has not been heard from for a second or two. This type of "hysteresis" is seen in a great number of perceptual phenomena in different sense modalities.

It appears, however, that a sudden change in the acoustic properties of the signal can reset the streaming mechanism more quickly than mere silence can. For instance, a sudden change in the frequency range occupied by the signal or the spatial location from which the sound seems to be coming can cause a segregated sequence to revert to being heard as unsegregated. Obviously the scene-analysis system treats these sorts of changes as indicating that it is now encountering a new sonic event.

The cumulative effect of exposure to a sequence that is alternating between frequency ranges has been explained in different ways. The explanation that I favor says that the auditory system is gradually building up evidence that the sequence actually contains different subsets of sounds with distinct properties and that they should be sorted

into separate streams. An alternative explanation of the cumulative effects that are seen in extended exposures to the sequence is that the integration into a single stream is accomplished by means of frequency-jump detectors. When the sequence continues to jump back and forth in frequency the detectors get tired out. At this point only shorter jumps can be followed and the sequence breaks into substreams. Although the two explanations look different they are not necessarily incompatible. Notice that the first is a functional one while the latter is stated in terms of physiology. It may be that the physiological events described by the second are serving the function described by the first.

Continuous sounds hold together better as a single stream than discontinuous ones do. This can be shown by comparing two sorts of sequences. In a discontinuous sequence, there is an alternation of high and low tones and all the tones are steady in pitch and separate from their neighbors. In a continuous sequence, each high tone is joined to its two adjacent low neighbors by a frequency glide. The continuous sequence holds together better than the discontinuous one. The coherence of continuous sequences can be interpreted in functional terms as a heuristic of the auditory system. This heuristic is equivalent to a bet that any sequence that exhibits acoustic continuity has probably come from a single environmental event.

The use of continuity does not necessarily imply that the auditory system tracks changes and predicts the properties of the next moment of sound. I do not think that the primitive process does this. However, there is good reason to believe that schema-based processes of integration do employ this strategy.

Summary of Factors Promoting Sequential Grouping Many of the factors that favor the grouping of a sequence of auditory inputs are features that define the similarity and continuity of successive sounds. These include their fundamental frequency, their temporal proximity, the shape of their spectra, their intensity, and their apparent spatial origin. These characteristics affect the sequential aspect of scene analysis. My description has seemed to imply that the things that group sequentially can be thought of as sounds. This was because the examples that I have given were all stated in terms of rather simple sounds rather than in terms of mixtures of sound. We find that the same factors serve to promote the sequential grouping of sound in mixtures, but, in this case, it is not whole sounds but parts of the spectrum that are caused to group sequentially. The resulting grouping helps the brain to create separate descriptions of the component sounds in the mixture.

Effects of Stream Segregation Generally speaking, the perceptual effects of sequential integration and segregation follow from their role in scene analysis. Auditory material that is assigned to the same stream has a much stronger tendency to be used together in any perceptual computation. Emergent properties of the sound are likely to be computed from within-stream elements. Therefore sequential patterns that involve elements of the same stream will be more easily perceived. This general description takes many specific forms.

It can be shown that our perceptual representations of any pattern tend to include material that is within a stream and to exclude material that is not in that stream. For example, if listeners are asked to recognize a pattern that is interleaved with distractor sounds, any factor that causes the distractors to fall into a separate stream will make the pattern easier to recognize. This is true regardless of whether the pattern that they are trying to hear is a familiar melody or an unfamiliar sequence that they have just been played as a "standard" and are holding in their immediate memories.

In some sense we are talking about camouflage and are saying that only within-stream material can camouflage a target. Even such a simple task as counting a rapid sequence of tones can be carried out more accurately when all the tones fall into a single stream.

So far most of the recognition research has used pitch differences to segregate the streams, but there is no reason that the other factors that I have listed as affecting sequential integration could not be used.

Temporal relations can also be more easily perceived when they involve elements that have been grouped sequentially by auditory scene analysis. For example, it is difficult to detect the order of all the sounds in a rapid repeating cycle when they fall into more than one stream. (If we want to create a clear demonstration of this fact, we have to design the task in such a way that it is impossible for a listener to get the right answer by considering only one stream at a time.)

Stream segregation can also affect the rhythm of a perceived sequence. For example, if we were to create a sequence of equally spaced tones in which every third tone was an octave higher than the others, it would divide into two streams. The higher-pitched one would have this rhythm:

—H—H—H. . .

The lower one would have this one:

LL–LL–LL–. . . .

A rhythm tends to be defined by sounds that fall into the same stream.

The perceived temporal overlap of sounds is also affected by segregation. If a rapid cycle of six tones consists of an alternation of three high and three low ones, all of equal duration, it splits perceptually into two streams. As you listen to the two streams you often get the impression that the high and the low cycles are running at different speeds. It is very difficult to judge whether the tones in the high stream are temporally overlapped with those in the lower one or not.

Our ability to detect a temporal gap between two tones seems to be affected by segregation even when the tones are not part of a longer sequence. It becomes harder and harder to judge the length of the gap as the frequency difference between the two adjacent tones increases. It is not absolutely certain whether this is really a result of stream segregation or is some other effect of frequency separation. To find out, we would have to see whether we could obtain the same effect if we replaced the frequency difference by some other factor that was known to affect stream segregation.

It appears that the factors that promote the grouping of auditory material act in a competitive way. For example, suppose we had a four-tone cycle, ABCD. If two of the tones, A and B, were separated in frequency from the other two, then A and B might form a separate stream of their own. Yet in a different four-tone cycle, ABXY, those same A and B tones might appear in two separate streams if one of these tones grouped more strongly with X and the other with Y. It appears that frequency proximities are competitive and that the system tries to form streams by grouping the elements that bear the strongest resemblance to one another. Because of this competition, we are able to create experiments in which we can capture an element out of a sequential grouping by giving it a better sound to group with.

This competition also occurs between different factors that favor grouping. For example, in a four-tone sequence ABXY, if similarity in fundamental frequency favors the groupings AB and XY, while similarity in spectral peaks favors the grouping AX and BY, then the actual grouping will depend on the relative sizes of the differences. If the differences in fundamental frequency are large while the differences in spectral peak positions are small, the fundamental frequency differences will control the grouping.

There is collaboration as well as competition. If a number of factors all favor the grouping of sounds in the same way, the grouping will be very strong and the sounds will always be heard as parts of the same stream. The process of competition and collaboration is simple to conceptualize. It is as if each acoustic dimension could vote for a grouping, with the number of votes it cast being determined by the

degree of similarity on that dimension and on the importance of the dimension. Then the streams whose elements were grouped by the most votes would be formed. Such a voting system would be valuable in a natural environment in which it is not guaranteed that sounds that resemble one another in only one or two ways will always have arisen from the same acoustic source.

Competition and capturing would not be possible if a sound could be in two streams at the same time. Giving a sound something better to group with would not remove it from the original stream. The fact that we can capture sounds out of streams implies that the brain has a bias against having sounds in two different streams at the same time. However, this bias toward "exclusive allocation" is not absolute. When the auditory system encounters complex spectra, it sometimes decides that two sounds with shared spectral components are present at the same time. If this decision is made, some of the spectral components are used to derive the properties of more than one perceived sound.

Any of the effects that I have mentioned can be used to measure the strength of stream segregation. Some effects, however, are not as useful in experiments because they are strongly influenced by factors other than primitive grouping. The most reliable indication that a stream has been formed is the exclusion of certain sounds from a perceived pattern even though listeners are trying to include them. It is not as valid a measure if they are attempting to exclude them. Their success in doing so may not be due to primitive preattentive scene analysis but to schema-governed processes of attention. Therefore, failure of inclusion is a better measure than success in exclusion.

The sequential grouping that is observed in stream segregation has two analogies in vision. The first is the tendency for visual perception to group elements that are near to one another in space. We often see the same groupings in the visual diagrams of patterns of sounds as we hear in the sounds themselves. Apparently, proximity has similar effects on grouping in vision and audition. A more dynamic analogy to stream segregation in vision is apparent motion. Chapter 1 showed how the two phenomena display very similar effects.

Not only do vision and audition show certain similarities to each other but events in vision can affect how sounds are perceived and vice versa. We may already be disposed at birth to relate vision to audition. A newborn will spend more time looking at a face that appears visually to be speaking the same words that it is hearing than one that is not. An example of the interrelationship is that the grouping of sounds can influence the grouping of visual events with which

they are synchronized and vice versa. For instance, suppose that two lights are too far apart to give good apparent motion when they are flashed in alternation (that is, they are not treated as parts of the same event). We can cause the perceived movement to be improved if the lights are synchronized with high and low tones, one light with each tone, provided that the two tones themselves are treated as parts of a single acoustic event. We can arrange this by placing them close enough together in frequency to fall into the same stream. It is not yet clear whether this particular form of coordination between vision and audition serves any useful purpose. Other forms of coordination between vision and audition have more obvious benefits. For example, the tendency to experience a sound as coming from a location at which visual events are occurring with the same temporal pattern (the so-called ventriloquism effect) can be interpreted as a way in which visual evidence about the location of an event can supplement unclear auditory evidence. The direction of influence is not just from vision to audition but in the reverse direction as well.

I have offered an explanation of stream segregation in terms of scene analysis. However, other forms of explanation have been offered by others. Some of them are physiological. One proposal is that overlapping hair cell populations in the inner ear must be stimulated by successive tones before a sequence can be integrated. When this condition is violated, separate streams are formed so as to group the tones that do conform to the requirement. A piece of evidence against this explanation is that under some circumstances tones from opposite ears can be grouped into the same stream. These tones are obviously not affecting the same hair cells. Another theory that I have already mentioned in this chapter is that a frequency-jump detector must register the transition between successive tones before the sequence can be integrated. With repeated frequency alternation, the detectors get tired out and can only follow shorter jumps. This theory has the advantage that it is the analogue of a motion-detector theory in vision that has been successful in explaining certain visual phenomena. Neither of these physiological theories can explain all the facts about the formation of auditory streams, but even if they could, they would not compete with the scene analysis explanation, which is functional rather than physiological in nature. They would merely describe the machinery by means of which scene analysis takes place.

However, functional theories have been offered that compete more directly with the theory of primitive preattentive grouping of acoustic evidence. These theories see the site of segregation and grouping as being within the attentional process itself. Attention is seen as trying to follow the changes in the stimulus. Rapid changes may ex-

ceed the capacity of attention to follow them. Sometimes the process that integrates successive sounds has been described as a filter that must pass all the sounds if they are to be incorporated into the same act of attention. The filter is conceptualized as being able to change its own setting with respect to the range of properties that it is tuned to (such as the range of frequencies that it will pass) but to be unable to change it very rapidly. Therefore it misses sudden changes in the sound. This is the explanation that is offered to explain why a sound is sometimes perceptually excluded from a stream. According to this view, only one stream of sound exists at a time, the one you are paying attention to. There is no such thing as a second grouping of perceptual evidence that is structured even if you are not paying attention to it. This is one important way in which this class of theory differs from the theory of primitive scene analysis. The latter says that links are formed between parts of the auditory evidence even though these parts may not currently be within the field of attention.

Another theory is similar to the attention-as-filter theory except that it sees the attention as being able to integrate a sequence when the changes in it can be anticipated as a result of prior learning. This tendency is seen as partially overcoming the inability to track sudden changes.

The Gestalt theory of grouping is similar to the idea of preattentive grouping. It sees the effects of similarity, temporal proximity, and continuity as innate principles that determine grouping. The idea of competitive grouping forces is also part of the theory. However, the Gestalt psychologists did not emphasize the relevance of these principles to the practical task of scene analysis.

Spectral Integration

I have summarized the facts about the sequential grouping of auditory evidence, but this is only a part of the story. In mixtures of sound the auditory system must decide which components, among those that are received concurrently, should be treated as arising from the same sound. This process was studied in simple experiments in which two concurrently presented pure tones, B and C, were alternated with a pure tone, A (see figure 1.16 of chapter 1). It was found that if B and C started and ended at the same time, they tended to be treated as two components of a single complex tone, BC, that was perceived as rich in quality. On the other hand, there was a tendency to treat B as a repetition of A whenever A was close in frequency to B. B seemed to be the object of a rivalry. When it was captured into a sequential stream with A, it was less likely to be heard as part of the complex

tone, BC. Conversely, when it was captured by C and fused with it, it was less likely to be heard as a repetition of A. It seemed that sequential grouping and spectral grouping were in a competition that served to resolve competing evidence concerning the appropriate grouping of sensory material.

Gliding tones can be captured out of mixtures too (see figure 2.17 of chapter 2). A gliding complex tone that is really a mixture of simultaneously gliding pure tones can have one of its components captured out of it by a preceding pure-tone glide.

Factors Influencing Spectral Integration If we look at a spectrogram of a mixture of sounds, as in figure 1.4 of chapter 1, we find that the spectral content arriving from one sound overlaps the components of the remainder of the sound both in frequency and in time. How can the auditory system know which frequency components to group together to build a description of one of the sounds? It seems that it does so by looking for correlations or correspondences among parts of the spectral content that would be unlikely to have occurred by chance.

One type of correspondence is between the auditory properties of different moments of time. A complex spectrum may have, embedded within it, a simpler spectrum that was encountered a moment earlier. That simpler spectrum might, for example, abut against the more complex one with no discontinuity. In such a case, it is reasonable to treat the part of the spectrum that matches the earlier one as merely a continuation of it and treat the rest of the later one as resulting from the addition of a new sound to the mixture. This could be referred to as the "old-plus-new" heuristic. It is this strategy that is observed in experiments in which a component is captured out of a complex tone by a preceding pure tone.

The grouping of a part of the current auditory input with earlier material depends on how similar they are. We know that at least two factors influence the similarity: frequency separation and (in the case of gliding components) the direction of frequency change. They also group more strongly if there are shorter silences separating them. Recall that these factors are identical to those that determine sequential grouping of simple tones. This leads us to believe that the old-plus-new heuristic is just another manifestation of the principles that control sequential grouping.

Another aspect of this heuristic is that the factoring out of the old spectrum from the current one creates a residual whose properties are heard more clearly. There is even some evidence that the auditory system uses the amplitudes of the various spectral components of the

earlier spectrum to decide not only *which* spectral components to subtract out but also *how much intensity* to leave behind at each frequency. This is a good strategy because the old and new sounds might have some frequency components that are the same. The subtraction (or a process roughly equivalent to subtraction) provides an estimate of the probable intensity of the frequency components of the sound that has been added to the first to create the complex spectrum.

We also have other ways of deciding which components, received at the same time, should be grouped to form a description of a single auditory event. Certain types of relations between these components can be used as clues that they should be grouped. The effect of this grouping is to allow global analyses of factors such as pitch, timbre, loudness, and even spatial origin to be computed on a set of sensory evidence that probably all came from the same event in the environment.

Some of the clues are based on the frequency relations between components. The first is their frequency separation. The further they are away from one another the less likely they are to be treated as parts of the same sound. Another fact is that partials that are more intense are easier to segregate from the spectrum. This may be because stronger tones better resist masking from nearby frequencies. The contribution of these two effects to scene analysis is not entirely clear, and they may be side-effects of other design principles in the auditory system.

However, there is an effect whose value is easy to appreciate. The scene-analysis system favors the grouping of partials that are harmonics of the same fundamental. This can be called the harmonicity principle. Its utility follows from the fact that when many types of physical bodies vibrate they tend to generate a harmonic spectrum in which the partials are all multiples (approximately) of the same fundamental. Instances include many animal sounds, including the human voice. Therefore if the auditory system can find a certain number of fundamentals that will account for all the partials that are present, then it is very likely that we are hearing that number of environmental sounds.

There are several effects of this grouping. One is that a pitch can be derived separately for each group of partials. This allows us to hear more than one pitch in a single spectrum. This grouping by harmonicity also explains why inharmonic spectra seem to have many pitches. A separate timbre can also be derived for each harmonic series, making it possible for us to segregate speech sounds (with their different timbres) when they have different pitches. The grouping also tends to cause the partials that are within the same group to

be fused perceptually. When this happens, it becomes impossible to hear the pitches of the individual partials. Sometimes an incomplete fusion of a set of partials may occur even though they form more than one harmonic series, provided that many of the partials occur in more than one of the series. This is why a tone will fuse so strongly with one that is an octave higher. All of the partials of the higher tone will be matched by the even-numbered partials of the lower one.

Two factors that affect spectral integration are described by the Gestalt principle of common fate. The Gestalt psychologists discovered that when different parts of the perceptual field were changing in the same way at the same time, they tended to be grouped together and seen to be changing as a group because of their common fate. A visual example can be made by drawing two clusters of dots, each on a separate transparent sheet. When the two are superimposed, we see only one denser set of dots. However, if the two sheets are moved in different patterns, we see two sets of dots, each set defined by its own trajectory of motion.

Common fate in audition can be defined in terms of correlated changes in the frequencies of different partials or in their amplitudes. It will be convenient to take the human voice as an example. Variations in the pitch of a voice are represented acoustically by similar changes in all frequency components (parallel changes on a log-frequency scale). When the pitch rises, not only does the fundamental frequency go up but all the harmonics go up by the same proportion too. It is plausible to believe that this correlated change, if it could be detected auditorily, could tell us that the changing partials all came from the same voice. The auditory system could group all such correlated changes and hear only one changing sound.

There is evidence to suggest that two types of frequency change (or modulation) are used for this purpose. One is micromodulation, the tiny fluctuations of the pitch of human voices that occur even when speakers think that they are holding a steady pitch. A slightly larger version of such modulation occurs in singing where it is called vibrato. The other type of frequency modulation is the slow kind that occurs when we voluntarily vary the pitch of our voice in a smooth way, as we do, for example, when we raise our pitch at the end of a question. This sort of change is called portamento in music. The synchronization of micromodulation or of slow modulation in different parts of the spectrum seems to cause those parts to be treated as parts of a single sound. However, it is not yet clear whether these effects can be fully accounted for by an alternative explanation. This account argues that parallel frequency changes allow partials to maintain their harmonic relations to one another over time and that it is these sus-

tained relations, rather than the change itself, that promotes the integration of the partials.

Another version of common fate in audition occurs when the auditory system detects synchronized amplitude changes in different parts of the spectrum. Just as with frequency modulation, this can occur on finer or grosser scales. The finer-scale modulation is not really a property of the physical signal itself but occurs within the auditory system. It happens when we listen to complex harmonic sounds such as the human voice. Because of the way in which our peripheral auditory systems filter the incoming sound, rapid fluctuations in intensity occur within the different auditory neural channels that respond to bands of frequencies in the voice. The fluctuations are periodic and have the same period in every channel; this happens to be the period of the fundamental of the voice. Experimental evidence supports the idea that this common neural periodicity can promote the integration of sensory evidence derived from different spectral regions.

We can also observe a grosser sort of correlated amplitude change in different parts of the spectrum. It is a property of the signal itself and occurs when sounds begin and end. All the frequency components derived from a single sound tend to start and stop at the same moment; those derived from different sounds tend to do so at different moments. This can be used to partition the set of frequency components derived from a mixture of sounds.

It seems that we ought to be able to pick out the auditory components contributed by a certain environmental event by watching the event and correlating visual changes with auditory ones. While it appears very probable that we can do this, the experimental evidence on this point is very indirect and more research is needed. We know that we look more closely at a talker's mouth in a noisy room, but we do not know what type of process uses this information. There is reason to believe that the recognition of a speech sound uses a schema that coordinates both the sound itself and the specific movements of the speaker's face that occur when that sound is made. It is not clear whether there exists, in addition, a more primitive process that notices visual-auditory correlations and uses them simply to partition the auditory sense material.

One of the most powerful strategies for grouping spectral components is to group those that have come from the same spatial direction and to segregate those groups that have come from different directions. That is why the individual instruments in an ensemble performance sound so much clearer in a stereophonic recording than in one that is monophonic. However, there is a requirement that has to be met before the auditory system can group spectral components by

their spatial origin. It first has to be able to assign an independent estimate of spatial origin to each separable band of frequencies in the spectrum. Results from physiological experiments on animals and perceptual experiments on humans suggest that these independent estimates are indeed derived by the auditory system. For example, spatial cues that suggest two locations of origin can cause a band of noise to split perceptually into two different frequency bands. A location can be specified for each band by delaying the components of that band in one ear relative to the other.

Although it seems logical to treat components that come from the same direction as having come from the same sonic event, this strategy is not infallible. Different events can occur close together in space, or along the same line radiating outward from the body of the listener. Even when the events themselves are at quite distinguishable locations, the sensory evidence for their positions can be distorted by reflection of the sound waves from nearby surfaces, by the presence of bodies interposed between the ear and the sound source, and by asymmetrical masking of the evidence at the two ears by louder sounds. This may be why spectral organization does not depend too strongly on spatial cues. A person can do a creditable job at segregating concurrent sounds from one another even when listening to a monaural recording. Spatial evidence is just added up with all the other sorts of evidence in auditory scene analysis.

So far I have described how spatial information affects perceptual organization. But the influence can flow in the opposite direction as well. Because spatial cues are often unreliable, the auditory system seems to average different spatial estimates to arrive at the perceived location of an event. However, this strategy would be useless if it did not know how many sounds there were. It would be harmful to average the spatial estimates that had been derived from energy coming from different unrelated events. Therefore the system must group the evidence before averaging the estimates of location. As an example of how this works, we can set up a situation in which two different signals appear to come from different directions. If we then use correlated micromodulation in the two signals to tell the auditory system that they are one and the same event, it will derive only one (diffuse) spatial estimate for the whole sound. Another example is the use of correlated visual information to correct the perceived location of a sound (the ventriloquism effect).

The scene-analysis process uses the history of a signal to correct momentary spatial estimates. For example, if a pure-tone signal is split and sent to both ears in phase and with equal intensity, we will hear a tone coming from the center. However, if the tone at one ear is

turned up and down in intensity fairly rapidly and abruptly, we will hear a pulsing tone at that ear and a steady tone at the other. The alternative perceptual solution, hearing a single tone move repeatedly from a central position to the side of the head receiving the louder tone, is not favored by the auditory system. It appears that an across-ear version of the old-plus-new heuristic concludes, instead, that there are two sources of sound. It decides that the balanced signal results accidentally from a continuation of the steady signal into the moments when a pulsing signal reaches its maximum. It partitions the balanced energy into a continuing (old) and an added (new) signal. Instead of one perceived location there are two.

We see, then, that there are more cues to the location of a sound-producing event than are usually mentioned in textbooks. We do not use only those cues that are based on the fact that our ears are on different sides of our heads and have a certain shape. We also use the fact that sound-producing events tend to persist over time, to move only slowly in space, and to give rise to sounds that have a coherent inner structure.

There are undoubtedly other factors that help a listener to partition the auditory evidence that arrives at a particular moment. It is possible that textural features (for example, ones that describe the temporal irregularity in the sound) may play a role in partitioning irregular spectra such as those we get when we tear a piece of paper, drag objects, or walk on crunchy snow. No research has been done on the partitioning of spectra that derive from combinations of such sounds. Yet it is worth noting that in the natural world these sounds are more numerous than the regular tones and noises that we have studied in the laboratory.

We also are ignorant of the role that rhythm plays in partitioning mixtures. Is it easier to decompose a mixture in which a component is changing in a regular repetitive fashion than one in which the changes are irregular? If so, what is the mechanism by which it is done, a primitive partitioning process or one that depends on our ability to guide our attention by rhythmic schemas?

When the scene-analysis process assigns different sensory components to the same analysis, we say that they are fused and no longer separately audible. There is another research context in which acoustic components become less audible in the presence of other ones. This is the research on masking. What is the relation between masking and fusion?

Masking and fusion differ in the tasks used to measure them. A sound is deemed to be masked if there is no way to tell whether it is present or absent in a mixture of sounds. It need not be audible as a

separate sound. On the other hand, it is considered to be fused if it is not audible as a separate sound even if you can tell whether it is on or off by some change that it induces in the perceived quality of the spectrum.

One similarity between masking and fusion is that, in both cases, some components of a complex auditory mixture lose their ability to be heard individually. Another similarity is that many of the same factors that influence the audibility of components in experiments on auditory scene analysis also affect masking. Variables that help to segregate one acoustic component from others also prevent that component from being masked by the others. For example, if a target sound is micromodulated, it is harder to mask it by a sound that is not. It is also harder to mask it by a sound that does not start in synchrony with it or by one that comes from a different direction in space.

The masking effects of a band of frequencies (let us call it band A) can also be reduced by causing band A to group with a different band of frequencies (band B) that is too far away from the target to mask it. The grouping is induced by synchronizing the amplitude fluctuations in the two bands, a manipulation that affects spectral integration. This effect is called "comodulation release from masking."

The similarities in the factors that influence masking and fusion suggest that physiological mechanisms that have been evolved to serve the purpose of scene analysis are involved in the phenomenon of masking.

Any perceived property of an incoming array of sound is computed from a subset that the brain has selected from that array. The purpose of computing a property is to describe a meaningful quality, that is, a quality of a sound that has been created by some distinct happening in the world such as a speaker's voice. If the subset of sense data selected for the computation is too small, the computation may yield separate qualities for the individual acoustic components generated by the event (say individual harmonics in the voice) rather than some quality for the whole sound created by the event. I refer to the former properties as "partial" properties. If, on the other hand, the array that is used for the computation is too large, the perceived quality will represent an array that is really an accidental mixture of the acoustic energy from more than one event—for example, the sum of two vowels from different voices. I call this kind of quality chimeric. Arranging it so that the computed qualities are meaningful, rather than partial or chimeric, is the job of scene analysis. There is some evidence that there is a bias toward integration in audition. Properties will be computed on the whole incoming array of sound unless there is specific evidence that the array should be subdivided into subsets.

Results of experiments and of informal observations show that separate pitches, timbres (e.g., perceived roughness), vowel identities, and even locations can be computed separately on the streams created by scene analysis. If this were not possible we would be able to hear only one sound at a time and the qualities of that sound would always represent the sum of all ongoing events.

The Continuity Illusion and Contralateral Induction Principles of auditory scene analysis, particularly the subdivision of the array of sense data received at a particular moment, can explain two phenomena encountered in auditory research. One is the continuity illusion and the other is contralateral induction.

The continuity illusion was described in chapters 1 and 3. If a short segment of an ongoing sound is deleted and replaced by a much louder sound that has the right spectral content, the softer sound is perceived as continuing right through the louder one. The scene-analysis explanation is that the auditory system has taken part of the sensory evidence contributed by the louder sound and assigned it to the percept that represents the softer sound. This has been interpreted as a compensation for masking, since the louder sound would have masked the softer one even if it had been there.

The examples range from the simplest, in which the soft sound is a steady pure tone and the louder one is a louder version of that same tone, to the most complex case, in which the softer sound is speech and the louder one is a noise burst. In the latter example, the speech is heard as being complete. This is called phonemic restoration.

The explanation for the continuity illusion can be broken into two parts. One explains how the brain of the listener decides whether or not the soft sound is continuing behind the louder (the "whether" question). The other explains how the brain chooses what the content of the missing portion is (the "what" question). Only the "whether" question is decided by general scene-analysis principles.

The process of deciding whether a sound A has continued through an interruption by another sound B seems to be governed by some fundamental rules. They can be stated in terms of the labeling of the parts of the stimulus that is given in figure 3.22 of chapter 3. A1 and A2 are the parts of the softer tone, A, that precede and follow the interruption. B is the louder tone that replaces the deleted segment of A.

The continuity illusion involves the partitioning of the sensory stimulation received at time B (the period when B is on) into two bundles. One is interpreted as a continuation of sound A. The other is interpreted as a second sound that comes on suddenly, and its acoustic

properties are perceived as consisting of ones that, when added to those of A, would yield the total stimulation received at time B. The scene-analysis system has concluded that A1 and A2 are not separate sounds but parts of the same one. This interpretation aids in the recognition of A, since the recognition process will try to identify a single long sound with properties of both A1 and A2 and not two sounds with separate sets of properties. When A1 and A2 are really parts of one sound, this will be an effective strategy.

The rules that govern the decision as to whether to restore A by partitioning the evidence supplied by B are as follows:

The "no discontinuity in A" rule. There should be no evidence that B is actually covering up a silence between A1 and A2 rather than covering the continuation of A. This means that there should be no evidence that A actually shuts off when B starts or turns on again when B finishes. There must be no audible silent gaps between A1, B, and A2, and there should be no changes in the amplitude of A1 or A2 near the boundary of B.

The "sufficiency of evidence" rule. During B, some subset of the neural activity in the auditory system should be indistinguishable from activity that would have occurred if A had actually continued. This implies that B should be sufficiently louder than A to provide adequate stimulation in the neural frequency channels normally stimulated by A. If B is a different frequency from A, then B will have to be of an even greater loudness so that the neural spread of excitation can supply a sufficient amount of activity in the frequency channels that normally signal A's presence. When a familiar signal (such as a spoken word) is being recognized by the schema-based recognition system, the latter will supply a hypothesis for what the missing part of A is. In this case, the stimulation will be required to contain the neural evidence that is normally activated by the missing part of A.

The "A1-A2 grouping" rule. There should be evidence that A1 and A2 have actually come from the same source. This means that the heuristics for sequential grouping would have put them into the same stream even if they had been separated by a silence instead of by B. If A1 and A2 do not fit well into the same stream, the restoration will not be favored. The rule allows the missing parts of two or more soft sounds, interrupted by the same louder sound, to be restored without confusion. The parts of each sound will be grouped into their own stream and a restored part will be computed for each stream.

The "A is not B" rule. The transition from A to B and back again should not be interpretable as sound A transforming into a new form, B, and then back again. If it is interpretable in this way, the listener

should not hear two sounds, one continuing behind the other, but simply one sound, changing from one form into another and then back again. The criterion for hearing interruption rather than transformation is probably whether or not the rate of change exceeds a critical value, thereby giving evidence for a discontinuity in the signal.

Contralateral induction occurs when a soft sound presented to one ear is accompanied by a loud "inducing" sound in the other ear. As a result, the perceived location of the softer sound is pulled toward the center of the body. The interesting part of this is that the inducing sound is not required to be the same as the softer sound. If it were, contralateral induction would just be an example of the well-known fact that the binaural intensity balance for a particular sound determines its perceived location.

An example of contralateral induction can be created by sending a pure tone to one ear and a noise burst to the other. It is found that in order to serve as an inducer the noise must stimulate neural frequency channels that correspond to those stimulated by the pure tone in the other ear. This can happen either because the inducer contains the needed frequency components or because it is loud enough to stimulate the corresponding frequency channels through spread of neural excitation. Contralateral induction is a case in which the scene-analysis system pulls out of the noise the frequency components that match those of the tone and interprets the matching components on the left and the right as evidence for a centrally located tone.

The effect resembles the continuity illusion in that a complex sound is decomposed into two parts: one that matches another sound and a residual. The matching part is treated as belonging to the sound that it matches. In both cases, the sound that has to be decomposed must be much louder than the other one so that enough stimulation is received in the corresponding neural channel to supply a reasonable match to the softer tone.

Contralateral induction is actually a particular way of hearing a complex binaural event. The louder sound is interpreted as masking one ear's share of the binaurally balanced energy that has arisen from a centrally located softer tone. We should be aware that this interpretation is not guaranteed to be correct. However, it is correct often enough that we notice it, and realize that it is an "interpretation," only when unusual conditions that we set up in the laboratory cause it to be incorrect.

The percept of contralateral induction is not the only way of interpreting binaurally matched stimulation. We can block it if we arrange

conditions appropriately. Here is a simple demonstration. Suppose we present a soft tone continuously to the left ear and a white noise continuously to the right. If we make the noise loud enough it pulls the localization of the tone to the center. This is the normal case of contralateral induction. Now we can eliminate the induction by simply pulsing the noise while holding the tone steady. If we do this, the scene-analysis system detects the fact that there are large amplitude changes in one ear that are not matched by any changes in the other, and this causes it to favor the interpretation of two streams of sound: a steady one on one side of the body and a pulsing one at the other. It refrains from centering the tone during the moments at which the noise burst is on, despite the fact that some of the right-ear stimulation could be used as a match for the stimulation on the left.

The examples of continuity that we have discussed show how the auditory system can resist momentary cases of masking. Masking occurs when there is no way to tell whether a part of the currently received sensory stimulation should be treated as a separate sound. (By analogy, we do not see a totally red piece of paper as containing a red spot at its center, since there is no border between the "spot" and the rest of the red.) Audition uses a number of methods to detect the fact that an embedded part of a complex mixture of stimulation should be extracted and interpreted as a separate sound. One type of clue is finding the sound in isolation before and after the complex mixture. Another is detecting a clearer, more isolated version of the sound in the other ear. A third is detecting a difference between the fundamental frequency of the embedded spectrum and the rest of the spectrum. In short, it can use any of the heuristics of the scene-analysis system.

Schema-Based Stream Segregation

Nature of Primitive and Schema-Based Organization

So far we have been focusing on primitive processes in scene analysis. These are assumed to establish basic groupings among parts of the sensory evidence so that the number and the qualities of the sounds that are ultimately perceived will be based on these groupings. The groupings are assumed to be based on rules that take advantage of fairly constant properties of the acoustic world, such as the fact that most sounds tend to be continuous, to change location slowly, and to have components that start and end together. However, the story of auditory organization would not be complete if it ended here. The experiences of the listener are also structured by more refined knowl-

edge of particular classes of signals, such as speech, music, machine noises, and other familiar sounds of our environment. Psychologists (and many computer scientists) argue that this knowledge is captured in units of mental control called schemas. Each schema incorporates information about one particular regularity in our environment. Regularity can occur at different levels of size and spans of time. So in our knowledge of language, for example, we would have a schema for the sound "a", one for the word "apple", one for the grammatical structure of a passive sentence, one for the pattern of give and take in a conversation, and so on.

We think that schemas become active when they detect, in the incoming sense data, the particular pattern that they deal with. Because many of the patterns that schemas look for (such as the structure of a sentence) extend over time, when part of the evidence is present and the schema is activated, it can prepare the perceptual process for the remainder of the pattern. I will refer to this preparation as schema-governed attention. A schema can be activated not only by the specific sense data that it has been constructed to recognize but by other schemas with which it is associated. So if you read the word "apple" you will be more prepared to read the word "fruit" as well.

Any natural situation in the world exhibits many regularities and therefore will activate many schemas. But they do not just fire off in parallel. Their job is to compose themselves together to create a consistent description of the world at that moment.

A schema claims certain portions of the sensory evidence and groups them to create the particular description that it is responsible for. In so doing, it acts like a scene-analysis process. The goal of this part of my discussion is to consider the relations between the primitive preattentive clustering of sensory input that I have described up to now and the schema-governed, attention-based construction of perceived objects. A particularly important question is how we might distinguish the contributions of these two hypothetical mechanisms to the grouping effects that have been studied in the laboratory.

There are cases in the research literature in which primitive grouping processes seem not to be responsible for the perceptual groupings. For example, when two simultaneous vowels are synthesized on the same fundamental frequency, start and stop at the same time, and emanate from the same spatial position, we know of no rules of primitive grouping that could partition the spectrum into two parts, one for each vowel. Nonetheless, we can often distinguish two vowels in such a mixture. We suspect that the schema for each vowel is picking out what it needs from the total spectrum rather than requiring that a partitioning be done by the primitive process.

Another example of the schema-based selection of sense data occurs in the phonemic restoration of a speech sound that has been masked by a sudden loud noise (we hear the speech continue through the noise). Apparently we select certain frequency components out of the noise and hear them as if they were the missing speech sounds. This selection must be accomplished by a process that expects particular sounds to be there. Presumably, that process is a schema that represents the sound pattern of a particular word.

The previous two examples show that schemas can select evidence out of a mixture that has not been subdivided by primitive scene analysis. There are also examples that show another capacity: the ability to regroup evidence that has already been segregated by the primitive process. For example, if we synthesize a two-formant speech sound in which each formant is constructed from harmonics related to a different fundamental frequency, listeners will have an unusual experience. They will hear two sounds, one corresponding to each related group of harmonics. Yet at the same time they will hear a single speech sound, the one conveyed by the full set of harmonics. It seems that the schemas that recognize speech sounds can, at least in some cases, put together evidence that has been partitioned by the primitive process.

Let me summarize the differences between the concepts of primitive and schema-based scene analysis. Primitive segregation employs neither past learning nor voluntary attention. It is present in infants and, therefore, probably innate. It partitions the sensory evidence by being sensitive to relations that indicate that parts of the input have come from different sound-generating events. These relations tend to be valid clues over wide classes of acoustic events. By way of contrast, the schemas that are involved in schema-based organization have been developed for particular classes of sounds. They supplement the general knowledge that is packaged in the innate heuristics by using specific learned knowledge.

Our voluntary attention employs schemas. For example, when we are listening carefully for our name being called out among many others in a list, we are employing the schema for our name. Anything that is being "listened for" is part of a schema. Therefore whenever attention is accomplishing a task, schemas are participating. This means that if we can find cases of attention, we can find instances of schema-based segregation. But how do we know when attention is being used? One mark of attention is that it involves a subjective experience of effort. Another is that the number of things that can be attended to at the same time is quite limited. When we see the factors

of effort and limited capacity operating in an auditory scene analysis task, we can infer the presence of attention.

This does not mean that the tasks that are used in experiments that seem to involve attention (and schemas) cannot be employed to demonstrate the influence of primitive grouping. The subjects in an experiment always have a conception of what they are listening for, and this "conception" is what we mean by a schema. However, the variables in the experiment may produce their results through their effects on primitive grouping. If the primitive process creates organizations that do not correspond to the grouping of evidence that the schemas need, the task of the schemas is made harder. The challenge to the theorist is to decide whether it is only one or both of these classes of processes that have been affected by the variables. This decision is made difficult by the fact that the two systems are likely to use the same sorts of sensory information. For example, if a tune that we are listening for is separated in frequency from distractor tones, this fact may affect both forms of organization. Primitive grouping would use the separation to put the tune and the distractors into separate streams. The schema-based process that was looking for the tune's notes might be more easily able to distinguish them from others that were far away in frequency than from nearby tones. Therefore we must look for two different patterns of causality that seem to correspond to the distinction between the primitive and the schema-governed process.

For example, we believe that regularity and familiarity of signals are dealt with by schemas. If we can find some sort of auditory organization that is independent of these factors, we have evidence for a primitive process that does not employ schemas.

The critical reader may ask why we should bother, in the first place, to distinguish schema-based segregation from primitive scene analysis. The answer is that there seems to be some empirical evidence that there are two classes of processes that show different patterns of causality. Let me mention some examples.

First, preliminary evidence shows that infants employ some of the grouping principles that I have called "primitive." Presumably the patterns of tones that were used to test them could not yet have been built into schemas.

Second, when listeners are presented with a sequence of alternating high and low tones, the effects of frequency separation and rate depend on what the listeners are trying to hear. If they are trying to hold the sequence together as a single stream, there is a strong segregating effect of the frequency difference between the high and low tones, and the effect is magnified by speed. However, if they are trying to listen

only to a substream (say the lower tones), there is almost no effect of frequency separation, and the small effect that exists is independent of speed. We can interpret the difference in the causal pattern as follows. When the attention-based process is opposing primitive streaming by trying to include tones that are being segregated by the primitive process, this becomes harder as the primitive segregation becomes stronger. However, when it is trying to select a subset of tones, it requires only a minimum frequency separation to do this and can succeed whether the targets are in their own stream or in a stream with other tones. In other words, it seems that attention-based segregation can easily subdivide a stream but has difficulty integrating material across streams.

It appears also that the two processes have different effects on perception. Primitive processes *partition* the sensory evidence whereas schema-based attentional ones *select* from the evidence without partitioning it. This means that the effects of primitive segregation are symmetrical. When it segregates high tones from low ones, we can listen more easily to either the high ones alone or the low ones alone. Similarly, when it separates two sets of sounds by spatial location, we can more easily listen to the ones on the right or the ones on the left. However, the effects of schema-based selection do not show this symmetry. When my own name is embedded in a mixture of sounds the fact that it is my name makes it easier for me to hear it in the mixture, but it does not make it easier for me to tell what the remainder of the mixture consists of. Schema-based selection often seems to use the evidence that it needs without removing it from the mixture. For these reasons, it may be possible to use the asymmetry of the partitioning to detect the presence of schema-based segregation.

The effects of speed may also distinguish the two types of processes. It appears that segregation based on variables such as differences in frequency, timbre, or location actually gets stronger as the sequence speeds up. But the ability to select by loudness or by familiarity gets worse. This suggests to me that the first group of variables is used by the primitive process but the latter group is not.

There is also some indication that there is a difference in the time span across which primitive grouping and schema-based integration operate. It appears that the schema-based process can look at relations over a longer time span than the primitive one can.

Tests of the Existence of a Primitive Process

If we are to distinguish the role of a primitive organizational mechanism in the total process of interpreting sound, we have to find a

mechanism that operates according to the description that I have just given. It must work independently of learning. It must not care about the regularity of the sequence and it must operate independently of attention. Let us look at these three issues, in turn, to try to determine whether the contribution of a primitive process can be detected.

It has been shown experimentally that learning can affect our ability to recognize a familiar tune whose notes have been interleaved with the notes of another one. If you tell the listeners the name of the tune to listen for, this will help them to hear it. However, if you tell them this name and then ask them to recognize the other tune, with which it is mixed, this information does not help them. This asymmetry shows that familiarity affects the schema-based process, rather than the primitive one.

Does Sequential Grouping Take Advantage of Trajectories? The question of whether a regular sequence of sounds (that is, where the sequence is predictable) is easier to recognize than an irregular one has also received some study. Auditory attention theorists have proposed that as we listen repeatedly to an auditory pattern we learn the regularities in it. This knowledge, because it allows us to anticipate the parts of the pattern before they occur, makes it possible for us to ready our attention and to integrate the sequence more easily into a coherent mental representation. The streaming phenomenon (that is, the exclusion of some of the tones in a sequence from an auditory stream) is interpreted, in this framework, as the failure of attention to follow the sequence. Therefore the regularity of the pattern of tones in a sequence, since it makes the readying of attention easier, ought to have very strong effects on the formation of streams.

This theory has received some confirmation from experiments on the listener's memory for regular and irregular sequences of tones. However, since the task of remembering is strongly affected by the schema-based process, it is not suitable for demonstrating the properties of primitive grouping.

Other experiments in which the listeners simply judged the number of streams without having to recognize or remember the patterns in them show quite different results. The regularity of the sequence does not affect the formation of streams. This suggests that streams are formed by a primitive process that is not affected by the predictability of the sequence.

One simple sort of predictable sequence is a regularly rising or falling sequence of pitches. We can call these simple trajectories. Can primitive scene analysis follow and segregate such trajectories from their acoustic contexts? Evidence both for and against such an ability

exists. Let us start with the evidence against it. When a sequence is made up by interleaving a descending trajectory of tones with an ascending one, as in figure 4.2 of chapter 4, it is hard for a listener to follow one of these right through the crossover point. If we try to listen to the descending one, we find that we can follow its first half down to the crossover point. Then we find ourselves listening to the latter half of the rising sequence instead. Our intentions seem to be defeated by some process that opposes them. I think this process is a primitive one that prefers to group tones that fall in the same frequency range rather than to link tones that fall along a simple trajectory. You can make it possible for the ear to follow one of the trajectories by giving its tones a different timbre from the remaining ones, but this manipulation is simply introducing timbre as a new basis for segregation. It is not strengthening any primitive trajectory-based segregation.

There are also some data that derive from the continuity illusion. They are found when we present listeners with a connected sequence of alternately rising and falling pure-tone glides as shown in figure 1.15 of chapter 1. Noise bursts are inserted in the place of the peaks of the pattern (where the frequency reaches a maximum and then begins to fall). When the auditory system restores the missing peak, we do not hear it as having the frequency of the missing one but as being at the highest frequency that is actually present (the highest one that has been spared by the deletion). This occurs despite the fact that the missing peak's frequency is predictable from the slopes of the non-deleted parts. Apparently there exists a primitive restoration mechanism that does not use the trajectory to predict the missing parts.

Another failure of trajectories to control scene analysis is found when you try to capture a tone out of a mixture or out of a sequence of tones by preceding this mixture or sequence by a series of captor tones. The capturing is often successful when the captors are at the same frequency as the target tone but typically is unsuccessful when the captors merely fall on a common trajectory with the target tone.

Now for the positive evidence for the use of trajectories in scene analysis. Some examples come from research on the illusion of continuity. When segments of pure-tone glides precede and follow a loud noise burst, the listener more easily hears this as a single glide continuing behind the noise when the segments before and after the noise line up on a common trajectory. This stimulus is different from the one in the previous example. In that one, the glides that preceded and followed the peak did not line up on a common trajectory; rather, they pointed to the same frequency peak. Apparently this pointing is not used by the system.

Other evidence that seems to suggest a role for trajectories in scene analysis is that when listeners are asked to report the order of tones in a sequence, the task is easier when the tones follow a rising or falling trajectory. I would like to interpret these results, however, not in terms of perceptual grouping but in terms of the ease of encoding a regular trajectory into memory. In general, the effects of regularity tend to be seen most clearly in tasks that involve memory.

However, even if sequences of tones that form a trajectory are found to form more integrated streams, and we find the integration to be perceptual in origin, there is still an explanation that does not require the auditory system to make specific use of the regularity in the trajectory. Think of what happens when the tones are *not* on a trajectory. The reason that stream segregation breaks up this sort of sequence is that tones often prefer to group with tones that are nearest to them in frequency rather than with those that are nearest in time. Yet notice that in a trajectory no such conflict can arise; the nearest neighbors are nearest on both dimensions. It may be this continuity, rather than the fact that the sequence follows a rule, that is responsible for the perceptual integrity of a trajectory.

I have noticed that trajectory-based organization increases with more exposures to the stimulus. This points to an involvement of learned schemas in the integration of trajectories. My personal conclusion is that the primitive Gestalt-like grouping process does not make use of the fact that a sequence may be rule-governed or may point to some future position of a sound in frequency or in time. When it appears that it does, either it is just making use of the frequency proximity of tones that are present in trajectories or we are looking at a process of schema-governed integration.

We may be misled by visual analogies. It may well be a primitive process that is responsible for the enhanced grouping of the parts of a line that fall on two sides of an occluding surface when the parts fall on a common trajectory. Primitive organization may also be responsible for the fact that when a moving form disappears behind a barrier, the perceived continuity of motion depends on whether the entering and exiting paths fall on a common trajectory. We must not be seduced by these cases. The regularities, in the world of objects, that justify these kinds of grouping in vision are not the same ones that affect trajectories of sound. For example, visible objects have physical inertia that keeps them moving on straight lines. Sounds have no inertia to keep their pitch changes moving along a simple trajectory. Therefore the auditory system may not have evolved analogous rules for the primitive integration of trajectories.

A common form of regularity in sound is rhythm. This observation prompts us to ask whether this form of regularity affects primitive grouping. Theorists who argue that streams are formed by the attentional process have proposed that since predictability is what allows sequences to be integrated, then rhythmic regularity is responsible for the formation of streams. However, experiments on temporally irregular sequences that contain high- and low-frequency tones have found that streams are formed in the same way as in rhythmic sequences. The task in these experiments has merely been to decide on the number of streams, not to recognize patterns embedded in them. The recognition of patterns may have a heavy schema component and may therefore be affected by temporal regularity or by any other form of regularity.

It is important to evaluate the view that primitive organization operates independently of attention. Not every researcher believes it. Testing the idea is difficult because attention operates in most tasks. One approach to circumventing this problem is to manipulate the primitive grouping of sounds that we are trying to *exclude* from attention. If we could obtain evidence that they were actually organized, then we might conclude that organization does not require attention. Some results suggest that such organization exists, but they are not conclusive. Perhaps the most suggestive evidence comes from experiments in which the attention of the listener is controlled in different ways in the same experiment. It is found that the primitive factors of frequency separation have different effects in the two cases. Assuming that the primitive process has divided a sequence of sounds into two streams, it is easier for attention to further subdivide one of these streams than to reintegrate the divided streams. It is the conflict between attention and the primitive grouping process that reveals the existence of the latter.

Another approach to testing for the existence of a primitive preattentive process it to look for the formation of coherent residuals. I have assumed that when a schema extracts what it needs from a mixture of sense evidence, it does not create partitions that serve to bracket the excluded evidence as well as the included evidence. In other words, it selects rather than partitions. One could test this assertion by designing experiments that presented particular auditory patterns to listeners repeatedly so that they formed schemas for these patterns. Then extra sounds could be added to these patterns as distractors and the subjects asked to recognize the familiar patterns despite the distractors. Then one could test to see whether the sounds that were not used by the schema would be, as a consequence of exclusion, more

likely to group with one another than if they had not been excluded by a schema.

Primitive Auditory Organization in Music

Role of Primitive Organization in Music

Both primitive scene analysis and complicated schemas play a role in our perception of music. Since the primitive process is the subject of this volume, let us consider its role.

Traditionally music is described as having a horizontal and a vertical dimension. These relate to the horizontal and vertical dimensions of a sheet of music. The horizontal one is concerned with the sequence of sounds that defines melody and the vertical one is concerned with the relations between simultaneous sounds that define harmony. Both these dimensions require primitive organization. The horizontal one is affected by sequential organization and the vertical one depends on simultaneous organization.

Perceptual organization plays a different role in music than in natural environments. In everyday life, its purpose is to segregate the streams of sound that are created by specific sound sources and to treat mixtures as accidental. If this always succeeded in music, instruments would never be able to blend to create new timbres or to define melodies that were carried in a succession of notes from different instruments. Music must defeat the stream-segregation tendencies (or at least work with them) to achieve its goals. Musical perceivers must perceive those organizations that are part of the architecture of the music itself rather than perceiving the individual pieces of hardware that are employed to produce the sound. They must hear fictional sources of sounds that have qualities that emerge from the set of sounds being grouped.

Fusion and segregation must therefore be carefully controlled in music. Their control has traditionally been carried out by rules of thumb, evolved over the course of musical history, rather than by an explicit understanding of principles of perceptual organization. However, a clear scientific understanding could contribute to a more exact theory of orchestration, particularly in relation to the newer musical forms for which pragmatic rules have not yet been evolved. For example, when composers generate musical sounds with a computer, they do not have to use distinct acoustic sound sources analogous to the instruments of the orchestra. The choice of whether to structure the sound as a collection of "instruments" is a decision that has to be made explicitly.

Sequential and simultaneous organization help to create many aspects of musical experience. Sequential grouping is the foundation of rhythm and of many aspects of melodic form, whereas simultaneous grouping is involved with such experiences as chord perception, timbre, consonance, and dissonance.

One form of segregation occurs between the musical piece that we are hearing and any accompanying nonmusical events, such as coughs or static. There is a second type of segregation between the different parts of the piece of music itself. This sort of organization has to be strong in some ways and weak in others. The segregation has to be strong enough for the listener to perceive each line of melody with its own distinct timbre. At the same time it must not be so strong as to keep us from perceiving the musical relations between the parts. The best solution would be to organize the music into a hierarchical form (parts within larger parts). Think of the way in which the fingers on a hand are perceived: distinct and yet united to create a larger form. We believe that perception is structured in this way, but we do not know whether these hierarchical structures are formed by primitive or by schema-based processes. I rather suspect that it is the latter. Since this book is primarily about the primitive process that is employed in all auditory perception, we will not be saying much more about either this hierarchical organization or the many forms of organization (scales, chord relations, and so on) that are specific to music.

Melody

Melody is the horizontal dimension of music. It is governed by principles of sequential organization. We saw earlier that when a sequence of tones contains large frequency transitions that occur in short periods of time, the sequence does not remain perceptually coherent. Music has adjusted itself to this fact. Small transitions in the fundamental frequencies (or pitches) of notes are much more common than large ones in traditional Western music, and when the speed has to be made very fast (as in trills, grace notes, and other ornamentation) the pitch jumps are made very small.

Rapid alternations of high and low tones are sometimes found in music, but composers are aware that such alternations segregate the low notes from the high. Transitions between high and low registers were used by the composers of the Baroque period to create compound melodic lines—the impression that a single instrument, such as a violin or flute, was playing more than one line of melody at the same time. These alternations were not fast enough to cause compulsory segregation of the pitch ranges, so the experience was ambi-

guous between one and two streams. Perhaps this was why it was interesting.

The segregation of high and low notes does not seem to be sensitive to musically significant relations such as the octave. A rapid succession of alternating high and low notes spaced an octave apart will segregate just as strongly from one another as notes that are not separated by this exact interval. The octave relation does not hold them together. This suggests that the formation of musical streams is strongly affected by a very primitive form of grouping that knows nothing about octaves.

Experiments have been done on the listener's ability to recognize melodies whose notes are interleaved with distractor tones. If the distractors fall into the same pitch range as the melody, the recognition is very hard. The further the distractors are in pitch from the melody, the easier the recognition becomes. In general, many of the findings with musical material have paralleled the results found in the laboratory with cycles of tones.

The transformations in pitch that occur over time define melodic form. For example, a rising sequence of tones is perceived as a rising "gesture." Unless the tones that define the beginning and end of the upward transition are perceived as part of the same stream, no sense of upward transition will be experienced. So transformations must be packaged within streams. Such a requirement makes sense in nature. Transformations signal important changes in the behavior of a sound source. The rise in intensity of a set of footsteps may tell us that the walker is drawing nearer. On the other hand, if each of a series of sounds derives from an event unrelated to the others, the intensity relation between them may be meaningless. Only if a single source (represented in our mind as a single stream) is creating the acoustic changes are they likely to really define a meaningful transformation in the world. We have presumably developed a perceptual system that looks for meaningful transitions within streams. Musical transformations have to be content to live within such a system.

My comments have so far been restricted to changes in pitch. But changes in timbre can also affect the integrity of a musical sequence. Sudden repeated changes in timbre can fragment our perception of a sequence of notes. This fact has been put to musical use in the technique of "klangfarbenmelodie," a rapid sequence of shifting timbres. More commonly, shifts in timbre are used to delineate musical units of larger sizes and to have their effects on phrasing. They are neither sudden enough nor frequent enough to affect the primitive coherence of the musical line.

Some notes in a line of music seem to define the overall form of the phrase whereas others, usually shorter (for example, grace notes), seem to serve as ornamentations that are appended to the form-bearing notes and seem to be subordinate to them. They seem to group with the note to which they are subordinate (let us call it the "anchor note") so as to create a larger event that has more complexity than a single note. The Gestalt psychologists called this relation "phenomenal dependency." Often the dependency is between successive notes played by a single instrument, but in ensembles it can be created between notes arising from different instruments. In this case the resulting complex auditory event can have qualities that a note from neither of the component instruments would have.

The tight perceptual binding between the sounds that generate these emergent events is accomplished by the Gestalt-like principles of primitive scene analysis. For example, the dependent note must be very close to the anchor note in frequency and time. Otherwise the two will not group to form a larger event. Ideally, both the time intervals and the pitch intervals between the sounds that are to form the event should be less than those between the notes that precede and follow them. The choice of which note is the dependent one and which the anchor depends on factors such as duration, intensity, and rhythm, the anchor typically being longer, louder, and on a major beat. The dependency can also be affected by the pattern of adjacent tones in ways that are predictable from Gestalt principles.

When component sounds are grouped sequentially to form a larger sonic event their own individual properties are to some extent lost. The loss is not as severe as with simultaneous sounds that are fused, but it is still noticeable. The perceived pitches of dependent tones tend to be altered or to become less distinct. This tends to happen less to the beginning and end tones than to the interior tones of rapid phrases. It also has less effect on tones that are the highest or lowest in their phrase.

I have emphasized the primitive processes that affect dependency, but schema-based ones that pertain to the musical style can also be involved. Dependent notes tend to be ones that are defined as unstable in the musical system, and they tend to resolve to stable anchor tones.

Timbre

Timbre can also play a role in the sequential organization of music. Timbre changes are rarely large enough and rapid enough to cause the line of melody to split into parallel streams, as in a "compound melodic line." Yet they could be used in that way. If the variations in

pitch were small and those of timbre were large, we know from laboratory studies that timbre could control stream segregation. Often, as I have pointed out in the previous section, timbre is used as a sequential glue for musical phrases. Playing a phrase in a timbre different from the ones preceding and following it can help it to be perceived as a unit in the music. A textbook example of this technique can be seen in Webern's orchestration of the Ricercar from Bach's *Musical Offering*, in which the phrases are carved out almost surgically by timbre. On the other hand, if timbre is made to change repeatedly and rapidly within a phrase, the sequence becomes fragmented. This technique is used in hockets in which notes from different instruments follow in rapid succession. It also occurs in Webern's use of klangfarbenmelodie.

In using timbre to outline segments of the music, the composer is exploiting the scene-analysis principles that audition has evolved for arriving at accurate descriptions of natural sound-producing events. When there is a sudden change in timbre, it is usually valid to infer that some new event has begun. When the timbres change continuously, it is more likely that a single event is changing in some way. We perceive the change as a transformation that defines a "happening." If we knew what the dimensions of timbre were, we could use movements along them to define transformations that could serve as structural elements in a piece of music. This would be a stronger use of timbre than simply using it to color in the forms that have already been defined by pitch changes.

But what are the dimensions of timbre? There have been a number of different approaches to this question. One has been to ask people to rate the similarity between different pairs of sounds and then to try to infer how many dimensions could account for these judgments. Three dimensions have been found: the brightness of the spectrum, the bite of the attack, and the simplicity of the behavior of the harmonics over time. These are very general qualities of sound.

Another approach has been to look for general dimensions of timbre in the acoustic dimensions that define the differences between spoken vowels. The dimensions that have been chosen correspond to the frequencies of the lowest two prominent peaks (formants) in the spectrum of the sound. Music has been created in which timbral shapes were formed by the changes of the sounds along these dimensions.

A third approach is to try to synthesize imitations of natural sounds in which spectral details are eliminated. If these details do not matter, an imitation will be judged as an acceptable example of the timbre of the original sound. This method of "analysis by synthesis" is ex-

pected to identify the important dimensions of timbre by a process of elimination.

Other researchers have argued that our brains are not built to hear sounds in the abstract, but to form descriptions of environmental events. Therefore the important dimensions of timbre will correspond to possible changes in the actual physical objects that produce the sound. For example, a simple change in the force with which a person strikes a gong could have many different effects on the way in which the amplitudes of different spectral components evolve over time. Yet a change in the striking force might be a simple perceptual dimension for a listener. If this were true, it would imply that the brain has implicit models of the physical systems that can produce sound and can use them to "read off" the changes in the physical system from changes in the sound.

An understanding of the primitive organization of timbre by the brain would open the way for its use as a carrier of form in music.

So far the discussion has treated timbre as some property of the evolving spectrum of a sound. Yet auditory information does not come to us in the form of separate sounds. We receive an array of auditory properties simultaneously from an unknown number of sources. Therefore scene analysis is responsible for timbre. The identity of the component sounds is known as a result of the partitioning that is done by scene analysis and the partitioning allows us to hear different timbres at the same time. This means that the use of timbre in music depends on a practical understanding of the principles of auditory fusion and segregation. When composers understand these principles, they can use instruments as generators of auditory features and use the rules of spectral grouping to form new "orchestral timbres" as clusters of these properties. The instrument becomes analogous to a color on the palette of a painter, and the perceived instruments that result from this blending are entirely fictitious.

The control of spectral integration plays an important role in music. Musicians may present musical sounds together for many reasons. At one extreme they may want the sounds to fuse and generate a global timbre. At the other, they may wish to create a polyphonic texture in which two or more distinct lines of sound can be heard. It seems likely that all of the factors that have been found in the laboratory to affect spectral integration have already been used in music for this purpose. If the scientists succeed in making these factors explicit, it will probably not change musical practice very much. However, it may provide a rational basis for a theory of orchestration. If based on basic principles of auditory organization, such a theory could be independent of particular musical styles.

Let us look at some of the ways that music uses scene-analysis principles. As an example we could examine how soloists can be kept distinct from their accompaniments. Take the use of harmonic relations, for example. Since the auditory system looks for different sets of harmonic series and segregates them, one method available to a soloist to maintain perceptual distinctness is to be producing pitches that are not the same as those produced at the same time by the accompaniment. Another is to produce nominally identical pitches that are slightly mistuned relative to the accompaniment. This occurs automatically when the soloist's notes carry some vibrato. Since the soloist's vibrato is not likely to be phase-locked with the vibrato of the accompaniment, the note will frequently be at a slightly different pitch from any in the background. The wide vibrato of an opera singer is an example.

Another technique is to play or sing a pitch that is higher or lower than any in the accompaniment. The actual frequency band occupied by the energy produced by the soloist can also be somewhat different from that of the accompaniment. For example, singers learn to produce the *singing formant*, a peak in the singer's spectrum that occurs in a spectral region that tends not to contain much power from accompanying instruments.

Onsets and offsets are very important too. Soloists can employ a "rubato" style to minimize the synchronization of their note onsets with those of the rest of the ensemble. A tight synchronization of the other players will assist in segregating the soloist who is using this technique. In general, if the soloist tries to be different from the ensemble on any acoustic variable, it helps to have the other instruments of ensemble be as close together on that variable as possible.

Even spatial location can be used to segregate the soloist. The major advantage of stereophonic recording over monaural is not that it gives a sense of space (although it does that too) but that it allows the individual instruments and voices to be more distinctly heard. As a consequence, features of the music that would be blurred in a monaural recording can be distinguished. Some attempts have been made, even in live performances, to overcome the loss of definition of musical lines in dense textures (lines that are close together in pitch) by placing the players at quite different locations with respect to the audience. This seems to be quite effective.

It is interesting to look at the role of primitive scene analysis in counterpoint. Polyphonic music is a style in which there are two or more melodic lines running in parallel. Counterpoint is the technique for composing the lines. In polyphonic music, the parts must not be either totally segregated or totally integrated. If they were totally

segregated there would be no overall coherence to the composition; if they were totally integrated there would be only one line of melody. Principles of auditory organization are employed to help achieve the level of vertical organization that is wanted at each point in the piece. Although schema-based principles, particular to the musical style, are also used, I will restrict my discussion to the primitive ones.

The segregation of one line from another can be improved by making sure that each one has a strong sequential organization within itself. This is controlled, in part, by the size of the pitch changes between one note and the next. Small steps favor sequential integration. Another requirement, if the lines are to remain distinct, is that they should be well separated in pitch and should not cross. Otherwise, as we have seen in corresponding laboratory examples, the perceived streams will not maintain a correspondence with the musical parts.

Segregation can also be improved by the weakening of fusion between corresponding notes of the different parts. Common fate between the parts (a factor that increases fusion) can be avoided by prohibiting synchronous onsets and offsets of the notes in different parts. This can be accomplished if the parts are given different rhythms. (These differences in rhythm will be appreciated, however, only when the parts end up being well segregated overall.) Also, when changes in pitch occur in two parts at the same time, they should not be parallel changes. We should avoid synchronous combinations of notes that have good harmonic relations, such as the octave (frequency ratio of 2:1) or the fifth (ratio of 3:2).

Traditional polyphonic music did not use spatial separation to keep the lines distinct, perhaps because this would have broken down the perceptual unity too much. However, present-day composers have employed this device.

To obtain vertical integration just violate these rules. Make large leaps from one pitch to the next within a part so as to weaken the sequential integration. Use harmonious combinations, such as the octave, and move the parts in parallel. Start and stop notes at the same time in the different parts.

Composers who worked in the polyphonic style were even able to control dissonance by the use of methods that exploited auditory scene analysis. I would like to distinguish two types of dissonance—psychoacoustic dissonance and musical dissonance. Psychoacoustic dissonance is the sense of roughness or unevenness that occurs when certain combinations of simultaneous tones are played. This sort of dissonance is not defined by the musical style. If two tones are spaced by a perfect fifth (seven semitones, or a ratio of 3:2) they sound

smooth or consonant when played together, whereas if they are spaced by a tritone (six semitones, or a ratio of 45:32) they sound rough. The roughness is caused when the partials of the two tones beat at a large number of unrelated rates. This does not happen with consonant combinations of tones. Another feature of psychoacoustically consonant combinations is that they seem to blend better than dissonant ones.

Musical dissonance, on the other hand, is a more cognitive experience. In many musical styles, certain combinations of simultaneous sounds are treated as stable and others as unstable. The resting points in the music will fall on the stable combinations. Listeners will tend to experience unstable combinations as points of tension or dissonance in the music and stable ones as points of rest. Unstable combinations in a piece will tend to resolve to stable ones. By the terms musical consonance and dissonance, I am referring to this stability and instability. In the Western musical tradition, the tones treated as unstable tend to be the very ones that are heard as psychoacoustically dissonant. This is probably not an accident. The combinations used as resting points were probably chosen because they were perceived as having a smoother quality. Although such a choice is not compulsory, it is easy to see why it was made.

Many composers of polyphonic music wanted to be freer to use the combinations of tones that were viewed as dissonant. This led them to try to control the dissonance of the experience. They did so by certain techniques of composition that exploited the principles governing primitive stream segregation. The experience of psychoacoustic dissonance is reduced when the notes whose combination is dissonant are made to fall into different perceptual streams. This is an example of the fact that, as the Gestalt psychologists pointed out, perceived qualities belong to organized perceptual units rather than existing for their own sake. It appears that the mere registration of incoherent beating by the auditory system is not sufficient to cause the listener to experience dissonance. If the tone combination that generates the beating is interpreted as an accidental co-occurrence of unrelated events, the dissonance is not assigned to the mental description of any of these events. Somehow it gets lost.

The techniques that these composers employed to control the fusion of the potentially dissonant tones were varied. The tones were not allowed to start and stop at the same time. They were captured into separate sequential streams by preceding them by tones close to them in pitch or by capturing them into smooth trajectories or into repetitive sequences.

Another approach to segregation that can be used when there are many tones playing at the same time is to cause different groups to form separate simultaneous chords in which the internal relations are consonant and to try, at the same time, to dissociate these chords from one another perceptually. The technique of "polytriads" takes this approach. The triads, taken as units, are prevented from grouping with one another by separating them well in pitch range and keeping their onsets and offsets asynchronous. Furthermore, good harmonic relations within each triad can set up separate harmonic frameworks that help to keep the streams distinct in perception. It is also useful to change the pitches of the notes inside the triad in parallel so that the continued harmonic relations can assist the integration.

Not every method that could minimize the experience of dissonance has been tried. We might decorrelate the vibrato in the potentially dissonant tones, or might separate them in space. All this presupposes that we want to suppress the experience of psychophysical dissonance. If we wanted to enhance it, and if we could use the kind of precise control over the sound that computer-generated music offers, we could exactly synchronize onsets and offsets, vibrato, and so on.

Any sort of perceptual feature that depends on isolating some of the patterns in music can be influenced by primitive grouping. We can take the example of rhythm. Rhythms are relationships between auditory events in the same stream. We can show this by the example of the polyrhythm. A polyrhythm is generated by superimposing the pulses from two different rhythms going on at the same time—say three pulses per second and four per second. If the two sets of pulses are not segregated by primitive factors (for example, if they are carried by tones of the same pitch) the listener will hear only the complex pattern generated by the sum of the two rhythms. However, if they are segregated, say by moving them apart in pitch, the two component rhythms will be easily heard.

The purpose of setting out these examples in the field of music has been to show that many of the principles of voice leading and orchestration can be understood as taking advantage of the natural tendencies of the auditory system to group sounds. A knowledge of such principles cannot prescribe musical goals, but it is possible that it could provide a principled foundation for the most basic level of musical organization. It may be interesting for musicians to find out that these principles are the same as those that allow a person to cross a busy street without getting killed or to carry on a conversation in a noisy room.

Auditory Organization in Speech

The next topic is the role that primitive auditory scene analysis plays in the perception of speech. Discovering its contribution is made more difficult by the fact that special-purpose schemas are heavily involved. So sorting out the relations between primitive organization and the organization imposed by speech-sound schemas is not easy. Some theorists argue that special schemas for recognizing the basic sounds of speech are innate in the human species. For the purposes of the present discussion it is not necessary to take a position on this issue. The existence of speech-sound schemas, innate or learned, makes it harder to uncover the contribution of primitive organization.

The "cocktail party" problem, as it is called, refers to the difficulty that a listener has in following one voice in a mixture of conversations. Research on this problem has shown that the difficulty is reduced if the target voice has some qualities that distinguish it from the others. These include pitch and spatial location and possibly voice quality and rate. (The predictability of later from earlier parts is also important, but since we are focusing on primitive organization, we will not discuss this factor.)

The statement that distinctive qualities allow us to attend to one voice obscures an important point. If we can perceive two or more separate locations, pitches, or timbres in the incoming signal, we must have already done some scene analysis. These qualities are the product of a partitioning of the auditory evidence. Therefore, if they assist in the sequential organization (the tracking of a single voice over time), they must have been made available as the result of an organization of the simultaneously present auditory components.

I want to briefly review first the sequential organization of speech and then its simultaneous organization.

Sequential Organization of Speech Sounds

The sequential integration of speech must operate on both shorter and longer time scales. On the shorter scale, the successive parts of a single word must be integrated. The success of word recognition depends on such sequential integration. Here is an example. We can create a laboratory version of the phrase "say chop" in which a short silence before the "ch" tells the listener that it is a "ch" and not a "sh". Yet if the voice changes from male to female between the two words, the male's voice will be heard as pronouncing "say" and the female's as pronouncing "shop". The critical silence is not given a phonetic significance because it is no longer a within-voice event.

Here is another example. It is said that the rhythm of a sentence is very important for its recognition. Yet rhythm is computed across the events that lie within a single perceptual stream. Without the stream-forming process, the rhythm could not be used.

Some aspects of this short-term integration are easy to account for by using the principles that I introduced earlier, but some aspects are not so easily explained. It is easy to account for the integration of the successive parts of a single vowel. Its properties tend to change very continuously. But it is harder to account for the integration of sounds that have very different qualities, such as "s" and "a". Phonemes (such as "s" or "a") occur at the rate of 10 or more per second in fluent speech. Yet we know that at 10 tones per second, an alternation of a noise burst and a tone will segregate into two streams. In what way is speech different? Are speech sounds integrated by schemas that make them immune to primitive organization? Probably not. When speech is speeded up by a factor of 2.5 or more it has been observed to segregate into substreams. Perhaps the coherence of normal speech happens because of, rather than in spite of, its acoustic structure.

To compare the sequential integration of speech sounds with other sounds, researchers have made up short repeating loops of vowel sounds or syllables. The listener's ability to integrate these cycles is much worse than for sequences that occur in natural speech. In the cycles, qualitatively different subsets of speech sounds tend to be heard in separate streams. In one experiment, syllables (such as "bee") that started with a stop consonant tended to segregate from isolated vowels (such as "oo") that were in the same loop. This suggests that the sudden onset of a stop-consonant syllable was treated as an important feature of that sound. The role of suddenness of onset as a feature defining the similarity of events has not yet been studied in the sequential integration of nonspeech sounds.

The stream segregation of cycles of vowels depends on similarities of pitch and on the speed of the sequence. The effects are the same as those that occur with nonspeech cycles. Segregation is greater with larger pitch differences and higher speeds.

The effects of pitch-based segregation have been studied with ordinary speech as well. Listeners were required to shadow a running speech sample (that is, to repeat it as it was being said). They received two different prose passages, one to each ear, and were asked to shadow whatever came into one ear (say the left). Occasionally the two passages would be swapped between the ears. Frequently, listeners made errors by continuing to track the same message rather than the desired ear. It was shown that this had happened for two reasons. The first was a tendency to follow the message that continued the same

conceptual content. This was clearly schema-based. But the second was a tendency to follow the same pitch contour over time. When the message suddenly switched ears, it would continue the contour that was begun before the switch, and this caused the listeners to follow the voice to the opposite side of the head. From this it appears that pitch continuity can hold a voice together even in the presence of a sudden change in location.

The converse of this fact is also true. Pitch discontinuities can break up a voice even when other factors favor continuity. A sudden pitch change occurring halfway through a synthesized word will make it sound as though a second voice has suddenly interrupted it. Here is another example. A formant pattern may sound like the syllable "wa" when the pitch is held constant, but if the pitch is suddenly changed in the middle of the "w", it sounds as if a first voice ends in the sound "oo" and a second begins with the sound "ba". The last part of the "w", presented with the second pitch, is interpreted as the brief transition of formants that defines a "b" rather than simply as part of the longer one that defines a "w". We can conclude that pitch continuities in the voice are important in holding it together as a perceptual unit.

The previous facts apply to voiced sounds, which, like tones, have pitch. Yet there are many sorts of consonants, such as "s" and "t", that have no pitch. How are they integrated with vowels? We know that when tones and noise bursts are spliced together into a repeating cycle, the tones and noises will segregate from one another and form separate streams. Yet this does not happen with natural connected speech.

Let us take an example of unrelated sounds in speech. Some African languages contain clicks whose position in a word can determine the word's meaning. This means that the perceptual stream must incorporate the clicks. Yet when a mechanical click is superimposed on a sentence, our perceptual systems segregate it from the speech sounds and we cannot decide on its exact position relative to them. What is the difference between a speech click and a superimposed mechanical click? When the click is produced in speech, it results from a movement that not only produces the click but affects the other adjacent sounds in the word. For example, the voicing will be briefly interrupted and the formants will undergo frequency transitions as the vocal tract moves into position to articulate the consonant. This synchronization of the click with changes in voicing may tie it to a definite location in the pattern of voicing. The situation is different with a mechanical click. If it is superimposed arbitrarily on a sentence, its occurrence will probably not be synchronized with

other changes and there will be no cues for integration. Quite to the contrary, there will probably be clues for continuity that travel right through the click. The same considerations apply to the integration of other sorts of consonants, such as "s", "t", or "ch", with vowels.

We do not know whether the synchrony of the changes is used by a primitive integrating mechanism or by speech-specific schemas. It is natural to conclude, when our perception deals differently with speech and nonspeech signals, that the difference must be due to speech schemas, but we should not draw such a conclusion before exploring the possibility that the effect may be due to primitive auditory grouping.

The idea that formant continuities can have an important effect on sequential integration has been supported by research findings. If a repeating cycle of vowels contains transitions connecting the corresponding formants of successive vowels, the sequence will be heard as much more coherent than it would be without these transitions. This is similar to what happens with a cycle of alternating high and low pure tones. When the tones are connected by frequency transitions, the sequence is less likely to split perceptually into high and low streams. Continuity of formants can help to overcome discontinuities in other properties of the speech. When there is a sudden shift in the fundamental frequency from that of a male voice to that of a female in the middle of a word, the word is more likely to be heard as a unit when the pattern of formant trajectories travels smoothly through the point of change.

We have mentioned two speech features—pitch and formants—whose acoustic continuity is important for the integration of speech. Another is spatial continuity. We rarely think of this sort of stability, but our auditory systems use it just the same. Its effects can be observed most easily when it is violated. For example, if we switch speech repeatedly back and forth between the ears, intelligibility suffers.

There is another factor that prevents natural speech from suffering from the same sorts of segregation as occur in repeating cycles of nonspeech sounds. There is a difference between a sentence and a repeating cycle of unrelated sounds. In the cycle, the same class of sound will repeat at regular intervals, close together in time. The tendency to group with its own class will compete with the tendency to integrate with its true sequential neighbors. Furthermore, the tendency that favors the segregation of different classes will grow stronger with more repetitions. Speech is different. There is no regular repetition of classes of sounds. Occurrences of similar noise bursts, such as "t" and "ch", may be spaced by seconds, and when

closer together may not be repetitive. However, if you repeatedly recycle a word formed of two classes of sound, such as "sissy", it will eventually segregate into two streams, one for the vowels and the other for the "s" sounds. This will happen more readily if the formant transitions from the "s" sounds to the vowels are spliced out. We must recall that the purpose of research with cycles is to force particular grouping tendencies to larger-than-life proportions. This means that the research on speech cycles has been successful in identifying factors involved in sequential integration and segregation, but it cannot accurately predict the actual amount of segregation in nonrepetitive material.

Simultaneous Organization of Speech Sounds

When we encounter speech, its sounds rarely occur alone. Therefore we face the problem of collecting those parts of the simultaneous jumble of auditory properties that define a single voice. This is the problem of simultaneous organization. The problem of extracting the sounds of a particular voice from a mixture is not independent of the problem of integrating them over time. If the auditory system can get a clear exposure to a single voice at one moment, it can select matching components from a mixture that occurs an instant later. This is an instance of the old-plus-new heuristic that we encountered in our study of the perceptual organization of nonspeech sounds. The auditory system can also use cues that depend directly on the relationships among simultaneously present components.

Much of the research on simultaneous organization has been concerned with how we decide how to allocate the components of the incoming spectrum to different voices. One important factor is fundamental frequency (experienced as pitch). Although it is only voiced sounds, such as vowels, that have a fundamental, they form a large enough part of the speech signal to make fundamental frequency very significant. The auditory system appears to look for one or more fundamentals that can account for as many of the frequency components in the spectrum as possible, and then to allocate all the components related to the same fundamental to the same stream. One experiment on this topic used speech spoken in a monotone. Two messages spoken on different pitches were much easier to distinguish than two spoken on the same pitch. The same effects of pitch differences have been obtained using pairs of synthesized vowels mixed together.

In normal speech the pitch is always changing. These changes are very important in the segregation of voices. Two voices are not likely to be following exactly the same pitch contour at the same time.

Therefore the accidental fit of the harmonics of two voices to a common fundamental is unlikely to persist for more than an instant.

Another important factor is location in space. Different spatial origins make voices easier to distinguish.

Differences in time of onset and in pulsations of intensity are important as well. When a voice rises suddenly in intensity, this increase is likely to be true of a number of its components. This shared property can tie them together.

Research on voice perception has tried to discover which properties of acoustic components occurring at the same time will tie those components together to define a voice. In this research there is usually only a single voice present and the perceptual choice is between integrating all of its components or missing some of them. Generally the need for manipulation of the features requires that the voice be synthesized. Different formants of the same voice are often given different properties and then the research tries to determine whether they resist integration. This kind of research has often produced paradoxical results. When the formants are given different properties, multiple sounds are heard, but these seem to be heard as a single speech phoneme or syllable.

The similarities in the formants that have been looked at are common spatial origin, common fundamental frequency, grouping with antecedent sounds, and asynchrony of onsets and offsets.

To look at the role of common spatial origin, different formants from the same synthetic voice have been sent to the opposite ears of listeners. If the other features of the formants match, the listeners will hear only a single sound and will integrate the formants to hear the phoneme that results from the combination. However, if the other features differ in some way, the listeners may hear two distinct *sounds* but only one phoneme or syllable. Whereas the left- and right-ear formants seem to be segregated to yield the experience of two sounds in space, the phonetic interpretation includes the information from both.

A similar effect is obtained when the formants differ in the fundamental frequency of their partials. Two sounds are heard, but the formants are integrated to derive a phonetic interpretation. There is one exception to this general pattern. When a number of formants are present and the listeners can draw different phonetic interpretations by grouping them in different ways, those with the same fundamental tend to be grouped together for phonetic purposes.

We have to explain why differences in raw acoustic properties will not affect phonetic integration unless there are alternative phonetic interpretations based on different groupings of formants. I think that

the success in achieving a phonetic integration of the formants when there is no competition is based not on raw acoustic factors but on schemas for speech sounds that can select the auditory evidence that they need. However, when several of these schemas are activated equally strongly, the choice of which one gets which piece of evidence may be determined by primitive scene analysis. This conclusion agrees with two facts: (a) the segregation of a mixture into two whole voices by their acoustic properties usually makes it easier to recognize the words; but (b) the segregation of the formants in a single voice often does no harm. In the case of two voices, many formants and harmonics will be present; presumably, different schemas will generally be competing for the same evidence and the partitioning of the data will resolve the competition.

So far we have thought of formants as being properties of a spectrum that are immediately available to perception. This is not the case. When we encounter a spectrum, it surely has peaks, but the position of the center of each peak may not be the product of only a single environmental sound. It may be determined by the superimposition of the partials from two or more environmental sounds. Yet auditory perception is not thwarted by this superimposition. It has ways of finding the peaks in the speech portion of the mixture. Experiments have shown that if primitive scene analysis decides that one of the partials in a spectral peak actually comes from a different source than the others, the estimated frequency center of the formant will change and this will alter the perceived speech sound. The decision to exclude the partial can depend on the fact that it is heard as a continuation of an earlier sound, or that it does not fall into the same harmonic series as the other partials in the spectral peak, or that it is too intense relative to the other partials. Why do schemas not resist the segregation of parts of a formant in the same way that they resist the segregation of whole formants from one another? Perhaps because it is not the role of schemas to decide what is or is not a formant. The defining of the formants in the spectrum may be the job of lower-level processes of auditory organization.

Duplex Perception and the Problem of Exclusive Allocation

The next topic is the principle of exclusive allocation. I am referring to the tendency to allocate the perceived features of a stimulus to one or the other of the perceived objects or events in the perceptual field but not to more than one. A visual example can be seen in figure 1.6 of chapter 1. In this drawing, a given contour is seen either as the

edge of the vase or as the edge of one of the faces, but never as both at once.

We have encountered similar examples in audition. A tone can be prevented from grouping with another one if it is captured by a better partner. This capturing would be impossible if a tone could be part of two unrelated organizations at the same time. It seems as if information is allocated exclusively to one organization or another. This might be a way of guaranteeing the most parsimonious perceptual description of the sensory input.

Despite the positive examples, we encounter cases in speech perception in which spectral information is used twice, once to define a speech sound and a second time to define a nonspeech sound. It has been claimed that this means that speech perception does not depend on primitive scene analysis. I would like to examine this issue by focusing on one example of the double use of evidence—duplex perception of speech.

The phenomenon can be produced in the following way. The first step is to synthesize a pair of syllables, say "da" and "ga", in such a way that only the initial transition of the third formant distinguishes them. Let us call this the *F3 transition*. Let us call the part common to the "da" and "ga" the *base*. Now let us present the base to one ear, say the left, and only the distinguishing F3 transition to the other ear. The two parts are temporally aligned as they would be in the full syllable. If we listen to this stimulus, we have an unexpected experience. At the ear that has the isolated transition sent to it we hear a chirp, and at the ear of the base a full "da" or "ga" syllable (depending on which F3 transition was sent to the other ear). We need the F3 transition to distinguish "da" from "ga", so we must have phonetically integrated the information from the two ears. Yet we heard two sounds, one at each ear. This means that we must have segregated the information from the two ears.

The experience is called duplex perception because the sensory evidence from the F3 transition is used twice. It contributes to the perception of the correct syllable at the left of the head and is also heard as a chirp at the right. Some theorists have used this duplexity of perception to argue that two independent systems have been activated: a speech-perception system and an "auditory" system. These systems have been conceptualized as being distinct in both a functional and a biological sense. The speech-perception system is seen as carrying out its own analysis without any dependence on the auditory system. The need for two systems to explain the effect seems to be justified by the belief that information cannot be used twice within each system. In other words, the rule of exclusive allocation must

hold within each one. Furthermore, it is argued, phonetic interpretation is not built on the results of scene analysis; otherwise the primitive segregation into two sounds would have prevented the phonetic integration.

My reason for thinking that this interpretation is wrong and that speech perception actually takes advantage of primitive scene analysis is as follows. We know that when voices are mixed, the recognition of speech *is* affected by factors such as spatial separation and pitch difference. If this fact does not imply that speech perception takes advantage of primitive organization, it must mean, instead, that the speech system itself contains methods for the segregation of voices by these factors, and the methods must respond in exactly the same way to these stimulus factors as primitive analysis does. It is simpler to suppose that it uses the existing multipurpose machinery that has been evolved to do primitive grouping.

Nonetheless, phonetic integration is remarkably unaffected by the acoustic relations between the base and the F3 transition. The loss in phonetic recognition from splitting the signal to the two ears instead of sending it all to one is significant, but not large. Research has found that the parts presented to different ears can also have different fundamentals or be started asynchronously. Although each of these differences reduces phonetic integration a little, it remains fairly strong. Even the frequency alignment of the tail end of the F3 transition with the continuation of the same formant in the base is not critical. The recognition schema for the syllable seems to be able to pull together the information it needs despite strong tendencies for primitive processes to separate it.

Since the theoretical implications that have been drawn from duplex perception of speech depend on the belief that the rule of exclusive allocation must inevitably hold true within a single perceptual system, I want to examine whether this belief is valid. If it is not, then duplex perception may not have to be seen as the conflict between two systems. Let us examine, then, the conditions for obtaining violations of the rule of exclusive allocation.

Separating the formants and sending them to different ears does not always produce duplex perception. In a two-formant synthesis, if each formant remains whole and the two are sent to opposite ears, as long as the two formants match well acoustically, not only is there phonetic integration but only one sound is heard. It seems that other acoustic differences must be added to the difference in spatial location before the listener will hear two sounds. When there are very strong cues that only a single sound is present it affects both the phonetic integration and the sounds-in-space organization.

Furthermore, presentation to separate ears is not necessary for producing the duplex effect. You can play both formants to a single ear; but if the formants have different fundamentals, the listener will hear two distinct sounds, yet at the same time will integrate the formants phonetically. It does not appear that any one acoustic difference between parts of the signal is essential for hearing two sounds, as long as there is a sufficiently large difference. Whether two sounds are heard or not, the formants will be integrated into a single speech percept as long as there is no competition among speech schemas for the use of particular formants.

Violations of exclusive allocation are not restricted to speech. We can play the highest and lowest notes of a musical chord to one ear and the middle note to the other. Let us set it up so that the middle note tells the listener whether the chord is major or minor. Many listeners will report hearing a full chord at the ear that gets the two notes and an additional note at the other ear. The full chord will be experienced as either major or minor, depending on the identity of the isolated note.

Another case of violation of exclusive allocation can be obtained with environmental sound. A recording of a metal door closing can be filtered to obtain a low-frequency and a high-frequency part. The low part, played alone, sounds like a wooden door closing and the high part in isolation sounds like a rattle. These parts are sent to opposite ears. The listeners will hear a metal door at the ear that receives the low part and an extra rattle at the other ear. The high-frequency sound is used twice, once in combination with the low to produce a metal door and separately, as well, to produce a rattle.

Other cases of duplex perception occur. When there are conflicting cues to the spatial location of a sound, sometimes two sounds will be heard. Also, when a partial is captured from a complex spectrum by a preceding copy of that partial played in isolation, the partial will be heard out from the spectrum, but the spectrum will still have some of the quality contributed by that partial.

We see from these examples that the violation of the rule of exclusive allocation need not involve the simultaneous use of the speech system and the nonspeech auditory system. Nor need it result from the use of two biologically and functionally distinct systems.

While the previous examples show that the violation *need not* involve the independent actions of two biological systems, they do not rule out the possibility that in the particular case of duplex perception of speech this sort of independent action is actually the cause. Therefore we have to look more closely at the assumption that the analyses

done by the speech-recognition system are not at all constrained by primitive auditory organization.

The first thing to remember is that the recognition of a speech sound when parts of it are sent to opposite ears is less accurate than when all parts are sent to the same ear. This, in itself, implies some constraint from auditory grouping. However, a stronger argument against the idea that the speech system can ignore primitive groupings comes from an experiment on duplex perception in which a perceptual grouping was shown to affect the recognition of the speech sound. In this experiment, the base was sent to the left ear and a formant transition (let us call it the "critical" transition) to the right as before. Together they made either the syllable "ga" or "da". But a series of additional formant transitions, short ones identical to the critical one, were played to the right ear before and after the syllable. These extra transitions grouped with the critical one and captured it into a stream of chirps at the right ear. This had the effect of reducing its contribution to the identification of the syllable. However, when the extra transitions were in a frequency range that was different from that of the critical one, they did not capture it into their stream and did not reduce its contribution to the phonetic identification of the syllable.

The violations of exclusive allocation seem to occur only in cases where two different sounds could be occurring at the same time. I do not know of cases that occur in the sequential grouping of pure tones. In cases of strictly sequential grouping, a tone seems to be required to be in only one stream at a time. It is only in cases of complex spectra that two percepts can be based on the same evidence. Perhaps the prohibition of the double allocation of evidence is weaker in complex spectra, where sounds might be overlapping in time. After all, sounds are transparent, so that a given spectral component at one ear could actually have resulted from the superimposition of components arising from more than one environmental event. Allocating the component to only one of these sonic events would be a mistake.

Here is an example. Suppose two sources of sound are active at the same time, one at my left and the other at my right. Suppose, in addition, that they both have a harmonic at 1,000 Hz and that the 1,000-Hz energy is balanced at my two ears. If the localization of sounds were determined by the balance of energy at the left and right ears, then I should hear a separate 1,000-Hz sound straight ahead of me. Such disembodied sounds should occur very frequently just by chance. But we never hear them. Instead, we allocate the energy to the larger sounds with which they have good acoustic relations. Such relations could include starting at the same time or fitting into the

same harmonic series. In the present example it is probable that the 1,000-Hz harmonic is allocated to both the left-side and the right-side sonic events.

How is the auditory system to know whether a piece of sensory evidence should or should not be allocated to two percepts? A reasonable guess is that it uses basic acoustic relations (such as spatial relations, harmonic relations, and frequency continuity) to assess how well the properties of the local piece of sensory evidence fit together with other pieces of evidence. If these relations strongly favor the allocation of the piece of evidence to one stream rather to another, it is assigned in an exclusive manner; however, if there is ambiguity, the evidence is allocated to both.

Ideally, the scene-analysis system should *share out* the evidence rather than assigning it twice. For example, if the isolated transition in the duplex perception of speech is used in completing the base, its perception as a chirp should be weaker, the energy having been divided between the two percepts. Actually, we do not know for sure that the energy from the transition *is* doubly allocated rather than being divided up. If it turns out that it is not shared out in some proportional way, this may be because it is too hard to estimate how much energy to allocate to each percept.

The multiple use of local bits of sensory data is not unusual in human perception. Many examples are found in vision. If we are looking at a shiny table top, the color of a small bit of the visual field will contribute (together with other bits, of course) to our perception of the color of the surface, to its glossiness, to the color of an object reflected in it, and to the color and compactness of the source of illumination. As the Gestalt psychologists were fond of pointing out, it is not true that the perceiver's knowledge about a local property of an object in the world is derived exclusively from the sensory properties of the corresponding local part of the visual field.

I have presented some examples to show that sensory evidence need not always be allocated in an all-or-nothing fashion. But if such examples can be found, why were we so convinced earlier that the rule of exclusive allocation was valid? It appears that this conviction was based on many confirming examples in audition and vision. Two examples that we can focus on are the ambiguous vase-faces example illustrated in figure 1.6 of chapter 1 and the capturing of tones from a sequential pattern that was shown in figure 1.7 of that chapter.

The argument that I have given up to this point seems to show that exclusive allocation occurs when the local sensory evidence fits in strongly with one organization and not with another. When the fit is more ambiguous, the evidence can be used twice. However, this

argument does not work well for the vase-faces drawing. The shared contour between the vase and a face must have about equally strong visual relations connecting it with both organizations. Otherwise the figure would not be ambiguous. Yet despite this ambiguity, the information is allocated in an exclusive fashion. It is either the edge of a face or the edge of the vase—never both at the same time.

The mutual exclusion of interpretations in the vase-faces drawing probably does not originate with a primitive analysis of the strengths of the connections among the components of the evidence. It may be a result of rules that govern the building of certain specific kinds of perceptual interpretations. There are two mutually contradictory interpretations of the spatial relations in the vase-faces picture. When an edge is seen as the outer contour of a face, we are interpreting the face region as being in front of the other region (the vase region). On the other hand, when we see the edge as defining an outer contour of the vase we are seeing the vase as in front of the other region (the faces region). It may be the contradictory interpretations of two regions that are being prevented (A in front of B and B in front of A), not merely the allocation of the edge to both regions. In general, exclusive allocation may be enforced by rules that prevent contradictions in the building of perceptual "descriptions."

We are left with a two-part account of perceptual organization. A first process, primitive auditory scene analysis, lays down links of different strengths between parts of the sensory evidence. This process uses the sequential and simultaneous relations that I have described earlier. Then a second process builds descriptions, using schemas (definitions of regularity) that may be either innate or learned. This process is strongly governed by requirements for consistency of interpretation. For example, a given simple sound cannot have two pitches at the same time. (It could have distinct *parts* with different pitches, but then it would no longer be the whole sound but its parts that were simple and were obliged to obey the rule.)

This line of reasoning can be applied to cases in which a tone is captured out of a sequence by other tones in an all-or-nothing fashion. I would argue that the all-or-noneness does not result from the fact that it now fits in better with the capturing tones. That fact just alters the strength of the links. It results from a rule that says that a tone cannot be part of two streams at the same time. This rule is part of a description-building system.

The two-part theory draws an important distinction between sensory evidence and perceptual descriptions. Sensory evidence is the raw input for which the descriptions must account. Its components may be shared out and used in multiple ways to build different aspects

of the descriptions as long as the descriptions account appropriately for the evidence. Descriptions, on the other hand, are required to be definite. A thing is this and not that, here and not there. If it is both here and there, it must be two things, and so on. The *things* of descriptions and the *components* of sensory evidence are two different sorts of entities and subject to different kinds of rules.

This theory can be applied to duplex perception of speech. At the first stage of the perceptual analysis of the dichotic stimulus, the acoustic relations set up fairly weak links between the sensory evidence arising from the two ears. Then schemas begin to build definite descriptions. I would propose that it is at this stage that both the schemas concerned with voices and those that describe sounds in space come into play. Neither is prior. The sounds-in-space system builds two descriptions, one for a low sound on the left and another for a high one at the right. It does so because the primitive links favor this interpretation. For the speech schemas, the subdivision favored by the primitive links is overcome by the fact that the combination of evidence from the two ears is favored by the fact that the combination is acceptable as a syllable of speech. The sharing of the right-ear evidence is permitted because of the transparency of sound, a fact about the world that often requires such sharing of evidence to take place if descriptions are to be arrived at correctly.

Next the sounds-in-space description must be composed with the speech interpretation to determine a location and other properties for the speech. Disembodied speech is apparently not a preferred percept. Why the system chooses to identify the speech with the left-side sound rather than treating it as a property of a compound left-and-right-side sound is not clear. Perhaps it is because the left-side sound has more of the qualities of a voice.

Directions for the Future

There is no chance that the reader will have arrived at this point in the discussion with the sense that anything is settled. Even if one accepts the general framework of auditory scene analysis, there still remains the problem of filling it out. I have assumed that the auditory system has a set of methods by which it untangles the mass of input and interprets it into a description of distinct sources of sound with their own separate properties. However, only certain experimental approaches and choices of sounds have been explored. Other broad avenues are untouched. Even within those areas in which there has been some research, many issues are unresolved and the detailed knowledge that could resolve them is not yet available. Let me sug-

gest, following the outline of the previous chapters, what some of these unknowns are.

On the topic of *sequential grouping*, a fairly general question is what sorts of similarities will affect the sequential grouping of frequency components. In order to answer this question we will have to decide whether the same process is responsible for two kinds of sequential grouping. The first is the grouping of already fused spectra to create a stream of tones or sounds. This first kind of grouping is seen in the streaming phenomenon. The second type is the sequential grouping that serves to decompose a complex spectrum into simpler ones by capturing components out of it. It would be simpler to believe that the principles are the same, but not enough studies have been done comparing them. For example, does the time interval between a capturing tone and its target harmonic in a larger spectrum affect their grouping in the same way that the time interval between pure tones does in the streaming experiment?

Other questions on this issue are also puzzling. For example, we know that a sequence of *already formed* tones can be grouped by their timbre. However, when it was a question of capturing a subset of components out of a complex spectrum, this could not be done by the use of timbre. We could not argue that the subset's timbre resembled that of an earlier component, because the computation of its timbre would depend on its having already been segregated from the rest of the spectrum. Spectral components do not have timbres; only "sounds" do. It is more reasonable to suppose that after the properties of spectral components have caused them to be grouped into distinct sounds, and global properties (such as timbre) have been computed for them, these computed properties are able to serve as the basis for another level of grouping (such as by timbre).

Even when puzzles about the level of grouping are not involved, we do not know which properties of sounds will cause primitive scene analysis to group them with others. Will spectrally shaped noises have a tendency to group with tones whose spectra have peaks in the same places? The answer to this question is important if we are to understand the perceptual integration of speech. What about the common environmental sounds, with their myriad of qualities: rasping, tinkling, grating, clanking, thumping, and so on? Can every nameable quality affect how sounds will group? We must hope that it is not so and that a limited number of measurable properties will affect their grouping. Otherwise the prediction of grouping from physical measurements may be impossible.

There are also questions about the *formation of units* in the sound stream. Do sudden rises in amplitude tell the auditory system that a

new event is beginning and that the computation of new global properties such as pitch and timbre should be started?

Many issues concerned with the segregation and integration of *simultaneous auditory components* also remain unresolved. For instance, does the parallel frequency modulation of partials actually improve their spectral integration? Is this an example of the Gestalt principle of common fate or is it just that the parallel movement (on a log frequency scale) preserves the "good" frequency relations between the harmonics over time? This question could be studied with inharmonic partials. In these sounds the maintaining of "bad" frequency relations over time by frequency modulation would not be expected to assist the integration.

We also do not know what evidence is used by the auditory system to decide that a particular harmonic is too loud to be incorporated into a larger spectrum, but we know that the system does so. Given the fact that a natural spectrum, such as found in the human voice, has many peaks and valleys, what mechanism decides that a harmonic is too loud to simply be a spectral peak?

There are still many unanswered questions about how the perceived qualities of a spectrum can be affected when scene analysis partitions it. For example, we know that more than one pitch can be derived at the same time when more than one harmonic series is detected. But we do not know whether the computation of pitch can be altered if we induce the partitioning of the spectrum by factors *other than* harmonic relations. For example, consider a spectrum, A, consisting only of every third harmonic of a certain fundamental. It sounds an octave and a fifth higher than spectrum B, which has all the harmonics of the specified fundamental. Suppose we rapidly alternate the two spectra. Would A capture its corresponding harmonics out of B so that the pitch of A was also heard as part of the B tone? If so, it would mean that the local pitch computation during B had been affected by stream segregation. What other global properties can be affected by auditory organization?

We also need further research on how the spatial separation of sources allows the spectra from those sources to be segregated. Even if the auditory system is capable of getting an independent estimate of the spatial origin of a large number of spectral bands, are these bands narrow enough to permit the separation of sources whose harmonics are interlaced in the spectrum? What are the limits of location-based segregation?

We also need to find out to what extent perceptual fusion and *masking* can be accounted for by the same mechanisms. Despite the differ-

ences in their definitions, is the difference between the factors that influence them merely a matter of degree?

Is it really possible to decide whether fusion is or is not the default condition of the spectrum? Does the auditory system always need specific evidence to override this default?

The *perceived continuity* of soft sounds through louder masking ones deserves more study. In particular, we must test the assumption that before continuity can be heard the auditory system must decide that the soft sound that enters the masking noise and the one that exits from it are the same sound. How much of this decision is determined by primitive links set up by similarity relations between the entering and exiting sounds and how much is based on a model of what the sound is and how it might have changed during the time that it was not clearly heard? It is possible that we can use the phenomenon of perceived continuity through a masking sound as a way of discovering how our brain expects a nonspeech environmental sound to change over time. For example, many sources of sound generate a richer spectrum when more energy is forcing the vibration. The voice, the trumpet, and the violin exhibit this kind of correlation between intensity and richness. Is knowledge of this correlation built into our auditory system as a general rule for integration? If it is, then a sound that violated the correlation (entering a masking noise burst at low intensity, but rich in harmonics, and exiting louder but purer) should be less likely to be heard as continuous than one that obeyed the correlation.

The questions that have been raised by considering the *role of schemas* in perception are many, but there is one central one. We do not know for sure whether there actually are two separate phases of scene analysis, one based on primitive grouping and using regularities in the auditory input that are shared by broad classes of sounds, and a second that builds detailed and coherent descriptions of the sound, employing schemas that incorporate knowledge about specific domains such as speech, music, or specific environmental sounds. If the two exist, they are sure to be present in every experiment. What research could separate their effects? What properties distinguish them: the effects of learning, the involvement of attention, the use they make of complex rhythmic patterns, the ability to remove the effects of certain sensory evidence from a mixture so that the residual can be more easily recognized?

How do schemas and primitive organization interact? When can schemas ignore the boundaries created by primitive organization?

The consideration of *music* presents us with issues that go unnoticed in other forms of sound. The basic one is that it is not sufficient to

think of the auditory system as organizing its input into separate simple sounds, each arising from a separate source. A piece of music is heard as separate from the coughs of the audience, yet it often has a part-whole structure with several lines going on within it. Is the primitive process that segregates the lines from one another exactly like the process that segregates the music from other sounds? What maintains the perceptual integrity of the music as a whole? Is the hierarchical structure maintained purely by musical schemas or by primitive grouping principles as well? (We should recall that a nonmusical sound, such as the one produced by a machine, can also have a part-whole structure.)

It is important for musicians to be able to predict whether a sequence of sounds will be experienced as a progression or a transformation. Is this determined entirely by our learning to expect the later parts when we hear the earlier ones or does the progression have to satisfy certain acoustic requirements so that the primitive organizing system will integrate it?

When will we hear two timbres at the same time in musical sound? Can segregation prevent us from hearing qualities of the unpartitioned spectrum, such as psychoacoustic dissonance? The research to really answer these questions has not yet been done. If it were carried out, it might make possible a science of orchestration.

There are several issues in *speech perception* that need to be resolved. What holds sequences of speech sounds of different categories together (such as fricatives and vowels)? For example, do the formants present in unvoiced sounds play a role in keeping them integrated with their voiced neighbors? Also, do changes in the voiced spectrum reveal the exact position at which a consonant is inserted? If so, is the integration of the vowels and consonants accomplished by primitive scene analysis or by speech-specific schemas?

We know that the depth of the spectral valleys between formants is not very important when only a single synthesized voice is played, but does the energy in the valleys play a role in holding the voice together as a coherent sound when other sounds are present?

How do the listener's mental model of speech, general models of sounds, and low-level grouping rules based on acoustic properties function in speech perception? How could these different types of constraints interact smoothly in a cooperative fashion? When the mental model accepts only some of the acoustic components that are present as being part of a speech sound, does the nonacceptance of some of them cause them to be formed into a separate stream that can be heard as an extraneous sound? What does it take to camouflage speech sounds (as opposed to masking them)? Does it require other

speech sounds, or can nonspeech sounds do it? In general, how do low-level properties and mental models (both for speech and for other types of sounds) cooperate in effecting the perceptual result?

This volume has looked at the part played by the Gestalt-like processes of primitive grouping in perception. However, the need to account for certain grouping processes in speech, such as duplex perception, has made it clear that an understanding of primitive grouping will take us only so far in understanding auditory perception. We have to go on to study the processes that use the sensory evidence to build descriptions.

There are several important questions about description building that have attracted some research but deserve to be pursued with greater effort. One is the question of whether some features of signals can be excluded from awareness when they cannot be given a solid home in a description. We have encountered some apparent examples in this volume. We found that in listening to music, the listener's awareness of the psychophysical dissonance between co-occurring notes could be suppressed if primitive grouping led to the notes being perceived in separate streams. The evidence came from examining musical practice and scholarship, not from direct experimentation. The latter need to be done.

There are issues in the formation of even simple tone sequences that may require an understanding of the roles of consistency and contradiction in the building of descriptions. For example, when a sound consisting of several spectral components is followed by a lower-frequency version of itself we will typically hear a single tone moving downward in pitch. Yet if one of the components of the first spectrum was at the same frequency as one of those in the second, you might expect the pair of matching components to be heard as a repeating pure tone. When will this happen? In general, when does a sound act as a whole in grouping with later sounds and when do its spectral components try to find matches, either individually or in groups? Does the decision among these percepts use principles that are outside the scope of primitive scene analysis? What is the role of the consistency of the derived interpretation in determining the choice of percept?

Other illusions—such as Deutsch's octave illusion, in which properties derived from different environmental events are assigned to the same perceived sound—also seem to require an explanation in terms of a description-building process.

Practical Applications

In the preceding chapters, I have not considered in any detail the practical applications of our understanding of auditory scene analysis. However there are many.

An important area of application is the automatic recognition of sounds by computers. Many of the methods that are currently in use, for example, in speech recognition, deteriorate badly when extraneous sounds are present. Incorporating a primitive scene-analysis stage into the recognition process might allow the systems to resist being derailed by these sounds. Some beginnings have been made by using a few features of a voice, such as its fundamental or its spatial location, to track it through a mixture, but no computational system has so far attempted to implement the full range of heuristics described in the earlier chapters.

Similar issues arise in automatic music transcription by computers. Before an incoming acoustic signal can be accurately transcribed into musical notation, the computer must decide how many instruments are playing at the same time and which instrument is playing which note. These decisions require the incoming spectrum to have been decomposed to derive a number of separate pitches and timbres. Only when such a capability is present can more "musical" properties of the sound, such as rhythm, tonality, and so forth, be derived. Current attempts have segregated the parts by using only the harmonic relations in the spectrum and the tendency of individual parts to take small steps. As a result they have been able to separate only simple mixtures of sounds. The difficulty in using a multiplicity of acoustic relations all at the same time seems to be due to the large computational load that is involved in handling a lot of information about alternative groupings of spectral components.

Another area of application is in designing signals for the workplace. It should be evident that the perceptual segregation of signals in work environments (for example, the segregation of instructions or of warning signals) is critical for the safe and effective execution of the activity. If a warning signal is perceptually grouped with an environmental sound or with another warning to create a global sound with new qualities, it will not be responded to appropriately. As advances in artificial intelligence give us machines that both talk and listen to us, the importance of knowledge about auditory scene analysis will increase for psychologists and engineers concerned with ergonomics.

A more detailed understanding of auditory perception should also make it easier for us to understand how people differ in their auditory perceptual capacities. In our laboratory we have noticed consistent differences between people. In one experiment, different cues for the

grouping of sounds—sequential, spectral, and spatial relations—were pitted against one another. Different people with normal hearing tended to resolve the conflict in different ways. That is, they heard different kinds of illusions. But the same person would resolve the contradiction in the same way on different occasions. This led me to suspect that different people give different weights to the various clues that point to the correct organization of the auditory evidence. In good listening environments, all the clues point in the same direction and people, despite their differences, generally come up with the same answers. But when conditions deteriorate, this no longer happens. Perhaps an individual can be characterized by a profile that describes the weights that he or she assigns to different clues. This might enable us to predict the kinds of failures that might be expected in different listening situations or with different types of hearing loss or different sorts of hearing aids for that person.

It is possible, finally, that certain types of neurological conditions that lead to a deficit in being able to deal with the order of sounds, or with mixtures of sounds, may be understandable as damage to the system that performs primitive organization.

Another point concerns computer models of human auditory perception. I have come to the conclusion after writing these chapters that we are now in a substantially different position than we were some years ago. At that time I decided that it was too early to attempt to construct explicit computational models of primitive scene analysis. But we have now reached the point where we have a good appreciation of many of the kinds of evidence that the human brain uses for partitioning sound, and it seems appropriate to begin to explore the formal patterns of computation by which the process could be accomplished. Unfortunately, as a psychologist, I do not personally have the skills and knowledge that are needed in this endeavor. I hope that some researchers who do have them will, after reading these chapters, accept the challenge of constructing an explicit model.

The level at which the research reported in this volume attacks the problem of the perception of sound is somewhere between the levels of auditory psychophysics and auditory pattern recognition. I find this level attractive because it deals with the issue of pattern but is, in some sense, content-free and therefore quite general. It is relevant to questions of everyday life, and yet the questions are often answerable in acoustic terms; we are never very far away from the sound.

A few final observations: What seems remarkable to me in rereading this volume is that many things we take as self-evident, such as the coherence of a single note or of a single voice, are perceived through a process of grouping. Another nonintuitive idea is that

many qualities that seem to be the automatic products of simple acoustic properties also depend on grouping. These include pitch, loudness, timbre, location, and dissonance. Finally, I am struck by the realization that the processes of audition that can accomplish the grouping and use it to derive these experiences must be doing so in time periods that we have to measure in milliseconds.

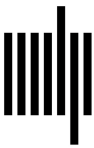
Auditory Scene Analysis

The Perceptual Organization of Sound

By: Albert S. Bregman

Citation:

Auditory Scene Analysis: The Perceptual Organization of Sound
By: Albert S. Bregman
DOI: 10.7551/mitpress/1486.001.0001
ISBN (electronic): 9780262269209
Publisher: The MIT Press
Published: 1994



The MIT Press

First MIT Press paperback edition, 1994

This book was set in Bembo
by Asco Trade Typesetting Ltd. in Hong Kong
from computer disks provided by the author,
and was printed and bound in the United States of America.

Library of Congress Cataloging-in-Publication Data

Bregman, Albert S.
Auditory scene analysis: the perceptual organization of sound /
Albert S. Bregman.

p. cm.
"A Bradford book."
Includes bibliographical references.
ISBN-13: 978-0-262-02297-2 (hc. : alk. paper)—978-0-262-52195-6 (pbk. : alk. paper)
ISBN-10: 0-262-02297-4 (hc. : alk. paper)—0-262-52195-4 (pbk. : alk. paper)
1. Auditory perception. I. Title.
[DNLM: 1. Auditory Perception. WV 272 B833a]
QP465.B74 1990
152.1'5—dc20
DNLM/DLC
for Library of Congress 89-14595
CIP

10 9 8 7